nature biomedical engineering

PEER-REVIEW INFORMATION

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

Corresponding author: Brandon Fornwalt

Editorial note

This document includes relevant written communications between the manuscript's corresponding author and the editor and reviewers of the manuscript during peer review. It includes decision letters relaying any editorial points and peer-review reports, and the authors' replies to these (under 'Rebuttal' headings). The editorial decisions are signed by the manuscript's handling editor, yet the editorial team and ultimately the journal's Chief Editor share responsibility for all decisions.

Any relevant documents attached to the decision letters are referred to as **Appendix #**, and can be found appended to this document. Any information deemed confidential has been redacted or removed. Earlier versions of the manuscript are not published, yet the originally submitted version may be available as a preprint. Because of editorial edits and changes during peer review, the published title of the paper and the title mentioned in below correspondence may differ.

Correspondence

Sat 30/05/2020
Decision on Article nBME-20-0869

Dear Dr Fornwalt,

Thank you again for submitting to *Nature Biomedical Engineering* your Article, "Automated analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality". The manuscript has been seen by four experts, whose reports you will find at the end of this message. You will see that the reviewers have good words for the work, and that Reviewers #1 and #4 raise a number of technical criticisms that we hope you will be able to address. In particular, we would expect that a revised version of the manuscript provides:

* Evidence of any improvements in performance when expert cardiologists are aided by the outcomes of the deep-learning algorithm.

* Extended occlusion studies, to identify any additional anatomical features contributing to pathology-specific mortality risk.

* Evidence that the training process has been designed to minimize overfitting, and that the outcomes are not influenced by any correlation of mortality with the number of videos per patient.

* Thorough methodology, including model architectures and all optimizer and model-specific hyperparameters (for reproducibility purposes).

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We hope that you will be able to resubmit the manuscript within 15 weeks from the receipt of this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*. Because of the COVID-19 pandemic, should you be unable to carry out work in the near future we advise that you reply to this message with a revision plan in the form of a preliminary point-by-point rebuttal to the comments from all reviewers that also includes a response to any points highlighted in this decision. We should then be able to provide you with additional feedback.

We hope that you will find the referee reports helpful when revising the work, which we look forward to receive. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

The paper describes the application of deep neural networks for learning spatio-temporal predictive feature representations from echocardiography videos that are associated with augmenting the prediction of 1-year all-cause mortality. There has been a recent interest in using 2D and 3D neural network architectures and deep learning for addressing important problems in cardiology. The proposed manuscript follows this trend and considers an interesting and relevant application of deep learning to predict all-cause mortality directly from electrocardiography. The developed model demonstrated a superior performance compared to four different cardiologists. The authors conclude that their approach has the potential "to significantly improve clinical prediction models by incorporating large unstructured data such as videos of the heart."

Specific comments
1) A recent study looked at clinical variables and EF, plus 57 additional echocardiographic measurements to develop machine learning models to predict mortality (Samad et al JACC Cardiovasc Imaging. 2019 Apr;12(4):681-689). The highest prediction model had AUC of 0.89 which is superior to that observed by authors. Have authors attempted to directly develop their model on echocardiographic measurements or adding them in addition to the 3D CNN model?

2) Authors state that 812,278 videos were collected from a total of 34,362 patients i.e, ~24 videos per patient. Does multiple video inputs from a given patient influence model performance and/or accuracy in predicting the outcome i.e., all-cause mortality? The risk of overfitting when using DNN models is large, simply because these models are highly parameterized.

3) Resizing images to the same size without deforming patterns is a major challenge. Authors reported that images were resized using cropping or zero-padding to match the common dimension among the view group. While zero-padding has been reported to have no effect on the accuracy, cropping poses the risk of missing the features or patterns that appear in border areas. Why did or did not authors consider using scaling as the reasonable choice to resize images (the larger ones)? Did cropping involved regions of beam-

formed images?

4) It is unclear which architecture (i.e., 2D CNN+LSTM, 2D CNN+GAP, 3D-CNN & 3D CNN+GAP) was selected and used in each of the 24 views. Was one architecture preferred over the other in certain view types (e.g., apical vs parasternal etc.)? CNN architectures with GAP layers are useful in gathering information on what sections of the video were weighted more in the final decision.

5) Was the increase in average AUC observed in “All views + EHR” model significantly different from all echocardiography view model? Figure 2 suggests changes across the different models i.e., DNNs and cardiologists may not be statistically significant.

6) What is the rationale for selecting 10x10x10 voxels in the occlusion experiments? Could this be reason behind most impactful regions being concentrated in the lower risk patients?

Other comments:
7) “While these numerous sources of data offer the potential for more precise and accurate clinical predictions, humans have limited capacity for data integration in decision making” This sentences can be contested, usually in a busy interventional proecdures, well trained humans can multitask with images (videos), pressure graphs and audio signals of heatbeats presented simultaneously during complex cardiological procedures. This sentence does not do justice to complex abilities by just presenting the complexity of video. The time of 15-20 minutes for this images is also not accurate, highly trained physicians can read multiple loops of echo (4 videos at a time- not uncommon to present 4 cardiac sequence videos simultaneously during stress echocardiography. Authors are not being accurate in their descriptions.

8) Authors used all cause mortality over 1 year, more details should be provided. How was this determined? Were all patients followed over exactly one year? Did author determine the type of mortality? Echo would be more useful to determine cardiac mortality, was this determined?

9) Please consider plotting the aggregated cardiologist’s accuracy score in Figure 2.

10) In particular, the statistical tests/methods employed to test model accuracies are not sufficiently transparent and lacks rigor. The software/packages used for performing the statistical and survival analysis should be reported.

Reviewer #2 (Report for the authors (Required)):

This is a very interesting paper that shows some of the potentials of deep learning methods for survival prediction. The manuscript is timely, uses a very large training dataset, with a fairly comprehensive approach in terms of statistical analysis of results, showing results from a number of experiments performed to demonstrate performance comparison.

In general, the paper is very well written and results are convincing. I would like to however point out that given potential for media and community attention associated with AI, the scientific community has to be very careful to ensure that the messaging here should not be about AI outperforming experts. Many experiments to-date which show such results, including those in this paper, simply isolate a physician from the patient chart, show them purely image data in a laboratory environment, and ask the physicians to predict or diagnosis X. This is not a very realistic or fair comparison.

In fact, the title of this paper itself is talking about AI improving the prediction. The messaging to me should be that if a cardiologist uses AI along with their own assessment, their prediction improves. As such, a study similar to whether human+AI would outperform both AI and human is warranted.

In terms of technical novelty, the paper uses some established network architectures. The literature is currently filled with many groups tackling to automate echocardiography data analytics. The authors do a great job citing some of those. Perhaps some others I would like to suggest are:

https://www.nature.com/articles/s41586-020-2145-8?proof=trueMay%2F
https://www.ncbi.nlm.nih.gov/pubmed/31993508
https://link.springer.com/article/10.1007/s11548-019-01954-w
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8910601
https://www.ncbi.nlm.nih.gov/pubmed/30515886

https://link.springer.com/chapter/10.1007%2F978-3-030-30493-5_24
https://www.tandfonline.com/doi/abs/10.1080/21681163.2019.1650398

One should also account for observer variability across labels used to train the AI models. Some literature are emerging including those below, which may help authors formulate future expansion of their work. These could be possibly cited in the discussion:

https://ieeexplore.ieee.org/abstract/document/8932548
https://link.springer.com/chapter/10.1007/978-3-030-32245-8_77

Reviewer #3 (Report for the authors (Required)):

I think overall this is a good paper regarding machine learning analysis
of echocardiograms. The authors have used raw (annotation free) echos to make mortality prediction and compared their neural network performance with image derived features, clinical variables and cardiologists.

Reviewer #4 (Report for the authors (Required)):

Summary
The authors used machine learning of a large echocardiographic dataset to predict all-cause mortality. They hypothesized that a DNN could learn spatiotemporal features of > 800,000 raw echocardiography videos in 34,362 patients could predict outcomes models better than traditional clinical indices. Their developed model predicted mortality at 1 year with an AUC of 0.83 (0.83-0.84) in 5-fold cross-validation using images from all views and EHR data. This was superior to the Pooled Cohort Equation and the Seattle Heart Failure risk score. They found that machine learning based on the parasternal long axis echocardiographic view performed superior to cardiologist predictions using 10 EHR variables, with AUCs of 0.84 (0.81-0.87) vs. 0.68 (0.64-0.71), respectively. They further compared performance to four cardiologist predictions using echocardiographic videos to discriminate which demographic populations would have mortality, and the DNN outperformed cardiologists with accuracy of
82% vs. 66, 70, 73, and 76% respectively. Finally, they evaluated the model for predicting 1-year mortality against the Seattle HF score in a population with either HFrEF or HFpEF and found that the DNN outperformed the Seattle model with AUC of 0.76 (0.74-0.77) vs. 0.70 (0.68-0.71).

Degree of Advance
The authors claim that “using video data also increases technical complexity and thus initial efforts to apply deep learning to echocardiography have focused on ingesting individual images rather than full videos.” However, this effort has recently been published [Ouyang et al, Nature 2020. PMID: 32269341]

Implications
This study adds to the growing evidence base that ML algorithms incorporating data from different sources can improve clinical risk stratification and prediction of adverse outcomes over human experts presented with the same information. This work sets the stage for prospective or other clinical utility studies as decision support tools. The exploration of video analysis for outcome determination is novel and the authors should be congratulated on this accomplishment.

Major technical
1. The authors should clarify which model architectures were used for each view that contributed to the “full” XGboost model. The following information may improve the presentation:
a. A table of final ‘best’ architectures used for each view
b. A table of one example view (ie., PLAX) that shows differences in test characteristics between different architectures (to see how much architecture contributes to success)
c. A figure showing overall architecture schematic that incorporates all views and clinical variables for the “DNN model (full)”.
2. It is not clear how videos were distributed between patients – for an average of ~20 per patient. Were some patients over and others under-represented? Was this distribution uniform in training, validation and test?
3. Why use cross-validation cohorts rather than a traditional training/validation/test (hold out) cohort? This latter approach may be more robust and reduce the risk of overfitting.

4. Cardiologists' reads are not clinically relevant. Reading blindly and independently from 21 variables alive or dead at 1 year plus an echocardiographic image is not in accordance with clinical practice. In making treatment decisions on mortality, the cardiologist has interview, physical exam, the EHR, imaging of multiple varieties etc. These intangibles likely improve the ability of the human to determine appropriateness of a given therapy. This must be emphasized in the study, since the dataset presented to the cardiologists is artificially limited.
5. The training level and experience of the cardiologists is critical yet not clear.
6. The low risk occlusion experiments seem to highlight the mitral annulus and apparatus in all views. This is important because diastolic function is assessed by annular velocities in practice, which changes significantly with age. These areas also calcify with age. It may stand to reason that low risk images will lack significant mitral annular calcification and will have increased annular velocities.
7. On the high-risk side, there may be several variable features that would portend mortality in different images for different pathologies, such as abnormalities within the heart like wall motion abnormalities in myocardial infarction patients or outside of the heart such as pericardial effusions/pleural effusions or adipose tissue at the probe tip.
8. Could the authors address if these areas evaluated by blanking identify novel echo findings that are relevant?
9. Are outside institution data being identified for generalizability?
10. How close is this work to being assessed prospectively?

Minor technical
1. Lines 88-20 are not clear and may be interpreted that the PLAX was combined with EHR data to create a model that predicted 1-year mortality rather than what I believe is intended: that the PLAX model without EHR data was compared to cardiologists with EHR data.
2. Lines 100-102 do not clearly state that the cardiologists and DNN were presented with echocardiogram videos.
3. Why use balanced dataset in final performance evaluation?
PLAX + 10 EHR variables (how were tehse chosen?)
Statistics protocols or materials

Citations
Ouyang et al, Nature 2020. PMID: 32269341

Mon 28/09/2020
Decision on Article NBME-20-0869A

Dear Dr Fornwalt,

Thank you for your revised manuscript, "Computer-assisted analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality", which has been seen by the original Reviewers #1, #2 and #4. In their reports, which you will find at the end of this message, you will see that Reviewers #2 and #4 are now happy with the work, and that Reviewer #1 is concerned about model underperformance given the stated limitations, arguing as well that the claim that the model improves the performance of cardiologists may need to be qualified. I hope that in a further revised version of the work you can address this reviewer's points.

As before, when you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

I look forward to receive a further revised version of the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

Thank you for revising your work and incorporating several revision changes we have requested. I remain concerned about two aspects of your study which in my opinion is not well addressed.

1. Lack of adequate external validation. Although you have used Ouyang's Stanford data-set, the use of EF as a surrogate of mortality prediction is not convincing in the figure presented and not equivalent to an external validation. While I understand that getting an external model with labeled outcome is a tough task, the generalizability of the model in a different institute/ region would require more confirmation.

2. The occlusion experiments are puzzling and at present the information shown are not yielding any tangible knowledge about how the neural networks are performing. The cardiologist experiments where these data are presented is also awkwardly designed since a cardiologist would not be able to depend on this information on improving his ability to discriminate a patient's risk. Moreover, the goal of echo examination is to diagnose a disease and the severity of the disease, have the authors attempted grouping patients eg. Heat failure, valvular heart disease, hypertensive heart disease and then looked for the performance for increasing severity of the disease and how the model behaves versus the physicians grade the disease. In the absence of a tangible understanding claiming the task performed by the physicians is inferior to that performed by the machine is problematic. In my opinion this section is controversial and the experiments are not performed in depth to arrive at the
stated conclusions.

3. Authors suggests that the lower AUC than the previous work than one observed by the model is due to lower sample size, then in most instances the previous model would be used, the incremental value of the current model is unclear. This limitation suggest need for more work and should be included in the main

conclusions and the abstract.

Reviewer #2 (Report for the authors (Required)):

Thank you for detailed response and new experiments. I recommend acceptance.

Reviewer #4 (Report for the authors (Required)):

The authors have improved the manuscript and addressed all major concerns.

Mon 09/11/2020
Decision on Article NBME-20-0869B

Dear Dr Fornwalt,

Thank you for your revised manuscript, "Computer-assisted analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality". Having consulted with Reviewer #1 (whose comments you will find at the end of this message), I am pleased to say that we shall be happy to publish the manuscript in *Nature Biomedical Engineering*, provided that the points specified in the attached instructions file are addressed.

When you are ready to submit the final version of your manuscript, please upload the files specified in the instructions file.

For primary research originally submitted after December 1, 2019, we encourage authors to take up transparent peer review. If you are eligible and opt in to transparent peer review, we will publish, as a single supplementary file, all the reviewer comments for all the versions of the manuscript, your rebuttal letters, and the editorial decision letters. **If you opt in to transparent peer review, in the attached file please tick the box 'I wish to participate in transparent peer review'; if you prefer not to, please tick 'I do NOT wish to participate in transparent peer review'.** In the interest of confidentiality, we allow redactions to the rebuttal letters and to the reviewer comments. If you are concerned about the release of confidential data, please indicate what specific information you would like to have removed; we cannot incorporate redactions for any other reasons. If any reviewers have signed their comments to authors, or if any reviewers explicitly
agree to release their name, we will include the names in the peer-review supplementary file. More information on transparent peer review is available.

Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

I think the limitations are well stated, the manuscript is acceptable in its current form.

Rebuttal 1

Dear Brandon,
Thank you again for submitting to Nature Biomedical Engineering your Article, "Automated analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality". The manuscript has been seen by four experts, whose reports you will find at the end of this message. You will see that the reviewers have good words for the work, and that Reviewers #1 and #3 raise a number of technical criticisms that we hope you will be able to address. In particular, we would expect that a revised version of the manuscript provides:

- *Evidence of any improvements in performance when expert cardiologists are aided by the outcomes of the deep-learning algorithm.*
- *Extended occlusion studies, to identify any additional anatomical features contributing to pathology-specific mortality risk.*
- *Evidence that the training process has been designed to minimize overfitting, and that the outcomes are not influenced by any correlation of mortality with the number of videos per patient.*
- *Thorough methodology, including model architectures and all optimizer and model-specific hyperparameters (for reproducibility purposes).*

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.
Please follow the following recommendations:

- *Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).*
- *If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).*
- *If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).*
- *Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.*
- *The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).*
- *Provide the rebuttal to the reviewer comments and the cover letter as separate files.*

We hope that you will be able to resubmit the manuscript within 15 weeks from the receipt of this message. If this is the case, you will be protected against

potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at Nature Biomedical Engineering. Because of the COVID-19 pandemic, should you be unable to carry out work in the near future we advise that you reply to this message with a revision plan in the form of a preliminary point-by-point rebuttal to the comments from all reviewers that also includes a response to any points highlighted in this decision. We should then be able to provide you with additional feedback.
We hope that you will find the referee reports helpful when revising the work, which we look forward to receive. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

Answer:
We appreciate the opportunity to revise the paper, and addressing the thoughtful comments provided by the reviewers and editors has significantly improved our work through substantial revisions as detailed below. In particular, we have resolved the following issues emphasized by the editor:

- We conducted an additional cardiologist survey designed to evaluate the change in cardiologists' prediction performance when assisted by the DNN model, see Table 2 and Figure 2c. This analysis provides evidence of potential performance improvements, as all cardiologists improved their accuracy and the aggregated score improved to 0.78 AUC.

- We extended the occlusion experiments by selecting 80 cases from the initial experiment and providing our cardiology colleagues with predicted, true outcomes, and occlusion maps for patients with and without history of myocardial infarction and heart failure. The four cardiologists who reviewed the samples all concluded that they could not identify patterns that would help them better assess the patients. This finding suggests to us that, while this occlusion experiment approach was rational and may still prove successful, there are many variables involved (e.g., the size of the occlusion mask, the handling of temporal features) that must be optimized to prove that our approach will contribute to the current clinical practice. Such extensive optimization is, we believe, beyond the scope of this work, so we have not attempted additional occlusion experiments.

- In the response to the second comment of Reviewer #1 (Rev 1.2), we listed the measures that we took to avoid overfitting. We also added the supplementary section "Note on Echocardiography View Missingness:" to clarify that the model did not know when a video was missing because of a prior imputation step, thus it could not be influenced by the number of available videos in an echocardiography study. To confirm this, we tested whether the performance was being affected by the number of available

videos for each patient and we found no evidence of performance difference among the tertiles with low, medium, and high numbers of available videos.

- We included Figure 11, which in addition to Tables 5, 6, 7, and 8 contain all necessary details to reproduce the proposed models in the paper.

Below, we display the reviewer's comment in italics font followed by our answer. We highlighted extracts of the paper in blue.

Reviewer # 1

The paper describes the application of deep neural networks for learning spatio-temporal predictive feature representations from echocardiography videos that are associated with augmenting the prediction of 1-year all-cause mortality. There has been a recent interest in using 2D and 3D neural network architectures and deep learning for addressing important problems in cardiology. The proposed manuscript follows this trend and considers an interesting and relevant application of deep learning to predict all-cause mortality directly from electrocardiography. The developed model demonstrated a superior performance compared to four different cardiologists. The authors conclude that their approach has the potential "to significantly improve clinical prediction models by incorporating large unstructured data such as videos of the heart."

Rev 1.1:

1) A recent study looked at clinical variables and EF, plus 57 additional echocardiographic measurements to develop machine learning models to predict mortality (Samad et al JACC Cardiovasc Imaging. 2019 Apr;12(4):681-689). The highest prediction model had AUC of 0.89 which is superior to that observed by authors. Have authors attempted to directly develop their model on echocardiographic measurements or adding them in addition to the 3D CNN model?

Answer:

We thank the reviewer for pointing out our previous work, [22]. The reported performance for 1-year mortality prediction in [22] was 0.85 AUC, 95% CI [0.84, 0.85], for a dataset of 331,317 echocardiography studies. In the current study, and for the same set of inputs, we report an AUC of 0.81, 95% CI [0.80,0.82] with a cross-validation set of significantly fewer (42,095) echocardiography studies.

The drop in performance is explained by the reduced sample size. To demonstrate it, we repeated the cross-validation experiment 100 times by randomly sampling 42,095 echocardiography studies (equal to our cross-validation set size, see Table 4) from the same set of 331,317 studies as in [22] and collecting the performance estimates. The average AUC across the 100 iterations was 0.812, 95% CI [0.810, 0.814], which matches the report performance in our limited dataset.

The DICOM file that contains an echocardiography study occupies approximately 1GB of disk storage. Retrieving all studies in [22] would have required

330TB. At the time of this study, we did not have access to that large amount of storage, thus we limited our sample size to meet our hardware capabilities. We added supplementary section "Comparison of cross-validation results with Samad et. al:" to report the comparison analysis and mentioned this limitation in the main text of the paper.
In the present paper, we extend [22] by adding video data to the input set. The "All Views+EHR" model incorporated video and tabular data (same tabular inputs as in [22]) and yielded the highest AUC (0.84, 95% CI [0.84,0.85]). For comparison, the "EHR" model proposed in [22] yielded an AUC of 0.81 for the same limited subset, see Figure 1.
To clarify the concern raised by the reviewer, we now explicitly state the "All Views+EHR" model performance in paragraph 6 of the main text:

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].

Rev 1.2:
2) Authors state that 812,278 videos were collected from a total of 34,362 patients i.e, ∼24 videos per patient. Does multiple video inputs from a given patient influence model performance and/or accuracy in predicting the outcome i.e., all-cause mortality? The risk of overfitting when using DNN models is large, simply because these models are highly parameterized.

Answer:

There are two sources of multiple videos per patient:

1. Multiple echocardiography studies taken from the same patient. The cross-validation set contains 42,095 studies from 34,362 patients.

2. Multiple videos within a single echocardiography study. A study may contain up to 24 videos each from a unique view.

In case the reviewer refers to item 1, we generated the cross-validation set by randomly sampling at the patient identifier level and expanding the evaluation of the model performance to all available studies from the selected patients. Therefore, as the reviewer points out, the performance estimate in a test fold could be influenced by the presence of a patient with multiple echocardiography studies. When enforcing a single study per patient (randomly selecting a study per patient) rather than using all available studies in the test folds, the average AUC for the "EHR" and "All Views + EHR" models increased from 0.806 to 0.817 and 0.841 to 0.850 respectively. We choose to report the AUC with the lower AUC estimate as it is the most conservative approach.
In case the reviewer refers to item 2, the average number of videos in the echocardiography studies from patients that survived beyond a year was significantly higher ($p < 0.01$) than in those that did not by a single video, 19.4 vs 18.3. Yet, the number of videos in a echocardiography study did not affect the model performance. To demonstrate this, we divided the echocardiography studies in each test fold into three groups that contained a low, medium, and high number of videos within the study, see Supplementary Table 5. Conducting an ANOVA

test, we did not find evidence that at least a group has a significantly different mean ($p = 0.12$) across the five folds.
To clarify this point, we added the supplementary section:

> **Note on Echocardiography View Missingness:**
>
> In clinical practice, different specific views may be acquired during an echocardiogram as ordered by a cardiologist. The number of views obtained from the patient therefore may hold the potential to embed information on the patient likelihood of dying within a year. When deployed, a mortality risk model built from echocardiography video data should be robust to these potential video missingness patterns.
> To enable support and robustness to missing videos, we incorporated an imputation step prior to the classification step, see Figure 11. Thus, the classification step should not learn from missingness patterns. In order to confirm the robustness of the proposed model, we searched for evidence a relationship between the number of available videos and model performance. We grouped studies on each test set of the cross-validation folds into tertiles. Then, we tested the null hypothesis that at least one of the tertiles had a significantly different mean AUC. The tertile ranges were (0,19], (19,21], and (21,24], where the average AUC for each group was 0.83, 0.86, and 0.85, respectively. Conducting an ANOVA test, we did not find evidence that at least one group had a significantly different mean ($p = 0.12$).

and summary statistics in paragraph 3 of the Image Collection and Preprocessing section as follows:

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

We should also note that, as shown in Figure 1, using multiple echocardiography video views ("All views") yielded higher performance (AUC: 0.83, 95% CI [0.83,0.84]) for predicting 1-year mortality than any single view, which ranged in performance from AUC of 0.70 to 0.80. We agree with the reviewer that training multiple videos at a time would require large networks that could be prone to overfitting. We took several preventive measures as listed below:

a. One view at a time: The DNNs were trained independently for each view resulting in several small networks. While a joint model has the potential to improve performance, a limiting factor is that not all views were collected for all patients. An exclusion criteria that removes patients based on the view availability would bias the sampled population. Moreover, GPU RAM limitations (32 GB) made it challenging to fit all 24 views within a single batch.

b. Low parameter design: We designed neural network architectures with a

low number of parameters as described in the last paragraph of the Neural Network Architectures section.

> 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 x 5 folds x 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 days. Deep learning models typically consist of millions of parameters, for example the Inception model has 25M parameters [50] and ResNet more than 40M parameters [51], 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, see Tables 5, 6, 7, and 8.

c. Early Stopping: Following the recommendations presented in [52], we also monitor the binary cross-entropy loss on a partition of the training set—referred to as the internal validation set—to avoid over training the model as described in the Cross-validation Procedure section:

> 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 10 epochs, we stopped the training and recovered the model weights at the minimum validation loss [52].

d. Dropout Layers: We have included aggressive dropout regularization [5] layers (50%) to each of the neural network designs as shown in Tables 5, 6, 7, and 8.

e. Separate train and test sets by patient: We conducted cross-validation experiments where the train set was separate from the test set and no patient was shared between them. We observed that the training set performance was close to the test set performance, which suggests that overfitting was not occurring (i.e. if overfitting was occurring, the training set performance would be significantly higher than the test set performance). For our largest models, "All views+EHR", the AUC on the training set was below 0.87 across all five folds.

f. Raw videos: We used the purest video data form. The raw video did not contain any artificial markings on which the model could learn and potentially overfit.

Rev 1.3:

3) Resizing images to the same size without deforming patterns is a major challenge. Authors reported that images were resized using cropping or zero-padding to match the common dimension among the view group. While zero-padding

has been reported to have no effect on the accuracy, cropping poses the risk of missing the features or patterns that appear in border areas. Why did or did not authors consider using scaling as the reasonable choice to resize images (the larger ones)? Did cropping involved regions of beam-formed images?

Answer:
We appreciate the reviewer's concern. We preferred not to scale the videos and preserve the natural anatomical proportions which we believe are important to identify a healthy heart. We also note that the cropping and padding was often minimal compared to the original image size.
We added the following clarification to the Image Collection and Preprocessing subsection:

> We note that the image size normalization (cropping and padding) had a minimal effect on the video because the standard echocardiography views center the anatomical region of interest. For example, we only cropped and padded more than 6 rows on less 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 (see Figures 4 and 5).

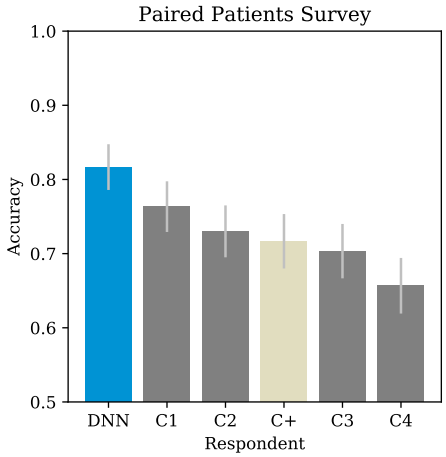
Rev 1.4:
4) It is unclear which architecture (i.e., 2D CNN+LSTM, 2D CNN+GAP, 3D-CNN & 3D CNN+GAP) was selected and used in each of the 24 views. Was one architecture preferred over the other in certain view types (e.g., apical vs parasternal etc.)? CNN architectures with GAP layers are useful in gathering information on what sections of the video were weighted more in the final decision.

Answer:
For each fold of the cross-validation experiment, we trained all architectures, picked the one with the highest AUC on the internal validation set, and applied it to the test set. There is no guarantee that one of the architectures was consistently preferred over others. We added the chosen architecture for each view and fold in the supplementary Table 2.
In summary, among the 24 video models and 5 folds (120 trainings) the 3D architecture was chosen 56 times (34 and 22 for the kernel sizes 3 and 5), the 3D+GAP architecture was chosen 53 times (39 and 14 for the kernel sizes 3 and 5), the 2D+GAP architecture was chosen 9 times and the 2D+LSTM was chosen 2 times.

Rev 1.5:
5) Was the increase in average AUC observed in "All views + EHR" model significantly different from all echocardiography view model? Figure 2 suggests changes across the different models i.e., DNNs and cardiologists may not be statistically significant.

Answer:
The "All views+EHR" model yielded marginally better AUC than all the views combined (see "All views" model in the cross-validation set in Figure 1).

Thanks to the reviewer comment, we identified a mistake on the error bars in Figure 2b. We amended the error, which was displaying twice the size of the intended 95% CI, and updated the figure.



The error bars now align to the reported statistical inference results, where the model yielded significantly improved performance compared to three out of four of the cardiologists. We should clarify that the "All views+EHR" and "All views" models were not directly compared to cardiologists. Instead, we built a separate model that used the same limited input set (PL DEEP view + 10 EHR variables) as the cardiologists to predict one-year mortality. We state that the performance of this model was superior to that of the cardiologists in paragraph 8 of the main text:

> The DNN model yielded an AUC of 0.84, 95% CI [0.81, 0.87], while the aggregated cardiologist score yielded an inferior AUC of 0.68, 95% CI [0.64, 0.71], see Table 1 and Figure 2a.

Rev 1.6:

6) What is the rationale for selecting 10x10x10 voxels in the occlusion experiments? Could this be reason behind most impactful regions being concentrated in the lower risk patients?

Answer:

The reviewer's insight may be correct. This is an exploratory analysis that we undertook to try to identify whether the DNNs were picking up on anatomical features. There is no precedent on the choice of this occlusion region, so it served as a starting point informed on being small enough to be anatomically selective (wall thickness and movement) while being large enough to cover the aggregation of multiple CNN filter responses. The largest CNN kernel consisted of 5x5x5 pixels.

The reviewer may be right that regions on high risk patients may respond differently to variations on the occlusion box size. However, it is beyond the scope of this work to do an exhaustive search in trying to optimize the appropriate size of the occlusion boxes. Interpreting these maps remains a challenge and we hope to study it more in future publications.

We added the following paragraph to the manuscript main body to note the

challenge of interpreting the occlusion maps.

> the interpretation of the findings of the DNN remains challenging, and this is a current problem facing all DNN models [46]. While we did use occlusion maps to show evidence that the DNN was affected by anatomically appropriate regions of the heart, we can only present empirical findings on a case by case basis without definitive general interpretation.

Rev 1.7:
Other comments:
7) "While these numerous sources of data offer the potential for more precise and accurate clinical predictions, humans have limited capacity for data integration in decision making" This sentences can be contested, usually in a busy interventional proecdures, well trained humans can multitask with images (videos), pressure graphs and audio signals of heatbeats presented simultaneously during complex cardiological procedures. This sentence does not do justice to complex abilities by just presenting the complexity of video. The time of 15-20 minutes for this images is also not accurate, highly trained physicians can read multiple loops of echo (4 videos at a time- not uncommon to present 4 cardiac sequence videos simultaneously during stress echocardiography. Authors are not being accurate in their descriptions.

Answer:
We thank the reviewer for the insight. It was not our intention to diminish the complexity of this task or the abilities of our clinical colleagues. We consulted with the clinician co-authors on this point and refined the first paragraph of the main text as follows:

> 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, computed tomography) and other diagnostics (e.g. electrocardiograms). While these numerous sources of data offer the potential for more precise and accurate clinical predictions, humans may have limits to recognition of echocardiographic patterns otherwise discerned by integration of data into complex decision making that is assisted by computers [1]. Hence, there is both a need and a substantial opportunity to leverage technology, such as machine learning, to manage this abundance of data and ultimately provide intelligent computer assistance to physicians [2, 3, 4].

Rev 1.8:
8) Authors used all cause mortality over 1 year, more details should be provided. How was this determined? Were all patients followed over exactly one year? Did author determine the type of mortality? Echo would be more useful to determine cardiac mortality, was this determined?

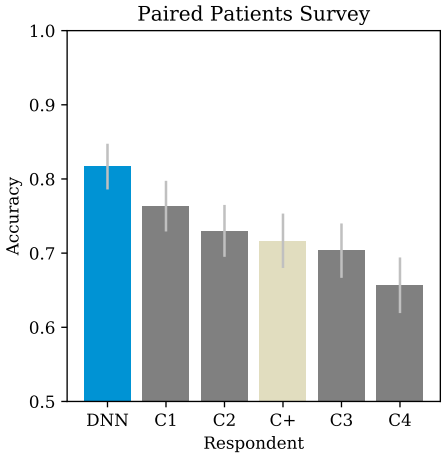
Answer:

We followed patients starting from the echocardiography study date until the death date or most recent encounter in our system. Our records are cross-referenced monthly against national death index data to ensure accuracy of the death date for each patient.
We did not determine the cause of death. We expect that most deaths are related to heart disease in our clinical population, however, we currently do not have a feasible approach to confirm the cause of death on all patients. Deriving cause of death from large EHR datasets in retrospect is a particularly challenging and potentially impossible task.
We wrote additional details on the label definition on paragraph 3 of the Electronic health records data preprocessing section.

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

Rev 1.9:
9) Please consider plotting the aggregated cardiologist's accuracy score in Figure 2.

Answer:
We added the aggregated cardiologist accuracy as the C+ beige bar for the paired survey in Figure 2b. For the Individual Patients Survey in Figure 2a, the aggregated cardiologist score is shown in the grey line.



Rev 1.10:
10) In particular, the statistical tests/methods employed to test model accuracies are not sufficiently transparent and lacks rigor. The software/packages used for performing the statistical and survival analysis should be reported.

Answer:
We added a statistical analysis subsection to the methods.

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 version 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 with $\pm 1.96\sigma/\sqrt{5}$.
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 2.5^{th} and 97.5^{th} 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 version 0.11.1.

Reviewer # 2

This is a very interesting paper that shows some of the potentials of deep learning methods for survival prediction. The manuscript is timely, uses a very large training dataset, with a fairly comprehensive approach in terms of statistical analysis of results, showing results from a number of experiments performed to demonstrate performance comparison.

In general, the paper is very well written and results are convincing. I would like to however point out that given potential for media and community attention associated with AI, the scientific community has to be very careful to ensure that the messaging here should not be about AI outperforming experts. Many experiments to-date which show such results, including those in this paper, simply isolate a physician from the patient chart, show them purely image data in a laboratory environment, and ask the physicians to predict or diagnosis X. This is not a very realistic or fair comparison.

In fact, the title of this paper itself is talking about AI improving the prediction. The messaging to me should be that if a cardiologist uses AI along with their own assessment, their prediction improves. As such, a study similar to whether human+AI would outperform both AI and human is warranted.

Answer:

We thank the reviewer for the comments and agree strongly with that sentiment. To help emphasize this point, we conducted an additional survey where we presented the cardiologists with information extracted from the model to assess whether the model prediction helped them improve their performance (with vs. without that model-derived data). In summary, the cardiologists improved their sensitivities by 13% while maintaining specificity constant.

Motivated by the results of this new experiment, we modified the title to "**Computer-assisted** analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality" and the second paragraph of the abstract now reads:

> We trained neural networks to predict 1-year all-cause mortality using a clinically-acquired echocardiography dataset of 812,278 videos (34,362 patients). The neural network showed superior performance to 1) a machine learning model using 58 human-derived variables from the echocardiogram and 100 EHR-derived clinical variables, 2) the widely-used Pooled Cohort Equations, and 3) the Seattle Heart Failure score (measured in an additional 2,404 patients with heart failure who underwent 3,384 echocardiograms). Furthermore, cardiologists substantially improved their predictive performance when assisted by the neural network prediction, increasing sensitivity by 13% while maintaining specificity. These results highlight the potential of neural networks to improve clinical prediction models by incorporating large unstructured datasets such as videos of the heart.

We also added a description in the Cardiologist survey section as follows:

> We presented the same 600 patients twice. First, we showed the individual sample as in Figure 6 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.
In order to avoid incremental performance changes while the cardiologists progressed through the survey, we presented them, prior to 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 four 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 5 examples that fell into each of the four quadrants of the confusion matrix. Figure 8 shows the interface for the model assisted portion of the third survey, where we added a "Machine Prediction" row and a occlusion map video.

And, we reported the results in the main body text as follows:

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 reduced less than 1% on average, see Table 2 and Figure 2c.

Rev 2.1:
In terms of technical novelty, the paper uses some established network architectures. The literature is currently filled with many groups tackling to automate echocardiography data analytics. The authors do a great job citing some of those. Perhaps some others I would like to suggest are:
https://www.nature.com/articles/s41586-020-2145-8?proof=trueMay
https://www.ncbi.nlm.nih.gov/pubmed/31993508
https://link.springer.com/article/10.1007/s11548-019-01954-w
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8910601
https://www.ncbi.nlm.nih.gov/pubmed/30515886
https://doi.org/10.1007/978-3-030-30493-5_24
https://www.tandfonline.com/doi/abs/10.1080/21681163.2019.1650398

Answer:
We thank the reviewer for providing additional literature. We added the references to these articles in multiple places throughout the paper. The new references are listed as [34, 35, 33, 31, 32, 36]. The study cited in [18] was in the original text.

Rev 2.2:
One should also account for observer variability across labels used to train the

AI models. Some literature are emerging including those below, which may help authors formulate future expansion of their work. These could be possibly cited in the discussion:
https://ieeexplore.ieee.org/abstract/document/8932548
https://link.springer.com/chapter/10.1007/978-3-030-32245-8_77

Answer:
We agree with the reviewer that uncertainty in labels is problematic, however the focus on mortality helps to minimize the problem because all-cause mortality is a definitive binary outcome not subject to interpretation.
To clarify the reviewer's point, we added a paragraph to the section "Electronic health records data preprocessing":

> 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.

Reviewer # 3

I think overall this is a good paper regarding machine learning analysis of echocardiograms. The authors have used raw (annotation free) echos to make mortality prediction and compared their neural network performance with image derived features, clinical variables and cardiologists.

Reviewer # 4

The authors used machine learning of a large echocardiographic dataset to predict all-cause mortality. They hypothesized that a DNN could learn spatiotemporal features of > 800,000 raw echocardiography videos in 34,362 patients could predict outcomes models better than traditional clinical indices. Their developed model predicted mortality at 1 year with an AUC of 0.83 (0.83-0.84) in 5-fold cross-validation using images from all views and EHR data. This was superior to the Pooled Cohort Equation and the Seattle Heart Failure risk score. They found that machine learning based on the parasternal long axis echocardiographic view performed superior to cardiologist predictions using 10 EHR variables, with AUCs of 0.84 (0.81-0.87) vs. 0.68 (0.64-0.71), respectively. They further compared performance to four cardiologist predictions using echocardiographic videos to discriminate which demographic populations would have mortality, and the DNN outperformed cardiologists with accuracy of 82% vs. 66, 70, 73, and 76% respectively. Finally, they evaluated the model for predicting 1-year mortality against the Seattle HF score in a population with either HFrEF or HFpEF and found that the DNN outperformed the Seattle model with AUC of 0.76 (0.74-0.77) vs. 0.70 (0.68-0.71).

Degree of Advance

The authors claim that "using video data also increases technical complexity and thus initial efforts to apply deep learning to echocardiography have focused on ingesting individual images rather than full videos." However, this effort has recently been published [Ouyang et al, Nature 2020. PMID: 32269341]

Implications

This study adds to the growing evidence base that ML algorithms incorporating data from different sources can improve clinical risk stratification and prediction of adverse outcomes over human experts presented with the same information. This work sets the stage for prospective or other clinical utility studies as decision support tools. The exploration of video analysis for outcome determination is novel and the authors should be congratulated on this accomplishment.

Answer:

We appreciate the comments made by the reviewer. The paper has been expanded to address the comments made by the reviewer.

About the degree of advance, we should note that the previous (not yet peer-reviewed) versions of our work date back to 26/11/2018 with arXiv identifier arXiv:1811.10553 which was prior to the Ouyang study cited above. Still, we understand the concern and thus removed the claim. We now included the paper by Ouyang et al. [34] in our discussion.

The last sentence of the fourth paragraph in the main text now reads:

> Initial efforts using DNNs have focused on using individual images [29, 14, 30] and full video models to estimate hand-crafted measurements [31, 32, 33, 34, 35, 36]. We investigate an approach that directly estimates patient outcomes from raw videos.

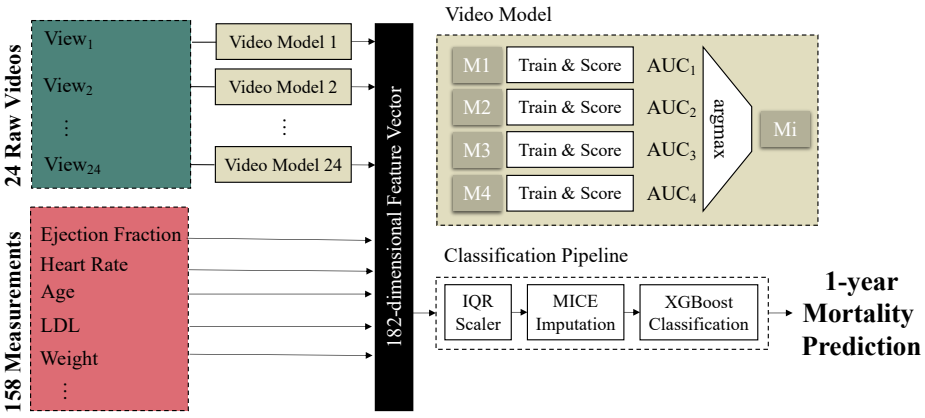
Rev 4.1:

Major technical

1. The authors should clarify which model architectures were used for each view

that contributed to the "full" XGboost model. The following information may improve the presentation:
a. A table of final 'best' architectures used for each view
b. A table of one example view (ie., PLAX) that shows differences in test characteristics between different architectures (to see how much architecture contributes to success)
c. A figure showing overall architecture schematic that incorporates all views and clinical variables for the "DNN model (full)".

Answer:
a. We added supplementary Table 2, where we show the architecture chosen for each view at each fold. Unfortunately, there was no commonly preferred architecture, although the 3D architectures (3D CNN and 3D CNN+GAP) were generally preferred to 2D architectures.
b. We added supplementary Table 3, where we show the performance obtained in the train, validation, and test sets of each fold for the PLAX example view (called PL DEEP in the paper).
c. We added Figure 11 to show the overall architecture.



Rev 4.2:
2. It is not clear how videos were distributed between patients – for an average of 20 per patient. Were some patients over and others under-represented? Was this distribution uniform in training, validation and test?

Answer:

The cross-validation set included 42,095 studies from 34,362 patients, that is 1.2 studies per patient. An echocardiography study consisted of multiple views, but not all views were available on all studies. There was 19.2 videos available per study. Therefore, the number of videos that relate to a patient is the number of videos available on each of the echocardiography studies that the patient underwent.
We added Supplementary Table 4 where we show the distribution of number of studies per patient:

# Studies	# Patients
1	28611
2	4344
3	1013
4	269
5	86
6	25
7	12
9	1
8	1

And, Supplementary Table 5, where we show the distribution of number of views per study:

Tertile	# Views	# Studies	Total
I	1	445	15,892
	2	376	
	3	110	
	4	88	
	5	159	
	6	212	
	7	68	
	8	102	
	9	174	
	10	359	
	11	607	
	12	240	
	13	235	
	14	293	
	15	493	
	16	886	
	17	1,596	
	18	2,947	
	19	6,502	
II	20	8,879	13,187
	21	4,308	
III	22	8,790	13,016
	23	2,973	
	24	1,253	

We refer Reviewer #4 to the response to Reviewer #1 comment 2 (Rev 1.2) for details on how each of this distribution of video representation affects performance. In summary, enforcing a single video per patient inflates the model AUC performance estimate, while being unaffected by the number of videos available in a study. We choose to report the AUC with the lower AUC estimate as it is the most inclusive and conservative approach.

Due to the random selection of the five folds in our cross-validation experiment, we expect the distribution to follow that of the entire set.

Rev 4.3:
3. Why use cross-validation cohorts rather than a traditional training / validation / test (hold out) cohort? This latter approach may be more robust and reduce the risk of overfitting.

Answer:
By using multiple folds in a cross-validation experiment we are able to measure the uncertainty of the performance estimations. We reported performance both in the cross-validation and on hold out cohorts. The model trained on the 42,095 studies from the cross-validation cohort was tested on the held-out survey (600 studies) and the Heart Failure cohort (3,384 studies).
We listed the measures we took to avoid overfitting in the response to Reviewer #1 comment 2, Rev1.2.

Rev 4.4:
4. Cardiologists' reads are not clinically relevant. Reading blindly and independently from 21 variables alive or dead at 1 year plus an echocardiographic image is not in accordance with clinical practice. In making treatment decisions on mortality, the cardiologist has interview, physical exam, the EHR, imaging of multiple varieties etc. These intangibles likely improve the ability of the human to determine appropriateness of a given therapy. This must be emphasized in the study, since the dataset presented to the cardiologists is artificially limited.

Answer:

We agree with the reviewer that the survey does not reflect clinical standards and cardiologists would otherwise have access to a more complete assessment of the patient. We do note that similarly, the model had the same limitations i.e. access to only a limited dataset. We chose to present a limited dataset in order to have an equal comparison between the model and cardiologist, to make efficient use of the cardiologists time and to provide a baseline performance for a generic patient cohort (survey set). Future work needs to incorporate ways of inputting this more extensive data into the models so that we can directly compare to a more real-life situation as the reviewer mentions. We now added an additional note on the last paragraph of the Cardiologist Survey subsection:

> We acknowledge that none of these surveys was designed to represent normal clinical practice, and prediction performance in a "real world" setting is likely enhanced through access to the full medical record, physical exam, etc. 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.

Rev 4.5:
5. The training level and experience of the cardiologists is critical yet not clear.

Answer:
We added a sentence to the 8^{th} paragraph of the main section to clarify this point.

> We used the survey set to evaluate the performance of four expert cardiologists, three Core Cardiovascular Training Statement (COCATS) level 3 and one level 2 in echocardiography. The cardiologists were independently and blindly asked to determine whether each patient would be alive or dead at 1 year following the echocardiogram.

Rev 4.6:
6. The low risk occlusion experiments seem to highlight the mitral annulus and apparatus in all views. This is important because diastolic function is assessed by annular velocities in practice, which changes significantly with age. These areas also calcify with age. It may stand to reason that low risk images will lack significant mitral annular calcification and will have increased annular velocities.
7. On the high-risk side, there may be several variable features that would portend mortality in different images for different pathologies, such as abnormalities within the heart like wall motion abnormalities in myocardial infarction patients or outside of the heart such as pericardial effusions/pleural effusions or adipose tissue at the probe tip.
8. Could the authors address if these areas evaluated by blanking identify novel echo findings that are relevant?

Answer:

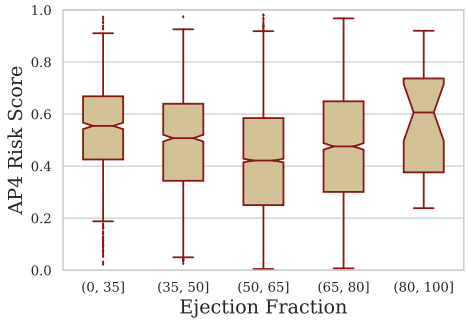
The reviewer presents reasonable speculation, but we are unfortunately unable to prove or disprove these hypotheses without substantial additional work that is beyond the scope of the current paper. We did add a new experiment (as described above in response to Reviewer #2) where cardiologists were presented with 80 occlusion map examples, and the cardiologists were unable to find patterns that helped them better discern between patients that would survive or not.
The interpretation of the occlusion maps remains extremely challenging and requires significant future work. We made note of this in the main text of the paper as follows:

> the interpretation of the findings of the DNN remains challenging, and this is a current problem facing all DNN models [46]. While we did use occlusion maps to show evidence that the DNN was affected by anatomically appropriate regions of the heart, we can only present empirical findings on a case by case basis without definitive general interpretation.

Rev 4.7:
9. Are outside institution data being identified for generalizability?

Answer:

To the best of our knowledge, there is no publicly available Echocardiography dataset with mortality outcomes. There is however a publicly available echocardiography dataset for segmentation with ejection fraction measurements [34], which can be used as a surrogate for mortality risk.
We applied our AP4 model trained on the entire cross-validation set and applied it to the 10,030 AP4 videos available within this public dataset. The predicted risk trend as a function of ejection fraction aligned with that reported in [40], as shown in Figure 1



We added a supplementary section reporting the results of this experiment.

Application to Stanford Dataset:

Ouyang et al. recently released a publicly available echocardiography video dataset with left ventricular ejection fraction (EF) measurements [34]. Even though our models were trained for mortality prediction, we applied the AP4 model trained on the cross-validation set to the 10,030 publicly available videos to report the mortality risk as a function of EF.
To make the size and frame rate compatible with our trained models, we padded the original 112 pixels width with zeros to a size of 150 pixels and cropped the height from 112 to 109 pixels. Finally, we resampled the frame rate to 30 frames per second with linear interpolation.
Using EF as a surrogate for mortality, we observed a trend similar to that reported in [40], see Figure 1, where the risk of mortality is lowest at normal EF ranges, 50% to 65%, and increases for both lower and higher EF ranges.

We believe that our health system, which consists of over 10 hospitals spanning north central Pennsylvania, would contain similar patient, acquisition hardware and technician heterogeneity to the general clinical US patient population and thus be generalizable.

Rev 4.8:
10. How close is this work to being assessed prospectively?

Answer:

We are currently assessing risk models based on tabular data only in Heart Failure patients (clinical trial identifier number NCT03804606 on *clinicaltrials.gov*) [23]. We are planning to add the Echocardiography video data soon to improve model accuracy, however, additional storage and compute hardware are needed in order to make this a reality, and the installation is currently in progress.

Rev 4.9:
Minor technical
1. Lines 88-20 are not clear and may be interpreted that the PLAX was combined with EHR data to create a model that predicted 1-year mortality rather than what I believe is intended: that the PLAX model without EHR data was compared to cardiologists with EHR data.
2. Lines 100-102 do not clearly state that the cardiologists and DNN were presented with echocardiogram videos.
3. Why use balanced dataset in final performance evaluation?
4. PLAX + 10 EHR variables (how were tehse chosen?)
5. Statistics protocols or materials
Citations Ouyang et al, Nature 2020. PMID: 32269341

Answer:

1. The PLAX model was indeed combined with the 10 EHR variables. This allowed us to compare the performance of the DNN model and cardiologist when presented with the same information. To clarify this point we reworded that sentence to:

   > 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 10 EHR variables to compare their performance with a model trained on the same input set.

2. We clarified this point with the change presented above

3. We designed a balanced dataset to be able to assess the predictive performance of both positive and negative cases equally. We also conducted a paired survey to control for mortality prevalence. We now clarify this point in the Cardiologist survey section as follows:

   > Following a sample size calculation (Pearson Chi-square 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% vs 70%), 80% power, a significance level of 5%, and an approximate 40% discordancy. The calculation (performed with Power Analysis Software PASS v15) showed that we needed at least 600 patients (300 alive, 300 deceased). Thus, we randomly sampled 300 positive and 300 negatives 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. Figure 6 shows the interface for the first survey. We showed the 10 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 based on sex, age (within 5 years) and left ventricular EF (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 in order to preserve the matching criteria. Thus, all positive cases were unique. Figure 7 shows the interface for the paired survey, where we showed the video and 10 EHR variables for two age-, sex-, and EF-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 Figure 6 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.

In order to avoid incremental performance changes while the cardiologists progressed through the survey, we presented them, prior to 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 four 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 5 examples that fell into each of the four quadrants of the confusion matrix. Figure 8 shows the interface for the model assisted portion of the third survey, where we added a "Machine Prediction" row and a occlusion map video.

4. We stated in the paper that the PLAX 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. Also, we state that the

10 EHR variables were chosen because they were reported as the most predictive of 1-year mortality in [22].

5. We added the Statistical Analysis section to the paper, see response to comment 10 of Reviewer #1 (Rev 1.10).

6. The Ouyang et. al citation has also been added, [34].

References

[1] Payne, J. W. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance* **16**, 366–387 (1976).

[2] Quer, G., Muse, E. D., Nikzad, N., Topol, E. J. & Steinhubl, S. R. Augmenting diagnostic vision with ai. *The 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. eemergency healthcare informatics. In Bronzino, J. D. & Peterson, D. (eds.) *Biomedical Signals, Imaging, and Informatics*, chap. 64 (CRC Press, 2015), 4th edn.

[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **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 transactions on pattern analysis and machine intelligence* **35**, 221–231 (2012).

[8] Karpathy, A. *et al.* Large-scale video classification with convolutional neural networks. In *Proceedings of the 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 (2017).

[11] Setio, A. A. A. *et al.* Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical 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 Digital Medicine* **1**, 9 (2018).

[13] Dou, Q. *et al.* Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical 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 Digital Medicine* **1**, 59 (2018).

[15] Van Woudenberg, N. *et al.* Quantitative echocardiography: real-time quality estimation and view classification implemented on a mobile android device. In *Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation*, 74–81 (Springer, 2018).

[16] Kusunose, K. *et al.* A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC: Cardiovascular Imaging* (2019).

[17] Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* **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 medical informatics and decision making* **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. *European heart journal* **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). *Journal of the American College of Cardiology* **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: Cardiovascular Imaging* (2018).

[23] Jing, L. *et al.* A machine learning approach to management of heart failure populations. *JACC: Heart Failure* (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* (2012).

[25] Cui, X. *et al.* Deformable regions of interest with multiple points for tissue tracking in echocardiography. *Medical image analysis* **35**, 554–569 (2017).

[26] Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine* 1–6 (2020).

[27] Gahungu, N., Trueick, R., Bhat, S., Sengupta, P. P. & Dwivedi, G. Current challenges and recent updates in artificial intelligence and echocardiography. *Current Cardiovascular Imaging Reports* **13**, 5 (2020).

[28] Horgan, S. J. & Uretsky, S. Echocardiography in the context of other cardiac imaging modalities. In *Essential Echocardiography: A Companion to Braunwald's Heart Disease*, 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. *Applied Soft Computing* **88**, 106049 (2020).

[31] Ge, R. *et al.* K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. *IEEE transactions on medical 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*, 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. *International journal of computer assisted radiology and surgery* **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 digital medicine* **3**, 1–10 (2020).

[36] Behnami, D. *et al.* Automatic cine-based detection of patients at high risk of heart failure with reduced ejection fraction in echocardiograms. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 1–7 (2019).

[37] Yadlowsky, S. *et al.* Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of internal medicine* **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 medical 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? *European Heart Journal* **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 Transactions on Medical 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*, 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. *Journal of the American College of Cardiology* **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. *The American journal of cardiology* **92**, 1350–1354 (2003).

[45] Kavalieratos, D. *et al.* Palliative care in heart failure: rationale, evidence, and future priorities. *Journal of the American College of Cardiology* **70**, 1919–1930 (2017).

[46] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

[47] Venugopalan, S. *et al.* Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729* (2014).

[48] Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate classification of echocardiograms using deep learning. *arXiv preprint arXiv:1706.08658* (2017).

[49] Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497 (2015).

[50] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).

[51] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

[52] Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the trade*, 55–69 (Springer, 1998).

[53] Buuren, S. & Groothuis-Oudshoorn, K. MICE: Multivariate imputation by chained equations in r. *Journal of statistical software* **45** (2011).

[54] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (ACM, 2016).

[55] Williams, B. A. & Agarwal, S. Applying the seattle heart failure model in the office setting in the era of electronic medical records. *Circulation Journal* **82**, 724–731 (2018).

Dear Brandon,
Thank you for your revised manuscript, "Computer-assisted analysis of echocardiographic videos of the heart with deep learning improves clinical prediction of all-cause mortality", which has been seen by the original Reviewers #1, #2 and #4. In their reports, which you will find at the end of this message, you will see that Reviewers #2 and #4 are now happy with the work, and that Reviewer#1 is concerned about model underperformance given the stated limitations, arguing as well that the claim that the model improves the performance of cardiologists may need to be qualified. I hope that in a further revised version of the work you can address this reviewers points.
As before, when you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.
I look forward to receive a further revised version of the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

Answer:

We appreciate the opportunity to clarify the comments raised by Reviewer #1. We have qualified our conclusion statements to align with the abstract where we state that we provided evidence that large unstructured datasets such as videos of the heart have the potential to improve clinical prediction models.
Below, we display the reviewer's comment in italics font followed by our answer. We highlighted extracts of the paper in blue.

Reviewer # 1
Thank you for revising your work and incorporating several revision changes we have requested. I remain concerned about two aspects of your study which in my opinion is not well addressed.

Rev 1.1:
1. Lack of adequate external validation. Although you have used Ouyang's Stanford data-set, the use of EF as a surrogate of mortality prediction is not convincing in the figure presented and not equivalent to an external validation. While I understand that getting an external model with labeled outcome is a tough task, the generalizability of the model in a different institute/ region would require more confirmation.

Answer:
We agree with the reviewer that the application of our AP4 model to the Stanford dataset does not fully qualify as external validation. Unfortunately, to our knowledge no other site has a set of raw echocardiography videos linked to one-year mortality to allow for a true external validation. We acknowledge this limitation in the main text and have removed the previous claim about the Stanford dataset showing "some evidence of external validation" as seen in the tracked changes version of the manuscript on page 13. The second limitation stated in the main text now reads:

> Second, though our data had inherent heterogeneity since they were derived from a large regional healthcare system with over 10 hospitals and hundreds of clinics, data from other independent healthcare systems with mortality outcomes will be required to fully assess generalizability.

Since it may be easier for other institutions to validate our model on their data rather than release such massive datasets that are nearly impossible to fully de-identify (for example ultrasound images sometime have "burned in" text with patient identifiers), we are also committed to facilitating replication as stated in our "Code and Data Availability Statement" section:

> ## Code and Data Availability Statement
>
> All requests for raw and analyzed data and related materials, excluding programming code, will be reviewed by our legal department 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 and materials that can be shared will be released via a material transfer agreement for non-commercial research purposes.

Rev 1.2:
2. The occlusion experiments are puzzling and at present the information shown are not yielding any tangible knowledge about how the neural networks are performing. The cardiologist experiments where these data are presented is also

awkwardly designed since a cardiologist would not be able to depend on this information on improving his ability to discriminate a patient's risk. Moreover, the goal of echo examination is to diagnose a disease and the severity of the disease, have the authors attempted grouping patients eg. Heat failure, valvular heart disease, hypertensive heart disease and then looked for the performance for increasing severity of the disease and how the model behaves versus the physicians grade the disease. In the absence of a tangible understanding claiming the task performed by the physicians is inferior to that performed by the machine is problematic. In my opinion this section is controversial and the experiments are not performed in depth to arrive at the stated conclusions.

Answer:

CNN interpretation is an important limitation of the deep learning field. Saliency and occlusion maps are commonly used to investigate whether the decision mechanism of a DNN is relatable to human expectations. Unfortunately, no further generalizations or conclusions can be derived from these types of experiments. For example, in skin cancer image classification tasks [10], the network pays attention to the damaged skin; for diabetic retinopathy screening tasks[47, 9], the models tend to focus mainly on hemorrhages and microaneurysms, and partially on exudates, which are associated with the disease; for intracranial hemorrhage detection [48], attention maps have been used to localize the abnormalities. These types of "proof of concept" attention experiments are unfortunately the best we can do and, as evidenced by references above, on par with other published high impact papers in the field. We modified the first limitation in the main text to clarify this point:

> 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 did use occlusion maps to show 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 hemorrhage 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.

The reviewer correctly states that the experiments presented in this paper are not sufficient to determine whether cardiologists could improve their performance by interpreting occlusion maps. When presented with examples of patients with Heart Failure and/or history of Myocardial Infarction, all four cardiologists noted that the occlusion maps were not helpful. Instead, the cardiologists relied on the predicted risk score to enhance their prediction ability. We have modified the following paragraph to address this, noting that the cardiologists relied on the prediction score:

> when presenting several examples of the occlusion maps to cardiologists, they anecdotally reported that they were unable to identify patterns that could help them better discern patient survival

outcomes. Instead, the cardiologists focused their attention on the model's predicted score to re-evaluate their initial assessment.

The reviewer's idea of grouping patients by diagnosis and exploring performance across different levels of disease severity is a good suggestion but still would not provide any generalizable insight into model interpretability as discussed above. Moreover, the amount of potential work and avenues to explore along these lines would in our opinion be the subject of 1 or more additional papers that are beyond the scope of the current work. In regard to physician performance, we have changed the main conclusions of the study to no longer focus on a cardiologist/model comparison, but instead on how the model complements physician performance. We modified the last paragraph of the main text to qualify our conclusion and align with the abstract:

> In conclusion, 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.

Rev 1.3:

3. Authors suggests that the lower AUC than the previous work than one observed by the model is due to lower sample size, then in most instances the previous model would be used, the incremental value of the current model is unclear. This limitation suggest need for more work and should be included in the main conclusions and the abstract.

Answer:

We provided evidence that large unstructured datasets such as videos of the heart have the potential to improve clinical prediction models, as we state in the abstract. The modified last paragraph of the main text now aligns with the abstract.

Finally, we showed that predicting mortality from raw videos provides superior performance to structured data derived from the echocardiograms for the largest dataset we could possibly fit within our hardware environment (42k raw echocardiograms). We also showed detailed evidence in comparison to our prior work (on structured data derived clinically from 331k echocardiograms) that the performance of the models should improve as we increase the sample size. Unfortunately, to increase our sample size to 331k raw echocardiograms would require storage capacity not currently available to us (petabytes of fast storage), and exclusive use of an expensive 16-GPU system likely for months of training.

References

[1] Payne, J. W. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance* **16**, 366–387 (1976).

[2] Quer, G., Muse, E. D., Nikzad, N., Topol, E. J. & Steinhubl, S. R. Augmenting diagnostic vision with ai. *The 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. eemergency healthcare informatics. In Bronzino, J. D. & Peterson, D. (eds.) *Biomedical Signals, Imaging, and Informatics*, chap. 64 (CRC Press, 2015), 4th edn.

[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **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 transactions on pattern analysis and machine intelligence* **35**, 221–231 (2012).

[8] Karpathy, A. *et al.* Large-scale video classification with convolutional neural networks. In *Proceedings of the 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 (2017).

[11] Setio, A. A. A. *et al.* Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical 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 Digital Medicine* **1**, 9 (2018).

[13] Dou, Q. *et al.* Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical 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 Digital Medicine* **1**, 59 (2018).

[15] Van Woudenberg, N. *et al.* Quantitative echocardiography: real-time quality estimation and view classification implemented on a mobile android device. In *Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation*, 74–81 (Springer, 2018).

[16] Kusunose, K. *et al.* A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *JACC: Cardiovascular Imaging* (2019).

[17] Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* **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 medical informatics and decision making* **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. *European heart journal* **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). *Journal of the American College of Cardiology* **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: Cardiovascular Imaging* (2018).

[23] Jing, L. *et al.* A machine learning approach to management of heart failure populations. *JACC: Heart Failure* (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* (2012).

[25] Cui, X. *et al.* Deformable regions of interest with multiple points for tissue tracking in echocardiography. *Medical image analysis* **35**, 554–569 (2017).

[26] Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine* 1–6 (2020).

[27] Gahungu, N., Trueick, R., Bhat, S., Sengupta, P. P. & Dwivedi, G. Current challenges and recent updates in artificial intelligence and echocardiography. *Current Cardiovascular Imaging Reports* **13**, 5 (2020).

[28] Horgan, S. J. & Uretsky, S. Echocardiography in the context of other cardiac imaging modalities. In *Essential Echocardiography: A Companion to Braunwald's Heart Disease*, 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. *Applied Soft Computing* **88**, 106049 (2020).

[31] Ge, R. *et al.* K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. *IEEE transactions on medical 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*, 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. *International journal of computer assisted radiology and surgery* **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 digital medicine* **3**, 1–10 (2020).

[36] Behnami, D. *et al.* Automatic cine-based detection of patients at high risk of heart failure with reduced ejection fraction in echocardiograms. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 1–7 (2019).

[37] Yadlowsky, S. *et al.* Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of internal medicine* **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 medical 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? *European Heart Journal* **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 Transactions on Medical 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*, 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. *Journal of the American College of Cardiology* **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. *The American journal of cardiology* **92**, 1350–1354 (2003).

[45] Kavalieratos, D. *et al.* Palliative care in heart failure: rationale, evidence, and future priorities. *Journal of the American College of Cardiology* **70**, 1919–1930 (2017).

[46] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

[47] Arcadu, F. *et al.* Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ digital medicine* **2**, 1–9 (2019).

[48] Lee, H. *et al.* An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering* **3**, 173 (2019).

[49] Venugopalan, S. *et al.* Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729* (2014).

[50] Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate classification of echocardiograms using deep learning. *arXiv preprint arXiv:1706.08658* (2017).

[51] Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 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 *Proceedings of the 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

[54] Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the trade*, 55–69 (Springer, 1998).

[55] Buuren, S. & Groothuis-Oudshoorn, K. MICE: Multivariate imputation by chained equations in r. *Journal of statistical software* **45** (2011).

[56] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 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. *Circulation Journal* **82**, 724–731 (2018).