






ORIGINAL RESEARCH

Deep Learning Algorithm for Automated Cardiac Murmur Detection via a Digital Stethoscope Platform

John S. Chorba , MD*; Avi M. Shapiro , PhD*; Le Le, MD, MPH; John Maidens , PhD; John Prince, DPhil; Steve Pham, MD; Mia M. Kanzawa, MD; Daniel N. Barbosa; Caroline Currie, BA; Catherine Brooks, MPH; Brent E. White , MD; Anna Huskin, RN, BS, BSN, CCRC; Jason Paek, BA; Jack Geocariss, BS; Dinatu Elnathan, RN, BS, BSN; Ria Ronquillo, MS; Roy Kim, BS; Zenith H. Alam , DO; Vaikom S. Mahadevan , MD; Sophie G. Fuller, BS; Grant W. Stalker, BS; Sara A. Bravo, MSN, RN, CCRP; Dina Jean, BS; John J. Lee, MD; Medeona Gjergjindrea, DO; Christos G. Mihos, DO; Steven T. Forman, MD; Subramaniam Venkatraman, PhD; Patrick M. McCarthy, MD; James D. Thomas, MD

BACKGROUND: Clinicians vary markedly in their ability to detect murmurs during cardiac auscultation and identify the underlying pathological features. Deep learning approaches have shown promise in medicine by transforming collected data into clinically significant information. The objective of this research is to assess the performance of a deep learning algorithm to detect murmurs and clinically significant valvular heart disease using recordings from a commercial digital stethoscope platform.

METHODS AND RESULTS: Using >34 hours of previously acquired and annotated heart sound recordings, we trained a deep neural network to detect murmurs. To test the algorithm, we enrolled 962 patients in a clinical study and collected recordings at the 4 primary auscultation locations. Ground truth was established using patient echocardiograms and annotations by 3 expert cardiologists. Algorithm performance for detecting murmurs has sensitivity and specificity of 76.3% and 91.4%, respectively. By omitting softer murmurs, those with grade 1 intensity, sensitivity increased to 90.0%. Application of the algorithm at the appropriate anatomic auscultation location detected moderate-to-severe or greater aortic stenosis, with sensitivity of 93.2% and specificity of 86.0%, and moderate-to-severe or greater mitral regurgitation, with sensitivity of 66.2% and specificity of 94.6%.

CONCLUSIONS: The deep learning algorithm's ability to detect murmurs and clinically significant aortic stenosis and mitral regurgitation is comparable to expert cardiologists based on the annotated subset of our database. The findings suggest that such algorithms would have utility as front-line clinical support tools to aid clinicians in screening for cardiac murmurs caused by valvular heart disease.

REGISTRATION: URL: <https://clinicaltrials.gov>; Unique Identifier: NCT03458806.

Key Words: auscultation ■ machine learning ■ neural networks ■ physical examination ■ valvular heart disease

The stethoscope is an iconic medical instrument nearly synonymous with Western medicine. It is easy to handle, it is inexpensive, and its use is universally accepted, even expected, as part of a physical examination. Yet, for the stethoscope to be a useful tool, it requires that the provider both hears

Correspondence to: John S. Chorba, MD, 600 16th St, Genentech Hall N514, MC2280, San Francisco, CA 94143. E-mail: john.chorba@ucsf.edu and Avi M. Shapiro, PhD, 1212 Broadway St, Suite 100, Oakland, CA 94612. E-mail: algorithms@ekohealth.com

*J.S. Chorba and A.M. Shapiro contributed equally.

Supplementary Material for this article is available at <https://www.ahajournals.org/doi/suppl/10.1161/JAHA.120.019905>

For Sources of Funding and Disclosures, see page 12.

Preprint posted on MedRxiv, April 20, 2020. doi: <https://doi.org/10.1101/2020.04.01.20050518>.

© 2021 The Authors and Eko Devices, Inc. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- A deep learning algorithm applied to digital heart sounds detects murmurs with similar accuracy to expert cardiologists.
- Applying the algorithm to heart sounds captured at the appropriate anatomic location identifies severe forms of aortic stenosis or mitral regurgitation.

What Are the Clinical Implications?

- Our results suggest that a murmur detection algorithm used with a digital stethoscope can serve as a clinical decision-support tool for the diagnosis of murmurs and valvular heart disease.
- By offloading the burden of “auscultation interpretation” from the provider, the algorithm could improve the utility of the auscultation in screening for severe forms of valvular heart disease.

Nonstandard Abbreviations and Acronyms

AS	aortic stenosis
MR	mitral regurgitation
VHD	valvular heart disease

and correctly interprets the diagnostic sounds of the patient. Although nearly all providers can perform the act of auscultation with minimal training, **interpretation of the heart sounds is difficult**, even for specialists. Interrater reliability in detecting a classic finding, the “murmur,” is fair at best ($\kappa=0.3\text{--}0.48$),^{1,2} and the ability to identify the underlying pathological feature is even worse.³ Moreover, these challenges are further exacerbated by a noisy and rushed environment, which is the norm in modern practice. Despite the paucity of data in this area, anecdotally, these conclusions ring true for a wide spectrum of medical providers. Because cardiac auscultation remains a cornerstone of the physical examination, diagnostic assistance of its interpretation could therefore be of great use.

Classic teaching of auscultation, and murmurs in particular, focuses on valvular heart disease (VHD). VHD is a major cause of mortality and reduced quality of life for tens of millions of patients worldwide.^{4–8} As life expectancies increase, so does the prevalence of VHD in elderly patients. Annual VHD fatalities have increased 2.8% each year in the United States since 1979 and are projected to double over the next 25 years.^{5,9} VHD can also manifest with a prolonged asymptomatic period, which can be dangerous if not identified. For example,

patients with asymptomatic severe aortic stenosis (AS) who do not undergo aortic valve replacement have an annual rate of sudden death of 3% to 13%.¹⁰

Echocardiography remains the gold standard for diagnosis of VHD, given its minimal physical risk and excellent test characteristics.¹¹ Yet, echocardiography requires both highly trained sonographers to acquire the data and cardiologists to interpret the images. Accordingly, echocardiography is expensive, with a total annual cost of \$1.2 billion for Medicare enrollees alone.¹² In addition, echocardiography requires a preexisting suspicion from the referring provider and may not be locally available for patients in medically underserved areas. Because VHD is associated with textbook auscultatory findings,¹³ cardiac auscultation can be a fast, familiar, and inexpensive tool to improve access to VHD screening, facilitate earlier detection of VHD, and reduce the need for echocardiography. We therefore investigated the use of an electronic stethoscope platform to develop a deep learning algorithm to identify cardiac murmurs.

Deep learning approaches have shown great promise in medicine, using radiologic studies¹⁴ and echocardiograms¹⁵ to develop interpretative algorithms, and can even translate auxiliary data, unintended to be part of the original data set, into useful information.¹⁶ Stethoscope sound analysis has recently led to applications in lung¹⁷ and heart¹⁸ sound classification. For example, an independently developed algorithm, focused on the binary distinction between pathologic and normal heart sounds, has tested favorably in a pediatric cohort.¹⁹ The **2016 PhysioNet/Computing in Cardiology Challenge** inspired a wide range of solutions by curating the largest public data set of normal and abnormal heart sounds.²⁰ However, deep learning approaches often require a large amount of ground truth labeled data. Most prior research, and the top performing submissions for this challenge, used traditional machine learning or shallow neural networks requiring hand-engineered features and manual tuning that made assumptions about temporal and spectral characteristics of heart sounds.^{21,22} These assumptions potentially limit their generalizability to widespread clinical use. A central contribution of the current work is an algorithm that learns the important features directly from the raw audio instead of them being prescribed. Because cardiologists are traditionally trained to identify VHD by auscultation, we hypothesized that a deep learning approach could perform similarly, if not better, than these specialty providers and assist in the diagnosis of VHD.

METHODS

Code and Data Availability Statement

Additional supporting data are available on request from the corresponding authors. Programming code

related to data processing and not subject to intellectual property or confidentiality obligations will be made available on request. All requests for raw and analyzed and related materials data will be reviewed by the corresponding authors and the Eko legal department to verify whether the request is subject to intellectual property or confidentiality obligations. Any data and materials that can be shared will be released via a Material Transfer Agreement. Patient-related data not included in the article were generated as part of a prospective clinical study (NCT03458806) and may be subject to patient confidentiality and institutional review board review.

Algorithm Development

Eko's heart murmur detection algorithm has been approved for US Food and Drug Administration 510(k) clearance²³ and is integrated with the Eko digital stethoscope and ECG software platform to assess heart sound recordings. The algorithm was trained on recordings from a Health Insurance Portability and Accountability Act (HIPAA)-compliant collection of 400 000 audio recordings from Eko CORE and DUO electronic stethoscopes. The training set consisted of 5878 deidentified audio recordings, totaling >34 hours from 5318 unique patients. Recordings were initially randomly selected from the first 60 000 collected in the cloud-based Eko database and then subselected to ensure sufficient murmur examples to train the model. The training data quality and patient population are thus representative of what we expect in actual clinical use. A fraction of the database was set aside for internal testing and tuning hyperparameters of the model. The validation set recordings used to measure classification performance of the trained algorithm are entirely separate and collected specifically for this purpose.

To complete the training set for a supervised learning problem, audio recordings and phonocardiograms were reviewed and labeled by one physician as 1 of 3 classes: heart murmur, no heart murmur, or inadequate signal. Recordings of lung sounds, noise, and human speech were examples of data labeled inadequate signal.

The neural network model used for phonocardiogram classification uses a ResNet²⁴ deep convolutional neural network architecture (Figure S1). Before being sent to the model, input recordings are filtered using an eighth-order Butterworth high-pass filter at 30 Hz and downsampled to 2000 Hz. The model consists of 34 layers, each made of a 1-dimensional convolution, rectified linear unit nonlinearity, batch normalization, dropout for regularization, and maximum pooling. Layers are linked by residual connections to facilitate training by allowing gradients to

propagate. The final output of the network consists of a fully connected layer followed with 3 outputs subsequently normalized to a probability distribution via a softmax function. The network was initialized with random weights and optimized using the ADAM optimizer.²⁵

The end-to-end algorithm makes a sequence of binary decisions to produce 1 of 3 possible outputs. First, it determines whether the recording is of sufficient signal quality for murmur classification, using the output from the neural network corresponding to "inadequate signal" as a measure of signal quality. If the signal quality is found to be below a prespecified threshold, then the recording is classified as "inadequate signal." Otherwise, the classifier then provides either a "heart murmur" or a "no heart murmur" output based on another set threshold. All model parameters and thresholds are fixed at training time.

To validate the end-to-end algorithm, separate expert clinicians annotated a test subset of 1774 recordings from 373 patients collected through the multisite clinical study. For the binary signal quality screening step, the algorithm output was compared with the annotations of signal quality qualitatively. For the final murmur detection step, algorithm output was compared with annotations of murmur presence using the measures of sensitivity and specificity.

To show that the algorithm detects murmurs associated with clinically significant VHD, such as AS and mitral regurgitation (MR), we compared murmur predictions for the clinical study participants with echocardiographic assessment of VHD. For AS, we compared recordings at either the aortic or the pulmonic positions, with aortic recordings preferred. For MR, we compared recordings at the mitral position only. A single recording was used for each subject, with all recordings required to have an algorithm output of "murmur" or "no murmur" (ie, adequate signal quality). For greater consistency, CORE recordings were preferred over DUO recordings because the former were more numerous. Recordings used for annotator metrics naturally required annotation, and for AS, this at times resulted in a pulmonic recording being assessed by the annotators while an aortic recording was assessed by the algorithm.

Clinical Study Design

We undertook a cross-sectional, multisite study of subjects presenting to the echocardiography laboratories and structural heart disease clinics at the Northwestern Memorial Hospital (Chicago, IL), University of California San Francisco Medical Center (San Francisco, CA), Los Alamitos Cardiology Clinic (Los Alamitos, CA), and Mount Sinai Medical Center (Miami, FL) to obtain paired electronic stethoscope recordings with clinical

echocardiography results. Inclusion criteria included age >18 years, a complete (ie, not limited) echocardiogram, and provision of **informed consent**. Because of the lower prevalence of severe VHD compared with normal hearts in subjects presenting to the echocardiography laboratories, we also prescreened potential subjects by chart review to preferentially enroll expected cases. The primary outcome measures were defined as the ability of the algorithm to differentiate either clinically significant AS or clinically significant MR from normal hearts, reported through a receiver operating characteristic (ROC) curve. The protocol was approved as a minimal risk study by each of the institutional review boards of the participating sites and registered on clinicaltrials.gov (NCT03458806). All patients gave written informed consent.

Sample Size Justification

For our validation set, we assumed an algorithm sensitivity and specificity of 0.9 for both the detection of clinically significant AS and MR, which were unknown at the time the study began. We estimated that a sample size of 110 subjects in each group (AS cases, MR cases, or structurally normal controls) would exceed a minimum threshold likelihood ratio of 5 with 95% confidence.²⁶ Because final echocardiography results were not known before auscultation, and enrollment of controls would likely exceed that of cases, we estimated that auscultation of 900 subjects would be required.

Stethoscope Recordings

Recordings of the phonocardiogram were performed by trained study personnel in a standardized manner in each subject's clinic or laboratory examination room at the study site. Each subject underwent 15-second recordings, while seated, at the 4 standard auscultation positions (aortic: second intercostal space, right sternal border; pulmonic: second intercostal space, left sternal border; tricuspid: fifth intercostal space, left sternal border; and mitral: fifth intercostal space, midclavicular line). A second attempt at recording was encouraged if real-time auscultation quality at a given position was poor. Subjects remained in the study database even if recordings were not obtained from all 4 positions. These recordings were obtained with the standard, clinically available Eko mobile application wirelessly connected first to the Eko CORE Stethoscope, then to the Eko DUO, which also records a single-lead ECG. Recorded phonocardiogram and ECG data were saved as 16-bit, 4000- and 500-Hz sampled WAV files, respectively, and were synced in real-time to a **Health Insurance Portability and Accountability Act**-compliant cloud storage location and sent to the algorithms for analysis. Auscultatory recordings were reviewed by the study investigators for quality control. At the time

of recording, study personnel performing auscultation were unaware of the final echocardiography reports.

Phonocardiogram Annotations

Using a custom-made web platform, expert annotators listened to heart sound recordings from a subset of the overall clinical study with headphones while viewing a plot of the phonocardiogram, and while blinded to the results of the algorithm and echocardiogram. **Expert annotators were cardiologists having completed fellowship training in cardiology and having at least 10 years of clinical cardiology practice**, and each received modest financial compensation. Annotations were performed on existing recordings while the phonocardiogram database was actively expanding, but because not all annotators were available after the entire database was collected, only a subset of the final database underwent complete annotation. Annotators assessed signal quality (on a 1–5 scale with defined rubric), murmur presence (true or false), and murmur grade (Levine scale 1–6²⁷). Because murmur grade was determined by recording only, murmur grades 4 through 6 were not used. To establish a single set of ground truth labels for a recording, we aggregated the responses of the 3 cardiologists. For murmur detection, we used a majority vote. For signal quality and murmur grade, which were encouraged but not required for annotation, we used the median of the responses if there were 3 and the lessor if there were only 2, which occurred for a small number of subjects.

Echocardiographic Data

Comprehensive transthoracic echocardiograms, including 2-dimensional, M-mode, and color and spectral Doppler imaging, were obtained as part of routine clinical care. Clinical echocardiograms and their reports followed American Society of Echocardiography guidelines.^{28,29} Reports therefore graded VHD as none, mild, moderate, severe, or critical, with additional borderline categories (eg, moderate to severe) also allowed. Cardiologists reading the echocardiograms were unaware of study participation and thus blinded to any auscultatory results. Because the reports directed the clinical care of the patients, we considered them as the “gold standard” for our study. The reports were deidentified, and a single report was associated with each subject at the study site. Most echocardiograms and phonocardiograms were captured on the same day, although echocardiograms within 1 month of recording were permitted, which occurred only for a small number of subjects presenting through the structural heart disease clinics. We defined “clinically important” or “significant” VHD cases as those graded moderate to severe or worse, for this level of disease would typically require an evaluation by a cardiologist

for possible procedural intervention. We allowed mixed VHD as cases provided that the disease severity at any other valve did not exceed that of the index valve. We defined controls as subjects free of valvular, structural, or congenital heart disease, with no valvular regurgitation or stenosis beyond trivial or physiologic severity.

Statistical Analysis

Data analysis and visualization were performed in Python using the standard packages numpy, pandas, seaborn, matplotlib, keras, and tensorflow. CIs were computed by bootstrap rather than approximations, which require assumptions about data distributions. To compare means, the Welch *t*-test was used. To compare proportions, such as sensitivity, on different data samples, the “N-1” χ^2 test was used. To assess interrater reliability, Fleiss’ κ was used.

RESULTS

Validation Study Population

Of the 962 subjects, 954 had sufficient phonocardiographic information for inclusion in the final analysis (Table 1 and Table S1). The patient population tended to be elderly, predominantly White, and nearly equally split in sex, consistent with cases of VHD seen mainly at academic medical centers. As expected, both AS and MR cases were older than their respective controls ($P<0.0001$ for each). Male sex was also more prevalent in the MR cases than in the MR controls ($P=0.0055$).

Murmur Detection Performance

We first compared algorithm output on the test subset of 1774 recordings with their annotated ground truth. Of this subset, the algorithm signal quality filter excluded 226 recordings from analysis. These “inadequate signal” recordings also had low annotator signal quality scores (Figure 1), showing that the algorithm does not prevent the analysis of potential murmurs when the recordings are clinically adequate. The remaining 1548 recordings, which constituted 87% of this test subset, received either a “murmur” or a “no murmur” output from the algorithm. Further murmur detection performance analysis of the algorithm was based on these 1548 recordings.

We then directly compared the algorithm’s murmur prediction with annotator defined ground truth (Table 2). Algorithm performance had a sensitivity and specificity for detecting murmurs of 76.3% (95% CI, 72.9%–79.3%) and 91.4% (95% CI, 89.6%–93.1%), respectively, and a positive predictive value of 86.6% (95% CI, 84.0%–89.3%) using the murmur prevalence (42.2%) from this test subset. The positive and negative likelihood ratios were 8.89 (95% CI, 7.35–11.08) and

0.259 (95% CI, 0.225–0.297), respectively. Individual annotators showed modest interrater agreement ($\kappa=0.478$), consistent with prior studies¹ and mirroring what would be expected in clinical practice.

We then evaluated whether certain patient, examination, or device characteristics affected algorithm performance. We looked first at murmur intensity, because in certain clinical contexts, softer murmurs can be less likely to indicate meaningful disease.^{2,30,31} When excluding murmurs of grade 1 intensity (annotator aggregated), sensitivity significantly increased to 90.0% ($P<0.0001$; Table 2).

Next, we evaluated whether algorithm performance differed on the basis of auscultatory position. Overall, these performances were similar, as evidenced by the overlapping CIs (Table 2). Notably, recordings at the pulmonic position were more numerous in this subset because fewer recordings at that position were removed by the algorithm as “inadequate signal.”

Because the recordings were made from 2 devices, the Eko CORE and the Eko DUO, we next evaluated whether algorithm performance differed on the basis of the specific device used. The sensitivity on CORE recordings was slightly higher than on those from DUO ($P<0.05$; Table 2), but was not statistically significant after controlling for signal quality (Table S2), an example of the well-known Simpson paradox.

Allowing the algorithm’s positive-negative decision boundary to vary, we then generated an ROC curve to illustrate the sensitivity and specificity tradeoffs (Figure 2). The US Food and Drug Administration–cleared murmur detection algorithm, however, operates at a single point on this curve (Figure 2, orange circle), with performance described above. Stratification of the ROC curve based on the annotator-aggregated murmur grade shows the improved characteristics of the algorithm with higher-grade murmurs (Figure 2, green line).

Valvular Disease Screening Performance

We then measured algorithm performance as a screening tool for VHD by comparing murmur predictions at the appropriate anatomic locations with a different gold standard: the clinical echocardiogram. First, we considered AS. Of the 954 eligible patients, we grouped 81 with AS labeled moderate to severe or greater as cases and 185 without structural heart disease as controls (Figure 3). As previously mentioned, this severity threshold for disease was chosen to include cases that would typically require further evaluation for possible mechanical intervention. We further removed 8 cases and 13 controls with “inadequate signal” classifications at both the aortic and pulmonic positions, giving a total of 73 cases and 172 controls (with

Table 1. Characteristics of Study Subjects

Characteristics	All Subjects	AS Cases	AS Controls	MR Cases	MR Controls
Total subjects	962	73	172	68	130
Age, mean±SD, y	65±15	73±11	56±15	64±12	55±14
<18, y	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
18–29 y	27 (2.8)	0 (0.0)	11 (6.4)	0 (0.0)	7 (5.4)
30–39 y	45 (4.7)	0 (0.0)	19 (11.0)	3 (4.4)	15 (11.5)
40–49 y	77 (8.0)	2 (2.7)	26 (15.1)	5 (7.4)	23 (17.7)
50–59 y	128 (13.3)	8 (11.0)	32 (18.6)	11 (16.2)	27 (20.8)
60–69 y	237 (24.6)	15 (20.5)	46 (26.7)	26 (38.2)	35 (26.9)
70–79 y	276 (28.7)	26 (35.6)	30 (17.4)	15 (22.1)	21 (16.2)
80–90 y	143 (14.9)	16 (21.9)	8 (4.7)	7 (10.3)	2 (1.5)
>90 y	28 (2.9)	6 (8.2)	0 (0.0)	1 (1.5)	0 (0.0)
Sex					
Women	450 (46.8)	33 (45.2)	93 (54.1)	21 (30.9)	67 (51.5)
Men	512 (53.2)	40 (54.8)	79 (45.9)	47 (69.1)	63 (48.5)
Race/ethnicity					
White	748 (77.8)	54 (74.0)	129 (75.0)	57 (83.8)	99 (76.2)
Black/AA	57 (5.9)	5 (6.8)	10 (5.8)	2 (2.9)	8 (6.2)
Asian	69 (7.2)	5 (6.8)	14 (8.1)	4 (5.9)	11 (8.5)
Hispanic/Latino	57 (5.9)	7 (9.6)	12 (7.0)	4 (5.9)	7 (5.4)
Other/unknown	31 (3.2)	2 (2.7)	7 (4.1)	1 (1.5)	5 (3.8)
Valvular disease					
Prosthesis	111 (11.5)	3 (4.1)	0 (0.0)	5 (7.4)	0 (0.0)
Aortic valve					
Regurgitation					
Mild	156 (16.2)	19 (26.0)	0 (0.0)	15 (22.1)	0 (0.0)
Moderate	48 (5.0)	6 (8.2)	0 (0.0)	3 (4.4)	0 (0.0)
Severe	9 (0.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Stenosis					
Mild	30 (3.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Moderate	30 (3.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Severe	82 (8.5)	73 (100.0)	0 (0.0)	1 (1.5)	0 (0.0)
Mitral valve					
Regurgitation					
Mild	225 (23.4)	19 (26.0)	0 (0.0)	0 (0.0)	0 (0.0)
Moderate	98 (10.2)	10 (13.7)	0 (0.0)	0 (0.0)	0 (0.0)
Severe	85 (8.8)	2 (2.7)	0 (0.0)	68 (100.0)	0 (0.0)
Stenosis					
Mild	21 (2.2)	3 (4.1)	0 (0.0)	0 (0.0)	0 (0.0)
Moderate	10 (1.0)	2 (2.7)	0 (0.0)	2 (2.9)	0 (0.0)
Severe	8 (0.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Pulmonic valve					
Regurgitation					
Mild	141 (14.7)	13 (17.8)	0 (0.0)	18 (26.5)	0 (0.0)
Moderate	13 (1.4)	0 (0.0)	0 (0.0)	1 (1.5)	0 (0.0)
Severe	2 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Stenosis					
Mild	2 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

(Continued)

Table 1. Continued

Characteristics	All Subjects	AS Cases	AS Controls	MR Cases	MR Controls
Moderate	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Severe	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Tricuspid valve					
Regurgitation					
Mild	305 (31.7)	19 (26.0)	0 (0.0)	25 (36.8)	0 (0.0)
Moderate	88 (9.1)	6 (8.2)	0 (0.0)	16 (23.5)	0 (0.0)
Severe	27 (2.8)	1 (1.4)	0 (0.0)	2 (2.9)	0 (0.0)
Stenosis					
Mild	2 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Moderate	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Severe	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

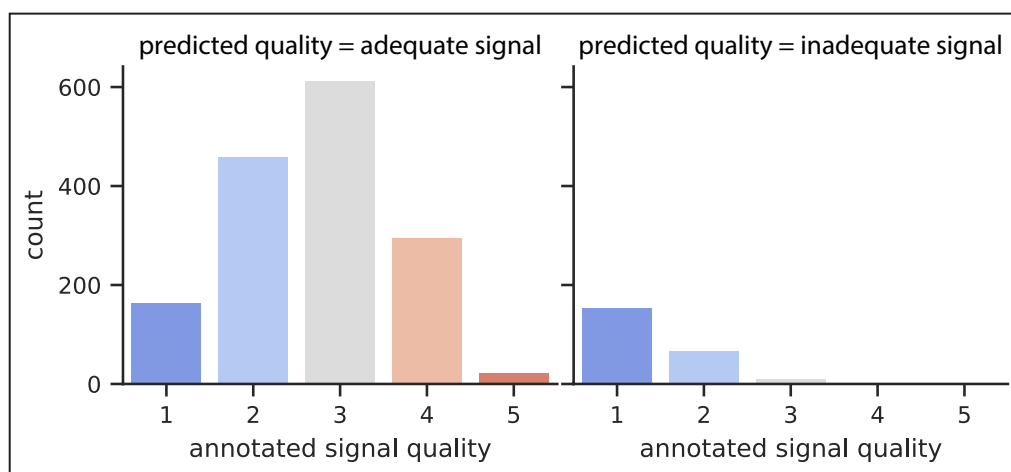
Data are given as number (percentage), unless otherwise noted. For brevity, intermediate grades of valvular disease severity are categorized as the higher grade (ie, moderate to severe included as severe). AA indicates African American; AS, aortic stenosis; and MR, mitral regurgitation.

40 cases and 82 controls annotated). The frequency of nondiagnostic signals was not significantly different between the controls (13 of 185) and the cases (8 of 81; $P=0.341$), suggesting that the signal quality classifier does not preferentially exclude disease. To maximize our sensitivity for detection, aortic position recordings were used for algorithm performance when they had adequate signal (and an annotation for annotator performance), and if not, pulmonic recordings were considered. We defined a positive test for AS as a “murmur” detected from the recording at either the aortic or pulmonic position as described, and a negative test where no murmur was detected.

For the detection of “clinically significant” AS, the algorithm operates with a sensitivity of 93.2% (95% CI, 86.9%–98.5%) and a specificity of 86.0% (95% CI, 80.9%–91.0%) (Table 3). Although the **commercially available algorithm** operates at a fixed point (Figure 4, orange circle), the ROC

curve illustrates the theoretical potential to tune these test characteristics to the appropriate clinical scenario. Overall, the murmur detection algorithm (Figure 4; area under the curve=0.952) compares favorably with the expert clinicians (Figure 4, green, red, and purple circles), whose performances on the annotated subset of 122 subjects fell just under the algorithm’s ROC curve.

We also screened for MR with the same murmur detection algorithm. Using our same overall patient cohort, and the same inclusion and exclusion criteria as for AS, except testing at the mitral location, we had 68 cases and 130 controls (with 29 cases and 62 controls annotated). At the mitral position, there were a greater number of “inadequate signal” recordings, which were statistically more common in controls than in cases ($P=0.0184$). For the detection of “clinically significant” MR, the algorithm operates with a sensitivity of 66.2% (95% CI, 54.7%–77.4%) and a specificity of

**Figure 1. Predicted and annotated signal quality.**

The plot on the right shows that the recordings predicted as “inadequate signal” by the algorithm have low signal quality, as assessed by the cardiologist annotators.

Table 2. Characteristics of Murmur Detection Algorithm

Confusion Matrix					
		Annotation (Ground Truth)		Total	
		Murmur	No Murmur		
Algorithm result					
Murmur detected		499	77	576	
No murmur detected		155	817	972	
Inadequate signal		28	198	226	
Total		682	1092	1774	
Test Characteristics					
	Recordings of Murmur (Total) (Inadequate Signal)*	Sensitivity (95% CI), %	Specificity (95% CI), %	LR Positive (95% CI)	LR Negative (95% CI)
Annotated grade					
All murmurs	499 (654) (28 [†])	76.3 (72.9–79.3)	91.4 (89.6–93.1)	8.86 (7.35–11.08)	0.259 (0.225–0.297)
≥2	406 (451) (17 [†])	90.0 (87.1–92.6)	91.4 (89.6–93.0)	10.5 (8.6–12.9)	0.109 (0.081–0.141)
Position					
Aortic	146 (382) (52 [†])	75.0 (68.5–81.4)	89.4 (84.6–93.4)	7.1 (4.9–11.4)	0.28 (0.21–0.35)
Mitral	126 (388) (86 [†])	71.2 (63.7–78.4)	92.7 (89.4–96.0)	9.7 (6.7–17.7)	0.31 (0.24–0.39)
Pulmonic	189 (450) (23 [†])	81.9 (76.6–86.9)	91.1 (87.3–94.6)	9.2 (6.4–14.7)	0.2 (0.14–0.26)
Tricuspid	113 (328) (65 [†])	75.4 (68.0–82.4)	92.4 (88.7–95.7)	10.0 (6.6–17.8)	0.27 (0.19–0.35)
Device					
CORE	401 (929) (153 [†])	77.6 (73.7–81.4)	91.2 (88.5–93.6)	8.8 (6.7–12.1)	0.25 (0.2–0.29)
DUO	175 (619) (73 [†])	73.2 (65.9–79.7)	91.6 (88.7–94.1)	8.7 (6.4–12.3)	0.29 (0.22–0.37)

Confusion matrix listed at top, with test characteristics stratified by annotated murmur grade, auscultation position, and auscultation device listed below. Under the heading of recordings, “murmur” indicates algorithm-identified murmurs, “total” indicates algorithm-analyzed recordings after removing inadequate signals, and “inadequate signal” indicates recordings labeled as inadequate signal by signal quality classifier. Test characteristics are computed after excluding inadequate signals from analysis. LR indicates likelihood ratio.

*represent inadequate signal recordings.

94.6% (95% CI, 90.4%–98.4%) (Table 3). The algorithm (Figure 5; area under the curve=0.865) compares similarly to the annotators, whose performances on the annotated subset of 91 subjects fall either along or below the algorithm’s ROC curve.

We also explored whether the signal quality classifier would bias our results by preferentially excluding lower-grade murmurs associated with cases of VHD. Importantly, of the few recordings anatomically corresponding to severe VHD but labeled as a grade 1 murmur by any annotator (9 in total: 4 for AS, and 5 for MR), the algorithm identified adequate signal and produced a correct “murmur” output for all. This suggests that the algorithm can still detect the softer murmurs indicative of clinically significant disease.

DISCUSSION

Our results suggest that the algorithm would be a useful decision support tool in detecting murmurs attributed to “clinically significant” VHD. To put this in perspective, in the elderly population, where the prevalence of surgically intervenable AS reaches 5%,³² a negative

test, carrying a negative likelihood ratio of 0.08, will nearly rule out the diagnosis, reducing its probability to <0.5%. Conversely, a positive test, with a positive likelihood ratio of 6.68, will increase disease probability to 26%. Because we compared cases with severe disease with disease-free controls in our study, a positive AS screening result in a separate population could also indicate nonsurgical AS (ie, AS of moderate severity or less). Even with this caveat, the test outcome is likely to influence clinical management in this common clinical scenario. Moreover, because the algorithm results are intended to be combined with a provider’s clinical interpretation, the overall accuracy of a clinical VHD diagnosis may be even higher.³³ In addition, to our knowledge, **our validation set represents the world’s largest adult echocardiogram-paired heart sound recording database.** Looking ahead, this database may facilitate the development of future algorithms to differentiate between innocent and pathologic murmurs, identify specific types of VHD, or correlate other cardiac pathological features to a patient’s auscultatory signature.

Our study has several limitations when applying the results to the clinic. First, we did not evaluate algorithm

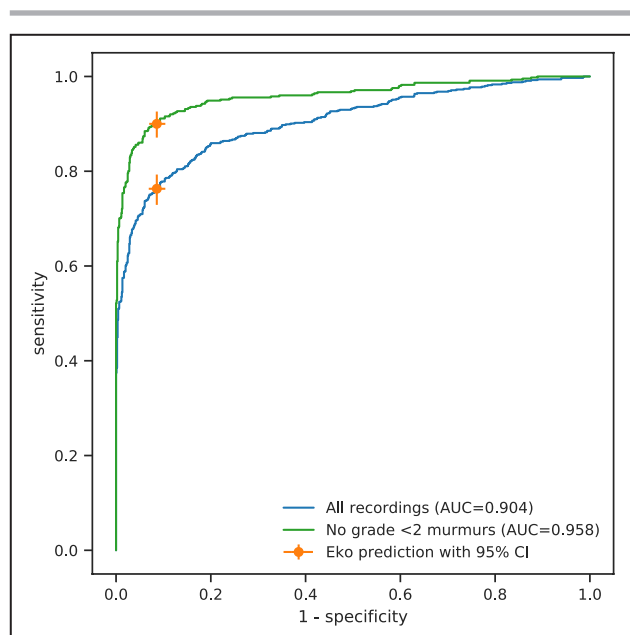


Figure 2. Performance of murmur detection algorithm. Receiver operating characteristic curves for all recordings (blue) and minimal intensity-filtered murmurs (green) are shown. Eko software operates with parameters yielding the orange marker. Error bars indicate 95% CIs. AUC indicates area under the curve.

implementation in direct clinical care. The Eko platform can be integrated into a delivery system's electronic medical record, but 15-second recordings are often longer than typical clinical practice. Thus, the ultimate effect on the length of the clinical encounter, and whether this translates to higher efficiency, lower costs, or improved outcomes all remain unknown. We plan subsequent studies to evaluate the effects of this technology on care delivery, because this was not tested here.

Second, our algorithm purposely does not interpret poor-quality heart sounds. An “inadequate signal” output does not rule out severe VHD. This does, however, mirror clinical practice, where examination findings, and auscultatory findings in particular, are often inconclusive. For any given patient, the test characteristics of the algorithm are best represented by excluding these nonevaluable results, because the provider sees the output of “inadequate signal,” rather than a test outcome. However, this could overestimate the true test characteristics when applied on a population level. Although extreme, applying an “intention-to-diagnose” approach, which groups all nondiagnostic results as incorrect outcomes, can identify the potential limits of such bias.³⁴ When doing so, the test characteristics are unsurprisingly worse, with sensitivity and specificity of 84% and 80%, respectively, for AS, and 57% and 69%, respectively, for MR. This analysis is not truly representative of the test, because the algorithm is extremely unlikely to

misclassify all nondiagnostic results, even if forced to make a decision. Nonetheless, this “intention-to-diagnose” analysis underscores the need to identify the predictors of nondiagnostic auscultation, which we hope to clarify in future studies. Noise cancellation software, for example, may help to reduce the number of these nonevaluable results, although this hypothesis requires further testing.

Third, we effectively performed a case-control study, and therefore the test characteristics we report may be influenced by the spectrum effect.³⁵ The high prevalence of disease in our cohort is likely to enrich for murmurs when compared with a general screening population, where the test characteristics could be different. We also compared a severe form of disease with healthy, normal controls. As a result, the specificity we report is likely higher than in a general population, because a case of mild AS or MR may well have a murmur detected by the algorithm and be labeled as disease. These events would be “false positives” for disease requiring surgical intervention. However, they should not affect the sensitivity, because neither the number of true positives nor false negatives would change. Because we anticipate the algorithm to be used primarily for screening purposes, where sensitivity is paramount, we suspect this particular bias to be of minimal clinical consequence. Ultimately, the results of the algorithm should be placed in the appropriate clinical context. Although a false positive from “mild” VHD, for example, would generally prompt further diagnostic testing, obtaining an echocardiogram to confirm the diagnosis and initiate disease surveillance may not be appropriate in every clinical situation.

Last, our validation set consisted primarily of patients presenting through both tertiary care centers and a community cardiology clinic diagnosed with severe VHD via standard transthoracic echocardiography. Thus, subjects requiring other diagnostics to confirm disease severity, such as dobutamine stress echocardiography for low-flow, low-gradient AS, or 3-dimensional transesophageal echocardiography for MR, were not captured. Similarly, our algorithm was developed on a training set from subjects seen in actual clinical practice. Although these subjects may well represent the US population, they may not be reflective of developing countries, where the prevalence and cause of VHD are different. Because these populations would likely benefit from an accessible and low-cost decision-support tool, like the one tested herein, further investigations are warranted.

Our results also illustrate several physiologic principles within cardiovascular disease. Although our algorithm evaluates for murmurs, auscultatory findings are much richer than this. AS is an excellent example, as the intensity of the A2 component of the second

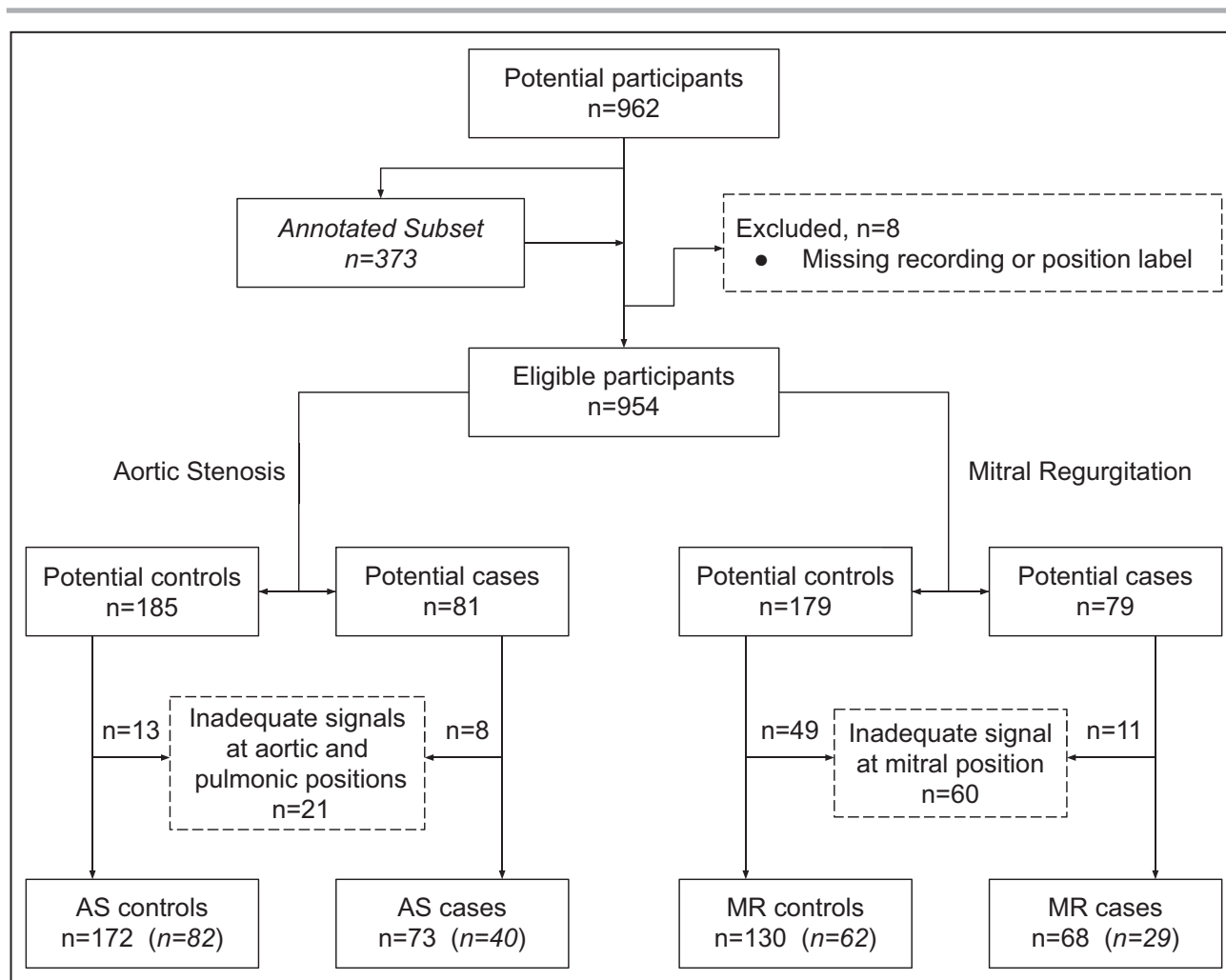


Figure 3. Flow of study participants.

We defined valvular heart disease cases as those graded moderate to severe or worse to encompass all levels of disease that could require timely intervention beyond serial monitoring. We defined controls as subjects free of valvular, structural, or congenital heart disease, with no valvular regurgitation or stenosis *beyond trivial or physiologic severity*. Potential participants included all enrolled subjects (ie, those with recordings). Eligible participants included only those with the appropriate data for analysis. Aortic stenosis (AS) was assessed by a single recording at either the aortic (preferred) or the pulmonic position, and mitral regurgitation (MR) was assessed by a single recording at the mitral position. Actual cases and controls were further filtered from potential cases and controls by removing subjects with “inadequate signal” at the corresponding anatomic locations by the signal quality classifier. Numbers listed *in italics* represent the subset of annotated recordings.

heart sound and the timing of the peak of the systolic murmur are considered indicators of disease severity.¹³ An extended algorithm inclusive of other predictors of VHD, beyond the presence of a murmur, may improve disease screening performance. In addition, both the cardiologists and the algorithm perform better in detecting AS than detecting MR. This may be attributed to AS having a more discernible auscultatory signature. Physiologically, MR produces a load-dependent murmur, with the severity of regurgitation dependent on minute-to-minute hemodynamics. Moreover, MR can be directional, and therefore may not manifest a murmur at a predefined auscultatory location. Additional recording positions

or physiologic maneuvers might also improve disease screening performance.

The algorithm tested herein addresses the need for an effective and accessible method to screen for murmurs and ultimately detect VHD. It is accurate and reliable, with comparable performance to that of an expert cardiologist, at least in the annotated subset of the overall data set we present herein. We anticipate that it would be particularly useful in hurried situations, such as rapid diagnosis in the emergency department or risk stratification for urgent noncardiac surgery, where minimizing the time to diagnostic test results, as well as the strain on providers, is particularly important. In this vein, we purposefully captured heart sounds in a real-world

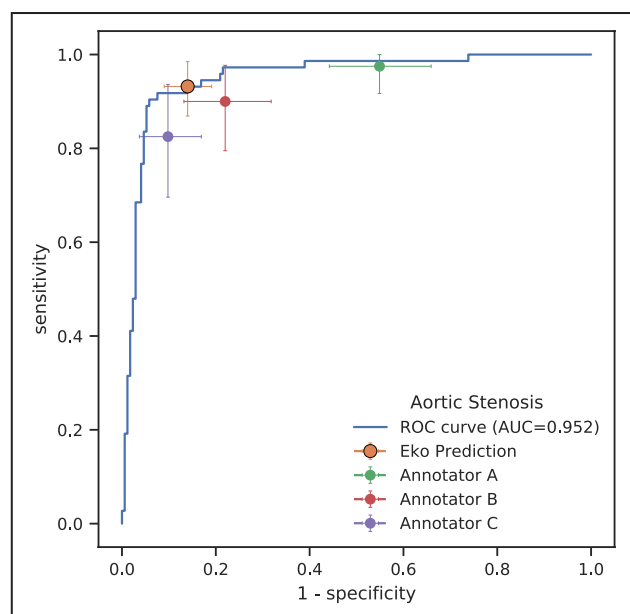
Table 3. Characteristics of Algorithm for VHD Screening

Variables	Subjects (Cases/Controls)	Sensitivity (95% CI), %	Specificity (95% CI), %	LR Positive (95% CI)	LR Negative (95% CI)
Aortic stenosis					
Algorithm	245 (73/172)	93.2 (86.9–98.5)	86.0 (80.9–91.0)	6.68 (4.82–10.37)	0.08 (0.018–0.155)
Annotator A	122 (40/82)	97.5 (91.7–100)	45.1 (34.1–55.8)	1.78 (1.46–2.2)	0.055 (0.0–0.2)
Annotator B	122 (40/82)	90.0 (79.5–97.7)	78.0 (68.2–86.8)	4.1 (2.78–6.84)	0.128 (0.029–0.27)
Annotator C	122 (40/82)	82.5 (69.6–93.6)	90.2 (83.1–96.3)	8.46 (4.77–23.01)	0.194 (0.069–0.338)
Mitral regurgitation					
Algorithm	198 (68/130)	66.2 (54.7–77.4)	94.6 (90.4–98.4)	12.3 (6.7–39.9)	0.357 (0.239–0.479)
Annotator A	91 (29/62)	82.8 (68.0–95.5)	64.5 (52.5–76.1)	2.33 (1.65–3.51)	0.267 (0.076–0.514)
Annotator B	91 (29/62)	69.0 (52.0–85.7)	82.3 (72.7–90.8)	3.89 (2.31–7.85)	0.377 (0.173–0.598)
Annotator C	91 (29/62)	58.6 (40.0–76.7)	87.1 (78.5–95.1)	4.54 (2.42–12.11)	0.475 (0.263–0.691)

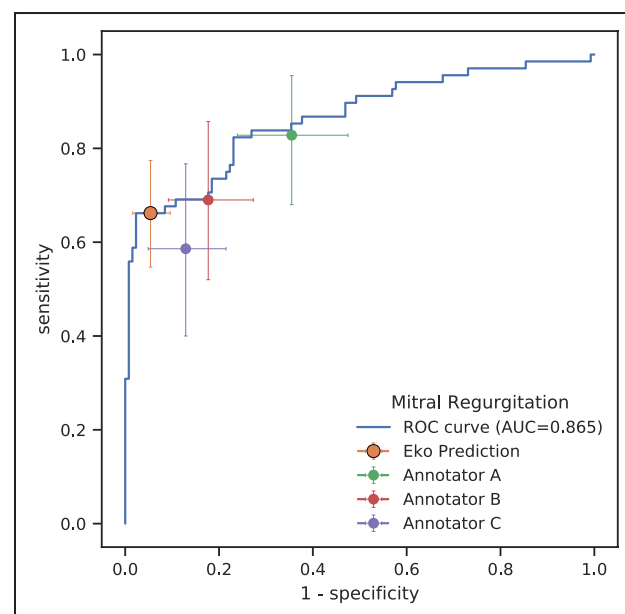
LR indicates likelihood ratio; and VHD, valvular heart disease.

clinical setting, rather than an artificial research environment, to enhance the generalizability of our findings. The considerable variability among highly trained clinicians we observed in our study also represents real-world practice, and the algorithm is well equipped to address this problem. Further potential benefits include enabling clinicians to detect VHD earlier and more consistently, and reducing morbidity and mortality because of earlier clinical intervention.³⁶ Because the algorithm operates at the point of care, requiring only cellular or Wi-Fi connectivity with the digital stethoscope and mobile platform,

it could serve as an affordable alternative to traditional echocardiography, which remains limited by cost, time, and access.³⁷ Although handheld echocardiography can also fill this role, it requires more advanced training than the simple capture of heart sounds with a stethoscope.³⁸ To the extent the algorithm accurately excludes severe VHD, it could render some echocardiograms moot, particularly those ordered to search for VHD that reveal normal hearts. Assuming this indication, and the result, each constitutes 10% of echocardiograms,³⁹ with 5 million studies performed yearly in the United States at

**Figure 4. Performance of aortic stenosis screening by murmur detection algorithm.**

The algorithm receiver operating characteristic (ROC) curve is shown in blue. Eko software operates at the orange marker. The performance of the individual cardiologists on the annotated subset of the overall data set is shown by the green, red, and purple markers. Error bars indicate 95% CIs. AUC indicates area under the curve.

**Figure 5. Performance of mitral regurgitation screening by murmur detection algorithm.**

The algorithm receiver operating characteristic (ROC) curve is shown in blue. Eko software operates at the orange marker. The performance of the individual cardiologists on the annotated subset of the overall data set is shown by the green, red, and purple markers. Error bars indicate 95% CIs. AUC indicates area under the curve.

a cost of \$1000 each,⁴⁰ this could translate to an annual cost savings of \$28 million nationally, even when applying the lower specificities from the intention-to-diagnose analyses for AS and MR. Moreover, any potential savings would be specific to this technology, as these echocardiograms are appropriate without another reliable way to exclude the pathological feature in question. In light of the recent and ongoing COVID-19 pandemic,⁴¹ this technology could also provide expert-level diagnostics through telemedicine, thereby limiting the transmission of a highly contagious disease. Furthermore, the digital stethoscope platform used herein could be extended to other auscultation findings, such as lung sounds. Overall, our study shows the promise of this tool as an adjunct to clinical care and illustrates the potential of it expanding into something even greater.

ARTICLE INFORMATION

Received October 26, 2020; accepted February 24, 2021.

Affiliations

From the Division of Cardiology, University of California San Francisco, San Francisco, CA (J.S.C., V.S.M., S.G.F., G.W.S., S.A.B., D.J.); Division of Cardiology, Zuckerberg San Francisco General Hospital, San Francisco, CA (J.S.C.); Eko, Oakland, CA (A.M.S., L.L., J.M., J.P., S.P., M.M.K., D.N.B., C.C., C.B., S.V.); Division of Cardiology, Bluhm Cardiovascular Institute, Northwestern University, Chicago, IL (B.E.W., A.H., J.P., J.G., D.E., P.M.M., J.D.T.); Los Alamitos Cardiovascular Medical Group, Los Alamitos, CA (R.R., R.K., S.T.F.); and Echocardiography Laboratory, Mount Sinai Heart Institute, Mount Sinai Medical Center, Miami Beach, FL (Z.H.A., J.J.L., M.G., C.G.M.).

Acknowledgments

We thank Stefanie Miller and Julie Petersen for help with subject recruitment. Author contributions: Conception and design: Drs Chorba and Maidens. Data acquisition: Dr Kanzawa, Dr Barbosa, C. Currie, C. Brooks, Dr White, A. Huskin, J. Paek, J. Geocaris, D. Elnathan, R. Ronquillo, R. Kim, Dr Alam, Dr Mahadevan, S.G. Fuller, G.W. Stalker, S.A. Bravo, D. Jean, Dr Lee, Dr Gjergjindreaj, Dr Mihos, and Dr Forman. Data analysis: Dr Shapiro. Data interpretation: Drs Chorba, Shapiro, Le, Maidens, Pham, White, Venkatraman, McCarthy, and Thomas. Software creation: Drs Shapiro, Maidens, Barbosa, Prince, and Venkatraman. Manuscript generation: Drs Shapiro, Le, and Chorba. Manuscript revision: Drs Chorba, Shapiro, Le, Maidens, Pham, Venkatraman, and Thomas.

Sources of Funding

This project was supported by National Institutes of Health R43HL144297 (to Drs Maidens and Chorba) and K08HL124068 (to Dr Chorba). Dr Thomas was supported in part by a grant from the Irene D. Pritzker Foundation. Eko sponsored and partially funded this study.

Disclosures

Dr Shapiro, Dr Le, Dr Maidens, Dr Prince, Dr Pham, Dr Kanzawa, D.N. Barbosa, C. Currie, C. Brooks, and Dr Venkatraman are employees of Eko. Dr McCarthy receives equity as a member of Eko's scientific advisory board. Dr Chorba is an unpaid advisor to Eko. The remaining authors have no disclosures to report.

Supplementary Material

Tables S1–S2
Figure S1

REFERENCES

1. Etchells E, Bell C, Robb K. Does this patient have an abnormal systolic murmur? *JAMA*. 1997;277:564–571. DOI: 10.1001/jama.277.7.564.
2. Dobrow RJ, Calatayud JB, Abraham S, Caceres CA. A study of physician variation in heart-sound interpretation. *Med Ann Dist Columbia*. 1964;33:305–308.
3. Mangione S, Nieman LZ. Cardiac auscultatory skills of internal medicine and family practice trainees: a comparison of diagnostic proficiency. *JAMA*. 1997;278:717–722. DOI: 10.1001/jama.1997.03550090041030.
4. D'Arcy JL, Prendergast BD, Chambers JB, Ray SG, Bridgewater B. Valvular heart disease: the next cardiac epidemic. *Heart*. 2011;97:91–93. DOI: 10.1136/hrt.2010.205096.
5. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. *Lancet*. 2006;368:1005–1011. DOI: 10.1016/S0140-6736(06)69208-8.
6. Osnabrugge RLJ, Mylotte D, Head SJ, Van Mieghem NM, Nkomo VT, LeReun CM, Bogers AJJC, Piazza N, Kappetein AP. Aortic stenosis in the elderly: disease prevalence and number of candidates for transcatheter aortic valve replacement: a meta-analysis and modeling study. *J Am Coll Cardiol*. 2013;62:1002–1012. DOI: 10.1016/j.jacc.2013.05.015.
7. Sliwa K, Carrington M, Mayosi BM, Zigiadias E, Mvungi R, Stewart S. Incidence and characteristics of newly diagnosed rheumatic heart disease in urban African adults: insights from the Heart of Soweto Study. *Eur Heart J*. 2010;31:719–727. DOI: 10.1093/eurheartj/ehp530.
8. Karthikeyan G, Mayosi BM. Is primary prevention of rheumatic fever the missing link in the control of rheumatic heart disease in Africa? *Circulation*. 2009;120:709–713. DOI: 10.1161/CIRCULATIONAHA.108.836510.
9. Coffey S, Cox B, Williams MJA. Lack of progress in valvular heart disease in the pre-transcatheter aortic valve replacement era: increasing deaths and minimal change in mortality rate over the past three decades. *Am Heart J*. 2014;167:562–567. DOI: 10.1016/j.ahj.2013.12.030.
10. Bhattacharyya S, Hayward C, Pepper J, Senior R. Risk stratification in asymptomatic severe aortic stenosis: a critical appraisal. *Eur Heart J*. 2012;33:2377–2387. DOI: 10.1093/eurheartj/ehs190.
11. Nishimura RA, Otto CM, Bonow RO, Carabello BA, Erwin JP 3rd, Guyton RA, O'Gara PT, Ruiz CE, Skubas NJ, Sorajja P, et al. AHA/ACC guideline for the management of patients with valvular heart disease. *Circulation*. 2014;2014:e521–e643. DOI: 10.1161/CIR.0000000000000503.
12. Virnig BA, Shippee ND, O'Donnell B, Zeglin J, Parashuram S. Trends in the use of echocardiography, 2007 to 2011: data points #20. Data Points Publication Series. 2011. Rockville, MD: Agency for Healthcare Research and Quality (US).
13. Mann DL, Zipes DP, Libby P, Bonow RO, Braunwald E. *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Philadelphia PA: Saunders; 2015. DOI: 10.1001/jama.294.3.376-a.
14. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Panykh OS, Geis JR, Pandharipande PV, Brink JA, Dreyer KJ. Current applications and future impact of machine learning in radiology. *Radiology*. 2018;288:318–328.
15. Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digit Med*. 2018;1:59.
16. VanRullen R, Reddy L. Reconstructing faces from fMRI patterns using deep generative neural networks. *Commun Biol*. 2019;2:193. DOI: 10.1038/s42003-019-0438-y.
17. Chamberlain D, Kodgule R, Ganelin D, Miglani V, Fletcher RR. Application of semi-supervised deep learning to lung sound analysis. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando, FL; 2016:804–807. DOI: 10.1109/EMBC.2016.7590823.
18. Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, Mark RG. Classification of normal/abnormal heart sound recordings: the Physionet/computing in cardiology challenge 2016. *Comput Cardiol*. 2016;43. DOI: 10.22489/cinc.2016.179-154.
19. Thompson WR, Reinisch AJ, Unterberger MJ, Schrieffer AJ. Artificial intelligence-assisted auscultation of heart murmurs: validation by virtual clinical trial. *Pediatr Cardiol*. 2019;40:623–629. DOI: 10.1007/s00246-018-2036-z.
20. Clifford GD, Liu C, Moody B, Millet J, Schmidt S, Li Q, Silva I, Mark RG. Recent advances in heart sound analysis. *Physiol Meas*. 2017;38:E10–E25. DOI: 10.1088/1361-6579/aa7ec8.
21. Dwivedi AK, Imtiaz SA, Rodriguez-Villegas E. Algorithms for automatic analysis and classification of heart sounds—a systematic review. *IEEE Access*. 2019;7:8316–8345. DOI: 10.1109/ACCESS.2018.2889437.

22. Potes C, Parvaneh S, Rahman A, Conroy B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Comput Cardiol*. 2016;43. DOI: 10.22489/cinc.2016.182-399.
23. 510(k) Premarket Notification.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV; 2016:770–778. DOI:10.1109/CVPR.2016.90
25. Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015. arxiv:1412.6980
26. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44:763–770. DOI: 10.1016/0895-4356(91)90128-V.
27. Levine SA. The systolic murmur. *JAMA*. 1933;101:436–438. DOI: 10.1001/jama.1933.02740310020005.
28. 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. *Eur Heart J Cardiovasc Imaging*. 2017;18:254–275. DOI: 10.1093/ehjci/jew335.
29. Zoghbi WA, Adams D, Bonow RO, Enriquez-Sarano M, Foster E, Grayburn PA, Hahn RT, Han Y, Hung J, Lang RM, et al. Recommendations for noninvasive evaluation of native valvular regurgitation. *J Am Soc Echocardiogr*. 2017;30:303–371. DOI: 10.1016/j.echo.2017.01.007.
30. Bonow RO, Carabello BA, Chatterjee K, de Leon AC, Jr FDP, Freed MD, Gaasch WH, Lytle BW, Nishimura RA, O’Gara PT, et al. ACC/AHA 2006 guidelines for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2006;2006:e84–e231. DOI: 10.1161/CIRCULATIONAHA.106.176857.
31. Attenhofer Jost CH, Turina J, Mayer K, Seifert B, Amann FW, Buechi M, Facchini M, Brunner-La Rocca HP, Jenni R. Echocardiography in the evaluation of systolic murmurs of unknown cause. *Am J Med*. 2000;108:614–620. DOI: 10.1016/s0002-9343(00)00361-2.
32. Eveborn GW, Schirmer H, Heggelund G, Lunde P, Rasmussen K. The evolving epidemiology of valvular aortic stenosis: the Tromsø study. *Heart*. 2013;99:396–400. DOI: 10.1136/heartjnl-2012-302265.
33. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, Lee EH, Kim EK. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health*. 2020;2:e138–e148. DOI: 10.1016/S2589-7500(20)30003-0.
34. Schuetz GM, Schlattmann P, Dewey M. Use of 3×2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. *BMJ*. 2012;345:e6717. DOI: 10.1136/bmj.e6717.
35. Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol*. 2009;62:5–12. DOI: 10.1016/j.jclinepi.2008.04.007.
36. Vukanovic-Criley JM, Criley S, Warde CM, Boker JR, Guevara-Matheus L, Churchill WH, Nelson WP, Criley JM. Competency in cardiac examination skills in medical students, trainees, physicians, and faculty. *Arch Intern Med*. 2006;166:610–616. DOI: 10.1001/archinte.166.6.610.
37. Guilford-Blake R. Tackling the rural access crisis: cardiologists will need an array of tools to meet patients’ needs. *Cardiovasc Bus*. 2018;12:28–32. <https://www.cardiovascularbusiness.com/topics/healthcare-economics/tackling-rural-access-crisis-cardiologists-will-need-array-tools-meet>. Accessed April 26, 2020.
38. Chamsi-Pasha MA, Sengupta PP, Zoghbi WA. Handheld echocardiography: current state and future perspectives. *Circulation*. 2017;136:2178–2188. DOI: 10.1161/CIRCULATIONAHA.117.026622.
39. Ward RP, Mansour IN, Lemieux N, Gera N, Mehta R, Lang RM. Prospective evaluation of the clinical application of the American College of Cardiology Foundation/American Society of Echocardiography appropriateness criteria for transthoracic echocardiography. *JACC Cardiovasc Imaging*. 2008;1:663–671. DOI: 10.1016/j.jcmg.2008.07.004.
40. Papolos A, Narula J, Bavishi C, Chaudhry FA, Sengupta PP. U.S. hospital use of echocardiography: insights from the nationwide inpatient sample. *J Am Coll Cardiol*. 2016;67:502–511. DOI: 10.1016/j.jacc.2015.10.090.
41. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, et al; China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727–733. DOI: 10.1056/NEJMoa2001017.

SUPPLEMENTAL MATERIAL

Table S1. Expanded Characteristics of Study Subjects. Percentage of total in parentheses, unless otherwise noted. For brevity, intermediate grades of valvular disease severity are categorized as the higher grade (*i.e.* moderate-to-severe included as severe). Actual cases and controls indicate the subset of potential cases and controls after filtering for ‘inadequate signal’ classifications at the appropriate anatomic location. Annotated cases and controls indicate the subset of actual cases and controls that underwent annotation by expert cardiologists. AS = aortic stenosis. MR = mitral regurgitation. SD = standard deviation. AA = African-American.

[illegible]

Prosthesis	111 (11.5%)	3 (3.7%)	3 (4.1%)	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (7.6%)	5 (7.4%)	4 (13.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Aortic Valve													
Regurgitation													
Mild	156 (16.2%)	22 (27.2%)	19 (26.0%)	14 (35.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	17 (21.5%)	15 (22.1%)	5 (17.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	48 (5.0%)	8 (9.9%)	6 (8.2%)	4 (10.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (5.1%)	3 (4.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	9 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Stenosis													
Mild	30 (3.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	30 (3.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	82 (8.5%)	81 (100.0%)	73 (100.0%)	40 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (1.3%)	1 (1.5%)	1 (3.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Mitral Valve													
Regurgitation													
Mild	225 (23.4%)	21 (25.9%)	19 (26.0%)	10 (25.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	98 (10.2%)	13 (16.0%)	10 (13.7%)	7 (17.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	85 (8.8%)	2 (2.5%)	2 (2.7%)	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	79 (100.0%)	68 (100.0%)	29 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Stenosis													
Mild	21 (2.2%)	4 (4.9%)	3 (4.1%)	1 (2.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (1.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	10 (1.0%)	3 (3.7%)	2 (2.7%)	2 (5.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (2.5%)	2 (2.9%)	1 (3.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	8 (0.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Pulmonic Valve													
Regurgitation													
Mild	141 (14.7%)	13 (16.0%)	13 (17.8%)	4 (10.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	19 (24.1%)	18 (26.5%)	6 (20.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	13 (1.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (2.5%)	1 (1.5%)	1 (3.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	2 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Stenosis													
Mild	2 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Severe	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Tricuspid Valve													
Regurgitation													
Mild	305 (31.7%)	21 (25.9%)	19 (26.0%)	10 (25.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	27 (34.2%)	25 (36.8%)	9 (31.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate	88 (9.1%)	8 (9.9%)	6 (8.2%)	3 (7.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	19 (24.1%)	16 (23.5%)	10 (34.5%)	0 (0.0%)	0 (0.0%)	

[illegible]

Table S2. Murmur Detection Algorithm Performance by Signal Quality and Recording Device. CORE and DUO sensitivities are only significantly different at $p < 0.05$ for recordings with signal quality equal to 2. CI = confidence interval.

Signal Quality	Device	Recordings	Sensitivity (95% CI)
1	CORE	35	50.0 (0.0–100.0)
	DUO	128	66.7 (0.0–100.0)
2	CORE	231	63.3 (48.5–76.9)
	DUO	227	52.6 (35.7–68.6)
3	CORE	408	70.9 (65.5–77.1)
	DUO	203	76.1 (68.4–84.4)
4	CORE	238	89.5 (84.4–93.8)
	DUO	56	82.9 (69.4–94.1)
5	CORE	17	100.0 (100.0–100.0)
	DUO	5	100.0 (100.0–100.0)

Figure S1. Schematic of Deep Learning Framework. See text (Methods) for details. ReLU = rectified linear unit. Batch Norm = batch normalization. Conv = convolutional layer.

