Generalizability and Quality Control of Deep Learning-Based 2D Echocardiography Segmentation Models in a Large Clinical Dataset

XIAOYAN Zhang
Geisinger Health
Alvaro E. Ulloa Cerna
Geisinger Health
Joshua V. Stough
Bucknell University
Yida Chen
Bucknell University
Brendan J. Carry
Geisinger Health
Amro Alsaid
Geisinger Health
Sushravya Raghunath
Geisinger Health
David P. vanMaanen
Geisinger Health
Brandon K. Fornwalt
Geisinger Health
Christopher M. Haggerty (chris.m.haggerty@gmail.com)
Geisinger Health

Research Article

Keywords: 2D echocardiography segmentation, deep learning, generalizability, quality control, segmentation uncertainty, convexity

Posted Date: November 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1097714/v1

Version of Record: A version of this preprint was published at The International Journal of Cardiovascular Imaging on February 24th, 2022. See the published version at https://doi.org/10.1007/s10554-022-02554-7.

Abstract

Use of machine learning for automated annotation of heart structures from echocardiographic videos is an active research area, but understanding of comparative, generalizable performance among models is lacking. This study aimed to 1) assess the generalizability of five state-of-the-art machine learning-based echocardiography segmentation models within a large clinical dataset, and 2) test the hypothesis that a quality control (QC) method based on segmentation uncertainty can further improve segmentation results. Five models were applied to 47,431 echocardiography studies that were independent from any training samples. Chamber volume and mass from model segmentations were compared to clinically-reported values. The median absolute errors (MAE) in left ventricular (LV) volumes and ejection fraction exhibited by all five models were comparable to reported inter-observer errors (IOE). MAE for left atrial volume and LV mass were similarly favorable to respective IOE for models trained for those tasks. A single model consistently exhibited the lowest MAE in all five clinically-reported measures. We leveraged the 10-fold cross-validation training scheme of this best-performing model to quantify segmentation uncertainty for potential application as QC. We observed that filtering segmentations with high uncertainty improved segmentation results, leading to decreased volume/mass estimation errors. The addition of contour-convexity filters further improved QC efficiency. In conclusion, five previously published echocardiography segmentation models generalized to a large, independent clinical dataset—segmenting one or multiple cardiac structures with overall accuracy comparable to manual analyses—with variable performance. Convexity-reinforced uncertainty QC efficiently improved segmentation performance and may further facilitate the translation of such models.

Introduction

Structural segmentation is an important step for interpreting 2D echocardiography, which is highly time-consuming and subject to significant inter- and intra-observer variability [1]–[6]. To overcome these limitations, several computer-aided methodologies, such as active shape models [7]–[10], level-sets [9], [10], and deep learning (DL)-based algorithms [11]–[14], have been developed to automatically segment cardiac structures in echocardiography images, with models based on DL particularly gaining increasing attention in recent years [15].

To date, most state-of-the-art DL models focus on segmentation of the left ventricular (LV) endocardium [12], [15]–[18], with only a few segmenting other cardiac structures such as the LV epicardium or the left atrium (LA), which could provide additional information for diagnosing and treating heart disease [19]. In 2018, Zhang et al. developed the first model that segments multiple chambers [20]. In 2019, Leclerc et al. published the CAMUS dataset for which LV endo- and epicardium, and LA endocardium were manually segmented [21]. This dataset greatly facilitated the development and improvement of multi-structural echocardiography segmentation models [19], [22]–[26], such as those trained with adversarial [19] or motion-segmentation co-learning strategy [25], [26]. The generalizability of these models to external, independent datasets was only partially tested on either the small CAMUS dataset (450 patients) [19] or the single-view EchoNet-Dynamic dataset with only LV endocardial contours [25]. Thus, the performance

of these multi-structural segmentation models within a large, independent, clinically-acquired echocardiography dataset remains unknown. Moreover, none of these models has been tested with automated quality control (QC) methods.

QC, as an important consideration for translating these AI segmentation models into potential clinical use, is mostly achieved by estimating aleatoric uncertainty which is described by the noise inherent in observations [27]–[29]. Unlike epistemic uncertainty which can be eliminated by training on big data or with data augmentation, aleatoric uncertainty can be formalized by a distribution over model outputs [27]. Common uncertainty modeling approaches require multiple segmentation predictions for a single input either through test-time augmentation [30] or by feeding into multiple models generated during training [31], [32]. To estimate the uncertainties, the differences between final/averaged segmentation and individual predictions are assessed at the pixel or image level [28], [29]. Previous studies showed that removing uncertain segmentations improved LV segmentation scores [28]. In addition to uncertainty, convexity has been used as a shape prior in image segmentation to improve model regularization and performance [33]. Its use as a quality score for segmentation, however, has not been tested yet.

Therefore, the main objective of the present study was to compare five state-of-the-art echocardiography segmentation models on a large (>47k studies), independent clinical echocardiography dataset comprising both apical two (a2c) and apical four (a4c) chamber views. Models were compared by their accuracy to assess any of five standard clinical measures: LV end-diastolic volume (EDV), LV end-systolic volume (ESV), LV ejection fraction (EF), LV wall mass (LVM), and maximal LA volume (LAV). Using the most generalizable segmentation model, we then tested a new method for measuring segmentation aleatoric uncertainty, i.e., using cross-validation (CV) model averaging, and evaluated the use of segmentation uncertainty combined with convexity as a QC. By doing so, the present study provides detailed evaluations for selection of segmentation models and paves the way for the development of an echocardiographic analysis pipeline that can be used in production with automatic QC.

Data And Methods

Datasets

The Institutional Review Board at Geisinger approved this retrospective study with a waiver of consent, in conjunction with institutional patient privacy policies. We randomly extracted 88,322 studies from the Geisinger Xcelera database (Philips Medical Systems) in the Digital Imaging and Communications in Medicine (DICOM) format which were collected between 1998 and 2020. Among the transthoracic echocardiography (TTE) videos with DICOM view labels, we identified 50,593 studies (37,704 unique patients) having both a2c and a4c videos longer than one heartbeat (online Fig. 1).

We extracted EDV, ESV, LVM, and LAV measures from the Xcelera database. While EDV, ESV, and LAV were computed using the bi-plane method of disks (MOD-bp) at Geisinger clinic, LVM was calculated using M-mode linear cube method. All these clinical values were estimated from a single or multiple heartbeat(s)

selected by cardiologists based on image quality. We also extracted physician-reported EF measurements, which were approximate values or ranges derived either qualitatively or through a 2D- or 3D-volume technique [34] and were not significantly different from MOD-bp EF derived from 2D echocardiograms. After excluding the studies without pixel scale information or clinical volume/mass measurements, we collected 47,431 studies from 35,826 unique patients containing 80,550 a2c and 85,975 a4c videos for segmentation and evaluation (online Fig. 1). The data characteristics are summarized in Table 1. Disease prevalence was based on either custom phenotypes [35]–[37] or models developed by the Electronic Medical Records and Genomics (eMERGE) Network [36], [38].

We also collected two publicly available 2D echocardiography datasets with manual segmentations, CAMUS [21] and EchoNet-Dynamic [16]. CAMUS contained 900 a2c and 900 a4c images at ED and ES from 450 unique patients [21]. EchoNet-Dynamic contained 10,025 a4c videos from 10,025 unique patients [16].

Segmentation

We deployed and evaluated five state-of-the-art segmentation models, including Zhang et al., Ouyang et al., Arafati et al., and Stough et al. 2D and 3D models (online Table S1), as follows:

A. Preprocessing

For all five models, after masking out the non-cone pixels, we reshaped the video frame dimensions with preserved aspect ratio using cubic interpolation to the target sizes and normalized/standardized pixel intensity per the model requirements (online Table S1). Since the Stough et al. 3D model required ED-ES clips, we identified the ED and ES frames based on the areas of LV segmentation produced by the Stough et al. 2D model using a peak-finding algorithm with a distance of 85% of cardiac cycle duration [16]. After cutting the whole videos into ED-ES clips, we converted the length of each ED-ES clip to 10 frames using trilinear interpolation. Although this preprocessing may be slightly different from the protocols described in the original codes/manuscripts, pilot studies did not detect any significant difference in their performance on Geisinger data.

B. Deployment

We fed preprocessed videos/clips into candidate models and obtained structural segmentations by identifying the maximum Softmax score output by each model. Since Stough et al. proposed to use the accumulative output from ten CV models which showed improved performance on CAMUS test data [24], [25], we accumulated ten sets of Softmax probabilities and identified the maximum score to produce a final prediction for both 2D and 3D models. To postprocess the segmentations generated by each model, we kept the largest region and filled any holes smaller than 128 pixels for each segmented structure using connected component analysis with scikit-image [21].

C. Evaluation

We estimated volume and mass from predicted segmentations using Simpsons modified MOD [24], [39] and computed the errors when compared to clinically reported measures. Within each study, we reported the median volume/mass estimation from all qualifying segmentation results (aggregating across multiple beats and videos). Specifically, after identifying ED and ES frames (see section Segmentation A) [16], we used LV endocardial segmentations from a single a4c view (MOD-sp4 method) or from a2c and a4c views (MOD-bp method) at ED and ES to estimate LV EDV, ESV, and EF. LAV was calculated similarly [4]. To estimate LVM, we computed the MOD-bp volume of the LV wall at ED as the difference between the epicardial and endocardial volumes and multiplied the value by myocardial density (1.05 g/ml) [6].

Segmentation QC

Adapted from state-of-the-art methods [29], we computed two image-level uncertainty measures using the ten sets of outputs from the ten 2D Stough et al. models:

$$\text{segmentation deviation } Seg\_dev = Mean(1 - DICE\{\bar{S}, S_i\}_{i=1\ldots10})$$

$$\text{segmentation variance } Seg\_CoV = Coefficient\ of\ Variance(DICE\{\bar{S}, S_i\}_{i=1\ldots10})$$

where $S_i, i \in \{1 \ldots 10\}$ were predicted segmentation samples obtained by identifying the maximum Softmax score output by each model for each pixel, and $\bar{S}$ was the mean predicted segmentation obtained by averaging the Softmax scores from ten models and then taking the maximum value. DICE was computed using the following equation [40]: $DICE = \frac{2|S_i \cap \bar{S}|}{|S_i|+|\bar{S}|}$. Segmentations with uncertainty larger than a threshold were excluded from downstream volume/mass estimation; this QC strategy was denoted as uncertainty-based QC.

We used 80% of CAMUS data to define the uncertainty threshold. To increase the incidence of bad segmentations, we augmented images 20 times by randomly rotating the image with the transducer as a pivot point, adding Gaussian noise, and applying intensity windowing [24]. This led 1,439 training images. We considered the predicted segmentation to be poor when the ground truth DICE (i.e., DICE between prediction and ground truth segmentation) <0.85 [29]. Pareto front curve, also called the tradeoff curve, which is usually used for optimizing bi-objective problems, was generated by plotting the percentage of poor segmentations remained after QC against the percentage of segmentations dropped by QC [29]. The ideal uncertainty threshold should remove the least number of samples (objective one) while dropping the largest number of poor segmentations (objective two).

We also computed convexity of the segmentation contour using a boundary-based method [41]. Specifically, the convexity score was calculated as $\frac{Per_2(R)}{Per_1(S)}$, where $Per_2(R)$ was perimeter of the minimal rectangle *R* surrounding the segmentation contour whose edges were parallel to the x and y axes, and $Per_1(S)$ was the sum of projections of the edges of segmentation contour *S* onto the x and y

axes [41]. As such, the convexity score ranged from (0, 1]; values approaching 1 represent oval or circular shape with smooth boundaries [41]. For LV wall segmentation, the convexity score was defined as the minimal value of LV endo- and epicardial convexity.

Convexity filters were added to reinforce the uncertainty-based QC by screening out segmentations that met any of the following criteria: 1) convexity <0.6 (shapes with multiple fragments, substantial indentations or protrusions according to the convexity rankings by Zunic and Rosin [41]); 2) Seg_dev >0.15 (equivalent to mean DICE <0.85); or 3) Seg_dev > a threshold (i.e., 0.039 for LV endocardium, 0.057 for LV wall, and 0.055 for LA endocardium as learned from the Pareto front curves in Fig. 3) and convexity <0.96 (around the 10th percentile observed for independent Geisinger and EchoNet-Dynamic data). This method was denoted as convexity-reinforced uncertainty QC.

We tested segmentation QC with the learned cutoffs on the CAMUS test set, EchoNet-Dynamic data, and Geisinger segmented studies. For the former two datasets, we compared changes in ground truth DICE before and after QC; since Geisinger data do not have ground truth segmentation, we compared changes in volume/mass estimation errors before and after QC.

Results

Segmentation

All five tested models showed median absolute errors (MAE) comparable with reported inter-observer errors (IOE) in segmenting LV endocardium [2]; however, only the Stough et al. 2D and 3D models and the Ouyang et al. model consistently produced errors lower than IOE for EDV, ESV, EF (Fig. 1). Among the three models that were able to segment the LV epicardium and estimate LVM, only the 2D and 3D models developed by Stough et al. were associated with MAE that were below the reported IOE [6], with the 2D model outperforming the 3D model (Fig. 1). As for LAV, the MAE associated with three of the four models that had the capability of segmenting LA endocardium (i.e., models developed by Stough et al. and Zhang et al.) were within one SD of the reported IOE [5] (Fig. 1). The Stough et al. 2D model exhibited the lowest LAV error (Fig. 1). Note that the IOE used in this study were the minimal relative errors that were estimated in previous studies using the same Simpsons MOD-bp method as we did in our study (Table 2) [2], [5], [6].

As shown in Bland-Altman density plots (Fig. 2), the Stough et al. 2D model exhibited consistently good performance in segmenting LV wall and LA endocardium with zero mean bias in LVM and LAV estimation. While EDV and ESV were underestimated especially at large volumes, the resultant LV EF prediction was associated with zero mean bias. Overall, for all the five clinical measures, the bias values clustered around zero.

Segmentation QC

Pareto front curves show that the two uncertainty scores, i.e., Seg_dev and Seg_CoV, were able to preferentially identify poor segmentations (Fig. 3). Seg_dev, however, exhibited higher efficiency with steeper slopes, especially for LV endocardium segmentation. For demonstration purposes, we selected the threshold closest in Euclidean distance to the origin point (0,0) for Seg_dev of each structure, i.e., 0.039 for LV endocardium (LV), 0.057 for LV wall (Myo), and 0.055 for LA endocardium (LA) (Fig. 3).

By removing segmentations with Seg_dev > cutoffs selected above, the mean ground truth DICE was improved for all three segmented structures (pre-QC vs post-QC: LV 0.945 vs 0.951, Myo 0.894 vs 0.903, LA 0.931 vs 0.940), as evaluated on augmented CAMUS test images (Fig. 4A). Notably, most of the poor segmentations with ground truth DICE <0.85 were removed by QC with Seg_dev. Specifically, QC removed 98.6%, 72.7%, and 84.3% of the poor segmentations, respectively, for LV endocardium, Myo, and LA endocardium (data not shown). In terms of identifying good segmentations (ground truth DICE ≥0.85) (Table 3), this uncertainty-based QC showed excellent precision on CAMUS test set with scores higher than 0.90 or approaching 1.0. This method removed 15%, 32%, and 19% segmentations with ground truth DICE ≥0.85 for LV endocardium, Myo, and LA endocardium, respectively, leading to slightly lower sensitivity scores (Table 3). The F1 scores were around 0.90 for LV and LA endocardium and approached 0.80 for Myo (Table 3).

Using the same LV Seg_dev cutoff, similar improvement in mean ground truth DICE with fewer poor segmentations was observed for LV endocardium in the independent EchoNet-Dynamic dataset (Fig. 4B). The mean ground truth DICE increased from 0.893 (pre-QC) to 0.909 (QC with Seg_dev) after removing 68% of poor segmentations (data not shown). QC performance metrics for identifying good segmentations were slightly lower compared to those observed for CAMUS test set (Table 3). We observed some disassociation between segmentation ground truth DICE and uncertainty measures, i.e., some segmentations with DICE <0.85 exhibited low uncertainty measures (Fig. 5C and 5D) while some segmentations with higher DICE exhibited high uncertainty values (Fig. 5A and 5B), especially for those with part of myocardium outside the image field. In both instances, segmentation convexity scores provided additional, independent insight (Fig. 5). Indeed, the addition of convexity to Seg_dev-based QC greatly increased the sensitivity score while slightly compromising the precision score, increasing the F1 score to 0.92 from 0.86 (Table 3). The percentage of good segmentations removed by QC dropped from 20–3% (Table 3). With convexity, Seg_dev-based QC only removed 20% of poor segmentations (data not shown).

We assessed the changes in absolute errors in volume and mass estimation before and after QC in Geisinger data using both Seg_dev and convexity-reinforced Seg_dev (Tables 4 and 5). With Seg_dev alone, the mean absolute errors for the reported clinical measures decreased by 3–15%, while the mean absolute errors of the removed studies ranged from 17–34%. However, in each case, a large proportion of studies (14–71%) were removed based on the specified QC thresholds. We again observed that many segmentations with significant uncertainty had high convexity scores (Fig. 6). Thus, we again observed that the use of QC with convexity-reinforced Seg_dev removed significantly less data compared to QC with Seg_dev alone (<10%; Table 5). Although the overall decreases in absolute errors observed after QC

with convexity-reinforced Seg_dev were lower (Table 4), the errors in the studies removed by this QC were much larger (Table 5). With both QC strategies, LVM estimation exhibited the largest loss of studies which was followed by EF estimation (Table 5).

Discussion

Overall, five published echocardiography segmentation models tested in the present study generalized well to a large external clinically-acquired dataset of 2D echocardiograms on most segmentation tasks, leading to volume and/or mass estimations with accuracy comparable to manual analyses [2], [4]–[6]. This clinically-acquired dataset involved a variety of cardiovascular disease conditions with a wide span of EF values. The comparable errors in LV EDV, ESV, and EF exhibited by the five models suggest a good adaptation to patients with different LV conditions. The striping pattern observed for EF Bland-Altman density plot was likely due to the tendency for the physician-reported EF to take discrete values (e.g., 60%) or ranges (e.g., 60–65% for which we used the mean value 62.5%) rather than naturally comprising a continuous representation, as the segmentation result does. Although mostly within the reported IOE range (i.e., 11 ± 12%) [5], the LAV errors were higher than the mean IOE for all of the four multi-structural segmentation models. This sub-optimal LA segmentation, as compared to LV segmentation, could be partially attributable to the high boundary-to-area ratio of LA. Another contributing factor could be the fact that LA was usually associated with more distortion/noise/motion, especially at the bottom region, since it was farther from the transducer. Moreover, for videos without focusing on LA, part of LA might be out of image.

The generalizable performance exhibited by four of these models (i.e., Stough et al., Ouyang et al., and Arafati et al. models) could be partially attributable to the fact that they were trained either on a very large dataset or with data augmentation [16], [19], [24], [25]. As for the Zhang et al. model, although it was trained with some data augmentation [20], the relatively small original training data likely restricted its generalizability, leading to modest performance as observed in the present study as well as on the CAMUS dataset [19].

Besides training data, model structure and training strategy also likely contributed to the different performance exhibited by the five models. While the Ouyang et al. model, which exhibited superior accuracy in segmenting the LV blood pool, leveraged atrous convolutions [16], Stough et al. employed CV model averaging with a simple U-net structure [24], [25]. In fact, models developed by Stough et al., especially the 2D frame-level model, stood out as superior in almost all segmentation tasks, suggesting CV model averaging as a robust method to generalize the segmentation models. Stough et al. also used appearance and shape co-learning strategies when training the 3D segmentation model [25]; the additional objective to enforce temporal coherency between ED and ES phases, however, may compromise the segmentation performance at individual ED and ES frames [25]. Moreover, the Stough et al. 3D model required ED-ES clips, which were generated based on frame-level segmentations; any errors at frame-level segmentation could propagate during 3D segmentation. All these factors could explain the sub-optimal accuracy exhibited by the 3D model, as compared to the Stough et al. 2D model. Similarly,

the balance between two objectives during adversarial training may compromise the segmentation task of the Arafati et al. model [19], contributing to its sub-optimal performance on this large external clinical dataset.

Additionally, the varied segmentation performance could partially arise from the training labels, i.e., the manual segmentations generated by different cardiologists. For example, compared to the precise LV and LA endocardial manual segmentations used to train the Arafati et al. model [19], the manual segmentations of the CAMUS and EchoNet-Dynamic datasets were conservative, especially at the apex and along the free wall of the LV [16], [21]. As a result, models trained on these two datasets, i.e., the Stough et al. 2D and 3D models and the Ouyang et al. model, tended to generate conservative segmentations, especially when the LV was enlarged with part of the myocardium outside the images, leading to underestimated LV volumes at ED and ES. However, the majority of segmentations generated, especially by the Stough et al. 2D model, clustered around zero bias for all five volume/mass estimations. These results, taken together, support the Stough et al. 2D model as the most generalizable segmentation model with great versatility, which could be leveraged for a production deployment in a clinical setting.

QC was achieved in this study by leveraging the 10-fold CV models trained by Stough et al. [24]. Segmentation uncertainty was easily obtained as a by-product when generating the final segmentation through accumulating the outputs of ten CV models. Our Seg_dev-based QC method efficiently removed the majority of poor segmentations for all three segmented structures, leading to slight increases in the mean ground truth DICE. It was not surprising to detect such minor increases in DICE score, particularly because the segmentation models already had superior performance before QC with less frequent failure (i.e., fewer segmentations with ground truth DICE <0.85). Moreover, the performance of our Seg_dev-based QC method on improving LV segmentation quality was comparable to the state-of-the-art results as evaluated on CAMUS and EchoNet-Dynamic datasets [28]. Like the other uncertainty methods [28], this method failed to flag some bad segmentations when all ten models performed consistently poorly. Moreover, although some final segmentations accumulated over ten models looked acceptable, they were dropped due to high uncertainty arising from the presence of low-contrast or invisible surrounding tissue in the images. This problem was more evident for LV wall segmentation. Significantly, the addition of convexity to Seg_dev-based QC greatly saved those segmentations with good convex shape and added more confidence in filtering out bad segmentations. In fact, the superior precision and sensitivity scores for picking up good segmentations from CAMUS and EchoNet-Dynamic datasets support our QC method, especially the convexity-reinforced uncertainty strategy, as an effective approach once appropriate cutoffs were set. This was further evidenced by the removal of large errors and the decreased absolute errors in downstream volume/mass estimation as shown in Geisinger data. However, it should be noted that there will always be a tradeoff between segmentation quality and number of studies excluded when defining a cutoff, and this cutoff may need to be adjusted depending on the deployment scenario of interest. The decision on QC cutoffs will be the prior step for the deployment of the Stough et al. 2D model with QC in the clinical setting.

One limitation related to our study was the lack of ground truth segmentation for the large Geisinger data. This restricted our evaluation to the estimation of volume and mass which was downstream of segmentation. The fact that the errors in volume/mass estimation were within human IOE lent sufficient support for the use of our model-generated segmentations to derive key clinical measures. Another limitation was the exclusion of a second external evaluation dataset which should be independent of any of the five models. However, in our pilot studies, we evaluated the performance of all the five models on CAMUS and part of EchoNet-Dynamic datasets using the same procedures described in this study (online Table S2). Overall, the Stough et al. 2D and 3D models outperformed the others in these studies. Finally, it is of great interest to evaluate a QC method on all the five models. But this will require either re-training these models in a similar ten-fold CV scheme on non-Geisinger datasets, or easy accessibility to multiple sets of model weights for each of the five models, which both are beyond the scope of the current study. Although aleatoric uncertainty can be estimated using test-time augmentation, it is tricky to choose an appropriate augmentation range.

In conclusion, all five state-of-the-art echocardiography segmentation models generalized well with good performance on most tasks within a large clinically-acquired echocardiography dataset. Stough et al. models, particularly the frame-level 2D model, exhibit the best performance in segmenting the three key left heart structures with accuracy comparable to manual analyses. The deployment of the proposed convexity-reinforced uncertainty QC method can improve the overall performance and enable real-time detection and correction of poor segmentations. Thus, incorporation of the Stough et al. 2D model and the proposed QC method into an echocardiographic analysis pipeline could potentially facilitate cardiac research and clinical diagnosis by providing efficient and accurate cardiac measurements. Further modifications to improve both segmentation models and post-segmentation analysis are possible and may help improve performance for both clinical and research applications.

Declarations

Funding: This work was supported by the PA Cure grant (SAP #4100079720), the Geisinger Clinic, and the Bucknell Geisinger Research Initiative.

Conflict of interest: Geisinger receives funding from Tempus for development of predictive models in cardiology; this work was not directly related to those efforts.

Author contributions: XZ, AUC, JS, SR, BF, and CH contributed to conception and design of the study. XZ performed the study, collected, and analyzed data, and wrote the manuscript. JS and YC contributed to model development. BC and AA provided clinical insights. DvM extracted data. All authors contributed to manuscript revision, read, and approved the submitted version.

Ethics approval: The Institutional Review Board at Geisinger approved this retrospective study with a waiver of consent, in conjunction with institutional patient privacy policies.

Consent to participate: Not applicable.

Consent to publish: Not applicable.

References

1. V. Mor-Avi *et al.*, “Real-time 3D echocardiographic quantification of left atrial volume: Multicenter study for validation with CMR,” *JACC Cardiovasc. Imaging*, vol. 5, no. 8, pp. 769–777, 2012, doi: 10.1016/j.jcmg.2012.05.011.
2. L. D. Jacobs *et al.*, “Rapid online quantification of left ventricular volume from real-time three-dimensional echocardiographic data,” *Eur. Heart J.*, 2006, doi: 10.1093/eurheartj/ehi666.
3. R. Hoffmann *et al.*, “Analysis of left ventricular volumes and function: A multicenter comparison of cardiac magnetic resonance imaging, cine ventriculography, and unenhanced and contrast-enhanced two-dimensional and three-dimensional echocardiography,” *J. Am. Soc. Echocardiogr.*, 2014, doi: 10.1016/j.echo.2013.12.005.
4. A. M. Keller, A. S. Gopal, and D. L. King, “Left and Right Atrial Volume by Freehand Three-dimensional Echocardiography: In Vivo Validation Using Magnetic Resonance Imaging,” *Eur. J. Echocardiogr.*, 2000, doi: 10.1053/euje.2000.0010.
5. V. C. C. Wu *et al.*, “Prognostic value of la volumes assessed by transthoracic 3d echocardiography: Comparison with 2d echocardiography,” *JACC Cardiovasc. Imaging*, vol. 6, no. 10, pp. 1025–1035, 2013, doi: 10.1016/j.jcmg.2013.08.002.
6. V. Mor-Avi *et al.*, “Fast measurement of left ventricular mass with real-time three-dimensional echocardiography: Comparison with magnetic resonance imaging,” *Circulation*, vol. 110, no. 13, pp. 1814–1818, 2004, doi: 10.1161/01.CIR.0000142670.65971.5F.
7. Y. Chen, F. Huang, H. D. Tagare, M. Rao, D. Wilson, and E. A. Geiser, “Using prior shape and intensity profile in medical image segmentation,” 2003, doi: 10.1109/iccv.2003.1238474.
8. Y. Chen, F. Huang, H. D. Tagare, and M. Rao, “A coupled minimization problem for medical image segmentation with priors,” *Int. J. Comput. Vis.*, 2007, doi: 10.1007/s11263-006-8524-2.
9. J. Y. Yan and T. G. Zhuang, “Applying improved fast marching method to endocardial boundary detection in echocardiographic images,” *Pattern Recognit. Lett.*, 2003, doi: 10.1016/S0167-8655(03)00121-1.
10. N. Lin, W. Yu, and J. S. Duncan, “Combinative multi-scale level set framework for echocardiographic image segmentation,” *Med. Image Anal.*, 2003, doi: 10.1016/S1361-8415(03)00035-5.
11. M. H. Jafari *et al.*, “A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data,” 2018, doi: 10.1007/978-3-030-00889-5_4.
12. G. Carneiro, J. C. Nascimento, and A. Freitas, “The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods,” *IEEE Trans. Image Process.*, 2012, doi: 10.1109/TIP.2011.2169273.

13. G. Carneiro, J. Nascimento, and A. Freitas, “Robust left ventricle segmentation from ultrasound data using deep neural networks and efficient search methods,” 2010, doi: 10.1109/ISBI.2010.5490181.
14. G. Carneiro and J. C. Nascimento, “Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, doi: 10.1109/TPAMI.2013.96.
15. C. Chen *et al.*, “Deep Learning for Cardiac Image Segmentation: A Review,” *Front. Cardiovasc. Med.*, vol. 7, no. March, 2020, doi: 10.3389/fcvm.2020.00025.
16. D. Ouyang *et al.*, “Video-based AI for beat-to-beat assessment of cardiac function,” *Nature*, 2020, doi: 10.1038/s41586-020-2145-8.
17. S. Moradi *et al.*, “MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography,” *Phys. Medica*, vol. 67, no. July 2019, pp. 58–69, 2019, doi: 10.1016/j.ejmp.2019.10.001.
18. M. Li, W. Zhang, G. Yang, C. Wang, and H. Zhang, “Recurrent Aggregation Learning for Multi-View,” vol. 6.
19. A. Arafati *et al.*, “Generalizable fully automated multi-label segmentation of four-chamber view echocardiograms based on deep convolutional adversarial networks,” *J. R. Soc. Interface*, vol. 17, no. 169, 2020, doi: 10.1098/rsif.2020.0267rsif20200267.
20. J. Zhang *et al.*, “Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy,” *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018, doi: 10.1161/CIRCULATIONAHA.118.034338.
21. S. Leclerc *et al.*, “Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography,” *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019, doi: 10.1109/TMI.2019.2900516.
22. V. Zyuzin, A. Mukhtarov, D. Neustroev, and T. Chumarnaya, “Segmentation of 2D Echocardiography Images using Residual Blocks in U-Net Architectures,” 2020, doi: 10.1109/USBEREIT48449.2020.9117678.
23. Y. Hu *et al.*, “AIDAN: An Attention-Guided Dual-Path Network for Pediatric Echocardiography Segmentation,” *IEEE Access*, vol. 8, pp. 29176–29187, 2020, doi: 10.1109/ACCESS.2020.2971383.
24. J. V. Stough, S. Raghunath, X. Zhang, J. M. Pfeifer, B. K. Fornwalt, and C. M. Haggerty, “Left ventricular and atrial segmentation of 2D echocardiography with convolutional neural networks,” 2020, p. 9, doi: 10.1117/12.2547375.
25. Y. Chen, X. Zhang, C. M. Haggerty, and J. V. Stough, “Assessing the generalizability of temporally coherent echocardiography video segmentation,” 2021, doi: 10.1117/12.2580874.
26. H. Wei *et al.*, “Temporal-Consistent Segmentation of Echocardiography with Co-learning from Appearance and Shape,” 2020, doi: 10.1007/978-3-030-59713-9_60.
27. M. S. Ayhan and P. Berens, “Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks,” 2018.

28. L. Dahal, A. Kafle, and B. Khanal, “Uncertainty Estimation in Deep 2D Echocardiography Segmentation,” *arXiv*, 2020.
29. M. Ng and G. A. Wright, “Estimating Uncertainty in Neural Networks for Segmentation Quality Control,” *32nd Conf. Neural Inf. Process. Syst. (NIPS 2018), Montréal, Canada*, no. Nips, pp. 3–6, 2018.
30. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks,” *Neurocomputing*, 2019, doi: 10.1016/j.neucom.2019.01.103.
31. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” 2017.
32. A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” 2017, doi: 10.5244/c.31.57.
33. L. A. Royer, D. L. Richmond, C. Rother, B. Andres, and D. Kainmueller, “Convexity shape constraints for image segmentation,” 2016, doi: 10.1109/CVPR.2016.50.
34. G. J. Wehner *et al.*, “Routinely reported ejection fraction and mortality in clinical practice: Where does the nadir of risk lie?,” *Eur. Heart J.*, 2020, doi: 10.1093/eurheartj/ehz550.
35. A. E. Ulloa Cerna *et al.*, “Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality,” *Nat. Biomed. Eng.*, 2021, doi: 10.1038/s41551-020-00667-9.
36. S. Raghunath *et al.*, “Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network,” *Nat. Med.*, 2020, doi: 10.1038/s41591-020-0870-z.
37. S. Raghunath *et al.*, “Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead Electrocardiogram and Help Identify Those at Risk of AF-Related Stroke,” *Circulation*, 2021, doi: 10.1161/circulationaha.120.047829.
38. L. Jing *et al.*, “A Machine Learning Approach to Management of Heart Failure Populations,” *JACC Hear. Fail.*, 2020, doi: 10.1016/j.jchf.2020.01.012.
39. E. D. Folland, A. F. Parisi, P. F. Moynihan, D. R. Jones, C. L. Feldman, and D. E. Tow, “Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques,” *Circulation*, 1979, doi: 10.1161/01.CIR.60.4.760.
40. A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Med. Imaging*, 2015, doi: 10.1186/s12880-015-0068-x.
41. J. Zunic and P. L. Rosin, “A new convexity measure for polygons,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, doi: 10.1109/TPAMI.2004.19.

Tables

Table 1 Characteristics of patients in Geisinger echocardiography cohort.

Measures	Mean (SD)	N
Age (years)	68 (16)	47,431
Sex (% Male)	52	47,431
BMI (kg/m^2)	31 (8)	47,273
DBP (mmHg)	71 (12)	47,030
SBP (mmHg)	128 (20)	47,033
Heart Rate (bpm)	76 (15)	46,696
EDV (ml; bp)	115 (53)	9,339
ESV (ml; bp)	60 (43)	9,033
EF (%)[#]	53 (13)	46,505
LAV (ml; bp)	67 (30)	23,939
LVM (g)	197 (71)	44,862

Notes:

*Data are mean (standard deviation or SD) except for 'Sex'.

[#]Physician-reported EF values.

Phenotypes	Prevalence (%)
Atrial Fibrillation (Afib) History	32
Coronary Heart Disease	37
Diabetes Mellitus	33
Heart Failure	31
Hypertension	76

Notes: prevalence was calculated based on 47,431 records for each phenotype (except Afib was based on 47,207 records).

Table 2 Reported inter-observer errors for five clinical metrics using 2D echocardiography.

Reference	Method	Statistic	EDV	ESV	EF	LVM	LAV	No. of Subjects
Jacobs et al. [2]	MOD-bp	AD(2SD)	19 (20)	24 (21)	14 (17)	-	-	10
Hoffmann et al. [3]	MOD-bp	Individual SD(95%CI)	19.7 (3.4)	34.5 (5.8)	14.3 (2.5)	-	-	63
Mor-Avi et al. [6]	MOD-bp	AD(SD)	-	-	-	37 (19)	-	21
Mor-Avi et al. [1]	area-length-bp	AD(SD)	-	-	-	-	12 (12)	92
Wu et al. [5]	MOD-bp	AD(SD)	-	-	-	-	11 (12)	25
Keller et al. [4]	MOD-bp	Individual SD	-	-	-	-	39	21

MOD-bp: Simpson's bi-plane method of disks; area-length-bp: bi-plane area-length method;

AD: absolute difference between 2 observers in percent of their mean;

Individual SD: standard deviation of difference between 2 observers in percent of their mean;

AD = $\sqrt{2}$ x Individual SD.

Table 3 Performance of QC methods for identifying good segmentations with DICE ≥0.85.

	Structure (No. of Good Segmentations)	Precision / Sensitivity / F1	Good Segmentations Removed (%)	QC
CAMUS test set	LV (7,055)	1.0 / 0.85 / 0.92	15	Seg_dev
	Myo (6,413)	0.95 / 0.68 / 0.79	32	
	LA (6,958)	0.99 / 0.81 / 0.90	19	
EchoNet-Dynamic	LV (17,131)	0.94 / 0.80 / 0.86	20	Seg_dev
		0.88 / 0.97 / 0.92	3	Seg_dev + convexity

Table 4 Summary of absolute errors in five clinical measures estimated before and after QC. Two QC strategies were tested, i.e., QC with Seg_dev and QC with Seg_dev + convexity.

	Mean Absolute Error (%)			Median Absolute Error (%)		
	Pre-QC	Seg_dev	Seg_dev + convexity	Pre-QC	Seg_dev	Seg_dev + convexity
EDV	20.2	18.0	19.1	16.4	15.5	16.2
ESV	25.1	21.5	24.0	19.5	17.6	19.2
EF	15.6	15.1	15.3	11.0	10.2	10.8
LVM	23.1	20.9	21.5	17.7	16.4	17.1
LAV	20.9	19.2	19.9	16.1	15.0	15.6

Table 5 Summary of absolute errors in five clinical measures for studies removed by QC. Two QC strategies were compared, i.e., QC with Seg_dev and QC with Seg_dev + convexity.

	Mean Absolute Error of Removed Studies (%)		Median Absolute Error of Removed Studies (%)		Studies Removed (%)	
	Seg_dev	Seg_dev + convexity	Seg_dev	Seg_dev + convexity	Seg_dev	Seg_dev + convexity
EDV	33	65	23	40	14	2
ESV	34	63	24	37	31	3
EF	17	22	12	16	49	7
LVM	24	38	18	26	71	9
LAV	29	36	22	27	14	4

Figures

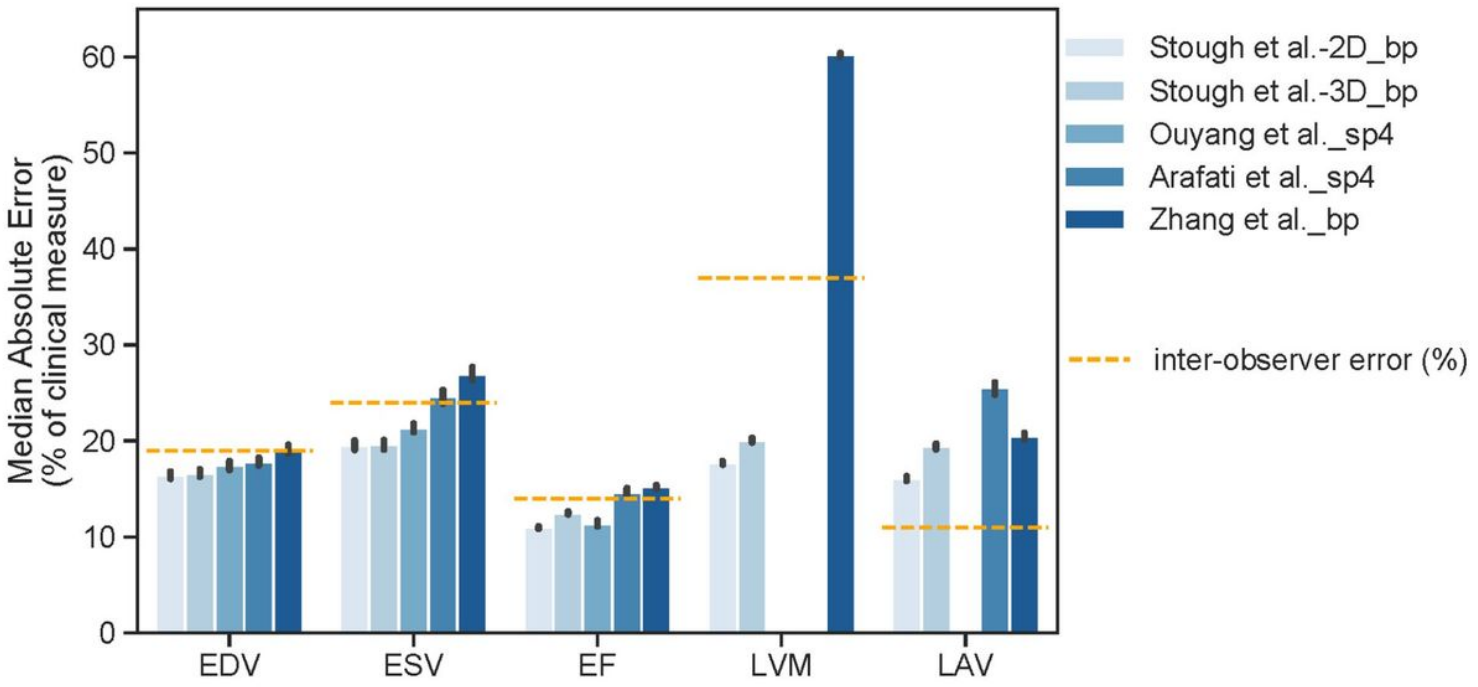

Figure 1

Comparison of five state-of-the-art segmentation models on Geisinger data. Segmentations generated by five models were used to estimate five clinical measures using Simpsons modified method of disk (MOD) (bp: bi-plane; sp4: single-plane from apical four chamber view). The median absolute errors (in percent of clinical values) were compared. Bars are median ± 95% CI. Orange dashed lines are the lowest inter-observer errors (%) reported for each measure [2-4].

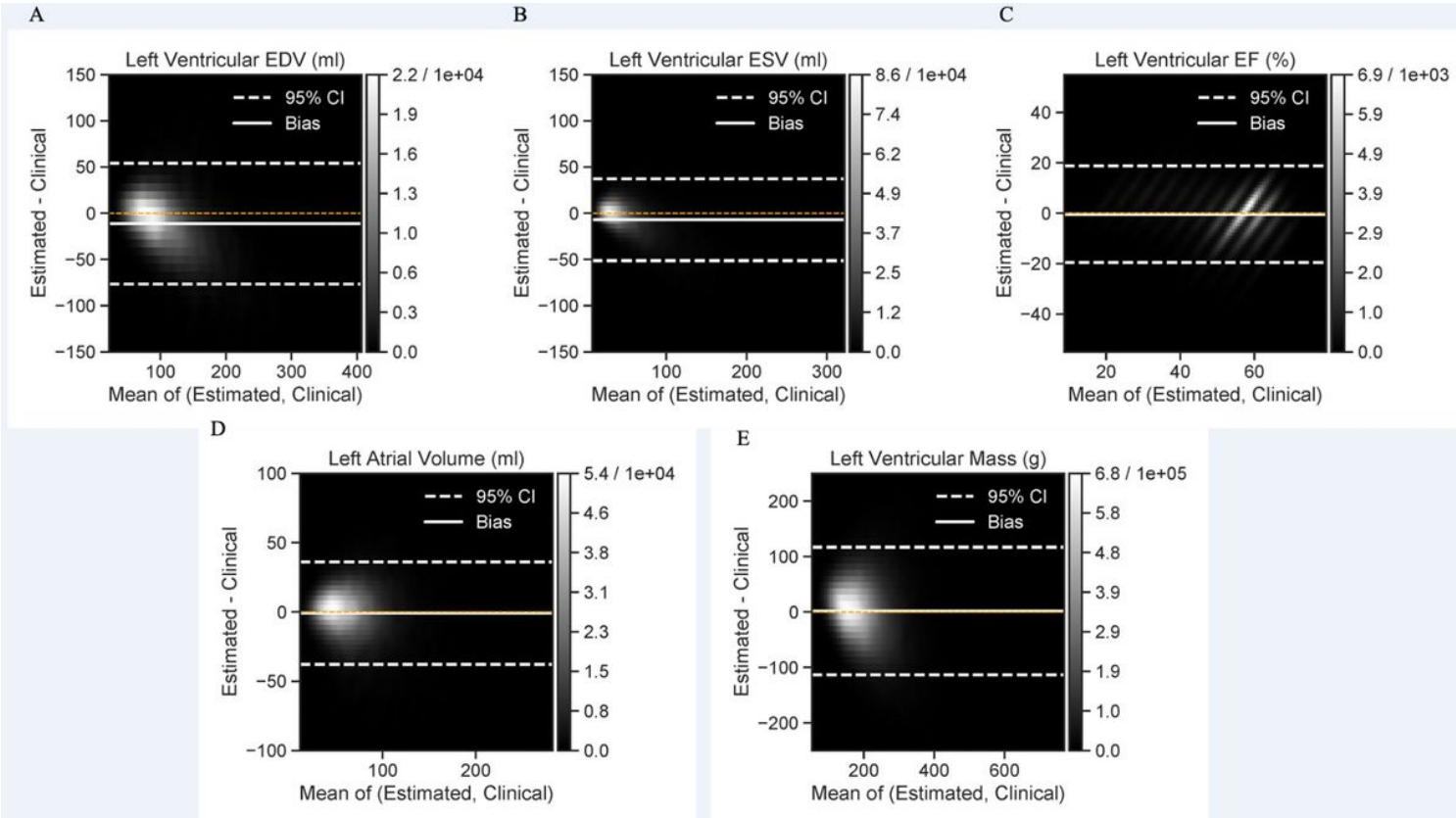

Figure 2

Error distribution of Geisinger volume and mass estimation associated with the Stough et al. 2D segmentation model. Bland-Altman density plots were generated for five clinical measures estimated using Simpsons modified bi-plane method of disk and segmentations output by the Stough et al. 2D model. White solid and dashed lines are mean bias and 95% CI. Orange dashed lines are 0 bias.

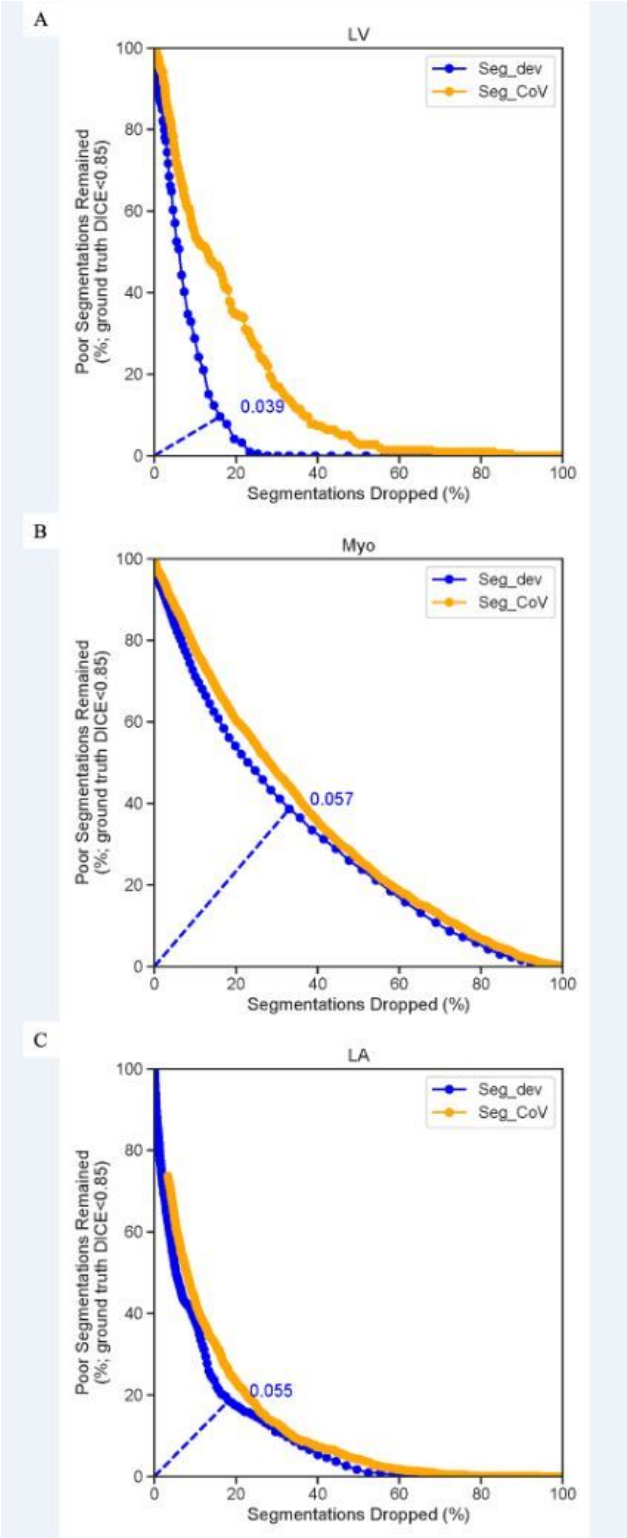

Figure 3

Pareto front curves for segmentation quality control (QC) obtained from augmented CAMUS training set. Percentage of poor segmentations with ground truth DICE <0.85 remained after QC was plotted, respectively, for LV endocardium (LV), LV wall (Myo), and LA endocardium (LA) against percentage of segmentations dropped by QC at different thresholds for two uncertainty measures. Blue dashed lines mark the cutoffs that were closest to the origin point (0,0) for uncertainty measure Seg_dev.

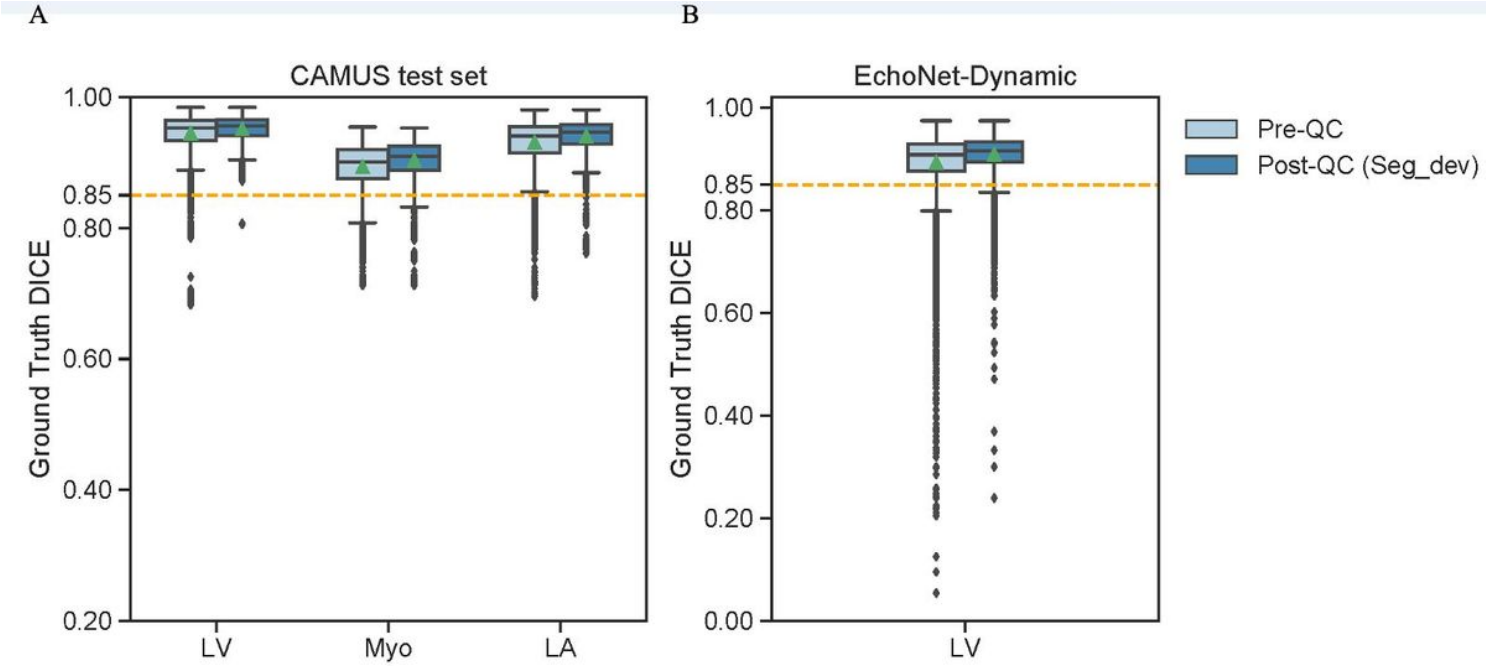

Figure 4

Box plots of ground truth DICE before and after segmentation quality control using Seg_dev. Distributions of ground truth DICE for LV endocardium (LV), LV wall (Myo), and LA endocardium (LA) were plotted (left panel: CAMUS test dataset; right panel: EchoNet-Dynamic dataset). Green triangles mark the mean DICE scores.

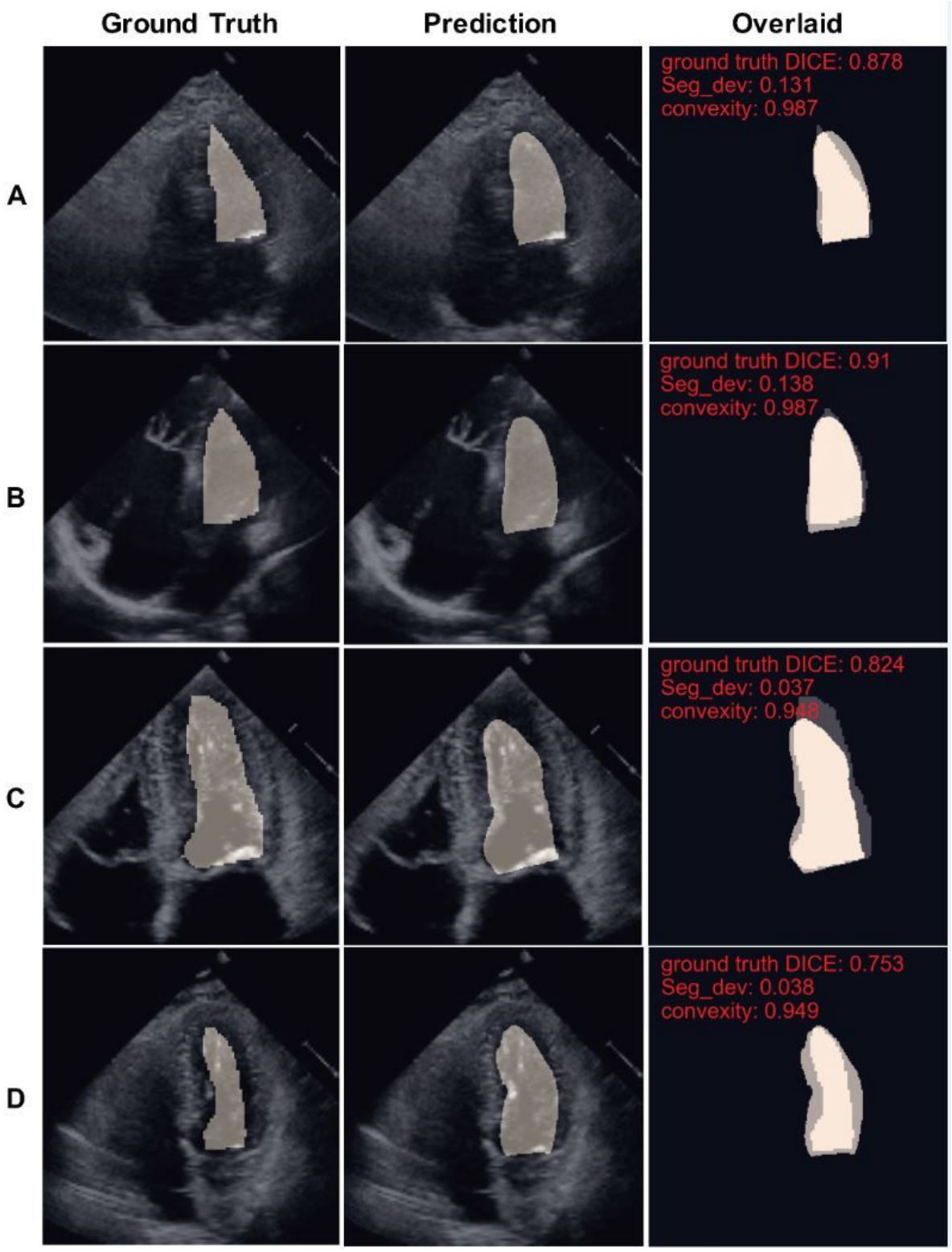

Figure 5

Examples of EchoNet-Dynamic segmentations with uncertainty and convexity measures. Representative LV endocardial segmentations with disassociated measures of ground truth DICE and Seg_dev were displayed. Ground truth was manual tracing; prediction was generated by the Stough et al. 2D model.

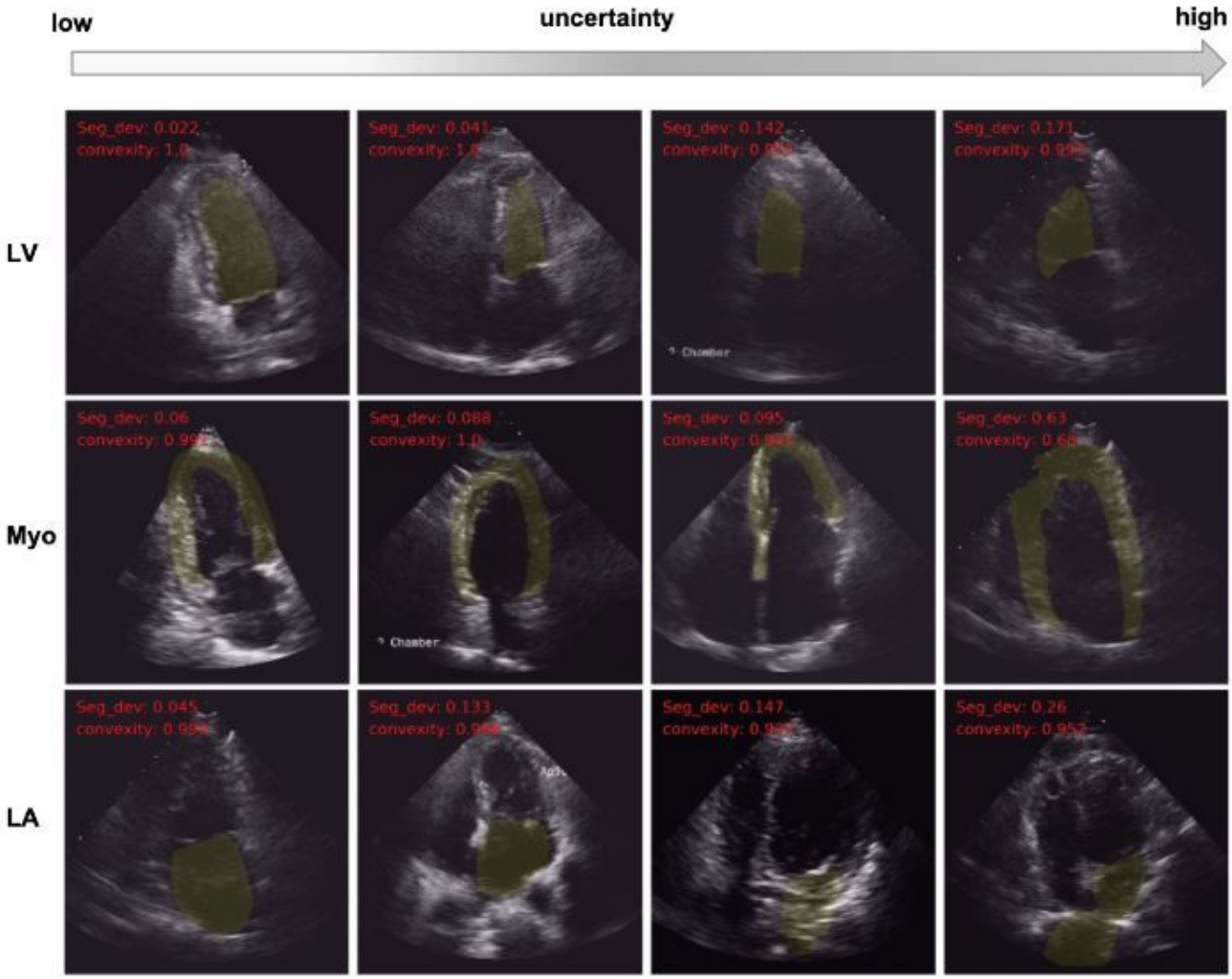

Figure 6

Examples of Geisinger segmentations with uncertainty and convexity measures. Representative segmentations with varying uncertainty and convexity were displayed for LV endocardium (LV), LV wall (Myo), and LA endocardium (LA). For LV wall, the convexity was measured as the minimal value of endocardial and epicardial convexity scores.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigure1.tiff
- SupplementaryInformation.docx