JACC: ADVANCES VOL. 3, NO. 9, 2024

ORIGINAL RESEARCH

IMAGING

Automated Assessment of Right Atrial Pressure From Ultrasound Videos Using Machine Learning

Dominic Yurk, PhD,[a,b] Joshua P. Barrios, PhD,[b,c] Elodie Labrecque Langlais, MS,[d,e,f] Robert Avram, MD, MS,[b,f,g] Mandar A. Aras, MD, PhD,[b] Yaser Abu-Mostafa, PhD,[a] Arun Padmanabhan, MD, PhD,[b,d,h,*] Geoffrey H. Tison, MD, MPH[b,c,i,*]

ABSTRACT

BACKGROUND Early recognition of volume overload is essential for heart failure patients. Volume overload can often be easily treated if caught early but causes significant morbidity if unrecognized and allowed to progress. Intravascular volume status can be assessed by ultrasound-based estimation of right atrial pressure (RAP), but the availability of this diagnostic modality is limited by the need for experienced physicians to accurately interpret these scans.

OBJECTIVES We sought to evaluate whether machine learning can accurately estimate echocardiogram-measured RAP.

METHODS We developed fully automated deep learning models for identifying inferior vena cava scans with rapid inspiration in echocardiogram studies and estimating RAP from those scans. The RAP estimation model was trained and evaluated using 15,828 ultrasound videos of the inferior vena cava and coupled cardiologist-assessed RAP estimates as well as 319 RAP measurements from right heart catheterization.

RESULTS Our model agreed with cardiologist estimates 80.3% of the time (area under the receiver-operating characteristic of 0.844) in a test data set, at the upper end of interoperator agreement rates found in the literature of 70 to 75%. Our model's RAP estimates were statistically indistinguishable from cardiologists' ultrasound-based RAP estimates ($P = 0.98$) when compared against the gold standard of right heart catheterization RAP measurements in a subset of patients. Our model also generalized well to an external data set of echocardiograms from a different institution (area under the receiver-operating characteristic of 0.854 compared to cardiologist RAP estimates).

CONCLUSIONS Machine learning is capable of accurately and robustly interpreting RAP from echocardiogram videos. This algorithm could be used to perform automated assessments of intravascular volume status. (JACC Adv. 2024;3:101192) © 2024 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

From the [a]Department of Electrical Engineering, California Institute of Technology, Pasadena, USA; [b]Division of Cardiology, Department of Medicine, University of California-San Francisco, San Francisco, California, USA; [c]Cardiovascular Research Institute, University of California-San Francisco, San Francisco, California, USA; [d]Gladstone Institutes, San Francisco, USA; [e]Department of Biomedical Engineering, École Polytechnique de Montréal, Montreal, Canada; [f]Heartwise (heartwise.ai), Montreal Heart Institute, Montreal, Canada; [g]Division of Cardiology, Department of Medicine, Montreal Heart Institute, University of Montreal, Montreal, Canada; [h]Chan Zuckerberg Biohub San Francisco, San Francisco, USA; and the [i]Bakar Computational Health Sciences Institute, University of California-San Francisco, San Francisco, California, USA. *Drs Tison and Padmanabhan contributed equally to this work and are considered to be as co-senior authors.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received April 1, 2024; revised manuscript received June 27, 2024, accepted July 9, 2024.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2024.101192

ABBREVIATIONS AND ACRONYMS

AUROC = area under the receiver operating characteristic curve

IVC = inferior vena cava

ML = machine learning

RAP = right atrial pressure

RHC = right heart catheterization

SFP = severe false positive

SFN = severe false negative

TTE = transthoracic echocardiogram

UCSF = University of California-San Francisco

Heart failure affects over 6 million Americans and is the leading cause of hospitalization in adults over 65, representing an annual financial burden of over 40 billion dollars on the health care system.[1] A common complication of heart failure is vascular congestion, or the accumulation of excess fluid in blood vessels due to impaired cardiac function. This leads to decreased fluid removal by the kidneys, volume overload, and elevated pressures within the heart. Patients with heart failure frequently experience exacerbations of vascular congestion which, if left unchecked, can lead to swelling of the extremities, difficulty breathing, renal failure, and hepatic congestion, which are associated with higher overall mortality.[2] Severe cases require hospital admission for adequate treatment. However, if vascular congestion is detected early, it can often be addressed in an outpatient setting through adjustment of oral diuretics and other medications.[3] Thus, methods to easily and accurately identify vascular congestion are of significant clinical interest.

A common method of assessing vascular volume status is the estimation of pressure in the right atrium of the heart, or right atrial pressure (RAP).[2] RAP is normally low (0-5 mm Hg), but in the setting of vascular congestion, it can be elevated to 10-30 mm Hg due to fluid accumulation in venous capacitance vessels.[4] The gold standard for RAP measurement is right heart catheterization (RHC), a procedure that involves introducing a flexible catheter tipped with a pressure transducer into the venous system and advancing it to the right side of the heart. This procedure yields highly accurate measurements of RAP but is invasive, carries procedural risks, requires specialized facilities, and is costly.[5] As such, it is only conducted in settings where the diagnostic benefits outweigh the associated risks. A common noninvasive alternative for quantitatively assessing RAP is ultrasound evaluation of the inferior vena cava (IVC) at its juncture with the right atrium. These scans are routinely collected as part of a standard transthoracic echocardiogram (TTE). RAP is estimated from IVC ultrasound videos using a process called the "sniff test." In this process, a TTE video of the IVC is recorded while the patient is asked to inhale sharply, creating a rapid change in thoracic pressure that temporarily compresses the IVC. In patients with normal RAP, the sniff will significantly collapse the IVC, while in patients with elevated RAP, the collapse will be attenuated or absent due to higher pressure inside the vessel (**Figure 1**).

The metrics used to estimate RAP from these scans were previously described [6] and are recommended by the American Society for Echocardiography and the European Association of Cardiovascular Imaging [7] as the standard for TTE-based RAP assessment. However, the application of these standards can be operator-dependent. Multiple studies have found substantial interoperator variability among medical trainees, fellows, and emergency physicians when collecting IVC ultrasound videos and estimating RAP, even after dozens of hours of training.[8-14] The most reliable assessments of TTE IVC videos are those of experienced cardiologists who regularly interpret these studies, but such specialists may not be uniformly available at all medical centers. If a machine learning (ML) model could be trained to automatically analyze IVC sniff tests with the proficiency of an experienced cardiologist, quick and robust RAP assessment could be made much more accessible, even outside of the context of a complete TTE study. We hypothesized that we could train a deep learning ML model to accomplish automated interpretation of IVC sniff test ultrasound videos collected as part of a routine TTE and compare its performance with both expert cardiologist assessments and the gold standard of RHC measurements of RAP.

METHODS

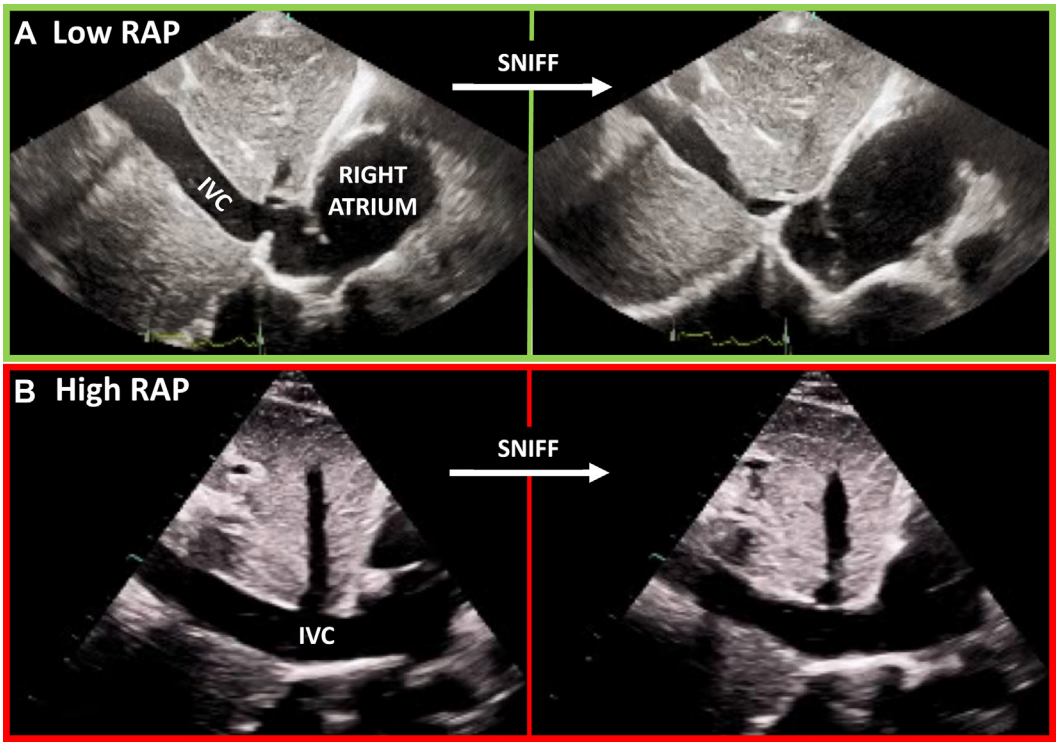
DATA SET AND DATA SCREENING. This work used Digital Imaging and Communications in Medicine format recordings of TTE studies obtained at the University of California-San Francisco (UCSF) between 2016 and 2020. TTE studies were excluded in their entirety for any of the following reasons:

- Multiple different RAP estimates were recorded for the same study.
- The study involved a stress test, pediatric or fetal patient, transesophageal or intracardiac ultrasound, or patient on a ventilator. None of these study types are representative of how a standard TTE sniff test would be evaluated.
- 99.7% of the remaining studies were conducted on 1 of 4 cardiac ultrasound machine models; the remaining 0.3% were excluded.

In addition, individual videos from studies were excluded based on the following criteria:

- The scan was <20 frames long. Such video clips are too short to contain a complete sniff test.

FIGURE 1 Examples of a Sniff Test



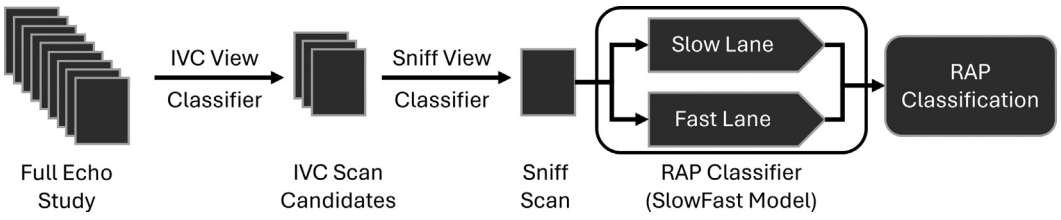
(A) An ultrasound view of the IVC and right atrium at a resting state and during a sniff. The high degree of collapse during the sniff indicates that this patient likely has a normal RAP. (B) A view of another IVC at rest and during a sniff. The low degree of collapse indicates that this patient likely has an elevated RAP. RAP = right atrial pressure; IVC = inferior vena cava.

- Physical pixel size was not recorded in the metadata. This made it impossible to measure IVC diameter in real units.
- Physical pixel size was in the lower or upper 5th percentile of pixel scale. This narrowed the total range of pixel scales from (0.002, 5.2) centimeters to (0.074, 0.168) centimeters. Extreme pixel sizes imply extreme ultrasound settings for parameters like scanning frequency and depth, which would generally not be used for viewing the IVC.
- Color Doppler mode was enabled. Sniff scans are rarely taken with Color Doppler enabled at UCSF, and Color Doppler signal in a small subset of data would likely confuse ML models. Identification of videos with Color Doppler was performed using a pretrained Doppler classification ML model from another study.

After exclusions, a total of 16,823 TTE studies with cardiologist RAP measurements remained. Each of these studies consisted of up to 200 individual ultrasound videos covering all areas of the heart, so two view classification ML models were trained to sort through these data and identify a single sniff test scan per study (**Figure 2**). The first identified all scans in the study showing any view of the IVC, and the second identified which of these IVC views was most likely to show a sniff test (see Appendix 2 for more details). This process eliminated 995 studies containing no recognizable IVC scan, leaving 15,828 studies collected from 11,869 patients (**Central Illustration**). All TTE studies had RAP estimates made by National Board of Echocardiography-certified UCSF cardiologists in the course of routine clinical care. In accordance with published standards,[7] these measurements were recorded as either 3 mm Hg (representing a range of 0-5 mm Hg), 8 mm Hg (5-10 mm Hg), or 15 mm Hg (10-30 mm Hg). Measurements of 3 mm Hg and 15 mm Hg represent normal and elevated RAP ranges, respectively, while a value of 8 mm Hg is considered indeterminate.

RHC DATA. We also interrogated a data set of RHCs performed at UCSF between 2012 and 2020. We identified 1739 paired TTE and RHC studies on the same patient occurring within a window of ±30 days

FIGURE 2 Data Flow Illustration



Going From a Full TTE Study to an RAP Classification Relies on a Series of 3 ML Classification Models. First, the IVC view classifier identifies scans in the sStudy Which Show the IVC. Second, the sniff view classifier identifies which of the IVC scans is most likely to Sshow a sniff test. Finally, the RAP classifier analyzes this sniff scan using the SlowFast architecture to determine an RAP estimate of 3, 8, or 15 mm Hg. ML = machine learning; TTE = transthoracic echocardiogram; other abbreviations as in Figure 1.

of each other. Although a smaller time separation window would decrease the chance that the patient's RAP changed between the 2 studies, it would substantially reduce the amount of paired data available for analysis. A window of 30 days balanced these considerations (see Appendix 3). If multiple RHC measurements were taken from the patient, the closest in temporal proximity to the TTE study was kept. Of the 1739 paired studies, 527 TTE studies were labeled with an RAP by a cardiologist. After applying our sniff identification procedure and manually screening to exclude scans without the IVC (views of the IVC without an obvious sniff were not excluded), 319 ultrasound videos remained. These studies represented our "golden" data for which we had matched sniff test ultrasound videos, the associated cardiologist RAP assessment based on this video, and a ground-truth RAP measurement from RHC. To avoid potential contamination, these scans were always kept in the test data set for all model training.

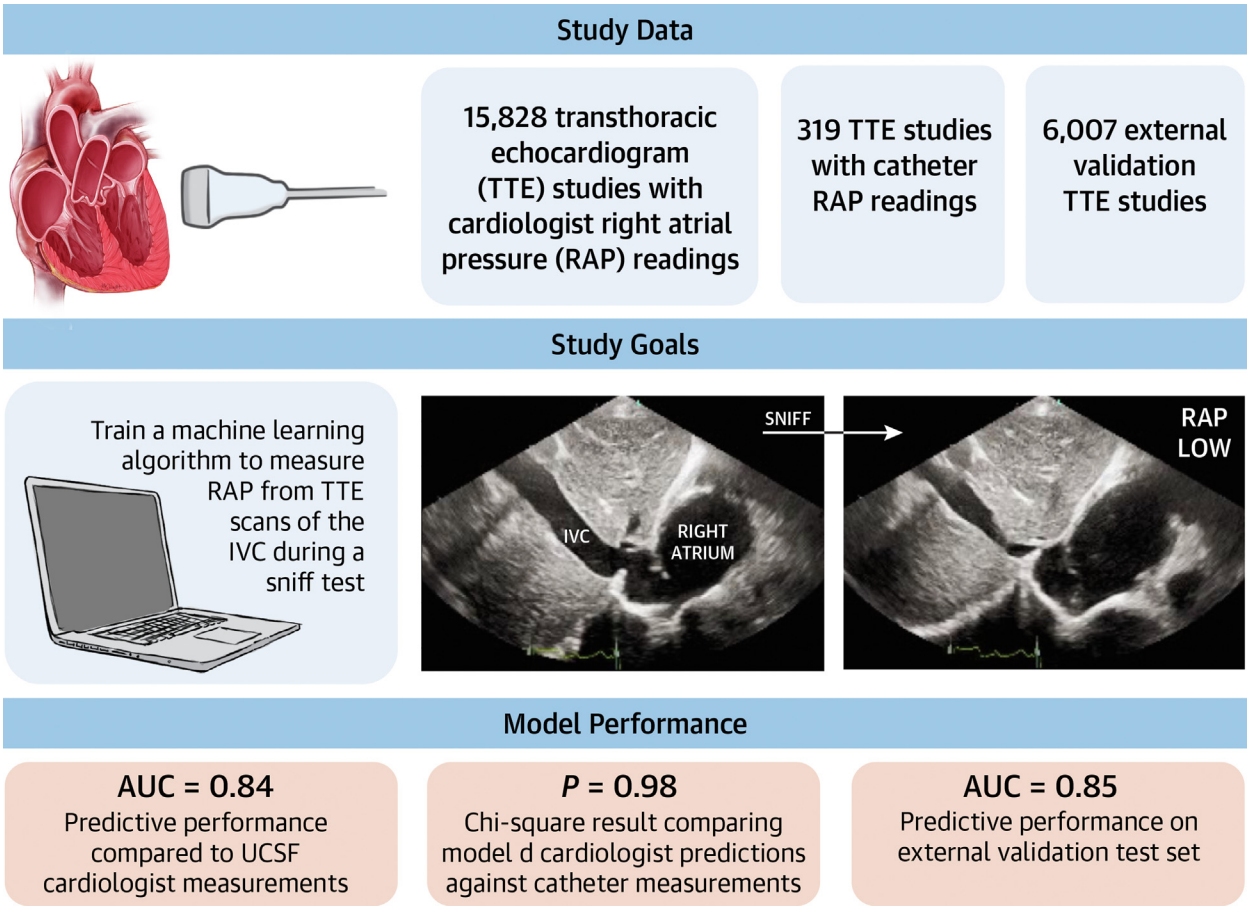
An additional consideration was how to determine the "accuracy" of cardiologist or ML RAP estimations (which are categorical) when compared to RHC measurements (which are continuous). Although a direct comparison could be made by binning the RHC values according to the ranges in Lang *et al*,[7] these ranges have some ambiguity; an RHC value of 5 mm Hg could be binned as "3" or "8," and an RHC value of 10 mm Hg could be binned as "8" or "15." As RHC measurements were reported as integers, these ambiguous situations were common. We elected to place these edge case readings into the lower of the 2 possible bins. Thus, an RHC value in the range [0, 5] was binned as "3," a value in the range (5, 10] was binned as "8," and a value in the range (10, 30] was binned as "15." This allowed for direct accuracy comparisons between any combination of outputs from RHCs, cardiologist estimations, and ML models.

EXTERNAL VALIDATION DATA. Our external validation data set consisted of 6007 TTE studies and cardiologists' interpreted RAP measurements obtained at the University of Montreal in 2022. The RAP measurements were performed according to standard TTE guidelines.[7] All studies were conducted on a model of ultrasound machine which was also present in the UCSF data set. Identical data exclusion and sniff scan selection procedures were applied to the Montreal data with the high-stringency IVC view classifier threshold (see Appendix 2), leaving 2,618 data points to use for model evaluation.

RAP CLASSIFICATION MODEL TRAINING AND EVALUATION. The 15,828 studies with a cardiologist-generated RAP estimate were divided into 12,664 training studies, 1,582 validation studies, and 1,582 test studies. Data were split randomly by patient, with the exception that all 319 studies with a matched RHC measurement were placed in the test data set. The RAP classification ML model used the SlowFast R50 architecture [15] with a length-3 output layer and softmax activation to generate probabilities for 3 classes: 3 mm Hg, 8 mm Hg, and 15 mm Hg. SlowFast uses a convolution backbone and works by analyzing data in 2 components: a "slow lane," which takes in every eighth video frame and is meant to identify static video features, and a "fast lane," which takes in all video frames and is meant to identify dynamic video features. Information from both lanes is synthesized to yield a final classification. This architecture makes intuitive sense for our application because RAP measurement relies on measuring one "slow" feature (IVC resting diameter) and one "fast" feature (IVC collapsibility).

The model was trained for 100 epochs using categorical cross-entropy loss with the Radam optimizer and an initial learning rate of 1e-3. The epoch that produced the best validation area under the receiver

CENTRAL ILLUSTRATION **Measuring Right Atrial Pressure From Transthoracic Echocardiogram Videos of the IVC Under Rapid Inspiration Using Deep Learning**



Yurk D, et al. JACC Adv. 2024;3(9):101192.

Measuring Right Atrial Pressure From Transthoracic Echocardiogram Videos of the IVC Under Rapid Inspiration Using Deep Learning Abbreviation as in Figure 1.

operating characteristic curve was saved as the best state. Regularization was performed via random data augmentation, batch normalization, dropout, and label smoothing. Training was performed on an NVIDIA RTX 6000 Ada graphics card (see Appendix 4 for more details).

Model performance was evaluated on the full test data set as well as 2 subsets: one consisting of the 319 studies with a coupled RHC measurement and one consisting of 554 "high-stringency" videos, which were scored highly by the IVC view classification ML model (see Appendix 2). Additionally, we evaluated model performance on an external data set from the Montreal Heart Institute consisting of 2,618 sniff test TTE videos and associated cardiologist RAP estimates. To permit uniform comparison of results between these data sets, all values for area under the receiver-operating characteristic (AUROC) curve, accuracy, and F_1 score reported below are calculated as a weighted average of the metric in each RAP class (3, 8, or 15 mm Hg) based on the frequency of that class in the overall UCSF data set. All metrics are accompanied by a 95% confidence interval, calculated using either a z-test for proportions or 1000-sample bootstrapping for AUROC and F_1. All reported P values were computed using 2-sided z-tests.

ETHICAL APPROVAL. This study was reviewed by the University of California-San Francisco Institutional Review Board, and the need for informed consent was waived. The external validation was reviewed and approved by the University of Montreal Institutional Review Board.

TABLE 1 Out-of-Sample Performance, Full Test Set

RAP Prediction Performance: Full Test Data Set				
Confusion Matrix	Predicted 3 mm Hg	Predicted 8 mm Hg	Predicted 15 mm Hg	Class AUROC
Cardiologist 3 mm Hg	952	167	29	0.859 (0.842-0.877)
Cardiologist 8 mm Hg	77	132	37	0.603 (0.567-0.637)
Cardiologist 15 mm Hg	31	44	113	0.889 (0.864-0.911)
Accuracy		77.3% (75.2-79.4)		
Severe false negative		16.5% (11.2-21.8)		
Severe false positive		2.5% (1.6-3.4)		
AUROC		0.823 (0.807-0.837)		
F_1 score		0.769 (0.752-0.784)		

The results of the echo-based RAP classification model evaluated on the full test data set of 1,582 scans. Green boxes represent correct predictions, the red box represents severe false negative (SFN) predictions, and the orange box represents severe false positive (SFP) predictions. SFN rate represents the percentage of patients classified as 15 mm Hg by cardiologists which the model classified as 3 mm Hg, and SFP represents the opposite error. Overall accuracy, AUROC, and F_1 values are computed as a weighted average from each class based on the frequency of that class in the data set. Values in parentheses represent 95% confidence boundaries.

AUROC = area under the receiver-operating characteristic; RAP = right atrial pressure.

RESULTS

DATA SET CHARACTERIZATION. Patients in the cohort were 51.8% female and ranged in age from 18 to 102 years old. RAP measurements were produced by 45 different physicians using 4 different models of cardiac ultrasound machines from 2 different manufacturers. The overall proportion of studies with cardiologist RAP estimates of 3 mm Hg, 8 mm Hg, and 15 mm Hg was 79.1%, 13.4%, and 7.5%, respectively. Within the test data set, the relative proportions were 72.6%, 15.5%, and 11.6%, respectively. This difference in distribution was because the test data set contained all studies with coupled RHC data, and patients who underwent RHC were more likely than the broader population who received a TTE to be in heart failure. Detailed summary statistics are provided in Supplemental Table S1.

MODEL PERFORMANCE. Comparison to cardiologist measurements. Model training and initial evaluation relied only on TTE videos of the IVC during a sniff test and cardiologists' interpretations of those videos. The model with the best validation performance, as measured by AUROC, was run on the test data set to gauge out-of-sample performance. The average frequency-weighted model AUROC was 0.823, and similarly frequency-weighted accuracy and F_1 values were 77.3% and 0.769, respectively (see Table 1). We also examined the rate at which patients given a measurement of 15 mm Hg by a cardiologist were labeled as 3 mm Hg by the model (a severe false negative [SFN]), as well as the rate at which patients given a measurement of 3 mm Hg by a cardiologist were labeled as 15 mm Hg by the model (a severe false positive [SFP]). These yielded rates of 16.5% and 2.5%, respectively.

Upon manual inspection, the full test data set was found to contain numerous scans in which the IVC was either noisy, obscured, or not present at all (Supplemental Figure S1). To address this, we created a high-stringency subset of the test data set which only kept the 35% of scans with the highest scores from the IVC view classification model. Upon manual

TABLE 2 Out-of-Sample Performance, High-Stringency Test Set

RAP Prediction Performance: High-Stringency IVC Test Data set				
Confusion Matrix	Predicted 3 mm Hg	Predicted 8 mm Hg	Predicted 15 mm Hg	Class AUROC
Cardiologist 3 mm Hg	328	44	4	0.879 (0.852-0.904)
Cardiologist 8 mm Hg	41	58	7	0.626 (0.575-0.677)
Cardiologist 15 mm Hg	4	30	38	0.917 (0.878-0.952)
Accuracy		80.3% (77.0-83.6)		
Severe false negative		5.6% (0.3-10.9)		
Severe false positive		1.1% (0.0-2.2)		
AUROC		0.844 (0.822-0.865)		
F_1 score		0.758 (0.729-0.784)		

Results of the echo-based RAP classification model evaluated on the high-stringency IVC test data set of 554 scans.

AUROC = area under the receiver-operating characteristic; IVC = inferior vena cava; RAP = right atrial pressure.

inspection, almost all scans passing this higher threshold seemed to be high-quality IVC videos. This filtering was blinded to any RAP information (either cardiologist estimates or model predictions) and should represent the model performance we could expect on higher-quality IVC data. Test performance from this data set is shown below in Table 2. Applying the high-stringency IVC threshold did not significantly change overall accuracy, AUROC, or F_1 score. However, it did lead to a significant drop in the SFN rate ($P = 0.02$) by almost two-thirds, from 16.5% to 5.6%. The SFP rate decreased with moderate significance, from 2.5% to 1.1% ($P = 0.09$).

Comparison to RHC measurements. To address the variability of RAP estimations from echocardiogram videos of the IVC, we evaluated model performance on a subset of 319 patients who had an RAP measurement obtained via RHC within 30 days of a cardiologist's TTE-based RAP estimate (see Methods). To evaluate the performance of both UCSF cardiologists and the ML model against the invasive gold standard, RHC measurements were converted from their original continuous values into categorical 3, 8, or 15 mm Hg bins. Table 3 below shows how RAP estimates from cardiologists and the ML model compared to measurements from RHC. While the model agreed with cardiologist predictions 77% of the time in the full test data set (see Table 1), its performance with respect to RHC results was effectively equivalent to that of the cardiologists, both in terms of overall accuracy and the SFP and SFN rates. Applying a categorical chi-squared test to compare the distribution of predictions between the cardiologists and echo-only ML model yielded $P = 0.98$, suggesting no statistically significant difference between the ML model and cardiologist interpretations.

External evaluation performance. As external validation of our RAP estimation ML model, we examined its performance in a data set of 2618 TTE studies collected in 2022 from the Montreal Heart Institute at the University of Montreal. Study and sniff video selection were performed using the same procedure as that used to generate the high-stringency UCSF test data set. This fully independent data set provided an opportunity to evaluate the generalizability of the model trained on UCSF data. Overall, model performance remained robust on the external data set (Table 4), with similar performance in overall accuracy, AUROC, and SFN rate as well as slight improvements in F_1 score ($P < 0.05$) and SFP rate ($P = 0.07$) when compared to the high-stringency UCSF test data set.

TABLE 3 Comparison to Right Heart Catheter

UCSF Right Heart Catheterization Data

RHC Value Range		Cardiologist Estimate of RAP From Echo		
	N	3 mm Hg	8 mm Hg	15 mm Hg
(0, 5) mm Hg	118	84	24	10
(5, 10) mm Hg	104	56	28	20
(10, 30) mm Hg	97	26	27	44
Accuracy		48.9% (43.4-54.4)		

ML Model

RHC Value Range		Model Prediction of RAP From Echo		
	N	3 mm Hg	8 mm Hg	15 mm Hg
(0, 5) mm Hg	118	78	30	10
(5, 10) mm Hg	104	50	35	19
(10, 30) mm Hg	97	27	25	45
Accuracy		49.5% (44.0-55.0)		

Comparison of RAP estimates made by both cardiologists and the ML model from echo IVC videos versus gold standard RAP measurements from right heart catheterization (RHC). The original continuous RAP values have been binned into 3 mm Hg, 8 mm Hg, and 15 mm Hg categories to allow for direct comparison with cardiologist and model outputs.

ML = machine learning; RAP = right atrial pressure; RHC = right heart catheterization; UCSF = University of California-San Francisco.

DISCUSSION

We present, to our knowledge, the first ML model capable of performing fully automated measurement of RAP from a TTE video of a sniff test. The model was trained on a large and varied data set and attained an AUROC of 0.823 (on the full test data set, $N = 1{,}582$) and 0.844 (on the high-stringency test data set, $N = 554$) compared to RAP measurements from experienced UCSF cardiologists. When both model and cardiologist RAP estimates from echoes were compared to RAP measurements obtained from the invasive gold standard RHC procedure in a subset of patients ($N = 319$), the ML model attained the same accuracy as the cardiologists and produced a statistically identical measurement distribution. The ML model also demonstrated strong generalization, showing robust performance on an independently obtained data set from another institution. Furthermore, the developed models are relatively lightweight, allowing for real-time inference on consumer graphics processing unit or even mobile devices (see Appendix 4.3).

While modern medical centers have widespread access to ultrasound equipment capable of IVC imaging, round-the-clock access to cardiologists for interpretation of these images for RAP assessment may be limited. Even at large medical centers with many cardiologists, it is not always feasible to immediately obtain expert TTE assessment for vascular volume status. The availability of a robust

TABLE 4 External Model Validation

Montreal Data				
Confusion Matrix	Predicted 3 mm Hg	Predicted 8 mm Hg	Predicted 15 mm Hg	Class AUROC
Cardiologist 3 mm Hg	1963	195	8	0.891 (0.875-0.906)
Cardiologist 8 mm Hg	118	151	13	0.631 (0.599-0.658)
Cardiologist 15 mm Hg	15	75	80	0.917 (0.893-0.937)
Accuracy	82.4% (80.9-83.9)			
Severe false negative	8.8% (4.6-13.1)			
Severe false positive	0.4% (0.1-0.6)			
AUROC	0.854 (0.842-0.866)			
F_1 score	0.805 (0.793-0.817)			

Model performance as compared to cardiologist RAP estimates, shown for the Montreal external test data set. Total accuracy, AUROC, and F_1 scores are still weighted by UCSF class frequencies to allow for direct comparison.

AUROC = area under the receiver-operating characteristic.

ML model would enable quick and noninvasive RAP assessment to be expanded beyond cardiology departments and democratized to anyone with access to an ultrasound machine.

Multiple prior studies have attempted to use ML to analyze ultrasound scans of the IVC and use it to predict parameters such as RAP [16] and fluid responsiveness.[17-19] However, these studies had access to limited data sets of at most 175 patients, hampering model training and leading to significant generalization concerns. Our data set of 15,828 individual TTE studies from 11,869 patients represents a nearly 100× increase in data set size over any similar previous study, enabling far more robust training and testing. Furthermore, the data set covered TTEs collected over a period of 4 years using 4 different models of ultrasound machines and interpreted by dozens of cardiologists, encompassing the variability inherent in real-world medical data. As a result, this study represents a unique opportunity to both train an accurate ML model and robustly evaluate its performance.

VARIABILITY OF HUMAN LABELING. When assessing ML model results, it is important to note that sniff test assessments are vulnerable to a variety of potential sources of error, such as varying IVC diameter at different points along its length or erroneous measurements due to misalignment of the imaging plane with the IVC.[20] Multiple studies have found substantial interoperator variability when assessing IVC dimensions,[8-14] and studies that specifically looked at interoperator agreement rates for RAP measurement or vascular volume status based on IVC scans found agreement rates of 70 to 75%.[11,12] Assuming that the conditions leading to this variability are broadly similar between different medical centers, our model performance may be nearing the performance limit set by underlying uncertainty in human labeled TTE data.

COMPARISON TO INVASIVE CATHETER MEASUREMENTS. The subset of our test data set that had matched RHC measurements provided a means of assessing model performance in a way that was independent of human interpreter variability. Alignment between cardiologist and RHC measurements in this data set compared favorably to the results of Magnino et al,[21] a study that was designed to compare sniff test evaluation to RHC measurements in controlled conditions (see Appendix 3). In comparison to RHC, the ML model attained almost identical accuracy and SFP/SFN rates to the cardiologist measurements. Furthermore, a chi-square test to compare the model and cardiologist prediction distributions indicated that they were statistically indistinguishable ($P = 0.98$). These results provide strong evidence that the ML model effectively replicated the performance of cardiologists in analyzing sniff tests. Furthermore, the robust performance of the model on the external test data set from the University of Montreal indicates that it was able to effectively apply clinical standards [7] in a generalizable manner, rather than focusing on some artifact unique to the UCSF data set.

ERROR MODALITY RATES. It is also worth considering that not all wrong answers carry the same clinical significance. If a model of this nature were deployed as a screening tool in a clinical setting, the most problematic error would be a SFN, ie, analyzing a scan from a patient who had an RAP in the 15 mm Hg range and placing them in the 3 mm Hg category. This would potentially result in a patient who needed immediate treatment for vascular congestion being evaluated as euvolemic and not directed to appropriate follow-up testing and treatment. Our model

results on the full test data set showed a SFN rate 16.5% and a SFP rate of 2.5%. While these rates could be problematic in clinical applications, we hypothesized that many of these misclassifications were due to poor-quality input videos or videos that did not actually show a sniff scan. When the test data set was restricted to videos that were scored highly by the IVC view classification model, the SFN rate dropped by almost two-thirds to 5.6% and the SFP rate dropped by over half to 1.1%, supporting our hypothesis. Outside of these SFN and SFP errors, the overall model performance did not change significantly between the full and high-stringency test data sets, indicating that the model was generally able to correctly classify poor-quality IVC scans. This is important for potential clinical utility, as even experienced sonographers struggle to obtain clear high-quality IVC views for some patients. We further highlight that this tool, like all diagnostic modalities utilized in clinical medicine, would be applied and interpreted in the entire clinical context of a patient presentation.

STUDY LIMITATIONS. There are several limitations to this work. As sniff scans were not labeled at the time of collection, we identified the appropriate video input from each full TTE study using view classification ML models. These models identified some videos that were not used by cardiologists to generate RAP measurements, which added noise to model training and evaluation. Using the high-stringency test data set mitigated this problem, but also excluded some valid sniff scan videos. In addition, the level of underlying variability in the cardiologist RAP measurements in this data set cannot be known, so the true performance limit of a model that "perfectly" replicates a cardiologist can only be estimated from variability levels found in other studies. Comparison to RHC provided us with a measure independent of human variability, but the size of this matched data set was limited. Furthermore, the RHC and TTE measurements were generally not taken on the same day, making it possible that the patient's RAP changed in the intervening time. Our selection of a 30-day allowable time window between the 2 studies was chosen to balance the changing RAP concern against data availability. As both the cardiologist and model evaluations were based solely on TTE data, any error due to a change in underlying RAP between the RHC and TTE measurements would be reflected equally in the cardiologist and model performance results. The study cohort also included relatively few patients with conditions like more-than-moderate tricuspid regurgitation or liver transplants which can impact the reading of sniff scans, so caution may be warranted in applying the model to these populations. Finally, the application of the current model as a clinical tool would be limited to medical settings with access to a high-end cardiac ultrasound device and a trained sonographer, as these were the conditions under which all training and evaluation data were acquired.

CONCLUSIONS

ML can be used to measure RAP accurately and robustly from ultrasound scans of the IVC. Future work will focus on evaluating and fine-tuning the ML models using data from low-cost point-of-care ultrasound devices, as well as using the IVC view recognition models that were developed to help guide novice ultrasound operators to acquire sniff scans on such devices. These developments would make vascular congestion screening possible in primary care offices or even home health settings, reducing the need for costly specialist referrals and improving the standard of care for heart failure patients.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Tison has received research grants from General Electric. Dr Avram has received speaker fees from Abbott, Boston Scientific, Boehringer-Ingelheim, and Novartis. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose. Dr Abu-Mostafa received research grants from three internal Caltech research endowments: the Merkin Institute for Translational Research (#13520291), The Sensing to Intelligence Initiative (#13520296), and the Gates-Grubstake fund (#101170). Dr Tison received support from the National Institutes of Health (Grants NHLBI K23HL135274, R56HL161475, and DP2HL174046). Dr Padmanabhan received support from the National Institutes of Health (Grant NHLBI K08HL157700), Michael Antonov Charitable Foundation, and Frank A. Campini Foundation. Dr Avram received support from the Fonds de la recherche en santé du Québec (Grant 312758), by CIFAR, by the Montreal Heart Institute Research Centre, the Montreal Heart Institute Foundation, and by the Des Groseillers-Bérard Research Chair. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

ADDRESS FOR CORRESPONDENCE: Dr Geoffrey H. Tison, University of California-San Francisco, 555 Mission Bay Blvd South, San Francisco, California 94158, USA. E-mail: geoff.tison@ucsf.edu. OR Dr Arun Padmanabhan, University of California-San Francisco, 555 Mission Bay Blvd South, San Francisco, California 94158, USA. E-mail: arun.padmanabhan@ucsf.edu.

PERSPECTIVES

COMPETENCY IN MEDICAL KNOWLEDGE: ML can automatically measure RAP from an ultrasound video of the IVC with similar accuracy to a cardiologist.

TRANSLATIONAL OUTLOOK: ML-based tools could be used to aid nonspecialized providers in quickly obtaining RAP measurements and in assessing risk of intravascular congestion. This may enable automated ultrasound-based assessments without needing to wait for a full echocardiogram study.

REFERENCES

1. Jackson SL, Tong X, King RJ, Loustalot F, Hong Y, Ritchey MD. National burden of heart failure events in the United States, 2006 to 2014. *Circ Heart Fail*. 2018;11:e004873.

2. Koratala A, Ronco C, Kazory A. Diagnosis of fluid overload: from conventional to contempo- rary concepts. *Cardiorenal Med*. 2022;12:141-154.

3. Ellison DH, Felker GM. Diuretic treatment in heart failure. *N Engl J Med*. 2017;377:1964-1975.

4. Hughes RE, Magovern GJ. The relationship between right atrial pressure and blood volume. *AMA Arch Surg*. 1959;79:238-243.

5. Hull JV, Padkins MR, El Hajj S, et al. Risks of right heart catheterization and right ventricular biopsy: a 12-year, single-center experience. *Mayo Clin Proc*. 2023;98:419-431.

6. Rudski LG, et al. Guidelines for the echocardiographic assessment of the right heart in adults: a report from the American society of echocardiography. *J Am Soc Echocardiogr*. 2010;23:685-713.

7. Lang RM, Badano LP, Mor-Avi V, et al. 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*. 2015;28:1-39.e14.

8. Rollas K, Kilicaslan B, Erden A, Kilicaslan I, Ortac Ersoy E, Akinci SB. The interrater reliability of inferior vena cava ultrasonography performed by intensive care fellows. *J Crit Intensive Care*. 2021;12:32-36.

9. Senthilnathan M, Kundra P, Mishra S, Velayudhan S, Pillai A. Competence of intensivists in focused transthoracic echocardiography in intensive care unit: a prospective observational study. *Indian J Crit Care Med*. 2018;22:340-345.

10. Bowra J, Uwagboe V, Goudie A, Reid C, Gillett M. Interrater agreement between expert and novice in measuring inferior vena cava diameter and collapsibility index. *Emerg Med Australas*. 2015;27:295-299.

11. Fields JM, Lee PA, Jenq KY, Mark DG, Panebianco NL, Dean AJ. The interrater reliability of inferior vena cava ultrasound by bedside clinician sonographers in emergency department patients. *Acad Emerg Med*. 2011;18:98-101.

12. Randazzo MR, Snoey ER, Levitt MA, Binder K. Accuracy of emergency physician assessment of left ventricular ejection fraction and central venous pressure using echocardiography. *Acad Emerg Med*. 2003;10:973-977.

13. Finnerty NM, Panchal AR, Boulger C, et al. Inferior vena cava measurement with ultrasound: what is the best view and best mode? *West J Emerg Med*. 2017;18:496.

14. Weekes AJ, Tassone HM, Babcock A, Quirke DP, Norton HJ, Jayarama K, Tayal VS. Comparison of serial qualitative and quantitative assessments of caval index and left ventricular systolic function during early fluid resuscitation of hypotensive emergency department patients. *Acad Emerg Med*. 2011;18:912-921.

15. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE; 2019:6202-6211.

16. Albani S, Pinamonti B, Giovinazzo T, et al. Accuracy of right atrial pressure estimation using a multi-parameter approach derived from inferior vena cava semi-automated edge-tracking echocardiography: a pilot study in patients with cardiovascular disorders. *Int J Cardiovasc Imag*. 2020;36:1213-1225.

17. Wshah S, Xu B, Bates J, Morrissette K. Deep fusion of ultrasound videos for furosemide classification. In: *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*. Kolkata, India: IEEE; 2022.

18. Blaivas M, Blaivas L, Philips G, et al. Development of a deep learning network to classify inferior vena cava collapse to predict fluid responsiveness. *J Ultrasound Med*. 2021;40:1495-1504.

19. Mesin L, Roatta S, Pasquero P, Porta M. Automated volume status assessment using inferior vena cava pulsatility. *Electronics*. 2020;9:1671.

20. Millington SJ. Ultrasound assessment of the inferior vena cava for fluid responsiveness: easy, fun, but unlikely to be helpful. *Can J Anaesth*. 2019;66:633-638.

21. Magnino C, Omedè P, Avenatti E, et al. Inaccuracy of right atrial pressure estimates through inferior vena cava indices. *Am J Cardiol*. 2017;120:1667-1673.

KEY WORDS artificial intelligence, deep learning, echocardiography, heart failure, vascular congestion

APPENDIX For supplemental information including tables and figures, please see the online version of this paper.