The International Journal of Cardiovascular Imaging
https://doi.org/10.1007/s10554-020-01981-8

ORIGINAL PAPER

# Automated estimation of echocardiogram image quality in hospitalized patients

**Christina Luong[1] · Zhibin Liao[2] · Amir Abdi[2] · Hany Girgis[1] · Robert Rohling[2] · Kenneth Gin[1] · John Jue[1] · Darwin Yeung[1] · Elena Szefer[3] · Darby Thompson[3] · Michael Yin-Cheung Tsang[1] · Pui Kee Lee[1] · Parvathy Nair[1] · Purang Abolmaesumi[2] · Teresa S. M. Tsang[1,4]**

Received: 5 June 2020 / Accepted: 19 August 2020


## Abstract

We developed a machine learning model for efficient analysis of echocardiographic image quality in hospitalized patients. This study applied a machine learning model for automated transthoracic echo (TTE) image quality scoring in three inpatient groups. Our objectives were: (1) Assess the feasibility of a machine learning model for echo image quality analysis, (2) Establish the comprehensiveness of real-world TTE reporting by clinical group, and (3) Determine the relationship between machine learning image quality and comprehensiveness of TTE reporting. A machine learning model was developed and applied to TTEs from three matched cohorts for image quality of nine standard views. Case TTEs were comprehensive studies in mechanically ventilated patients between 01/01/2010 and 12/31/2015. For each case TTE, there were two matched spontaneously breathing controls (Control 1: Inpatients scanned in the lab and Control 2: Portable studies). We report the overall mean maximum and view specific quality scores for each TTE. The comprehensiveness of an echo report was calculated as the documented proportion of 12 standard parameters. An inverse probability weighted regression model was fit to determine the relationship between machine learning quality score and the completeness of a TTE report. 175 mechanically ventilated TTEs were included with 350 non-intubated samples (175 Control 1: Lab and 175 Control 2: Portable). In total, the machine learning model analyzed 14,086 echo video clips for quality. The overall accuracy of the model with regard to the expert ground truth for the view classification was 87.0%. The overall mean maximum quality score was lower for mechanically ventilated TTEs (0.55 [95% CI 0.54, 0.56]) versus 0.61 (95% CI 0.59, 0.62) for Control 1: Lab and 0.64 (95% CI 0.63, 0.66) for Control 2: Portable; $p = 0.002$. Furthermore, mechanically ventilated TTE reports were the least comprehensive, with fewer reported parameters. The regression model demonstrated the correlation of echo image quality and completeness of TTE reporting regardless of the clinical group. Mechanically ventilated TTEs were of inferior quality and clinical utility compared to spontaneously breathing controls and machine learning derived image quality correlates with completeness of TTE reporting regardless of the clinical group.

**Keywords** Echocardiography · Artificial intelligence · Machine learning

## Introduction

Echocardiography is an accessible and effective modality for evaluation of cardiac structure and function, particularly at the point of care. Point of care cardiac ultrasound may be of greatest utility in critically ill patients. However, one would expect compromised image quality and diagnostic yield due to patient characteristics, respiratory status, and/or limitations in patient positioning. Little is known about image quality in point of care ultrasound and formal comprehensive transthoracic echo (TTE) remains the standard for cardiac imaging in this patient population. Despite TTEs

Christina Luong, Zhibin Liao and Amir Abdi are Joint first authors.

Purang Abolmaesumi and Teresa S. M. Tsang are Joint senior authors.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10554-020-01981-8) contains supplementary material, which is available to authorized users.

✉ Teresa S. M. Tsang
t.tsang@ubc.ca

Extended author information available on the last page of the article

Published online: 19 November 2020

being performed by expert scanners on full functionality platforms, the anecdotal yield of TTE in critically ill patients is thought to be poor. The quantification of image quality and diagnostic utility in this population would be an important reference point for point of care cardiac ultrasound. We sought to determine the feasibility of automated TTE quality assessment in three inpatient cohorts to estimate overall and view specific image quality and its relationship to the comprehensiveness of echo reporting.

Cardiac ultrasound has been a challenging modality for application of machine learning due to the high volume of heterogeneous data which includes numerous acquisition angles, video clips of variable duration, and Doppler interrogation. Our group previously demonstrated the accuracy of a machine learning model for automated computation of a quality score using recurrent neural networks [1, 2]. Our work is part of a growing body of research that demonstrates the capabilities of machine learning models for classification of complex cardiovascular data [3–11]. These platforms have the potential for application in clinical care and may augment future performance and reporting of cardiovascular imaging.

There were three primary objectives of this study: (1) Assess the feasibility of a machine learning model for echo quality analysis, (2) Establish the comprehensiveness of real-world TTE reporting by clinical group, and (3) Determine the relationship between machine learning image quality and comprehensiveness of TTE reporting. We hypothesize that TTEs in mechanically ventilated patients will attain lower machine learning quality scores as compared to controls and that there will be an association between image quality and completeness of reporting.

## Methods

### Clinical cohort and transthoracic echo case selection

Institutional review board approval was obtained for all aspects of this study. The TTEs of interest were all comprehensive studies completed on mechanically ventilated patients using the local high-end platform between 01/01/2010 and 12/31/2015 at a single tertiary care center. Studies limited by physical barriers (bandages) and scans by students or nonsonographers were excluded. We also excluded studies with undocumented body surface area (BSA), sex, or scanning sonographer (cohort summarized in Fig. 1). For each mechanically ventilated case TTE there were two TTEs done in spontaneously breathing controls that were matched for BSA (within 0.3m$^2$), sex, scanning sonographer, and date of study (within 2 years). The matched control groups were Control 1, TTEs on inpatients scanned in the lab ("Control 1: Lab") and Control 2, portable TTEs on spontaneously breathing patients who were not transferred to the lab ("Control 2: Portable"). The investigator involved for inclusion of matched controls was blinded to the outcome of proportion of parameters reported. The retrieval of images and assessment of echo quality was

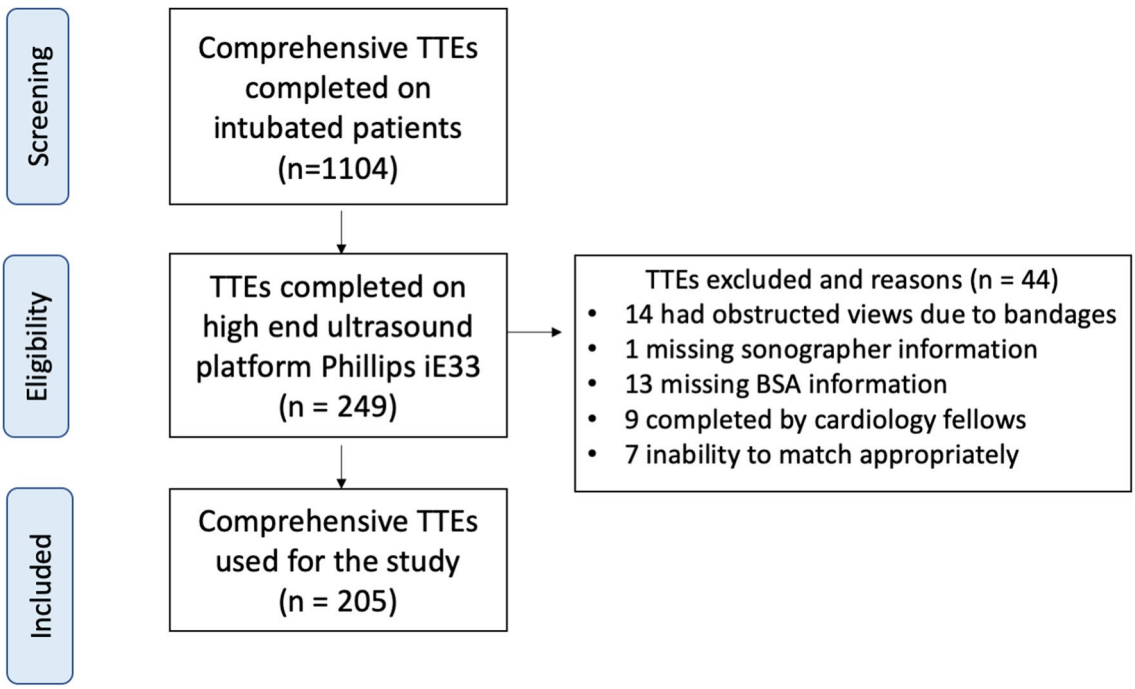


**Fig. 1** Summary of cohort selection. *BSA* body surface area, *TTE* transthoracic echocardiogram

conducted independently using the machine learning model without knowledge of the clinical report. The machine learning model was not trained for analysis of color or spectral Doppler; therefore, Doppler containing images from each echocardiogram were excluded from the study.

## Echocardiographic examination

All echocardiograms for this study were performed on the Philips iE33 platform using a commercially available phased array transducer (Philips S51 frequency 5–1 MHz). TTE studies were acquired and interpreted in accordance with the American Society of Echocardiography standards [12, 13]. All images were digitally recorded and analyzed using XCelera Version R3.1L1 3.1.1.422-2009. Standard parameters sought for all clinical TTEs included left and right ventricular dimension, left atrial volume, left ventricular (LV) ejection fraction, right ventricular (RV) systolic function (qualitative grade), diastolic function, and Doppler interrogation for hemodynamic assessments. These standard measurements were documented in the echocardiography database as part of the clinical report.

## Framework of model and predictive analysis

The machine learning model used was developed for automated view identification and image quality quantification using 16,772 randomly selected 2D TTE videos from 3157 patient studies stored on the regional Picture Archiving and Communications System. The TTEs used to train and validate the model were acquired using Philips iE33 and General Electric Vivid i ultrasound platforms. The majority of studies were completed by certified sonographers with a small proportion generated by cardiology residents and sonography students. During model development, image quality was manually graded for each clip by a level III echocardiographer with the following scale:

(1) 1.0 point if > 75% of the expected blood-tissue interface was clearly defined (view specific)
(2) 0.75 points if 51–75% of the expected blood-tissue interface was clearly defined (view specific)
(3) 0.5 points if 26–50% of the expected blood-tissue interface was clearly defined (view specific)
(4) 0.25 points if < 25% of the expected blood-tissue interface was clearly defined (view specific)

The criteria used was similar to other published methods [14–18]. The model architecture is summarized in Fig. 2.

A view classifier and regression model were unified inside a single deep learning model which was trained on the TTE videos. The model used for this study has been previously published [19] and simultaneously detected the view and estimated the quality of a given video. The deep learning framework was a composition of several functional modules. The first module was a DenseNet Convolutional Neural Network module [20, 21], which extracted image features from individual frames in the input video. The second module was a Long Short Term Memory (LSTM) module [22] which received the set of image features produced by the DenseNet

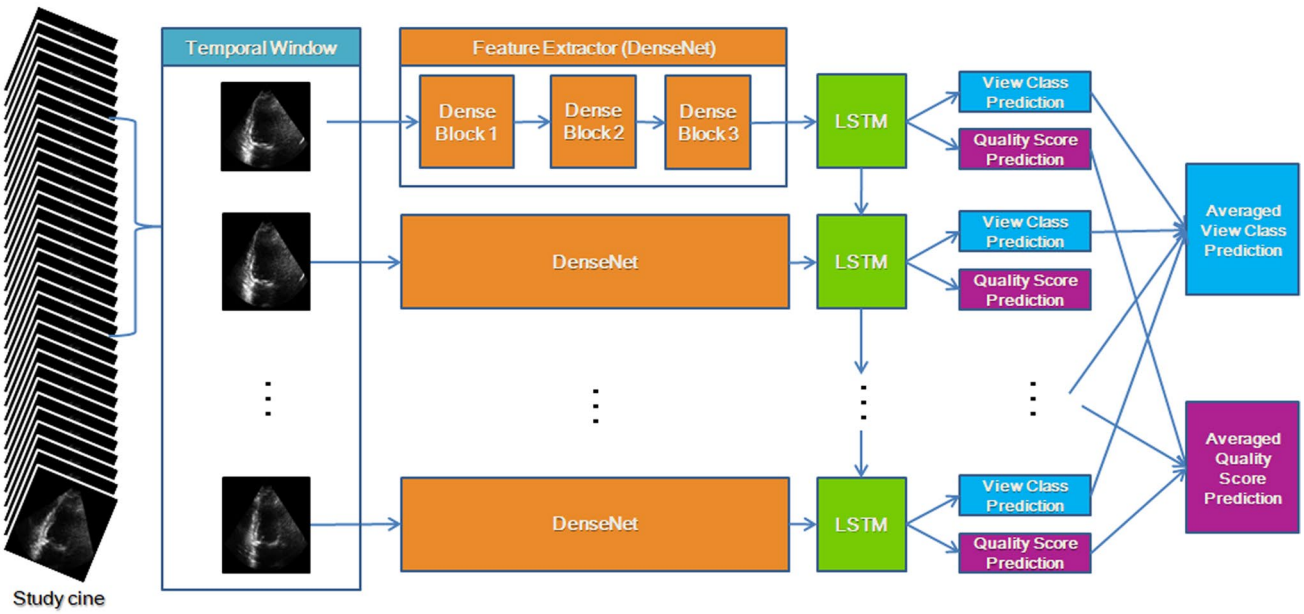


**Fig. 2** Architecture of the deep learning framework. Architecture of the deep learning model for simultaneous echo view plane classification and quality estimation. *LSTM* long short term memory model

module in a forwarding temporal order, generating a corresponding set of temporal dependent features. This new set of features was used by both the classification and regression modules to estimate respective view and image quality for every frame. Finally, the per-frame quantities were averaged by the number of frames in the video to produce an aggregated prediction for the entire video. This framework was optimized by minimizing a combination of a cross-entropy loss (for the view classification task) and a mean absolute difference error loss (for the quality estimation task), via a stochastic gradient descent-based optimization method. Additional details regarding the design of the deep learning model are included in the Supplemental material.

## Application of machine learning for echo quality analysis

Analysis of TTE data in the mechanically ventilated and control cohorts required extraction of 2D videos with semi-automated framing of each component image for machine learning analysis using a cropping program we developed for beam detection. Pixel intensities were rescaled from zero to one without additional intensity normalization. The nine views for analysis were: Parasternal long axis (PLAX), apical 2-chamber (AP2), apical 4-chamber (AP4), apical 3-chamber (AP3), parasternal short axis at the aortic valve level (PSAX-A), parasternal short axis at the mitral valve level (PSAX-M), parasternal short axis at the papillary muscle level (PSAX-PM), subcostal 4-chamber (SC4), and subcostal view of the inferior vena cava (IVC).

These clinical studies often included multiple videos for each view, therefore a maximum quality score was reported for each of the nine views. The maximum quality score for each view was analyzed as the best quality clip that would be used clinically. An overall mean maximum quality score was calculated for each TTE (combined maximum score for the nine views divided by nine). When a TTE did not have any clips for a particular view, this was considered a quality score of zero.

## Evaluation of diagnostic yield of inpatient transthoracic echo

The completeness of a TTE report, referred to as diagnostic yield for this study, was based on the proportion of standard parameters that were documented. The 12 parameters selected for this analysis reflect a combination of major and minor targets. Major targets pertain to core indications for cardiac point of care ultrasound [23] and included: pericardial effusion, LV systolic function (subjective grading), RV systolic function (subjective grading), and RV dimension. Minor targets were aortic sinus and ascending aorta dimension, left atrial volume, aortic valve structure, aortic valve regurgitation with color Doppler, mitral valve structure, mitral valve regurgitation with color Doppler, and estimated right ventricular systolic pressure using spectral Doppler of tricuspid valve regurgitation. These 12 parameters were measured or graded for all comprehensive TTEs unless the sonographer and interpreting physician were unable to make an appropriate measurement, in which case the field was left blank.

## Statistical analysis

### Demographic information

Demographic information was expressed as mean ± SD with inverse probability weighted statistical analysis. Between group differences for age and gender were assessed with a weighted rank test [24] and a weighted chi-square test, respectively. Between group differences for BSA was evaluated with a weighted F-test [25]. Categorical variables for TTE diagnostic yield were expressed as percentages and analyzed with a likelihood ratio test to separate conditional logistic regression models. The analyses were performed using R statistical software version 3.3.2. A p-value of $< 0.05$ was considered statistically significant.

### Quality scores

For each view, a linear regression model was fit with quality score as the response and patient group and the clinical characteristics BSA, sonographer, sex, age and interpreting physician as predictors. Inverse probability weighting was used to account for oversampling of ventilated patients in the matched analysis cohort, where it was assumed that lab patients constitute 45%, portable patients 45% and ventilated patients 10% of the true patient population. The allocated proportions were based on a convenience audit of studies over a 1-week period. Estimates and 95% confidence intervals for the mean maximum quality scores are presented.

The primary hypothesis is that image quality is worse in mechanically ventilated patients compared to the control groups, so t-tests were conducted to determine whether there were significant differences in the quality scores between (a) Control 1: Lab and Control 2: Portable groups, and (b) the mechanically ventilated group and the combined control groups (1: Lab and 2: Portable). The control groups were combined for testing to minimize the number of statistical hypothesis tests conducted. Population estimates were obtained by calculating the mean quality scores for each patient group at the sample means levels for all other predictors. The Holm adjustment was used to account for multiple comparisons.

### Diagnostic yield

For each of the three groups, a logistic regression model was fit with the presence or absence of all 12 diagnostic parameters as the outcome. Group (Control 1: Lab, Control 2: Portable, mechanically ventilated) and the factors used to match the TTEs (by sample selection: BSA, sonographer, sex; by statistical adjustment: age and interpreting physician) were considered predictors. Inverse probability weighting was used to account for the oversampling of mechanically ventilated patients in the matched analysis cohort. Following the hypothesis in the image quality analysis that diagnostic yield will be lower in the mechanically ventilated patients compared to the control groups, t-tests were conducted to determine whether there were significant differences in diagnostic yield between (a) Control 1: Lab and Control 2: Portable groups, and (b) the mechanically ventilated group and the combined control groups (1: Lab and 2: Portable). Population estimates were obtained by calculating the estimated proportion of diagnostic yield for each patient group at the sample means levels for all other predictors. The Holm adjustment was used to adjust for multiple comparisons.

#### Machine learning image quality and diagnostic yield

To assess the relationship between TTE quality score and diagnostic yield, an inverse probability weighted regression model was fit with diagnostic yield as the outcome and machine learning quality score as the predictor. Smoothed locally estimated scatterplot smoothing (LOESS) curves were fit for each group to compare the mean maximum machine learning quality score and diagnostic yield. A post-hoc cut-point of 0.5 for the mean maximum machine learning scores was identified from the curves. An inverse probability weighted logistic regression model was then fit to estimate the diagnostic yield when the quality score was below or above 0.5. The logistic regression model included group, machine learning quality score category, the interaction between group and machine learning quality score category, and clinical characteristics as covariates.

## Results

There were 205 mechanically ventilated TTEs between 01/01/2010 and 12/31/2015 that met inclusion criteria for analysis of diagnostic yield. Each case TTE was matched with a Control 1: Lab and Control 2: Portable TTE, totalling 615 TTEs for analysis of diagnostic yield. Patient selection was matched for BSA, sonographer, sex, and timing of scan. TTEs on males comprised 133 of the 205 studies (64.9%) in each of the three groups. The mean BSA of each group were similar by design: males $2.00 \pm 0.22 m^2$, females $1.76 \pm 0.20 m^2$. The mechanically ventilated cohort were significantly younger compared to the control groups, mean age $58.4 \pm 17.6$ years versus $67.1 \pm 16.9$ years (Control 1: Lab) and $63.2 \pm 12.6$ years (Control 2: Portable); $p < 0.00001$.

For the image quality evaluation, a proportion of subjects were excluded as their imaging data was not available for quality analysis by the machine learning model. After excluding samples with missing video data and their accompanying matches, there were 525 remaining matched TTEs for analysis (175 per group).

### Machine learning model performance

Based on evaluation of the machine learning model during development, the overall accuracy of the model with regard to the expert ground truth for the view classification was 87.0%. The absolute error for the image quality estimation for each view was $0.12 \pm 0.09$, using a validation set of 3078 clips.

### Machine learning prediction of mean maximum image quality scores

There were 14,086 videos analyzed with the machine learning model from the 525 TTE studies. The machine learning model predicted view classification and quality with a total analysis time of less than 15 min. The views included were as outlined in the methods section and encompass standard components for a comprehensive study. The overall mean maximum quality score for all nine views was significantly poorer for mechanically ventilated TTEs compared with either control group. The overall estimated maximum quality score for mechanically ventilated TTEs were 0.55 (95% CI 0.54, 0.56) versus 0.64 (95% CI 0.63, 0.66; Control 1: Lab) and 0.61 (95% CI 0.59, 0.62; Control 2: Portable), $p = 0.002$. Estimates and 95% confidence intervals for the mean maximum quality scores for each view are presented in Fig. 3 and Table 1. The decreasing trend of best quality for Control 1: Lab patients and worst quality for mechanically ventilated patients was present for all but the 2 subcostal views (SC4 and IVC) and PSAX-PM view. After adjusting for multiple comparisons, mechanically ventilated TTEs had worse quality scores than either of the two control groups for the following views: AP2, AP3, AP4, PLAX, PSAX-A and PSAX-M (Holm adjusted p-value $< 0.001$).

To ensure that the lower overall study quality scores for mechanically ventilated TTEs were not driven by missing views (designated scores of zero), we also examined the mean maximum quality scores using only available views without application of a penalty (Table 2). As with Table 1, these results demonstrate that the mechanically ventilated TTEs are of lower image quality as compared with the two control groups with the exception of the subcostal views.

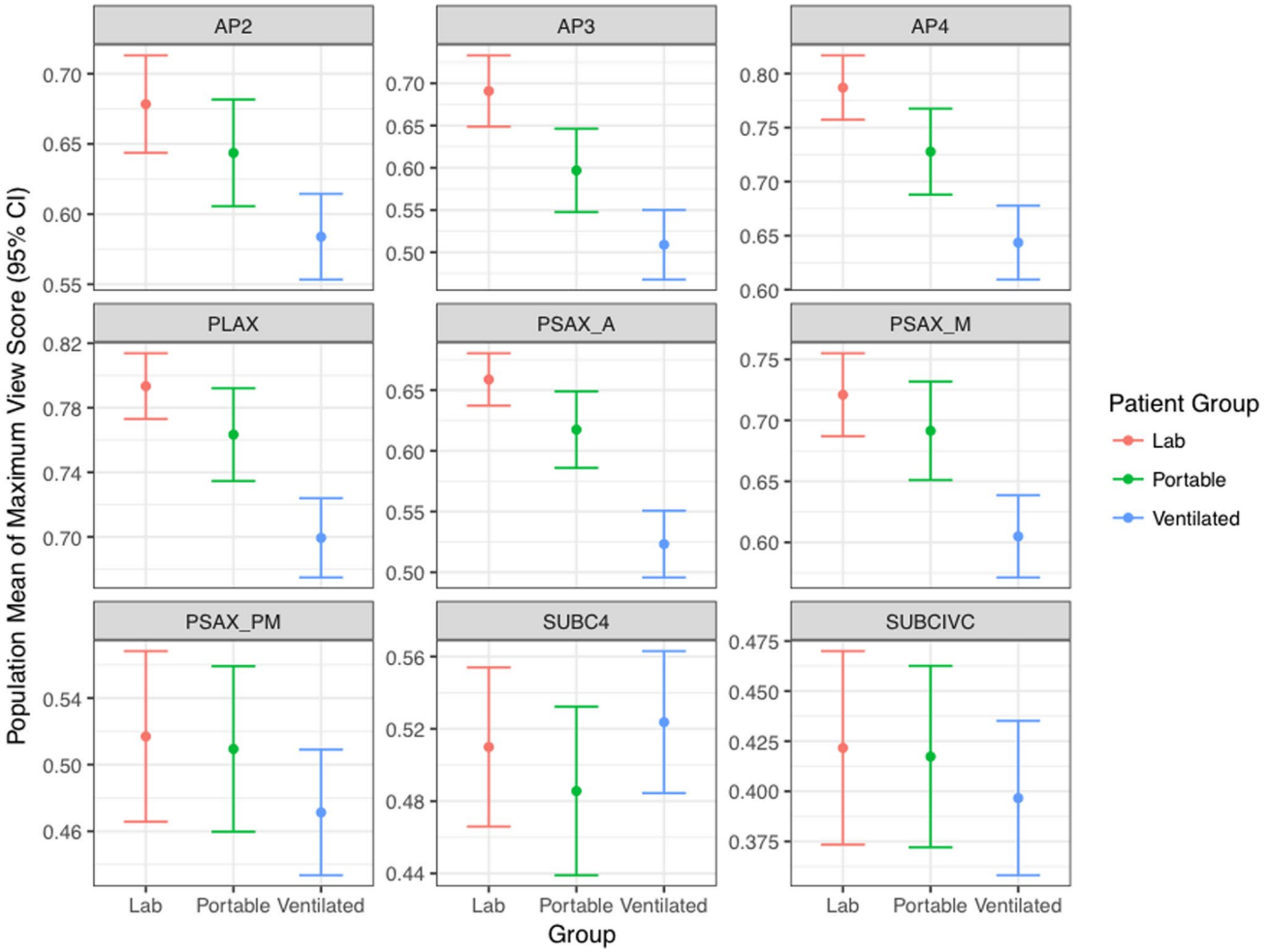


**Fig. 3** Mean maximum echo quality score by group. Population mean estimate and 95% confidence interval of mean maximum quality score by view and group. *AP2* apical 2-chamber view, *AP3* apical 3-chamber view, *AP4* apical 4-chamber view, *IVC* subcostal view of the inferior vena cava, *PLAX* parasternal long axis view, *PSAX-A* parasternal short axis, aortic valve level, *PSAX-M* parasternal short axis, mitral valve level, *PSAX-PM* parasternal short axis, papillary muscle level, *SC4* subcostal 4-chamber

**Table 1** Machine learning mean maximum quality scores by group and view

| Echo video view | Control 1: lab (95% CI) | Control 2: portable (95% CI) | Mechanically ventilated (95% CI) |
|---|---|---|---|
| PLAX | 0.79 (0.77, 0.81) | 0.76 (0.73, 0.79) | 0.70 (0.67, 0.72)* |
| AP2 | 0.68 (0.64, 0.71) | 0.64 (0.61, 0.68) | 0.58 (0.55, 0.61)* |
| AP4 | 0.79 (0.76, 0.82) | 0.73 (0.69, 0.77) | 0.64 (0.61, 0.68)* |
| AP3 | 0.69 (0.65, 0.73) | 0.60 (0.55, 0.65) | 0.51 (0.47, 0.55)* |
| PSAX-A | 0.66 (0.64, 0.68) | 0.62, (0.59, 0.65) | 0.52 (0.5, 0.55)* |
| PSAX-M | 0.72 (0.69, 0.75) | 0.69 (0.65, 0.73) | 0.60 (0.57, 0.64)* |
| PSAX-PM | 0.52 (0.47, 0.57) | 0.51 (0.46, 0.56) | 0.47 (0.43, 0.51) |
| SC4 | 0.51 (0.47, 0.55) | 0.49 (0.44, 0.53) | 0.52 (0.48, 0.56) |
| IVC | 0.42 (0.37, 0.47) | 0.42 (0.37, 0.46) | 0.40 (0.36, 0.44) |

*AP2* apical 2-chamber view, *AP3* apical 3-chamber view, *AP4* apical 4-chamber view, *IVC* subcostal view of the inferior vena cava, *PLAX* parasternal long axis view, *PSAX-A* parasternal short axis, aortic valve level, *PSAX-M* parasternal short axis, mitral valve level, *PSAX-PM* parasternal short axis, papillary muscle level, *SC4* subcostal 4-chamber

*Holm adjusted p-value < 0.001

**Table 2** Machine learning mean maximum quality scores by group and view, omitting missing values

| Echo video view | Control 1: lab (95% CI) | Control 2: portable (95% CI) | Mechanically ventilated (95% CI) |
|---|---|---|---|
| PLAX | 0.80 (0.78, 0.82) | 0.79 (0.76, 0.81) | 0.71 (0.69, 0.73)* |
| AP2 | 0.72 (0.69, 0.75) | 0.70 (0.68, 0.73) | 0.63 (0.61, 0.66)* |
| AP4 | 0.80 (0.77, 0.82) | 0.77 (0.75, 0.80) | 0.70 (0.68, 0.73)* |
| AP3 | 0.74 (0.71, 0.77) | 0.72 (0.69, 0.80) | 0.65 (0.62, 0.67)* |
| PSAX-A | 0.66 (0.64, 0.68) | 0.65 (0.62, 0.67) | 0.56 (0.53, 0.58)* |
| PSAX-M | 0.76 (0.73, 0.78) | 0.75 (0.72, 0.78) | 0.66 (0.64, 0.69)* |
| PSAX-PM | 0.70 (0.67, 0.73) | 0.68 (0.65, 0.71) | 0.61 (0.59, 0.63)* |
| SC4 | 0.63 (0.60, 0.67) | 0.62 (0.59, 0.66) | 0.63 (0.60, 0.65) |
| IVC | 0.63 (0.58, 0.67) | 0.61 (0.57, 0.66) | 0.64 (0.60, 0.67) |

*AP2* apical 2-chamber view, *AP3* apical 3-chamber view, *AP4* apical 4-chamber view, *IVC* subcostal view of the inferior vena cava, *PLAX* parasternal long axis view, *PSAX-A* parasternal short axis, aortic valve level, *PSAX-M* parasternal short axis, mitral valve level, *PSAX-PM* parasternal short axis, papillary muscle level, *SC4* subcostal 4-chamber

*Holm adjusted p-value < 0.001

## Diagnostic yield of inpatient echocardiograms

In mechanically ventilated patients, the diagnostic yield for minor parameters is lower compared to the combined control groups (Holm adjusted p-value < 0.001), however, there is no difference in reporting of major parameters. The proportion estimate for reporting of major parameters was 95.7% (95% CI 91.5%, 97.9%) in the control groups and 93.7% (95% CI 89.0%, 96.5%) in the mechanically ventilated group (p = ns). The population proportion estimate and 95% confidence interval for reporting of minor parameters was 92.6% (95% CI 97.7%, 95.7%) for the control groups and 84.5% (95% CI 78.4%,89.1%) for the mechanically ventilated group (p = 0.02).

## Machine learning image quality score and diagnostic yield

To examine the relationship between mean maximum quality score and diagnostic yield, a kernel smoother was fit for each group (Fig. 4, dotted lines). The smoothed lines follow the mass of data points. Higher mean maximum machine learning quality was associated with improved diagnostic yield of TTE. In this data, the relationship between diagnostic yield and quality score is relatively constant for quality scores above 0.5 in all groups, however, diagnostic yield declines in Control 1: Lab and mechanically ventilated patients with quality scores ≤ 0.5. A post-hoc cut-off of 0.5 mean maximum quality score was selected based on the observed relationship to test for significant differences in this data. Diagnostic yield was relatively constant for mean maximum quality scores ≥ 0.5 but dropped off in the Control 2: Portable and mechanically ventilated patients with quality scores < 0.5, so a cut-off of 0.5 was chosen. Overall

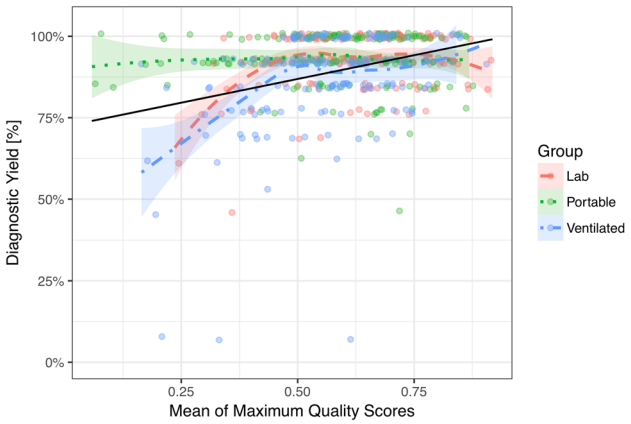


**Fig. 4** Relationship of echo quality and diagnostic yield. Relationship between machine learning mean maximum quality score and diagnostic yield. Higher quality score is positively correlated with diagnostic yield as shown with an inverse probability weighted linear regression (black line)

quality score and diagnostic yield were associated (inverse probability weighted linear regression line plotted in black in Fig. 4; p-value < 0.001).

Figure 5 demonstrates the relationship between mean maximum quality score as a categorical variable (cut-point 0.5) and diagnostic yield. Mechanically ventilated TTEs with quality scores ≤ 0.5 had lower diagnostic yield compared to the control groups quality scores > 0.5. In patients with mean maximum quality scores ≤ 0.5, the estimated diagnostic yield was 90.5% (95% CI 84.6%, 94.3%) in Control 1: Lab TTEs, 93.6% (95% CI 89.3%, 96.2%) in Control 2: Portable TTEs, and 84.1% (95% CI 77.1%, 89.3%) in mechanically ventilated TTEs. In samples with mean maximum quality scores of > 0.5, the estimates were 95.4%

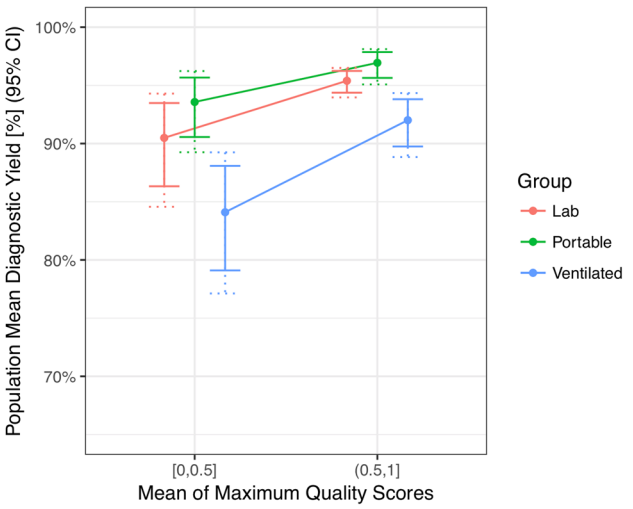


**Fig. 5** Relationship of echo quality and diagnostic yield. Relationship between machine learning mean maximum quality score as a categorical variable (cut-point 0.5) and diagnostic yield. The 95% confidence intervals and confidence intervals adjusted for multiple comparisons (dotted lines) are demonstrated for each group. Mechanically ventilated samples with quality scores ≤ 0.5 had lower diagnostic yield compared to the control groups quality scores > 0.5

(95% CI 94.0%, 96.5%) in Control 1: Lab TTEs, 96.9% (95.1%, 98.1%) in Control 2: Portable TTEs, and 92.0% (95% CI 88.9%, 94.3%) in mechanically ventilated TTEs. There is evidence of an association between diagnostic yield and machine learning quality scores ≤ 0.5 compared to scores > 0.5 by patient group (p = 0.02). After adjusting for multiple comparisons, there is evidence that the diagnostic yield is lower in mechanically ventilated TTEs compared to Control 2: Portable TTEs regardless of machine learning score category.

## Discussion

The principle findings of this study were: (1) a machine learning model was feasible for automated estimation of TTE image quality; (2) TTEs in mechanically ventilated patients are of inferior image quality compared to spontaneously breathing controls; and (3) TTEs with lower image quality were associated with reporting of fewer standard parameters.

This pragmatic study demonstrates that quality assessment of echocardiograms is achievable with a machine learning model and is associated with a clinically relevant outcome of diagnostic yield. There was rapid classification of quality making it a tool with implications for real time feedback and quality assurance. Such applications would be of benefit for point of care cardiac ultrasound when less experienced scanners and interpreters utilize ultrasound to guide triaging and patient management decisions.

### Assessment of echocardiographic image quality

Though there is no reference standard for the evaluation of echocardiographic image quality, several groups have proposed criteria focused on the percentage of endocardial border visualized. These scoring frameworks are practical with simple criteria based on scales from three to five points [14–18]. A previous machine learning publication defined echo image quality as the maximum probability of a view assignment. However, this metric does not necessarily represent conventional expert visual assessment of quality and is of unknown generalizability. Furthermore, the quality score is tied to view classification and is limited to several hundred TTEs that were used for training [11].

Although there are more sophisticated methods for quantifying TTE image quality [1, 26], we chose a 4-point scale to rapidly amass data for development of a robust machine learning model. Our results demonstrate the impact of image quality on the diagnostic utility of TTE in hospitalized patients and are consistent with other groups that show limitations in interpretation for poor quality studies in other clinical settings [27, 28]. Though suboptimal endocardial border definition can be improved with ultrasound enhancing agents, we chose to train and implement our model without contrast to better align with current use of point of care ultrasound. The devices used at the point of care are often not equipped for imaging with ultrasound enhancement and would be beyond the scope of practice for most front-line users.

### Application of machine learning model

Our study introduces the use of machine learning for analysis of an imaging database to assist in answering important and traditionally labor-intensive clinical questions. Areas of research that remain underexplored due to perceived lack of feasibility may be addressed with the use of validated and accurate machine learning models. We demonstrate the potential for machine learning models as an efficient platform for research.

This study provides a better understanding of the relationship between TTE image quality and interpretability and may have implications for use in cardiac point of care ultrasound. The machine learning model can be applied to existing ultrasound platforms for education and image optimization. The use of this tool at image acquisition by novice scanners has the potential to augment scanning performance through immediate feedback and can improve safety by tempering diagnostic confidence when image quality is suboptimal. This model may also be applied to image archiving

systems for quality assurance processes for TTE in standard echocardiography laboratories.

### Study limitations

Though our machine learning model has been shown to perform well, the labeled scoring thresholds on which it was developed is relatively simple. A machine learning model with data regarding valve structure, image depth, and gain would enrich the classification but would require a more resource intensive annotation process. The model would also be improved with labeled data from a number of readers and the inclusion of echo studies completed on a variety of ultrasound machine types to improve external validity. Due to the limited clinical information available, variables regarding comorbidities that would impact TTE quality, such as pulmonary disease, could not be integrated for statistical adjustment or sample matching.

The cohort of interest was limited to TTEs scanned only on the Philips iE33 platform. Restricting the analysis to a single model of machine improved the internal validity for comparisons between groups but reduce the generalizability of these results. Further external validation of this work should be completed at multiple sites using a variety of ultrasound vendors. Furthermore, the samples included in this study relied on the accurate designation of comprehensive or attempted comprehensive studies by the scanning sonographer which may introduce inconsistencies with human error. The age difference between groups may have been a confounder. This was partially addressed with statistical adjustment. However. It is important to note that there is evidence that echo image quality deteriorates with older age [29] and would bias our results toward the null. Therefore, the poorer image quality in younger, intubated patients likely represents a true difference compared with spontaneously breathing controls.

## Conclusions

We demonstrate that machine learning can provide efficient prediction of echocardiographic image quality and that this is correlated with clinically relevant data. We confirm that TTEs completed in mechanically ventilated patients are of inferior quality and diagnostic utility as compared to spontaneously breathing controls. Our machine learning platform is capable of rapid quality estimation that may transform teaching of echo scanning, quality assurance, and use of cardiac point of care ultrasound. This publication contributes to the growing body of work toward a goal of augmented acquisition and interpretation of cardiac ultrasound.

**Funding** Canadian Institutes of Health Research (CIHR). Natural Sciences and Engineering Research Council of Canada (NSERC).

**Data availability** The data will not be published publicly but can be made available for facilitation of review.Code availability The code will not be published publicly.

### Compliance with ethical standards

**Conflict of interest** The authors do not have any relevant conflicts of interest to disclose.

**Ethical approval** This study was reviewed and approved by the University of British Columbia Clinical Review Ethics Board.

**Consent to participate** This study was a retrospective review of anonymized images, therefore individual consent was not obtained.

**Consent for publication** The authors of this study consent to publication.

## References

1. Abdi A, Luong C, Tsang T, Jue J, Gin K, Yeung D, Hawley D, Rohling R, Abolmaesumi P (2017) Quality assessment of echocardiographic cine using recurrent neural networks: feasibility on five standard view planes. Paper presented at the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017, Lecture Notes in Computer Science, Springer, Cham
2. Abdi AH, Luong C, Tsang T, Allan G, Nouranian S, Jue J, Hawley D, Fleming S, Gin K, Swift J, Rohling R, Abolmaesumi P (2017) Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. IEEE Trans Med Imaging 36(6):1221–1230. https://doi.org/10.1109/TMI.2017.2690836
3. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP (2016) Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. J Am Coll Cardiol 68(21):2287–2295. https://doi.org/10.1016/j.jacc.2016.08.062
4. Madani A, Arnaout R, Mo M, Arnaout R (2018) Fast and accurate view classification of echocardiograms using deep learning. npj Digital Med. https://doi.org/10.1038/s41746-017-0013-1
5. Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, Gall W, Dudley JT (2016) Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. Circ Cardiovasc Imaging. https://doi.org/10.1161/CIRCIMAGING.115.004330
6. Asl BM, Setarehdan SK, Mohebbi M (2008) Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intell Med 44(1):51–64. https://doi.org/10.1016/j.artmed.2008.04.007
7. Katz DH, Deo RC, Aguilar FG, Selvaraj S, Martinez EE, Beussink-Nelson L, Kim KA, Peng J, Irvin MR, Tiwari H, Rao DC, Arnett DK, Shah SJ (2017) Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. J Cardiovasc Transl Res 10(3):275–284. https://doi.org/10.1007/s12265-017-9739-z

8. Carneiro G, Nascimento JC, Freitas A (2012) The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. IEEE Trans Image Process 21(3):968–982. https://doi.org/10.1109/TIP.2011.2169273
9. Smistad E, Østvik A, Haugen BO, L⊘vstakken L (2017) 2D left ventricle segmentation using deep learning. In: 2017 IEEE International Ultrasonics Symposium (IUS), 6–9 September, 2017, pp. 1–4
10. Khamis H, Zurakhov G, Azar V, Raz A, Friedman Z, Adam D (2017) Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Med Image Anal 36:15–21. https://doi.org/10.1016/j.media.2016.10.007
11. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC (2018) Fully automated echocardiogram interpretation in clinical practice. Circulation 138(16):1623–1635. https://doi.org/10.1161/CIRCULATIONAHA.118.034338
12. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, Lancellotti P, Muraru D, Picard MH, Rietzschel ER, Rudski L, Spencer KT, Tsang W, Voigt JU (2015) Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr 28(1):1–39.e14. https://doi.org/10.1016/j.echo.2014.10.003
13. Schiller NB, Shah PM, Crawford M, DeMaria A, Devereux R, Feigenbaum H, Gutgesell H, Reichek N, Sahn D, Schnittger I (1989) Recommendations for quantitation of the left ventricle by two-dimensional echocardiography. American Society of Echocardiography Committee on Standards, Subcommittee on Quantitation of Two-Dimensional Echocardiograms. J Am Soc Echocardiogr 2(5):358–367
14. Tighe DA, Rosetti M, Vinch CS, Chandok D, Muldoon D, Wiggin B, Dahlberg ST, Aurigemma GP (2007) Influence of image quality on the accuracy of real time three-dimensional echocardiography to measure left ventricular volumes in unselected patients: a comparison with gated-SPECT imaging. Echocardiography 24(10):1073–1080. https://doi.org/10.1111/j.1540-8175.2007.00525.x
15. Hoffmann R, Lethen H, Marwick T, Arnese M, Fioretti P, Pingitore A, Picano E, Buck T, Erbel R, Flachskampf FA, Hanrath P (1996) Analysis of interinstitutional observer agreement in interpretation of dobutamine stress echocardiograms. J Am Coll Cardiol 27(2):330–336
16. Kusunose K, Shibayama K, Iwano H, Izumo M, Kagiyama N, Kurosawa K, Mihara H, Oe H, Onishi T, Ota M, Sasaki S, Shiina Y, Tsuruta H, Tanaka H, Investigators J (2018) Reduced variability of visual left ventricular ejection fraction assessment with reference images: the Japanese Association of Young Echocardiography Fellows multicenter study. J Cardiol. https://doi.org/10.1016/j.jjcc.2018.01.007
17. Nagata Y, Kado Y, Onoue T, Otani K, Nakazono A, Otsuji Y, Takeuchi M (2018) Impact of image quality on reliability of the measurements of left ventricular systolic function and global longitudinal strain in 2D echocardiography. Echo Res Pract 5(1):27–39. https://doi.org/10.1530/ERP-17-0047
18. Medvedofsky D, Mor-Avi V, Byku I, Singh A, Weinert L, Yamat M, Kruse E, Ciszek B, Nelson A, Otani K, Takeuchi M, Lang RM (2017) Three-dimensional echocardiographic automated quantification of left heart chamber volumes using an adaptive analytics algorithm: feasibility and impact of image quality in nonselected patients. J Am Soc Echocardiogr 30(9):879–885. https://doi.org/10.1016/j.echo.2017.05.018
19. Van Woudenberg N, Liao Z, Abdi AH, Girgis H, Luong C, Vaseli H, Behnami D, Zhang H, Gin K, Rohling R, Tsang T, Abolmaesumi P (2018) Quantitative echocardiography: real-time quality estimation and view classification implemented on a mobile android device. In: Stoyanov D, et al. (eds) Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation. Springer, Cham, pp 74–81
20. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol 1
21. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML 37:448–456
22. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
23. Levitov A, Frankel HL, Blaivas M, Kirkpatrick AW, Su E, Evans D, Summerfield DT, Slonim A, Breitkreutz R, Price S, McLaughlin M, Marik PE, Elbarbary M (2016) Guidelines for the appropriate use of bedside general and cardiac ultrasonography in the evaluation of critically ill patients-part ii: cardiac ultrasonography. Crit Care Med 44(6):1206–1227. https://doi.org/10.1097/ccm.0000000000001847
24. Lumley T, Scott AJ (2013) Two-sample rank tests under complex sampling. Biometrika 100(4):831–842
25. Lumley T, Scott A (2014) Tests for regression models fitted to survey data. Aust N Z J Stat 56(1):1–14
26. Gaudet J, Waechter J, McLaughlin K, Ferland A, Godinez T, Bands C, Boucher P, Lockyer J (2016) Focused critical care echocardiography: development and evaluation of an image acquisition assessment tool. Crit Care Med 44(6):e329–335. https://doi.org/10.1097/ccm.0000000000001620
27. Hensel KO, Roskopf M, Wilke L, Heusch A (2018) Intraobserver and interobserver reproducibility of M-mode and B-mode acquired mitral annular plane systolic excursion (MAPSE) and its dependency on echocardiographic image quality in children. PLoS ONE 13(5):e0196614. https://doi.org/10.1371/journal.pone.0196614
28. Thaden JJ, Tsang MY, Ayoub C, Padang R, Nkomo VT, Tucker SF, Cassidy CS, Bremer M, Kane GC, Pellikka PA (2017) Association between echocardiography laboratory accreditation and the quality of imaging and reporting for valvular heart disease. Circul Cardiovasc Imaging. https://doi.org/10.1161/CIRCIMAGING.117.006140
29. Kitzman DW (2000) Normal age-related changes in the heart: relevance to echocardiography in the elderly. Am J Geriatr Cardiol 9(6):311–320

## Affiliations

**Christina Luong[1] · Zhibin Liao[2] · Amir Abdi[2] · Hany Girgis[1] · Robert Rohling[2] · Kenneth Gin[1] · John Jue[1] · Darwin Yeung[1] · Elena Szefer[3] · Darby Thompson[3] · Michael Yin-Cheung Tsang[1] · Pui Kee Lee[1] · Parvathy Nair[1] · Purang Abolmaesumi[2] · Teresa S. M. Tsang[1,4]**

[1] Division of Cardiology, University of British Columbia, Vancouver, BC, Canada

[2] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

[3] Emmes Canada, Burnaby, BC, Canada

[4] Diamond Health Care Centre, University of British Columbia, 9th Floor Cardiology, 2775 Laurel Street, Vancouver, BC V5Z 1M9, Canada