nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-020-00667-9

Check for updates

Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality

Alvaro E. Ulloa Cerna[1,2], Linyuan Jing[1], Christopher W. Good[3], David P. vanMaanen[1], Sushravya Raghunath[1], Jonathan D. Suever[1], Christopher D. Nevius[1], Gregory J. Wehner[4], Dustin N. Hartzel[5], Joseph B. Leader[5], Amro Alsaid[3], Aalpen A. Patel[6], H. Lester Kirchner[7], John M. Pfeifer[1,8], Brendan J. Carry[3], Marios S. Pattichis[2], Christopher M. Haggerty[1,3,9] and Brandon K. Fornwalt[1,3,6,9] ✉

Machine learning promises to assist physicians with predictions of mortality and of other future clinical events by learning complex patterns from historical data, such as longitudinal electronic health records. Here we show that a convolutional neural network trained on raw pixel data in 812,278 echocardiographic videos from 34,362 individuals provides superior predictions of one-year all-cause mortality. The model's predictions outperformed the widely used pooled cohort equations, the Seattle Heart Failure score (measured in an independent dataset of 2,404 patients with heart failure who underwent 3,384 echocardiograms), and a machine learning model involving 58 human-derived variables from echocardiograms and 100 clinical variables derived from electronic health records. We also show that cardiologists assisted by the model substantially improved the sensitivity of their predictions of one-year all-cause mortality by 13% while maintaining prediction specificity. Large unstructured datasets may enable deep learning to improve a wide range of clinical prediction models.

Imaging is critical to treatment decisions in most modern medical specialties and has also become one of the most data rich components of the electronic health record (EHR). For example, during a single routine ultrasound of the heart (an echocardiogram), approximately 10–50 videos (around 3,000 images) are acquired to assess heart anatomy and function.

In clinical practice, a cardiologist has limited time to interpret these 3,000 images within the context of numerous other data streams such as laboratory values, vital signs, additional imaging studies (radiography, magnetic resonance imaging, nuclear imaging and computed tomography) and other diagnostics (for example, electrocardiograms). While these numerous sources of data offer the potential for more precise and accurate clinical predictions, the ability of humans to recognize echocardiographic patterns unaided is limited compared with that afforded by integration of data into complex decision making assisted by computers[1]. Hence, there is both a need and a substantial opportunity to use technology, such as machine learning, to manage this abundance of data and ultimately provide intelligent computer assistance to physicians[2–4].

Recent advances in deep learning (deep neural networks (DNN)) technologies such as convolutional neural networks (CNN), recurrent neural networks, dropout regularization[5] and adaptive gradient descent algorithms[6], in conjunction with massively parallel computational hardware (graphics processing units (GPUs)), have enabled state-of-the-art predictive models for image, time-series and video-based data[7,8]. For example, DNNs have shown promise in diagnostic applications such as diabetic retinopathy[9], skin cancer[10], pulmonary nodules[11], cerebral microhaemorrhage[12,13] and aetiologies of cardiac hypertrophy[14]; yet, the opportunities presented by machine learning are not limited to such diagnostic tasks[2,15,16].

The prediction of future clinical events, for example, is a natural extension of machine learning in medicine. Nearly all medical decisions rely on accurate prediction. Indeed, diagnoses are primarily used to help predict both future clinical outcomes and what treatments or interventions will have a positive impact on those outcomes. Therefore, the prediction of future clinical events is a crucial task, and its complexity suggests that there is ample room for assistance from computer-based methods. Recent studies demonstrate this potential for prediction of in-hospital mortality[17,18] and palliative care referrals[19]. In cardiology, variables derived from the EHR have been used to predict 2–5 yr all-cause mortality in patients undergoing coronary computed tomography[20,21] and 5 yr all-cause mortality in patients undergoing echocardiography[22] and to directly estimate the benefit of treatments to prolong survival in patients with heart failure[23].

Notably, these initial outcome-prediction studies in cardiology exclusively used human-derived (that is, 'hand-crafted') measurements from imaging, as opposed to automatically analysing the raw image pixel data. While metrics derived from traditional image processing may ultimately improve clinical prediction performance[24,25], an approach that is unbiased with respect to human perception and

[1]Department of Translational Data Science and Informatics, Geisinger, Danville, PA, USA. [2]Electrical and Computer Engineering Department, University of New Mexico, Albuquerque, NM, USA. [3]Heart Institute, Geisinger, Danville, PA, USA. [4]Department of Biomedical Engineering, University of Kentucky, Lexington, KY, USA. [5]Phenomic Analytics and Clinical Data Core, Geisinger, Danville, PA, USA. [6]Department of Radiology, Geisinger, Danville, PA, USA. [7]Department of Population Health Sciences, Geisinger, Danville, PA, USA. [8]Heart and Vascular Center, Evangelical Hospital, Lewisburg, PA, USA. [9]These authors contributed equally: Christopher M. Haggerty, Brandon K. Fornwalt. ✉e-mail: bkfornwalt@geisinger.edu

pattern-recognition ability may provide even further predictive performance improvement. This concept was recently demonstrated by using deep learning to enhance prediction of all-cause mortality directly from electrocardiogram voltage–time traces, with superior performance compared with models using only hand-crafted measurements[26].

Thus, there is strong potential for an automated analysis to enhance the performance of predictive models by using all available image data rather than relying solely on human-derived and clinically inspired measurements. The potential benefit of this approach for echocardiography is further strengthened by the added availability of rich spatiotemporal (video) data, which are both widely used in the clinic[27] and have relatively low cost[28].

Initial efforts using DNNs have focused on using individual images[14,29,30] and full video models to estimate hand-crafted measurements[31–36]. In this Article, we investigate an approach that directly estimates individual outcomes from raw videos. We hypothesized that a DNN could learn spatiotemporal features from echocardiography video data to enhance clinical prediction of 1 yr all-cause mortality. We used a large video database (34,362 participants, 42,095 studies and 812,278 videos) linked to EHR data that included hand-crafted echocardiography-derived measurements (EDMs), additional clinical variables and individual outcomes to show that this hypothesis holds true. We also showed superior prediction performance of the DNN model compared with four cardiologists and two benchmark clinical risk models, the pooled cohort equations (PCE)[37] and Seattle Heart Failure (SHF) risk score[38]. Finally, we show that the result from the DNN model retains prognostic significance for nearly a decade, well beyond the 1 yr prediction for which it was trained.

Results

We first cross-validated the DNN model on our large clinically acquired echocardiography video database (812,278 videos). Independent models were trained for individual views (for example, parasternal long-axis and apical four-chamber views) and aggregated to form a feature vector that consisted of the outputs from individual view models. On average, using all echocardiography video views combined yielded higher performance (area under the receiver operating characteristic curve (AUC) = 0.83, 95% confidence interval (CI) (0.83, 0.84)) for predicting 1 yr mortality than using either 58 EDMs (AUC = 0.75, 95% CI (0.74, 0.76)) or the combination of the 58 EDMs and 100 additional clinical variables from the EHR, including relevant cardiovascular-related diagnoses, laboratory values, demographics and vital signs (AUC = 0.81, 95% CI (0.80, 0.82); Fig. 1). The largest model that combined all views and the 158 EHR-derived measurements yielded an AUC of 0.84, 95% CI (0.84, 0.85).

Models based on individual views ranged in performance from an AUC of 0.70–0.80, with the parasternal long-axis view (PL DEEP) producing the best individual performance. Finally, we calculated the PCE score[27], a clinical standard benchmark for future cardiovascular disease, for the same samples. The PCE score yielded an AUC of 0.64 (95% CI (0.62, 0.66)) for 1 yr mortality prediction, which was inferior to all tested DNN models.

Given this proof of concept from the cross-validation experiments for predicting 1 yr mortality from echocardiography videos with a DNN, we then re-trained the DNN models using all 812,278 videos from the cross-validation experiments, and evaluated performance on two new and distinct groups of participants. The first group was an independent set of 600 participants (survey set), balanced for the 1 yr mortality outcome (300 patients who survived for 1 yr after echocardiography and 300 patients who died within 1 yr). The second group was a cohort of 2,404 patients with heart failure (defined as 'definite' heart failure by eMERGE guidelines[39]) who underwent 3,384 echocardiograms.

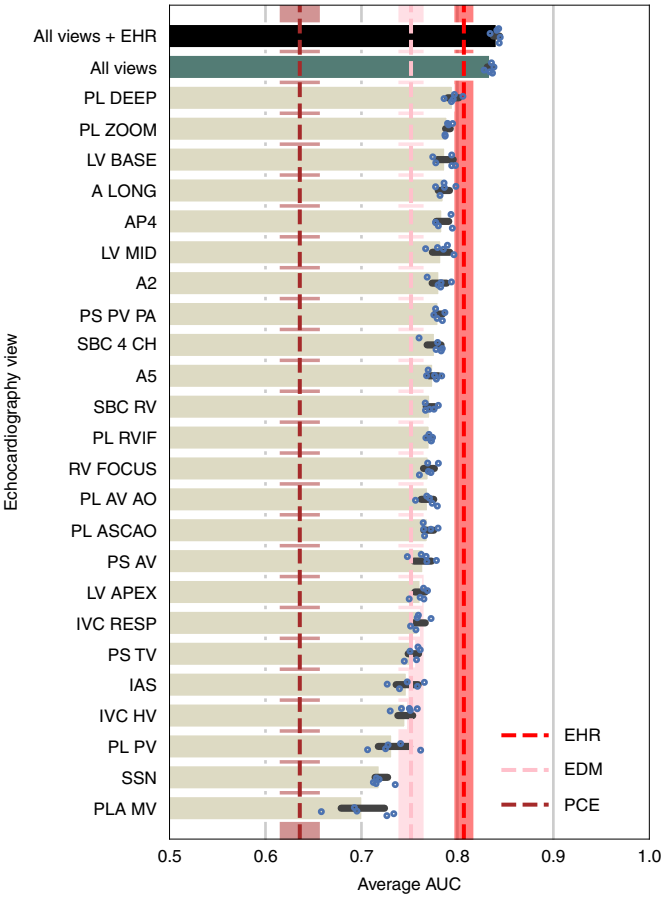

Fig. 1 | AUC across five folds for individual and collective echocardiography views. Average AUC across five folds for each individual echocardiography view (beige), all echocardiography views (green) and all views with 158 EHR features (black), including 58 EDMs and 100 clinical variables. The performance of an XGBoost model built from the 58 EDMs is shown in pink, and the performance of an XGBoost model built from the 158 total EHR features including the EDMs is shown in red. The performance of the PCEs is shown in brown. The black whiskers and the pink, red and brown shades denote the 95% CIs. Full descriptions of the abbreviated echocardiogram views are provided in Supplementary Table 6.

We used the survey set to evaluate the performance of four expert cardiologists, three of whom were Core Cardiovascular Training Statement (COCATS) level 3 and one with level 2 in echocardiography. The cardiologists were independently and blindly asked to determine whether each individual would be alive or dead 1 yr following the echocardiogram. For the sake of assessing the cardiologists' performances in an efficient manner, we presented a limited input set of a single video from the parasternal long-axis view (the highest-performing individual view) and ten EHR variables to compare their performance with a model trained on the same input set.

We constructed a risk score from the cardiologists' answers by aggregating the number of positive predictions (deceased within 1 yr) for each patient. The DNN model yielded an AUC of 0.84, 95% CI (0.81, 0.87), whereas the aggregated cardiologist score yielded an inferior AUC of 0.68, 95% CI (0.64, 0.71) (Fig. 2a and Table 1).

To further evaluate the performance of the DNN model compared with that of the cardiologists, we rearranged the survey set to show matched pairs. No individual-level feedback was provided to the cardiologists between experiments. In this second survey, the cardiologists and the model were presented with two studies at a time: one study was from an individual who died within 1 yr and

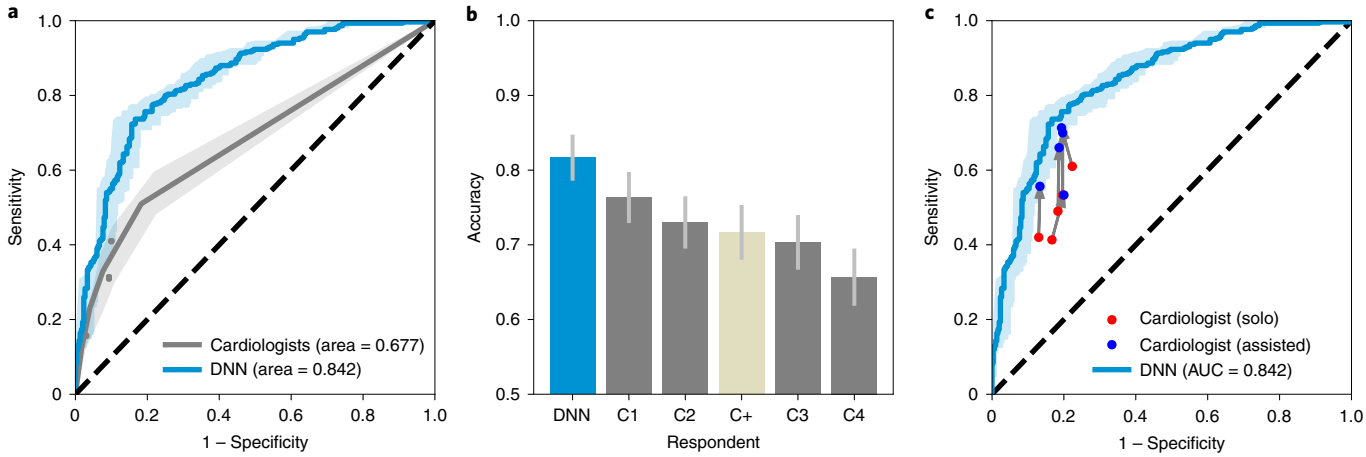

Fig. 2 | Survey results showing the performance estimates for the four cardiologists and the DNN model. a–c, The results correspond to the individual patient survey (**a**), the paired-patients survey (**b**) and the model-assisted survey (**c**). The DNN model is shown in blue and the cardiologists are represented in grey (dots for the individual operating points of each cardiologist and a line for the ROC of the aggregated score), except in **c** where the cardiologists' operating points are shown in red (without model assistance) and blue (with model assistance). In **b**, C+ shows the accuracy of the cardiologists by consensus. The bootstrapped 95% CI for each estimate is shown in shaded areas in **a** and **c**, and in grey whiskers in **b**. In **a** and **c**, the reference line (dashed line) corresponds to the performance of a random classifier.

the other was from a participant who lived beyond 1 yr after the echocardiogram. Both the cardiologists and the model were asked to select the individual from each pair with the higher chance of death within 1 yr. We matched 300 pairs by sex, age (within 5 yr) and left ventricular ejection fraction (LVEF) (within 10% absolute difference). This survey was designed to control for the outcome prevalence and directly measure discrimination performance. The DNN model yielded an accuracy of 82%, while the four cardiologists scored 66, 70, 73 and 76% (Fig. 2b and Table 1). We note that simple heuristics, such as selecting the older patient or the lower LVEF as the positive sample, resulted in 43% (131 samples) and 36% (108 samples) accuracy, respectively. Using a paired proportion test, the model yielded significantly higher performance than three out of four cardiologists after correcting for multiple comparisons ($P < 0.05/4$).

Next, we evaluated whether the cardiologists could improve their performance when assisted by the model. Similar to the first survey, we showed an individual study at a time, collected the cardiologist prediction, and then immediately presented the same study along with the machine prediction score. The aggregated cardiologist score AUC improved from 0.72, 95% CI (0.68, 0.76) to 0.78, 95% CI (0.74, 0.81) with assistance from the model predictions, which marginally overlaps with the DNN performance. In the survey, on average, the cardiologists correctly changed 10.3% of their predictions and incorrectly changed in 3.8% of their predictions. Sensitivity increased by 13%, while specificity decreased by less than 1% on average (Fig. 2c and Table 2).

The second group of individuals to which we applied our fully trained DNN model (all views plus EHR) was a cohort of 2,404 patients with heart failure (defined as 'definite' heart failure by eMERGE guidelines[39]) who underwent 3,384 echocardiograms. We chose this group of patients as an important additional clinical validation since heart failure is prevalent and costly and the management of heart failure relies heavily on survival prediction models such as the SHF risk score[38]. Within this cohort, the SHF score yielded an AUC of 0.70, 95% CI (0.68, 0.71), whereas the DNN model yielded an AUC of 0.76, 95% CI (0.74, 0.77). Notably, this superior performance of the DNN was observed for patients with both reduced (HFrEF) and preserved (HFpEF) LVEF (Table 3).

We computed predictions based on a mid-range threshold for the DNN model (0.5) and the SHF score (1.5) to discriminate between high and low risk. The range of scores was 0 to 1 for the DNN model and −1 to 4 for the SHF model[38]. Figure 3 shows the Kaplan–Meier survival curves stratified by the 1 yr mortality prediction labels, which we used to compute Cox proportional hazard ratios. The plots demonstrate that, despite the prediction being for 1 yr all-cause mortality, the result held long-term predictive power over the next 9 yr, with a hazard ratio of 2.9, 95% CI (2.6, 3.2) for the DNN model, compared with a hazard ratio for the SHF score of 2.2, 95% CI (2.0, 2.4). In particular, the DNN model maintained a higher negative predictive value compared with the SHF score (89% versus 83%, respectively), while maintaining the same positive predictive value (PPV) of 40%. This indicates that the DNN model better selected participants at low risk and may help to reduce unnecessary interventions on these lower-risk patients with heart failure.

There was no publicly available echocardiography dataset with mortality outcomes to serve as a truly external validation set for our work. However, we applied our apical four-chamber model to a large external echocardiography dataset with LVEF measurements[34], which we used as a surrogate for mortality risk. The predicted risk as a function of LVEF aligned with the reported trend of mortality outcomes versus LVEF[40] (Supplementary Information, 'Application to Stanford dataset').

Finally, we investigated what the DNN model was learning from the echocardiography videos. To do this, we occluded sample videos with 10 × 10 × 10 3D voxels and calculated the difference in the likelihood score that resulted from occluding that particular region. We show example results for parasternal long-axis (Fig. 4), apical two-chamber (Supplementary Fig. 8) and apical four-chamber (Supplementary Fig. 9) views. Since the results of the occlusion are videos, we displayed the first frame and overlaid red regions to denote significant changes in risk score (>2.5× s.d.) for at least 10 frames. We demonstrated these occlusion experiments for four patients with the highest prediction score who died within 1 yr (top row of figures) and four patients with the lowest prediction score who survived beyond 1 yr (bottom row of figures). These patients were selected from the test set of the first cross-validation experiment fold. Note that for the high-risk patients, the occlusion decreases the risk score, whereas for the low-risk patients, the occlusion increases the risk score. Generally, we observed that the regions with the largest effect on the risk score coincided with anatomically relevant regions of the heart, particularly the left atrium,

Table 1 | Performance summary for the two surveys

	Individual survey ($N=600$)					Paired survey ($N=300$)	
	AUC	Correct	Accuracy (%)	Sensitivity (%)	Specificity (%)	Correct	Accuracy (%)
C1	-	393	66 (62, 69)	41 (35, 47)	90 (87, 93)	229	76 (72 81)
C2	-	365	61 (57, 65)	31 (26, 37)	91 (87, 94)	197	66 (60 71)
C3	-	338	56 (52, 60)	16 (12, 20)	97 (95, 99)	211	70 (65 76)
C4	-	366	61 (57, 65)	31 (26, 36)	90 (87, 94)	219	73 (68 78)
C+	0.68 (0.64, 0.71)	376	63 (59, 67)	33 (28, 38)	92 (89, 95)	215	72 (67 77)
DNN	0.84 (0.81, 0.87)	465	78 (74, 81)	78 (73, 82)	77 (73, 82)	245	82 (77 86)

Predictive performance of each of the four cardiologists (C1–C4), the aggregated cardiologists score (C+), and the DNN model. Correct predictions, accuracy, sensitivity and specificity are shown in rounded percentages with their respective 95% CIs in parentheses.

Table 2 | Summary of performance of cardiologists C1–C4 in completing the individual survey ($N=600$) without (solo) and with (+DNN) computer assistance

	Correct		Change		Accuracy (%)		Sensitivity (%)		Specificity (%)	
	Solo	+DNN	+	−	Solo	+DNN	Solo	+DNN	Solo	+DNN
C1	416	456	65	25	69	76	61	71	78	81
					(66, 73)	(73, 79)	(55, 67)	(66, 76)	(73, 82)	(76, 85)
C2	392	442	80	30	65	74	49	66	82	81
					(62, 69)	(70, 77)	(43, 55)	(61, 71)	(77, 86)	(77, 86)
C3	387	427	65	25	65	71	42	56	87	87
					(61, 68)	(68, 75)	(36, 48)	(50, 61)	(83, 91)	(83, 91)
C4	374	400	39	13	62	67	41	53	83	80
					(58, 66)	(63, 70)	(36, 47)	(48, 59)	(79, 88)	(75, 85)
C+	401	451	71	21	67	75	53	70	80	80
					(63, 71)	(72, 79)	(48, 59)	(65, 75)	(76, 85)	(76, 85)

'Change' shows the number of times the cardiologists altered a false prediction (+) and a true prediction (−) using the machine score. 'C+' shows the performance of the aggregated cardiologists score. Accuracy, sensitivity and specificity are shown in rounded percentages with their respective 95% CIs in parentheses.

Table 3 | DNN model performance (and bootstrapped 95% CI AUCs) in heart failure cohort of 2,404 patients who underwent 3,384 echocardiograms, separated into HFrEF and HFpEF

	All ($n=3,384$)	HFrEF ($n=2,026$)	HFpEF ($n=1,357$)
SHF score	0.70 (0.68, 0.71)	0.70 (0.67, 0.72)	0.69 (0.66, 0.72)
DNN model (full)	0.76 (0.74, 0.77)	0.76 (0.74, 0.78)	0.75 (0.72, 0.78)

Benchmark comparison was made to the SHF score. Note that one study did not meet the criteria for either HFrEF or HFpEF because no LVEF was reported.

left ventricle and the mitral and aortic valve planes. These regions appeared to be more limited and localized in the videos from low-risk individuals, whereas in the videos from high-risk individuals they appeared to include surrounding anatomy; however, when presenting several examples of the occlusion maps to cardiologists, they reported that they were unable to identify patterns that could help them better predict patient survival outcomes. Instead, the cardiologists focused their attention on the model's predicted score to re-evaluate their initial assessment.

Discussion

We have shown the potential for neural networks to assist physicians with the clinical task of predicting 1 yr all-cause mortality. We used full, raw (annotation-free) echocardiographic videos to make predictions by learning from more than 812,278 clinically acquired echocardiography videos of the heart (50 million images). We showed that the ability of this DNN model to discriminate 1 yr mortality surpassed that of models using only image-derived and standard clinical measurements from the EHR as well as multiple existing clinical risk scores. Moreover, the DNN model enhanced the predictive performance of four trained cardiologists. This echocardiography video-based DNN model can therefore add value beyond a standard clinical interpretation.

We chose survival as an unambiguous clinical outcome to demonstrate feasibility for this initial work. Even when observer variability in echocardiography may exist for predicting human-defined outcomes[41,42], our focus on mortality labels allows us to minimize, if not eliminate, this challenge. Improving predictive performance may directly improve patient risk assessment prior to elective surgical procedures or influence therapy guidance for both primary and secondary prevention of cardiovascular disease in the outpatient setting. At the population level, an improved mortality risk model may enable health systems and insurance providers to better understand and optimally deploy resources to the population, as demonstrated previously using only EHR variables in patients with heart failure[23]. For heart failure in particular, methods for identifying candidates for advanced therapies such as cardiac transplant and implantation of durable mechanical support devices historically rely on mortality risk assessments based partly on peak oxygen consumption and invasive haemodynamics. Consideration for defibrillator placement in patients with heart failure is also predicated on a reasonable expectation of meaningful survival for more than 1 yr

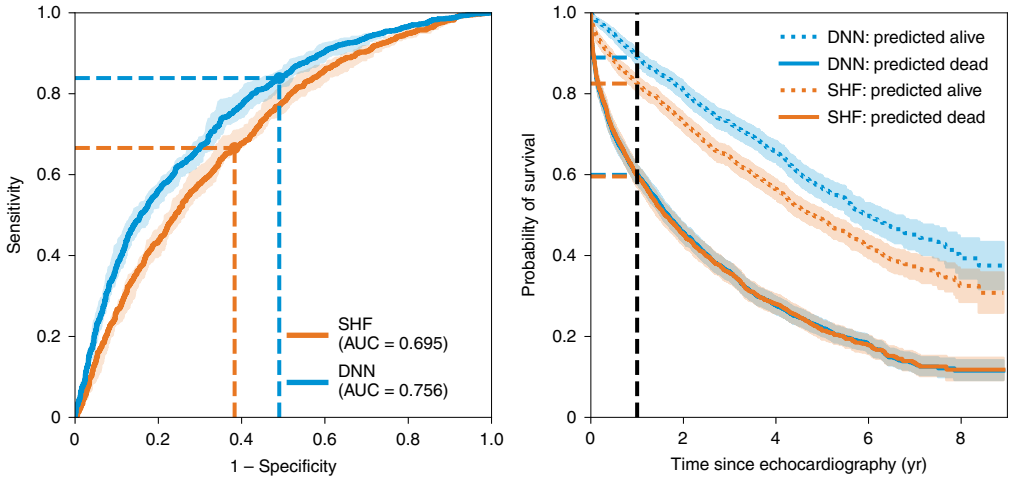

Fig. 3 | Performance comparison with the SHF score. Left: ROC curves for the SHF score (orange), the DNN model (blue), and their mid-range operating points ($N = 3{,}384$). Right: Kaplan–Meier curves for the outcome prediction from the operating points shown on the left, where the dotted and continuous lines denote groups predicted to be alive and dead, respectively. The shaded areas show the bootstrapped 95% CI for each line. The vertical black line marks the 1 yr point for which the DNN model was trained and the horizontal lines mark the 1 – positive predictive value and negative predictive value for both models at the given operating point.

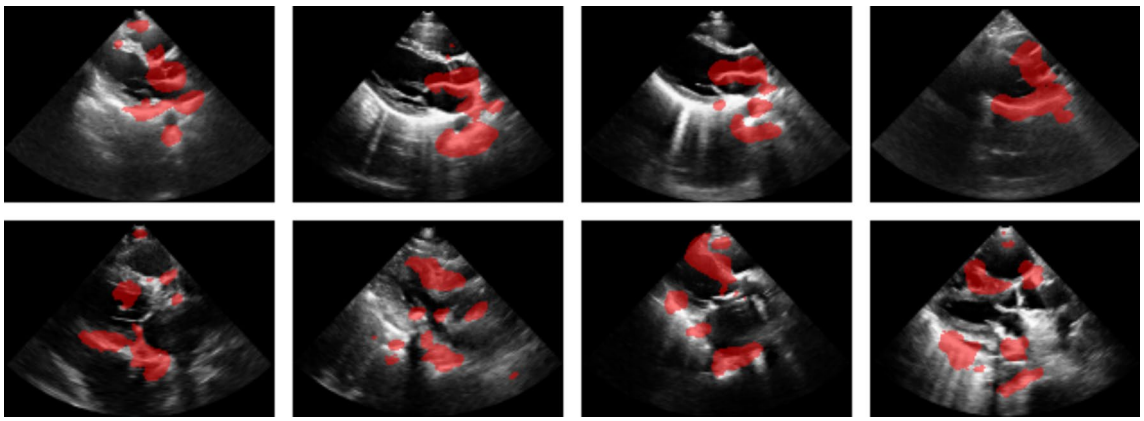

Fig. 4 | Occlusion results for the lowest and highest prediction risk scores for 1-yr mortality. Occlusion results for the four lowest (top) and four highest (bottom) prediction risk scores for 1-yr mortality obtained from a parasternal long-axis view. The red areas highlight regions that changed the risk score by more than 2.5×s.d. (of all changes) for at least 10 frames. The red areas are overlaid on the first frame of the video for anatomical reference.

(ref. [43]). Implementation of a more accurate mortality-based risk tool may have additive benefits[44]. Finally, estimation of 1 yr mortality is particularly important for planning the transition to palliative care and hospice[45]. Further research will be needed to evaluate the performance of neural network models to predict additional clinically relevant outcomes in cardiology, such as future hospitalizations or the need for major procedures such as valve replacement.

This work has several limitations. First, determining exactly what information within the echocardiograms that our model is using to make accurate predictions is challenging, and this is a general problem facing all DNN models[46]. We used occlusion maps to provide evidence that the DNN was affected by anatomically appropriate regions of the heart, similar to previous work in skin cancer classification[10], diabetic retinopathy screening[47] and intracranial haemorrhage detection[48]. Unfortunately, this standard approach is limited since only empirical findings can be shown on a case-by-case basis without definitive general interpretation. Future work on building interpretable DNN models without sacrificing performance is needed to fully address this limitation of the field. Second, although our data had inherent heterogeneity, since they were derived from a large regional healthcare system with more than ten hospitals and hundreds of clinics, data from other independent healthcare systems with mortality outcomes will be required to fully assess generalizability. Third, infrastructure limitations probably led to reduced model performance; that is, we were able to use less than 10% of available echocardiograms at our institution because using all echocardiograms would require substantial hardware (around 1 petabyte of fast storage attached to GPU systems). We show (in Supplementary Information, 'Comparison of cross-validation results with Samad et al.') that sample size is the primary explanation for the reported decreased prediction performance for a model trained from the 158 EHR-derived variables in the current study (AUC of 0.81 in 42,092 echocardiograms) compared with our previous work using an order of magnitude more data (AUC of 0.85 in 331,317 echocardiograms[22]; note that raw video data was not used in this previous study).

In summary, we have developed a methodology and architecture for extracting clinically relevant predictive information from medical videos with a deep neural network and subsequently provided evidence of the feasibility and potential of such predictive models. With the ongoing rate of technological advancement and the rapid growth in electronic clinical datasets available for training,

neural networks will augment future medical image interpretations with accurate predictions of important clinical outcomes such as mortality.

Methods

EHR data preprocessing. The Institutional Review Board approved this study with a waiver of consent, in conjunction with our institutional patient privacy policies. Our institutional echocardiography archives, as of January 2020, contained a total of 683,662 echocardiography studies from 305,282 unique patients collected over the previous 22 yr. We extracted all structured physician-reported EDMs ($n = 58$) from these studies. Furthermore, through our institutional phenomics initiative database, we linked these echocardiography measurements to patient demographics (3), vitals (5), laboratory (2) and problem list data (90) (International Classification of Diseases, Tenth Revision (ICD-10) codes) from our institutional EHR (Epic Systems; 1996–present). Supplementary Table 12 shows a list and description of all 158 EHR variables used in the study.

All continuous variables were cleaned to remove physiologically out-of-limit values (manually defined by a cardiologist), which were presumed to reflect input errors and set as missing. We identified 8 categorical variables in the echocardiography measurements that each reported 5 valvular regurgitation and stenosis severity levels (including not assessed) and converted them to 40 one-hot encoded binary variables. We also identified an ordinal variable reporting diastolic function, and coded it as −1 for normal, 0 for dysfunction (but no grade reported), and 1, 2 and 3 for diastolic dysfunction grades I, II and III, respectively. For non-echocardiography-derived measurements, such as low-density lipoprotein, high-density lipoprotein, blood pressure, heart rate (if not taken at the study), weight and height measurements, we retrieved the most recent past measurement, within a 1 yr window, relative to the echocardiogram acquisition date. We calculated the patient's age and survival duration from the date of the echocardiogram.

The patient status (dead or alive) was identified on the basis of the last-known living encounter or confirmed death date, which is cross-referenced monthly in our system against national death index databases. For labelling 1 yr mortality, a positive sample was defined as an echocardiography study within 1 yr of the patient's death date. A negative 1 yr mortality label was defined as an echocardiography study that occurred more than 1 yr before the death date (if deceased) or last-known physical encounter within our system (if alive). Studies without a death date or at least 1 yr follow-up (physical encounter) were excluded.

Image collection and preprocessing. An echocardiography study consists of several videos containing multiple views of the heart. Two clinical databases, Philips iSite and Xcelera, contained all echocardiograms collected at Geisinger. We used DCM4CHEE (v.2.0.29) and AcuoMed (v.6.0) software to retrieve a DICOM file for each echocardiography video.

The retrieved DICOM files contained an annotated video and a raw video from when the equipment was configured to store it. The raw video contained only the beam-formed ultrasound image stored in a stream-of-bytes format, whereas the annotated video contained annotations (such as the view name) on top of the raw video (Supplementary Fig. 1). We used raw videos for all analyses. Since the videos in raw format varied in frame rate across studies, we linearly interpolated all videos to 30 frames per second.

Along with the video data, the DICOM file included tags that labelled each video, indicating the specific image orientation in which it was acquired, which we refer to as a view. These view tags had slight variations across studies for the same type of view. For example, an apical four-chamber view could be tagged as 'a4', 'a4 2d', or 'ap4'. We visually inspected samples of each unique tag and grouped them into 30 common views (Supplementary Table 6).

For the entire cross-validation cohort, the average number of views available for negative samples was 19.4 and the interquartiles were 19 and 22. For positive samples, the average was 18.3 videos and the interquartiles were 18 and 22 videos per sample. The median number of videos was 20 for both positive and negative samples.

When a study had multiple videos from the same view, we selected the video with the longest duration.

Since each video from a view group could potentially have different dimensions, we normalized all videos to the most common row and column dimension pairs of its corresponding view. We cropped or padded each frame with zeros to match the most common dimensions among the view group, keeping the beam-formed image centred.

We note that the image-size normalization (cropping and padding) had a minimal effect on the video because the standard echocardiography views centre the anatomical region of interest. For example, we cropped and padded more than 6 rows on fewer than 3% of the PL DEEP videos, from which only 17 cases were cropped and the rest were zero padded. Generally, border areas do not contain features of interest (Fig. 4 and Supplementary Fig. 1).

Data selection. We systematically extracted echocardiography studies from our clinical imaging archives (acquired after February 2011) to research servers for this analysis, and subsequently retained only raw video data from these studies, where available. This extracted subset of the total clinical archive was divided into three distinct groups to conduct the experiments described above (the characteristics of each are described in Supplementary Table 1). In each case, follow-up beyond 1 yr or date of death within 1 yr was known.

(1) Cross-validation experiment: comprised 42,095 studies with 812,278 videos collected from 34,362 individuals, drawn without predefined patient-selection criteria from the clinical echocardiography archive.
(2) Cardiologist surveys: comprised 600 studies with 11,357 videos collected from 600 individuals, again taken from the unselected clinical data extraction but held out from the cross-validation experiment set and pre-specified to have balanced outcome labels (300 dead and 300 alive at 1 yr).
(3) Heart failure experiment: comprised 3,384 studies with 58,561 videos collected from 2,404 individuals, specifically selected from the clinical archive on the basis of the presence of heart failure—based on the 'definite' eMERGE algorithm[39] criteria—at the time of the echocardiogram.

The 42,095 studies in the cross-validation set are a subset of a previously published cohort[22] (Supplementary Information, 'Comparison of cross-validation results with Samad et al.', presents an analysis on performance differences).

Cardiologist survey. According to a previous study[22], the top 10 most predictive clinical (EHR) variables for 1 yr mortality following an echocardiogram are age, tricuspid regurgitation maximum velocity, heart rate, low-density lipoprotein, LVEF, diastolic pressure, pulmonary artery acceleration time, systolic pressure, pulmonary artery acceleration slope and diastolic function. These 10 variables contained more than 95% of the power for predicting 1 yr survival in 171,510 patients[22]. For the sake of assessing the cardiologists' performances in an efficient manner, we presented these top 10 variables as a summary of the patient's status as of the day of the echocardiogram. Along with these ten measurements, we also presented a parasternal long-axis video. This view is typically reported by cardiologists as the most informative summary view of overall cardiac health because it contains elements of the left ventricle, left atrium, right ventricle, aortic and mitral valves, and whether or not there is a pericardial or left pleural effusion, all within a single view.

Following a sample size calculation (Pearson chi-squared test) to estimate and compare prognostic accuracy between the cardiologists and the model, the cardiologists completed a survey set of 600 samples. We assumed a 10% difference in accuracy between machine and cardiologist (80% versus 70%), 80% power, a significance level of 5%, and an approximate 40% discordancy. The calculation (performed with Power Analysis Software PASS v.15) showed that we needed at least 600 patients (300 alive and 300 deceased). We therefore randomly sampled 300 positive and 300 negative studies that contained a parasternal long-axis view, ensuring that none of these patients remained in the cross-validation set.

The first survey presented one patient sample at a time and was designed to score the cardiologists' aggregated discrimination ability. Supplementary Fig. 2 shows the interface for the first survey. We showed the ten EHR variables in a table and two versions of the video, raw and annotated. The application then recorded the cardiologist prediction as they clicked on either the 'alive' or 'dead' buttons.

The second survey presented paired samples and was designed to assess the discrimination ability of each cardiologist while controlling for mortality prevalence. We prepared 300 pairs on the basis of sex, age (within 5 yr) and LVEF (within 10%). We paired all 300 positive cases to a negative case, where 213 negatives were unique and the remaining 87 pairs had to contain already-used negatives to preserve the matching criteria. Thus, all positive cases were unique. Supplementary Fig. 3 shows the interface for the paired survey, where we showed the video and ten EHR variables for two age-, sex- and LVEF-matched patients.

The third and last survey presented individual samples followed by the same sample with additional information extracted from the DNN model. We presented the machine score and occlusion maps to assess whether the inclusion of machine information could improve the cardiologist-aggregated score performance.

We presented the same 600 patients twice. First, we showed the individual sample as in Supplementary Fig. 2 and, immediately after, we showed the same sample with the calibrated risk score from the model and occlusion map. The cardiologists then either amended or reiterated their prediction.

To avoid incremental performance changes while the cardiologists progressed through the survey, we presented them, before taking the survey, with 80 examples with machine predictions, occlusion maps and true outcomes from the cross-validation set. The 80 examples were distributed in 4 groups of 20, grouped by history of heart failure only, history of myocardial infarction only, history of both, or history of neither. Each of the four groups were further split into five examples that fell into each of the four quadrants of the confusion matrix. Supplementary Fig. 4 shows the interface for the model-assisted portion of the third survey, where we added a 'machine prediction' row and a occlusion-map video.

We note that no individual patient-level response feedback was presented to the cardiologists between any surveys (to avoid confounding results of subsequent surveys from knowledge gained through prior surveys) and a minimum of 15 d elapsed between surveys for a given cardiologist.

We acknowledge that none of these surveys was designed to represent normal clinical practice, and prediction performance in a real-world setting is probably enhanced by access to the full medical record, physical exam and other factors. However, the surveys were designed to efficiently estimate a baseline performance when constrained to a limited input set, as well as the ability to enhance that baseline performance with the assistance of a DNN model.

Neural network architectures. We designed four different low-parameter architectures: (1) A time-distributed 2D CNN with long short-term memory (LSTM) (Supplementary Fig. 5 and Supplementary Table 2), (2) a time-distributed 2D CNN with global average pooling (GAP) (Supplementary Table 3), (3) a 3D CNN (Supplementary Fig. 6 and Supplementary Table 4) and (4) a 3D CNN with GAP (Supplementary Table 5). For simplicity, we abbreviate the four candidate architectures: 2D CNN + LSTM, 2D CNN + GAP, 3D CNN and 3D CNN + GAP.

The 2D CNN + LSTM consisted of a 2D CNN branch distributed to all frames of the video. This architecture was used for a video description problem[49], where all frames from a video belonged to the same scene or action. It is therefore assumed that static features would be commonly found by the same 2D kernels across the video. This assumption was put in practice for echocardiography view classification[50]. The LSTM layer aggregates the CNN features over time to output a vector that represents the entire sequence.

The 2D CNN + GAP approach exchanged the LSTM layers for the average CNN features as a time aggregation of frames. The GAP layer provided two advantages: it required no trainable parameters, saving 10,736 parameters from the LSTM layers, and it enabled feature interpretation. The final fully connected layer after the GAP provided a weighted average of the CNN features, which could indicate what sections of the video were weighted more in the final decision. The 3D CNN approach aggregated time and space features as the input data flowed through the network.

As opposed to the 2D CNN approach, a 3D CNN incorporated information from adjacent frames at every layer, extracting spatiotemporal dependent features which have also proven to be useful for video classification[7,51]. In a 3D CNN approach, a GAP layer reduced the fully connected layer input from the feature map size to the number of filters. Thus, the GAP layer also reduced the number of parameters from 641 to 17.

We defined the convolutional units of the 2D and 3D CNNs as a sequence of 7 layers in the following composition: CNN layer, batch normalization, rectified linear unit, CNN layer, batch normalization, rectified linear unit and Max Pooling (Supplementary Fig. 9). All kernel dimensions were set to three and Max Pooling was applied in a 3×3 window for 2D kernels and $3\times3\times3$ for 3D kernels. We also added four additional versions by increasing the kernel sizes from three to five pixels in all dimensions, resulting in a total of eight candidate video models per echocardiography view.

We chose a low-parameter design due to the high computational cost of the presented experiments and to reduce the chance of overfitting. To complete all experiments, we fit a total of 1,152 neural network models (24 views × 5 folds × 8 models for the cross-validations experiments plus 24 views x 8 models for the final versions) which fully occupied all 16 GPUs in our NVIDIA DGX-2 for approximately 40 d. Deep learning models typically consist of millions of parameters; for example, the Inception model has 25 million parameters[52] and ResNet more than 40 million parameters[53], rendering the computational cost to train such large networks as prohibitive and, given the performance demonstrated in our models, potentially unnecessary. Our largest model consisted of less than 20,000 parameters (Supplementary Tables 2–5).

We implemented these networks using the docker container tensorflow:19.08–py3 (available at nvcr.io/nvidia/ or via https://ngc.nvidia.com) with Python v.3.6.8, Tensorflow module v.1.14, and Keras module v.2.2.4-tf.

Cross-validation procedure. Using the cross-validation set described in Supplementary Table 1, we split the echocardiography studies into five folds, where, at each of the five iterations, a fold was used for testing and the rest for training. We enforced two constraints on the folds content: (1) studies from the same participant could not be present in more than one fold, and (2) each fold contained a similar positive prevalence (of 1 yr all-cause mortality) as the entire dataset. For each training set, we set aside one-tenth of the studies, with a balanced prevalence, as a proxy to the test set for internal validation.

As we trained the DNN, we evaluated the loss (binary cross-entropy) on the internal validation set at each epoch. If the internal validation loss did not decrease for more than ten epochs, we stopped the training and recovered the model weights at the minimum validation loss[54].

We trained all video architectures on all available views in the training set. For each view, we chose the architecture with the highest AUC in the internal validation set and used that model to report performance for that view in all subsequent experiments (a summary of the architectures chosen for each view is shown in Supplementary Table 8, and an example for the PL DEEP view is provided in Supplementary Table 9).

We concatenated EHR-derived features and video risk scores for each view to fit a classification pipeline composed of an interquartile range scaler, a multivariate imputation by chained equations[55] and an XGboost classifier[56]. This pipeline was fit at each training fold and applied to its corresponding test set to produce the output risk score.

Since the mortality prevalence in the overall dataset was imbalanced (14.6% of patients died within a year of the echocardiography study), we set the weights (w_i) for each class i as follows:

$$w_i = \frac{\text{Total number of samples}}{2\times(\text{Number of samples in class } i)} \quad (1)$$

All training was performed on an NVIDIA DGX-2 computer unit by independently fitting each model on each of the 16 available GPUs.

Statistical analysis. In all survival analyses, we used the time to death or last-known living encounter (censored) from the echocardiography study and the predicted labels to stratify the probability of survival for the Kaplan–Meier plots and Cox proportional hazard ratio analysis. The analysis was conducted using the lifelines Python package v.0.25.4. The thresholds for both the DNN and SHF models were chosen as the midpoint in the score range.

For the cross-validation experiment where we obtained an AUC estimate per fold, we reported the average across the 5 folds and 95% CI, computed as $\pm 1.96\sigma/\sqrt{5}$.

For the remaining experiments, where only a single AUC was available (heart failure and survey cohorts), we bootstrapped the AUC estimation for 10,000 iterations and reported the 2.5th and 97.5th percentiles as the 95% CI.

To report significant differences when comparing the predictive performance with the paired survey data, we conducted paired proportion tests on the number of correct answers out of the 300 samples. We conducted a total of four tests comparing each of the four cardiologists to the DNN model, hence the P-value corrected threshold of 0.05/4. For the statistical computations, we used the statsmodel package for Python v.0.11.1.

Implementation of the SHF score. We computed the SHF score as described[57], with the exception that systolic blood pressure, haemoglobin, percentage of white blood cells in the form of lymphocytes, uric acid, total cholesterol and sodium were defined as the most recent available measurement before (within 1 yr) or the day of the echocardiogram, instead of using a potentially closer measurement in the future. For predicting future events, we blinded both the DNN and SHF models to data collected after the date that the echocardiogram was acquired.

Heart failure subtype definition. Heart failure type (that is, HFrEF versus HFpEF) was determined for each sample using all previous available ejection fraction measurements up to 6 months prior to heart failure diagnosis as follows: (1) HFrEF if any LVEF ≤ 50%; (2) HFpEF if all LVEF ≥ 50%; or (3) no subtype was assigned if no LVEF was reported.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. All requests for raw and analysed data will be reviewed by the Legal Department of Geisinger Clinic to verify whether the request is subject to any intellectual property or confidentiality constraints. Requests for patient-related data not included in the paper will not be considered. Any data that can be shared will be released via a Material Transfer Agreement for non-commercial research purposes.

Code availability

All requests for code will be reviewed by the Legal Department of Geisinger Clinic to verify whether the request is subject to any intellectual property or confidentiality constraints. Any code that can be shared will be released via a Material Transfer Agreement for non-commercial research purposes under the Creative Commons Attribution NonCommercial–NoDerivatives 4.0 license. Code to reproduce Supplementary Fig. 10 is available at: https://github.com/alvarouc/geisinger-echo-mortality.

Received: 16 April 2020; Accepted: 24 November 2020;
Published online: 08 February 2021

References

1. Payne, J. W. Task complexity and contingent processing in decision making: an information search and protocol analysis. *Organ. Behav. Hum. Perform.* **16**, 366–387 (1976).
2. Quer, G., Muse, E. D., Nikzad, N., Topol, E. J. & Steinhubl, S. R. Augmenting diagnostic vision with AI. *Lancet* **390**, 221 (2017).
3. Jha, S. & Topol, E. J. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* **316**, 2353–2354 (2016).

4. Kyriacou, E., Constantinides, A., Pattichis, C., Pattichis, M. & Panayides, A. in *Biomedical Signals, Imaging, and Informatics* 4th edn (eds Bronzino, J. D. & Peterson, D.) Ch. 64 (CRC Press, 2015).
5. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
6. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
7. Ji, S., Xu, W., Yang, M. & Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2012).
8. Karpathy, A. et al. Large-scale video classification with convolutional neural networks. In *IEEE conference on Computer Vision and Pattern Recognition* 1725–1732 (2014).
9. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
10. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
11. Setio, A. A. A. et al. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE Trans. Med. Imaging* **35**, 1160–1169 (2016).
12. Arbabshirani, M. R. et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digit. Med.* **1**, 9 (2018).
13. Dou, Q. et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. *IEEE Trans. Med. Imaging* **35**, 1182–1195 (2016).
14. Madani, A., Ong, J. R., Tibrewal, A. & Mofrad, M. R. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digit. Med.* **1**, 59 (2018).
15. Van Woudenberg, N. et al. in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation* (eds Stoyanov, D. et al.) 74–81 (Springer, 2018).
16. Kusunose, K. et al. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC Cardiovasc. Imagin* **13**, 374–381 (2019).
17. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
18. Kwon, J.-m, Kim, K.-H., Jeon, K.-H. & Park, J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography* **36**, 213–218 (2019).
19. Avati, A. et al. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* **18**, 122 (2018).
20. Motwani, M. et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur. Heart J.* **38**, 500–507 (2016).
21. Hadamitzky, M. et al. Optimized prognostic score for coronary computed tomographic angiography: results from the confirm registry (coronary CT angiography evaluation for clinical outcomes: an international multicenter registry). *J. Am. Coll. Cardiol.* **62**, 468–476 (2013).
22. Samad, M. D. et al. Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. *JACC Cardiovasc. Imaging* **12**, 681–689 (2018).
23. Jing, L. et al. A machine learning approach to management of heart failure populations. *JACC Heart Fail.* **8**, 578–587 (2020).
24. Murillo, S. et al. Motion and deformation analysis of ultrasound videos with applications to classification of carotid artery plaques. In *SPIE Medical Imaging* (SPIE, 2012).
25. Cui, X. et al. Deformable regions of interest with multiple points for tissue tracking in echocardiography. *Med. Image Anal.* **35**, 554–569 (2017).
26. Raghunath, S. et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).
27. Gahungu, N., Trueick, R., Bhat, S., Sengupta, P. P. & Dwivedi, G. Current challenges and recent updates in artificial intelligence and echocardiography. *Curr. Cardiovasc. Imaging Rep.* **13**, 5 (2020).
28. Horgan, S. J. & Uretsky, S. in *Essential Echocardiography: A Companion to Braunwald's Heart Disease* (eds Solomon, S. D. et al.) 460–473 (Elsevier, 2019).
29. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**, 1623–1635 (2018).
30. Li, M. et al. Unified model for interpreting multi-view echocardiographic sequences without temporal information. *Appl. Soft Comput.* **88**, 106049 (2020).
31. Ge, R. et al. K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. *IEEE Trans. Med. imaging* **39**, 1690–1702 (2019).
32. Ge, R. et al. Echoquan-net: direct quantification of echo sequence for left ventricle multidimensional indices via global-local learning, geometric adjustment and multi-target relation learning. In *International Conference on Artificial Neural Networks* (eds Tetko, I. et al.) 219–230 (Springer, 2019).
33. Jafari, M. H. et al. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. *Int. J. Comput. Assist. Radiol. Surg.* **14**, 1027–1037 (2019).
34. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
35. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *npj Digit. Med.* **3**, 10 (2020).
36. Behnami, D. et al. Automatic cine-based detection of patients at high risk of heart failure with reduced ejection fraction in echocardiograms. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* https://doi.org/10.1080/21681163.2019.1650398 (2019).
37. Yadlowsky, S. et al. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann. Intern. Med.* **169**, 20–29 (2018).
38. Levy, W. C. et al. The Seattle Heart Failure model. *Circulation* **113**, 1424–1433 (2006).
39. McCarty, C. A. et al. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
40. Wehner, G. J. et al. Routinely reported ejection fraction and mortality in clinical practice: where does the nadir of risk lie? *Eur. Heart J.* **41**, 1249–1257 (2020).
41. Liao, Z. et al. On modelling label uncertainty in deep neural networks: automatic estimation of intra-observer variability in 2D echocardiography quality assessment. *IEEE Trans. Med. Imaging* **39**, 1868–1883 (2019).
42. Behnami, D. et al. Dual-view joint estimation of left ventricular ejection fraction with uncertainty modelling in echocardiograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Shen, D. et al.) 696–704 (Springer, 2019).
43. Yancy, C. W. et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **62**, e147–e239 (2013).
44. Lund, L. H., Aaronson, K. D. & Mancini, D. M. Predicting survival in ambulatory patients with severe heart failure on beta-blocker therapy. *Am. J. Cardiol.* **92**, 1350–1354 (2003).
45. Kavalieratos, D. et al. Palliative care in heart failure: rationale, evidence, and future priorities. *J. Am. Coll. Cardiol.* **70**, 1919–1930 (2017).
46. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
47. Arcadu, F. et al. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digit. Med.* **2**, 92 (2019).
48. Lee, H. et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat. Biomed. Eng.* **3**, 173 (2019).
49. Venugopalan, S. et al. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2015).
50. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *npj Digit. Med.* **1**, 6 (2018).
51. Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In *Proc. IEEE International Conference on Computer Vision* 4489–4497 (2015).
52. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (2016).
53. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (2017).
54. Prechelt, L. in *Neural Networks: Tricks of the Trade* (eds Montavon, G. et al.) 55–69 (Springer, 1998).
55. Buuren, S. & Groothuis-Oudshoorn, K. MICE: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, jss.v045.i03 (2011).
56. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
57. Williams, B. A. & Agarwal, S. Applying the Seattle Heart Failure model in the office setting in the era of electronic medical records. *Circ. J.* **82**, 724–731 (2018).

Acknowledgements

This work was supported in part by funding from the Pennsylvania Dept of Health (SAP 4100070267 and 4100079720) and the Geisinger Health Plan and Clinic. The content of this article does not reflect the view of the funding sources.

Author contributions

A.E.U.-C., C.M.H. and B.K.F. conceived the study and designed the experiments. A.E.U.-C. conducted all experiments. A.E.U.-C. and S.R. wrote the software for applying

deep learning to echocardiography videos. A.E.U.-C., L.J., D.P.v., D.N.H., J.D.S. and J.B.L. assembled the input data. H.L.K., G.J.W., M.S.P. and A.A.P. gave advice on experiment design. L.J., C.D.N., C.M.H. and B.K.F. manually audited the data for the cardiologist survey. C.W.G., A.A., J.M.P. and B.J.C. completed the surveys and provided insights on interpretability experiments. A.E.U.-C., C.M.H., M.S.P. and B.K.F. wrote the manuscript. All authors critically revised the manuscript.

Competing interests

Geisinger receives funding from Tempus for ongoing development of predictive modelling technology and commercialization. Tempus and Geisinger have jointly applied for a patent related to this work. None of the authors has ownership interest in any of the intellectual property resulting from the partnership.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41551-020-00667-9.

Correspondence and requests for materials should be addressed to B.K.F.

Peer review information *Nature Biomedical Engineering* thanks Partho Sengupta, Purang Abolmaesumi and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

nature research

Corresponding author(s): Brandon K. Fornwalt

Last updated by author(s): Nov 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Dcm4chee (JAVA DICOM command line toolkit, version 2.0.29) scheduled the moves to the Clinical Trial Processor, which stored the DICOM as files. Images in Philips iSite were retrieved using AcuoMed Batch Processing (version 6.0) to schedule the moves, and a Research PACS (DCM4CHEE v2.17) to receive the images.
Data analysis	We used pydicom v1.02 to read the binary data from the (200D, 3CF4) DICOM element. For data processing we used python v3.6, tensorflow v1.14, pandas v1.0.1, and statmodels v0.11.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. All requests for raw and analysed data will be reviewed by the Legal Department of Geisinger Clinic to verify whether the request is subject to any intellectual property or confidentiality constraints. Requests for patient-related data not included in the paper will not be considered. Any data that can be shared will be released via a Material Transfer Agreement for non-commercial research purposes.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We needed at least 600 patients (300 alive, 300 deceased), as indicated by a Pearson Chi-square test-sample-size calculation, to estimate and compare prognostic accuracy between the model and a cardiologist's predictions. We assumed a 10% difference in accuracy between machine and cardiologist (80% vs 70%), 80% power, a significance level of 5%, and an approximate 40% discordance. This was calculated using Power Analysis Software (PASS v15).
Data exclusions	Pre-established criteria for exclusion of data were: 1. Exclude echocardiography studies acquired before February 2011, without raw video data available, 2. Exclude videos from the echocardiography study that did not contain view label information, 3. Exclude patients without enough follow up, as we can not define the survival status. Additional criteria were imposed to remove rare echocardiography views (less than 5,000 samples).
Replication	We contained all our software dependencies in a Docker container. All code is version-controlled and experiments can be reproduced on hardware with NVIDIA V100 GPUs. Data used in this study was stored and backed up for future experimental reproducibility.
Randomization	The dataset was randomly divided in training and testing folds for cross-validation. The 600 studies for the survey set consisted of 300 randomly drawn samples from each group (survives or dies within a year of the echocardiography study).
Blinding	The cardiologists were blinded from actual patient-survival data, from each other's and from the model's performance, until the final release of the results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The sample in our study represents a patient population in a clinical setting. The average patient ages for the cross-validation survey and the two heart-failure datasets were 66 (+/– 16; s.d.), 68 (+/–16, s.d.), and 73 (+/–14, s.d.), respectively. The proportion of studies from male patients was 51%, 56%, and 56% respectively. Supplementary Table 1 provides a detailed description.
Recruitment	We collected samples from a clinical database. Potential sources of bias are the large prevalence of Caucasians (>95%) in our geographical area and the availability of raw echocardiography videos as inclusion criteria.
Ethics oversight	This retrospective study was approved by Geisinger's institutional Review Board and was performed with a waiver of consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.