EBioMedicine 73 (2021) 103613

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

Deep learning evaluation of biomarkers from echocardiogram videos

J Weston Hughes[a], Neal Yuan[b], Bryan He[a], Jiahong Ouyang[c], Joseph Ebinger[b], Patrick Botting[b], Jasper Lee[b], John Theurer[b], James E. Tooley[d], Koen Nieman[d,e], Matthew P. Lungren[e], David H. Liang[d], Ingela Schnittger[d], Jonathan H. Chen[d], Euan A. Ashley[d], Susan Cheng[b], David Ouyang[b,1,2,*], James Y. Zou[a,c,f,1,2,*]

[a] Department of Computer Science, Stanford University, Palo Alto, CA 94025
[b] Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, 90048
[c] Department of Electrical Engineering, Stanford University, Palo Alto, CA, 94025
[d] Department of Medicine, Stanford University, Palo Alto, CA, 94025
[e] Department of Radiology, Stanford University, Palo Alto, CA, 94025
[f] Department of Biomedical Data Science, Stanford University, Palo Alto, CA, 94025

ARTICLE INFO

Article History:
Received 11 July 2021
Revised 16 September 2021
Accepted 20 September 2021
Available online xxx

Keywords:
Deep learning
Artificial intelligence
Echocardiography

ABSTRACT

Background: Laboratory testing is routinely used to assay blood biomarkers to provide information on physiologic state beyond what clinicians can evaluate from interpreting medical imaging. We hypothesized that deep learning interpretation of echocardiogram videos can provide additional value in understanding disease states and can evaluate common biomarkers results.

Methods: We developed EchoNet-Labs, a video-based deep learning algorithm to detect evidence of anemia, elevated B-type natriuretic peptide (BNP), troponin I, and blood urea nitrogen (BUN), as well as values of ten additional lab tests directly from echocardiograms. We included patients (n = 39,460) aged 18 years or older with one or more apical-4-chamber echocardiogram videos (n = 70,066) from Stanford Healthcare for training and internal testing of EchoNet-Lab's performance in estimating the most proximal biomarker result. Without fine-tuning, the performance of EchoNet-Labs was further evaluated on an additional external test dataset (n = 1,301) from Cedars-Sinai Medical Center. We calculated the area under the curve (AUC) of the receiver operating characteristic curve for the internal and external test datasets.

Findings: On the held-out test set of Stanford patients not previously seen during model training, EchoNet-Labs achieved an AUC of 0.80 (0.79-0.81) in detecting anemia (low hemoglobin), 0.86 (0.85-0.88) in detecting elevated BNP, 0.75 (0.73-0.78) in detecting elevated troponin I, and 0.74 (0.72-0.76) in detecting elevated BUN. On the external test dataset from Cedars-Sinai, EchoNet-Labs achieved an AUC of 0.80 (0.77-0.82) in detecting anemia, of 0.82 (0.79-0.84) in detecting elevated BNP, of 0.75 (0.72-0.78) in detecting elevated troponin I, and of 0.69 (0.66-0.71) in detecting elevated BUN. We further demonstrate the utility of the model in detecting abnormalities in 10 additional lab tests. We investigate the features necessary for EchoNet-Labs to make successful detection and identify potential mechanisms for each biomarker using well-known and novel explainability techniques.

Interpretation: These results show that deep learning applied to diagnostic imaging can provide additional clinical value and identify phenotypic information beyond current imaging interpretation methods.

Funding: J.W.H. and B.H. are supported by the NSF Graduate Research Fellowship. D.O. is supported by NIH K99 HL157421-01. J.Y.Z. is supported by NSF CAREER 1942926, NIH R21 MD012867-01, NIH P30AG059307 and by a Chan-Zuckerberg Biohub Fellowship.

1. Introduction

Diagnostic medical testing provides insight into human physiology and disease conditions, with testing ranging from blood based biomarkers and genetics testing to imaging studies that provide deep insight into anatomy and changes over time [1–3]. Blood based

* Corresponding author.
E-mail addresses: david.ouyang@cshs.org (D. Ouyang), jamesz@stanford.edu (J.Y. Zou).
[1] Co-senior author
[2] To whom correspondence should be addressed.

https://doi.org/10.1016/j.ebiom.2021.103613

Research in context

Evidence before this study

In the last few years, deep learning algorithms applied to medical imaging have demonstrated the ability to identify clinical phenotypes and diagnoses beyond conventional use of imaging. Electrocardiogram (ECG) waveforms have been shown to identify signals for hyperkalemia, heart failure, anemia, age, sex, mortality, and other phenotypes not traditionally associated with ECGs. Echocardiograms, or cardiac ultrasounds, provide high temporal and spatial resolution videos of the heart and applying deep learning algorithms have shown the ability to estimate age and sex and predict mortality.

Added value of this study

This is the first study to our knowledge to use deep learning on echocardiograms to evaluate the values of numerous biomarkers, including both biomarkers related to heart health (troponin I and B-type natriuretic peptide) as well as systemic disease (hemoglobin for anaemia and blood urea nitrogen for kidney disease). This algorithm was validated in two separate healthcare systems to demonstrate robustness and to avoid bias or overfitting. Interestingly, the estimated B-type natriuretic peptide levels were more predictive of subsequent heart failure hospitalization than the actual biomarker levels.

Implications of all the available evidence

Our study suggests that important disease and physiology specific patterns are present in cardiac videos that might be invisible to the human eye but contain important diagnostic and prognostic information. The ability to identify patients with abnormal biomarkers has important implications for disease screening and patient management.

laboratory testing is a fundamental tool for disease diagnosis and management as changes in assayable biomarkers can be some of the earliest signs of physiological perturbations [5–7]. Despite the frequent utilization of both laboratory testing and medical imaging in routine clinical practice, the deeper connections between medical images and biomarker values are relatively underexplored [4]. It remains unknown whether routinely obtained imaging studies might contain information that can broadly estimate common biomarker values and more deeply inform clinicians about the patient condition.

Recent advances in Artificial Intelligence have shown that deep learning applied to medical images can identify phenotypes beyond what is currently possible by observation from human clinicians alone [8–11]. Such discoveries have spanned across a variety of imaging modalities in many medical specialties and have uncovered imaging correlates for a wide range of disease states, molecular signatures, and physiologic conditions [12–15]. Given that orthogonal and complimentary information is obtained from the many different forms of diagnostic testing, subtle associations and relationships can be missed in conventional clinical assessment.

Echocardiograms, or cardiac ultrasounds, are the most common form of cardiovascular imaging, combining rapid image acquisition, lack of ionizing radiation, and high temporal resolution to capture spatiotemporal information on cardiac motion and function [16,17]. Previous works have shown deep learning based assessment of echocardiograms can identify physiological state and hints of both systemic as well as cardiac diseases [9,18,19]. In the extremes, abnormal blood chemistry can influence cardiac function [20], and over time, structure, but it is unknown whether transient or subtle variations biomarkers are reflected in the physiologic state that can be extracted from medical imaging. A deep learning assessment of frequently obtained, no radiation, low cost, and information dense imaging, such as echocardiogram videos, could provide additional diagnostic information that alleviates the need for other invasive, costly, or burdensome forms of testing. This is the first demonstration that echocardiograms can be used to detect abnormal blood biomarkers through deep learning analysis of the ultrasound videos, and our artificial intelligence algorithms generalize imaging and biomarker results across healthcare systems.

2. Methods

2.1. Data sources, study population, and ethics

We used the Stanford Research Repository (STARR) and the Echocardiography Lab Database to identify the population of patients who received at least one lab test and one echocardiogram study at Stanford Healthcare. We selected 108,521 full transthoracic echocardiogram studies between January 1, 2006 and December 31, 2018. Of the 108,521 studies, 70,066 studies had associated biomarker information within 30 days between echocardiogram study and biomarker blood draw. An additional external test dataset of 1,301 videos from January 1, 2019 to June 30, 2019 with corresponding biomarker results was obtained from Cedars-Sinai Medical Center and processed using the same preprocessing pipeline and used as an external test dataset without further model training or fine tuning.

A single apical-4-chamber 2D gray-scale video was identified from each study and used to represent the study for mapping to laboratory values. Previously described methods [19] were used to preprocess echocardiogram videos to standard resolution and remove extra information outside of the ultrasound sector such as text, ECG and respirometer data, as well as identifying information. The 70,066 videos were split by patient into 59,434 videos for training, 5319 videos for validation, and 5313 videos for internal testing, such that the same patient never appeared in multiple splits of the data. If there were multiple videos from the same patient, we treated them as individual samples for training. During model training, the lab closest in time to each video was used as the training label, and videos were excluded from training if the patient did not have a corresponding video-laboratory value pair. This research was approved by the Stanford University (Protocol 43721) and Cedars-Sinai Medical Center (STUDY00001049) Institutional Review Boards.. This research was deemed minimal risk and a waiver of consent was obtained for retrospective chart review of deidentified medical information.

2.2. Outcomes for model estimation

The primary outcome of the study was the ability of the AI-enhanced echocardiogram video to identify patients with abnormal lab biomarkers. We chose the most relevant cardiac biomarkers (troponin I, BNP) as well as commonly obtained biomarkers (hemoglobin, white blood cell count, platelet count, sodium, chloride, BUN, creatinine, aspartate aminotransferase, and alanine aminotransferase) and biomarkers of relevant complementary disease states (hemoglobin A1c, and C-reactive protein) as targets for our study. Binary thresholds for model performance assessment were determined by the reference range of the particular laboratory's assay, and for biomarkers with significant variance, model training was performed on the logarithm of the result value (Supplemental Tables 1). Notably, the EchoNet-Labs estimate for BNP was trained on paired echocardiogram videos and NT-proBNP results from Stanford Medicine, and tested on BNP data from Cedars-Sinai, which uses a different assay. In addition to different patient demographics, date ranges of data acquisition, and geographic locations, the two institutions use different picture archiving and communication systems (Philips Xcelera and Scimage

PICOM PACS respectively) to obtain the echocardiographic data. In the validation and test sets, the same process was applied with the additional constraint that only labels acquired within 30 days of the echocardiogram were included. In the case of CRP, a window of 365 days was used to increase sample size.

2.3. Model development and training

Models were built using Python 3.8 and PyTorch 1.4. Extending on previous work [19], EchoNet-Labs uses a (2+1)D-ResNet consisting of 34 layers of alternating spatial and temporal convolutions in a ResNet structure [24]. We chose the same hyperparameter configuration as in previous work [19] and found that architecture choice (e.g. R3D and MC3) and temporal step size (e.g. 1/1, 1/2, or 1/4 the sampling rate of the original video) do not significantly affect results. All models were pretrained on the Kinetics-400 dataset [25]. Independent video regression models were trained for each lab value, taking as input a randomly selected $32 \times 112 \times 112$ sub-video and estimating the lab value. Training a single model to estimate all lab values decreased or did not change performance on the top performing lab values. Videos were augmented during training by randomly shifting by up to 12 pixels.

The models were trained to minimize the mean squared error between the estimated and true lab values. Model training used a stochastic gradient descent optimizer with an initial learning rate of 0.001, momentum of 0.9, and batch size of 20 for 45 epochs. The learning rate was decayed by a factor of 0.1 every 15 epochs. Estimating was set up as a binary classification task, detecting abnormal versus normal lab value, based on standard thresholds. For these biomarkers, clinicians recognize inherent heterogeneity on retesting and often make clinical decisions on whether broadly these biomarkers are either normal or abnormal. To understand model generalization, each model was evaluated on a held-out test set not used in any way during model development, from a set of patients completely disjoint from those used during training. Finally, for the four most successful biomarkers we report results on the Cedars-Sinai external validation dataset. For each lab, we report the AUC on the validation and test sets, with bootstrapped 95% confidence intervals. We additionally compute the positive predictive value (PPV), negative predictive value (NPV), recall, and F1 score on the test set, based on the optimal F1 cutoff on the validation set. To understand the impact of input sample size on EchoNet-Labs, we trained separate models with datasets at different sized subsets for each biomarker. Models were trained by randomly selecting 1000, 2000, 4000, and 8000 training examples for each model. We also explored training a single model to estimate all values through multi-task learning, but found for key lab values that training individual models performed better.

For all models, the weights from the epoch with the highest validation AUC was selected for final testing. Our final model averaged estimates across the entire echocardiogram video over all possible 32 frame sub-videos rather than randomly selecting one to account for potential variance between beats. We report area under the receiver operator characteristic curve (AUC) as the primary performance metric in Figure 2 and Supplementary Table 3. All confidence intervals are 95% confidence intervals generated by bootstrapping on the relevant test set. Estimating a single lab value with EchoNet-Labs, with all test-time augmentation, takes less than 5 seconds.

2.4. Model interpretation

To further understand the features needed to make classifications, we retrained models for anemia, BNP, troponin I, and BUN on differently ablated inputs. For each input ablation, we trained and tested on identically ablated data. To understand if motion-based features are necessary for classification, we trained and tested a model on a single randomly selected frame of each echo, repeated 32 times in a video to fairly compare to other 3D ResNet models. To understand if the motion of the left ventricle on its own is sufficient for classification, we trained and tested a model on a video of the segmented outline of the left ventricle generated by Echonet-Dynamic, with none of the original video data present. To understand if only the information in and around the left ventricle is sufficient to classify, we trained and tested a model with all data outside of a bounding box around the left ventricle obscured. To produce a video of just the left ventricle, we found the smallest bounding box which contained the left ventricle in all frames, expanded it by 5 pixels, and set all pixels outside of that region to 0.

2.5. Comparison to benchmark for event prognostication

One way a model might learn to estimate a biomarker value would be to identify features which are already known to be contained in echocardiogram data, and use those covariates as well as discrete demographic information to estimate the biomarker value. Age and sex have been previously shown to be associated from echocardiogram videos with high accuracy [9,19], and echocardiogram videos contain information about left ventricular ejection fraction, heart rate, and right ventricular systolic pressure. To determine if the model truly learned novel features, we trained logistic regression and XGBoost models using these demographics and echocardiography derived metrics to compare with EchoNet-Labs. To evaluate the prognostic value of estimated vs. actual biomarker values, single variable logistic regression was performed on the input biomarker value to predict 1-year heart failure hospitalization as identified by the electronic health record through admission ICD9 code.

2.6. Role of the funding source

No entity other than the authors listed played any role in the design of the study; the collection, analysis, or interpretation of the data; writing of the report; or in the decision to submit the paper for publication.

3. Results

We developed a deep learning framework, EchoNet-Labs, to answer whether medical imaging might be able to estimate biomarker values and whether these results generalize across different clinical settings and healthcare systems (Fig. 1). EchoNet-Labs is a convolutional neural network with residual connections and spatio-temporal convolutions that provides a beat-by-beat estimate for biomarker values. Extending our prior work on deep learning applied to echocardiogram videos [19], EchoNet-Labs incorporates both spatial and temporal information to perform both regression and classification tasks. To train and evaluate EchoNet-Labs, we curated a dataset of 70,066 echocardiogram videos from 39,460 patients at Stanford Medicine and 1,301 videos from 819 patients from Cedars-Sinai Medical Center. Echocardiogram videos from Stanford Medicine were preprocessed and curated for apical 4-chamber view videos and divided based on patient into 59,434 training, 5,319 validation, and 5,313 internal test examples. An additional dataset of 1,301 apical 4-chamber view videos from Cedars-Sinai Medical Center were never seen during model training and served as a hold-out external test set for this study (Table 1).

On the held-out test set from Stanford Medicine that was not previously seen during model training, EchoNet-Labs estimated biomarker values from echocardiogram videos with high sensitivity and specificity (Fig. 2). EchoNet-Labs achieved an area under the curve (AUC) of 0.80 (0.79-0.81) in detecting anemia (low hemoglobin), of 0.86 (0.85-0.88) in detecting elevated BNP, of 0.75 (0.73-0.78) in detecting elevated troponin I, of 0.74 (0.72-0.76) in detecting

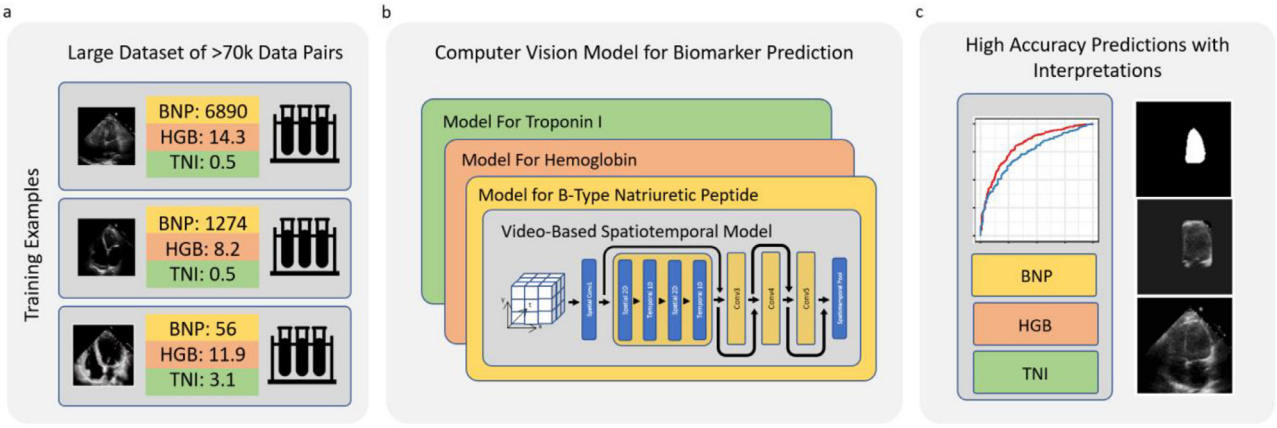

Fig. 1. Overview of EchoNet-Labs system and study design. a. A training dataset of over seventy thousand echocardiogram videos and paired biomarker values from the same patient were used to train a video-based AI system for estimation of laboratory values. b. Our deep learning based AI system used spatio-temporal convolutions to infer biomarker values from both anatomic (spatial) and physiologic (temporal) information contained with echocardiogram videos. c. To understand the relative importance of spatial and temporal information, ablation datasets removing texture, motion, and extracardiac structures were adopted to perform interpretations experiments.

elevated BUN, and up to 0.72 in detecting abnormalities in ten other common laboratory tests (Supplementary Table 2). EchoNet-Labs' also provides calibrated uncertainties for each of its estimates using Platt scaling (Supplementary Fig. 2).

To assess the cross-healthcare-system reliability of the model, EchoNet-Labs was additionally tested, without any tuning, on an external test dataset of 1,301 patients from Cedars-Sinai Medical Center. On the external test dataset, EchoNet-Labs achieved an AUC of 0.80 (0.77-0.82) in detecting anemia, of 0.82 (0.79-0.84) in detecting elevated BNP, of 0.75 (0.72-0.78) in detecting elevated troponin I, and of 0.69 (0.66-0.71) in detecting elevated BUN, which is similar to the model's accuracy on the Stanford test patients.

To assess the clinical utility of EchoNet-Labs generated lab values, we compared the performance of estimated BNP levels vs. true BNP levels for predicting future heart failure hospitalization within 1 year. Prediction of heart failure hospitalization using estimates BNP values achieved an AUC of 0.76 (0.72 – 0.80) while actual BNP values achieved an AUC of 0.71 (0.67 – 0.76), suggesting deep learning estimation of biomarker values can identify additional information about disease states not fully represented by the standard, often noisy, measurements of individual biomarkers. EchoNet-Dynamic's performance was superior to a model for predicting rehospitalization from standard echocardiographic parameters with an AUC of 0.55 (0.53-0.57), suggesting that there is additional prognostic information in the echocardiogram videos not fully captured by human interpretation or clinical biomarker assays.

To provide context for EchoNet-Labs' estimation results, we also trained logistic regression and XGBoost models to estimate each biomarker using demographics and standard quantitative metrics from echocardiograms (age, gender, race, ethnicity, heart rate, LVEF, RVSP,

Table 1
Baseline characteristics of the Stanford and Cedars-Sinai study participants.

	Stanford Total	Training	Validation	Test	CSMC External Test
Number of Patients	40,104	33,662	3,223	3,339	1,301
Number of echocardiogram studies	70,066	59,433	5,319	5,313	1,301
Demographics					
Age, years (SD)	60.1 (16.9)	60.0 (16.9)	60.6 (16.6)	61.0 (16.5)	70.3 (21.1)
Female, n (%)	31,134 (44.4)	26,329 (44.3)	2,358 (44.3)	2,447 (46.1)	536 (41.2)
Characteristics					
Heart Failure, n (%)	14,688 (21.0)	12,231 (20.6)	1,273 (23.9)	1,274 (24.0)	275 (21.2)
Diabetes Mellitus, n (%)	13,617 (19.4)	11,335 (19.1)	1,131 (21.3)	1,204 (22.6)	281 (21.6)
Hypertension, n (%)	26,074 (37.2)	21,795 (36.8)	2,144 (40.3)	2,217 (41.7)	430 (33.1)
Hyperlipidemia, n (%)	21,737 (31.0)	18,224 (30.7)	1,747 (32.8)	1,846 (34.7)	320 (24.6)
Coronary Artery Disease, n (%)	14,731 (21.0)	12,312 (20.7)	1,223 (23.0)	1,268 (23.9)	382 (29.4)
Renal Disease, n (%)	11,254 (16.1)	9.266 (15.6)	996 (18.7)	1,042 (19.6)	391 (30.0)
Biomarker					
B-type Natiuretic Peptide (BNP), n (% above threshold)	23,451 (0.70)	21,011 (0.69)	1,179 (0.73)	1,261 (0.76)	898 (0.70)
Hemoglobin, n (% above threshold)	58,048 (0.85)	50,256 (0.86)	3,912 (0.79)	3,880 (0.79)	1,226 (0.69)
Troponin I, n (% above threshold)	15,847 (0.63)	14,599 (0.63)	629 (0.64)	619 (0.60)	683 (0.71)
Blood urea nitrogen (BUN), n (% above threshold)	9,134 (0.18)	8,088 (0.18)	542 (0.22)	504 (0.20)	1,056 (0.29)
Creatinine, n (% above threshold)	12,221 (0.17)	10,212 (0.17)	991 (0.19)	1,018 (0.20)	NA
Alanine aminotransferase (ALT), n (% above threshold)	959 (0.01)	781 (0.01)	91 (0.02)	87 (0.02)	NA
C-Reactive Protein (CRP), n (% above threshold)	9,684 (0.54)	8,612 (0.54)	545 (0.55)	527 (0.61)	NA
Aspartate aminotransferase (AST), n (% above threshold)	970 (0.01)	779 (0.01)	93 (0.02)	98 (0.02)	NA
Sodium, n (% above threshold)	68,352 (0.99)	58,366 (0.99)	5,003 (0.97)	4,983 (0.97)	NA
White blood cell (WBC) count, n (% above threshold)	6,266 (0.10)	5,114 (0.09)	583 (0.12)	569 (0.12)	NA
Hemoglobin A1c, n (% above threshold)	9,083 (0.22)	8,004 (0.21)	539 (0.32)	540 (0.31)	NA
Alkaline Phosphatase, n (% above threshold)	4,776 (0.13)	3,991 (0.11)	359 (0.19)	426 (0.22)	NA
Platelet, n (% above threshold)	54,607 (0.80)	47,180 (0.81)	3,772 (0.77)	3,655 (0.75)	NA
Chloride, n (% above threshold)	21,318 (0.31)	18,169 (0.31)	1,617 (0.31)	1,532 (0.30)	NA

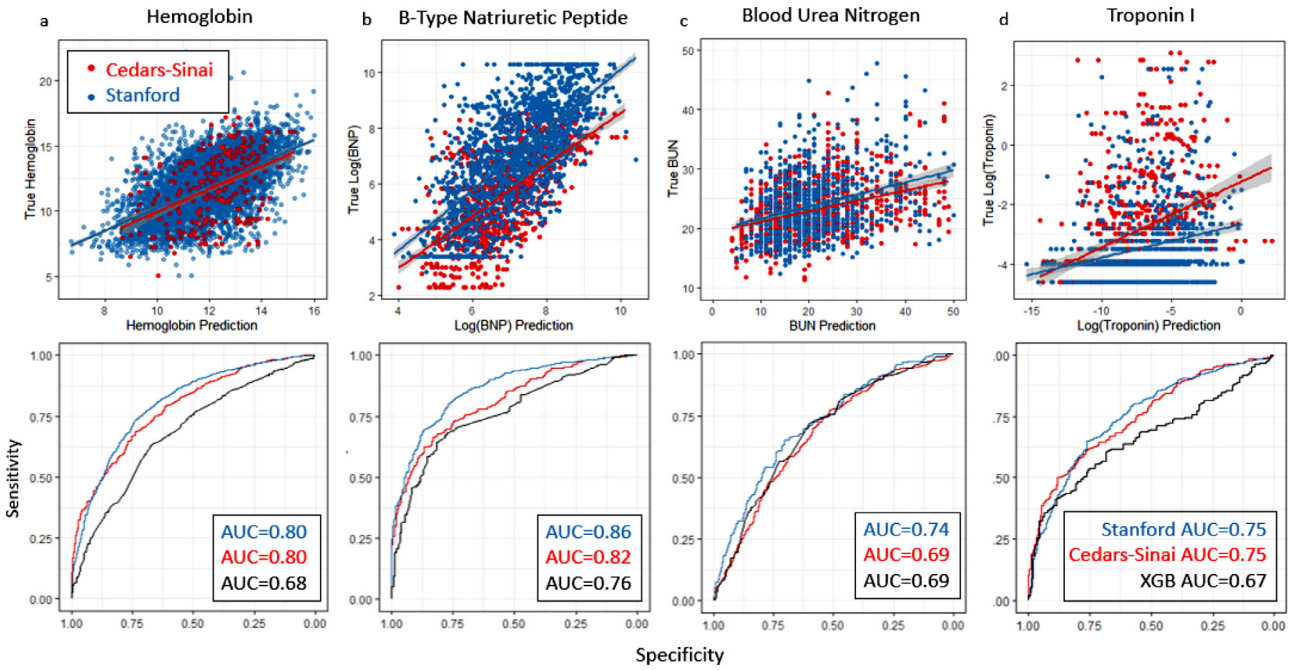

Fig. 2. Performance of EchoNet-Labs on Internal and External Test Datasets. a-d. Scatterplots (top) and receiver-operating characteristic (ROC) curves (bottom) for detection of abnormal (a) hemoglobin (Stanford n=3,880, CSMC n=1,226), (b) B-Type Natriuretic Peptide (1,261, 898), (c) Blood Urea Nitrogen (504, 1,056), and (d) Troponin I (619, 683). Blue points and curves denote to a held-out test set of patients from Stanford Medicine not previously seen during model training. Red points and curves denote to performance on the external test set from Cedars-Sinai Medical Center. Black curves denote a benchmark with XGBoost using demographics and echocardiogram features on the Stanford test set.

E/e', TVVel, LA volume, MV_E, A wave, e', E/A). This baseline models achieved AUC of 0.65-0.68, 0.77, 0.67, and 0.65-0.68 for detecting anemia, and abnormal BNP, troponin I, and BUN respectively, which is substantially lower than EchoNet-Labs' performance. This comparison suggests that EchoNet-Labs captures novel features in the videos beyond correlates of patient demographics and commonly annotated cardiac features.

We primarily evaluate the ability of EchoNet-Labs to classify lab values above or below clinically meaningful cutoffs because the most straightforward to clinically interpret. EchoNet-Labs estimation can also be thresholded at different cutoffs to classify different levels of the same lab value. To demonstrate this, we used the EchoNet-Labs hemoglobin model to detect severe anemia (hemoglobin below 10), moderate to severe anemia (hemoglobin below 13), and polycythemia (hemoglobin above 16). For all three tasks, AUCs were between 0.767 and 0.790 (Supplementary table 3).

To investigate which features are most relevant to EchoNet-Labs' estimation of each biomarker, we trained a series of models on various transformations of the input data to remove different types of information (Fig. 3). This demonstrates that both the motion of the ventricle in the absence of fine-grained pixel and texture, and the fine-grained pixel and texture in the absence of motion information,

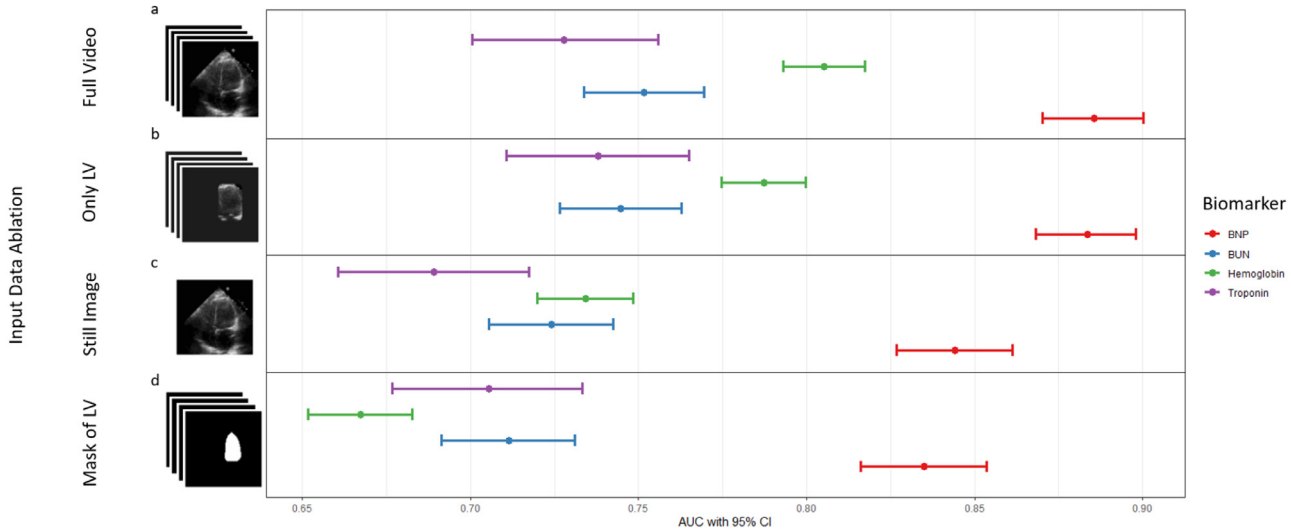

Fig. 3. EchoNet-Labs input ablations and impact on model performance. Experiments showing performance of models trained on ablated input data that hides specific information. (a) Results on standard video input. (b) Results with input where region outside of the left ventricle are obscured. (c) Results with removing temporal information with single frame input. (d) Results with removing texture information and showing location of the left ventricle. For each ablation setting, a separate model was trained on that type of ablated data to quantify the information content in the data. The width of each bar indicates the bootstrap 95% CI for each detection AUC. LV = Left Ventricle.

each contain a large amount of the information. EchoNet-Labs achieved high performance in detecting elevated BNP, Troponin I, and BUN based on video of only the region around the left ventricle (AUCs of 0.88, 0.74, and 0.74 respectively, compared to 0.89, 0.73, and 0.75 on the full video), suggesting information from that region alone might be sufficient for biomarker estimation. EchoNet-Labs' performance was slightly worse but still quite accurate when making estimation based on video of only the tracing of the left ventricle endothelium (AUCs of 0.84, 0.71, and 0.71), and based on a single randomly selected frame of video (AUCs of 0.84, 0.69, and 0.71). When evaluating results of detecting anemia, model performance depended more on texture information as performance was greatly limited by restricting input to the left ventricular border (AUC dropped from 0.81 to 0.67, versus 0.73 on single frame).

We performed sensitivity analysis with regard to training sample size to understand the effect of quantity of data (Supplementary Fig. 1). Upward trends were observed for all values as dataset size increased without a clear inflection point, suggesting that further growth in the dataset size could further increase accuracy. In particular, doubling the size of the datasets consistently leads to uniform increases, suggesting that partnering with other healthcare systems to produce multiplicatively larger datasets would lead to further gains in accuracy. Even with large sample sizes of up to 58,000 training examples, we do not see an inflection in improvement in performance, suggesting EchoNet-Labs can be improved with additional training examples.

To understand whether the model performed differently on different subgroups, we compared performance on the Stanford test set in racial, ethnic, and age subgroups, as well as subgroups based on normal, mildly reduced, and moderately-severely reduced LVEF (Supplementary Fig. 3). Not all patients had demographic information available, and subgroups are based only on patients confirmed to be members of that subgroup. We found that the model did not perform significantly worse on any racial or ethnic group, although in some cases data was limited leading to wide confidence intervals. The model was more accurate in detecting anemia in patients above the age of 60 and in detecting abnormal BNP in patients with moderately reduced LVEF, and was less accurate in detecting abnormal BNP in patients with normal LVEF and BUN in patients with moderately-severely reduced LVEF.

4. Discussion

EchoNet-Labs is a video-based deep learning algorithm that achieves state-of-the-art estimation of biomarkers from echocardiogram videos. Using 70,066 echocardiogram videos and paired biomarker results, EchoNet-Labs has high accuracy in detecting abnormal hemoglobin, BNP, troponin I, and BUN, and this performance was superior to a model using traditional risk factors. The model performance was robust to changing the clinical environment, and experiments degrading the input data show EchoNet-Labs incorporates both motion and texture based information for its assessment. The results of this study support a growing body of literature highlighting that deep learning analysis of medical imaging can identify correlative findings of systemic physiology that was previously thought to be only obtained from orthogonal diagnostic testing [8,23].

With deep learning models and model interpretation techniques, our study highlights the association between imaging phenotypes and biomarkers of both cardiovascular and systemic disease. Echocardiogram videos are commonly used to diagnose heart failure, which has a strong association with some biomarkers (e.g. BNP) and can help explain the strong performance of EchoNet-Labs for BNP. Similarly troponin I is most abundantly found in cardiac myocardium and is frequently used as a marker of myocardial injury and myocardial infarction. Our analysis was restricted to a corpus of apical-4-chamber view echocardiogram videos, which does not capture all the relevant structures seen on ultrasonographic study, and future work should evaluate the benefit of ensembling additional information and views to identify prognostic features. For example, the apical-4-chamber view does not contain the posterior wall and the addition of other echocardiographic videos can more fully capture the relationship with troponin I, particularly if capturing interpretable features such as posterior wall motion abnormalities. Such localization might be more tenuous for evaluation of hemoglobin and other biomarkers less associated with particular structures, however future experiments could identify which cardiac structures could explain strong associations with biomarkers.

Surprisingly, we show disease states and biomarkers not directly related to cardiovascular function can be readily estimated from echocardiogram videos. While critical illness can manifest with multi-system presentations (anemia of chronic disease and many types of cardiorenal syndrome), biomarkers most commonly associated with kidney function and cellular proliferation might be more strongly correlated with cardiac function than previously shown, extending prior work in other modalities that show medical imaging might have additional value in understanding the patient condition [21,22]. Biomarkers, such as hemoglobin and BUN, are associated with systemic disease but now shown convincingly to be estimated accurately both by imaging and electrical signals of the cardiovascular system [23]. How hemoglobin and BUN values are associated with cardiac motion has not been previously characterized. Our findings provide the first evidence that variation in these values are visually detectable in heart motion. The physiological response to anemia includes tachycardia and compensatory changes in cardiac function which could be picked up by deep learning models in the detection of abnormal hemoglobin. Improved understanding of the close relationships between imaging and laboratory testing can lead to further understanding of the relationship between imaging phenotypes and disease processes.

While laboratory testing is cheaper than echocardiogram studies, there might be multiple reasons and scenarios where deep learning generated lab values can be useful. First, we show that EchoNet-Labs estimates for BNP are more prognostic for heart failure admissions than from conventional lab testing, suggesting that the model could have denoising properties which deserves further exploration. Second, although laboratory testing is more common than echocardiograms, it is far from ubiquitous and requires phlebotomy. In our cohort of all echocardiograms, 35.4% of studies did not have any laboratory testing within 30 days. Additionally, for studies with any associated biomarkers, 26.7% of other biomarkers were not obtained at the same time. EchoNet-Labs can fill in the lab values for these patients, and facilitate downstream screening, longitudinal analysis and clinical follow-up.

Performance in deep learning estimates of biomarkers varied considerably by biomarker, with the highest AUCs for some biomarkers associated with cardiovascular disease (troponin I and BNP) while other blood chemistries had less dynamic range and were not able to be estimated confidently. Integrating echocardiograms and lab values can help inform the interpretation of both tests and in doing so provide an overall more accurate picture of disease. It may also help clarify how certain lab abnormalities might correspond to changes in cardiac structure and function. Additionally, our experiments suggest EchoNet-Labs can continue to be improved with additional training examples, which suggest a promising direction of further exploration. The ability of EchoNet-Labs to evaluate laboratory values from echocardiogram does not imply that a causal relationship between them, however it demonstrates that these different phenotyping modalities capture common patient and disease information not previously identified. Further work must be undertaken to understand how sex and age can be evaluated from fundoscopic imaging and biomarkers can be evaluted from echocardiogram videos.

If proven to be reliable, biomarker evaluation from fast, cheap imaging could be useful in numerous clinical contexts. In the emergency room, point-of-care echocardiography is already used to triage procedures and assist medical decision-making in medical emergencies. While laboratory testing requires phlebotomy and processing, often offsite, ultrasound is often readily available and rapidly attained, even in resource-limited settings. As a rapid adjunct to conventional testing, EchoNet-Labs can help stratify patients by risk or guide medical decision making in obtaining expensive laboratory testing when there is low clinical suspicion and low probability for abnormal testing results.

Contributors

JWH, DO, and JYZ designed the study. JWH, BH, and DO developed the neural network. **J**WH and DO did the statistical analysis. JWH and DO verified the underlying data. JWH, DO, and JYZ wrote the manuscript. JE, JB, JL, JT, DHL, JHC, and DO collected the data. All authors critically reviewed the final manuscript.

Data and code availability

All of the code for EchoNet-Labs is available at https://github.com/echonet/.

Declaration of Competing Interest

Dr Nieman reports grants from Siemens Healthineers, Bayer, HeartFlow Inc., personal fees from Siemens Medical Solutions USA, outside the submitted work. The other authors have nothing to disclose.

Acknowledgments

We thank Jessica Torres Soto, Pierre Elias, and Marco Perez for helpful feedback and discussion. J.W.H. and B.H. are supported by the NSF Graduate Research Fellowship. D.O. is supported by NIH K99 HL157421-01. J.Y.Z. is supported by NSF CAREER 1942926, NIH R21 MD012867-01, NIH P30AG059307 and by a Chan-Zuckerberg Biohub Fellowship.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2021.103613.

References

[1] Ashley EA, et al. Clinical assessment incorporating a personal genome. Lancet 2010;375:1525–35.
[2] Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148:1293–307.
[3] Jackson HW, et al. The single-cell pathology landscape of breast cancer. Nature 2020;578.
[4] Ai T, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020;296:E32–40.
[5] Grossman DC, et al. Screening for prostate cancer: US preventive services task force recommendation statement. JAMA 2018:319.
[6] Bibbins-Domingo K, et al. Screening for colorectal cancer: US preventive services task force recommendation statement. JAMA 2016:315.
[7] Owens DK, et al. Screening for HIV infection: US preventive services task force recommendation statement. JAMA 2019;321.
[8] Poplin R, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2018;2:158–64.
[9] Ghorbani A, et al. Deep learning interpretation of echocardiograms. NPJ Digit. Med. 2020;3:10.
[10] Dauvin A, et al. Machine learning can accurately predict pre-admission baseline hemoglobin and creatinine in intensive care patients. NPJ Digit. Med. 2019;2:116.
[11] Avram R, et al. A digital biomarker of diabetes from smartphone-based vascular signals. Nat. Med. 2020;26:1576–82.
[12] Attia ZI, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019;394:861–7.
[13] He B, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng. 2020;4:827–34.
[14] Kwon J-M, et al. Artificial intelligence for detecting mitral regurgitation using electrocardiography. J. Electrocardiol. 2020;59:151–7.
[15] Ko WY, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J. Am. Coll. Cardiol. 2020;75.
[16] Papolos A, Narula J, Bavishi C, Chaudhry FA, Sengupta PP. US hospital use of echocardiography: insights from the nationwide inpatient sample. J. Am. Coll. Cardiol. 2016;67:502–11.
[17] Douglas PS, et al. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 appropriate use criteria for echocardiography. A report of the American college of cardiology foundation appropriate use criteria task force, American society of echocardiography, American heart association, American society of nuclear cardiology, heart failure society of America, heart rhythm society, society for cardiovascular angiography and interventions, society of critical care medicine, society of cardiovascular computed tomography, society for cardiovascular magnetic resonance American college of chest physicians. J. Am. Soc. Echocardiogr. 2011;24:229–67.
[18] Zhang J, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation 2018;138:1623–35.
[19] Ouyang D, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252–6.
[20] Hunter RW, Bailey MA. Hyperkalemia: pathophysiology, risk factors and consequences. Nephrol. Dial. Transplant 2019;34.
[21] Attia ZI, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circ. Arrhythm. Electrophysiol. 2019;12:e007284.
[22] Raghunath S, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat. Med. 2020. doi: 10.1038/s41591-020-0870-z.
[23] Kwon J-M, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Health 2020;2:e358–67.
[24] Tran D, et al. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2018.
[25] Carreira J, Zissmerman AQuo Vadis. Action recognition? A New Model and the Kinetics Dataset. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*; 2017.