European Heart Journal - Digital Health (2024) **5**, 77–88
https://doi.org/10.1093/ehjdh/ztad069

ORIGINAL ARTICLE

Population data–based federated machine learning improves automated echocardiographic quantification of cardiac structure and function: the *Automatisierte Vermessung der Echokardiographie* project

Caroline Morbach [1,2,]*, Götz Gelbrich[3,4], Marcus Schreckenberg[5], Maike Hedemann[1], Dora Pelin[1], Nina Scholz[1], Olga Miljukov[3], Achim Wagner[6], Fabian Theisen[1], Niklas Hitschrich[5], Hendrik Wiebel[5], Daniel Stapf[5], Oliver Karch[6], Stefan Frantz [1,2], Peter U. Heuschmann[3,4,†], and Stefan Störk [1,2,†]

[1]Department Clinical Research and Epidemiology, Comprehensive Heart Failure Center, University Hospital Würzburg, Am Schwarzenberg 15, D-97078 Würzburg, Germany; [2]Department of Medicine I, University Hospital Würzburg, Oberdürrbacherstr. 6, D-97080 Würzburg, Germany; [3]Institute of Clinical Epidemiology and Biometry, University of Würzburg, Josef-Schneider-Str. 2, 97080 Würzburg, Germany; [4]Clinical Trial Center, University Hospital Würzburg, Josef-Schneider-Str. 2, 97080 Würzburg, Germany; [5]TOMTEC Imaging Systems GmbH, Freisinger Str. 9, 85716 Unterschleissheim, Germany; and [6]Service Center Medical Informatics, University Hospital Würzburg, Josef-Schneider-Str. 2, 97080 Würzburg, Germany

Received 25 April 2023; revised 8 September 2023; accepted 3 October 2023; online publish-ahead-of-print 15 November 2023

Aims

Machine-learning (ML)-based automated measurement of echocardiography images emerges as an option to reduce observer variability. The objective of the study is to improve the accuracy of a pre-existing automated reading tool ('original detector') by federated ML-based re-training.

Methods and results

Automatisierte Vermessung der Echokardiographie was based on the echocardiography images of $n = 4965$ participants of the population-based Characteristics and Course of Heart Failure Stages A–B and Determinants of Progression Cohort Study. We implemented federated ML: echocardiography images were read by the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the University Hospital Würzburg (UKW). A random algorithm selected 3226 participants for re-training of the original detector. According to data protection rules, the generation of ground truth and ML training cycles took place within the UKW network. Only non-personal training weights were exchanged with the external cooperation partner for the refinement of ML algorithms. Both the original detectors as the re-trained detector were then applied to the echocardiograms of $n = 563$ participants not used for training. With regard to the human referent, the re-trained detector revealed (i) superior accuracy when contrasted with the original detector's performance as it arrived at significantly smaller mean differences in all but one parameter, and a (ii) smaller absolute difference between measurements when compared with a group of different human observers.

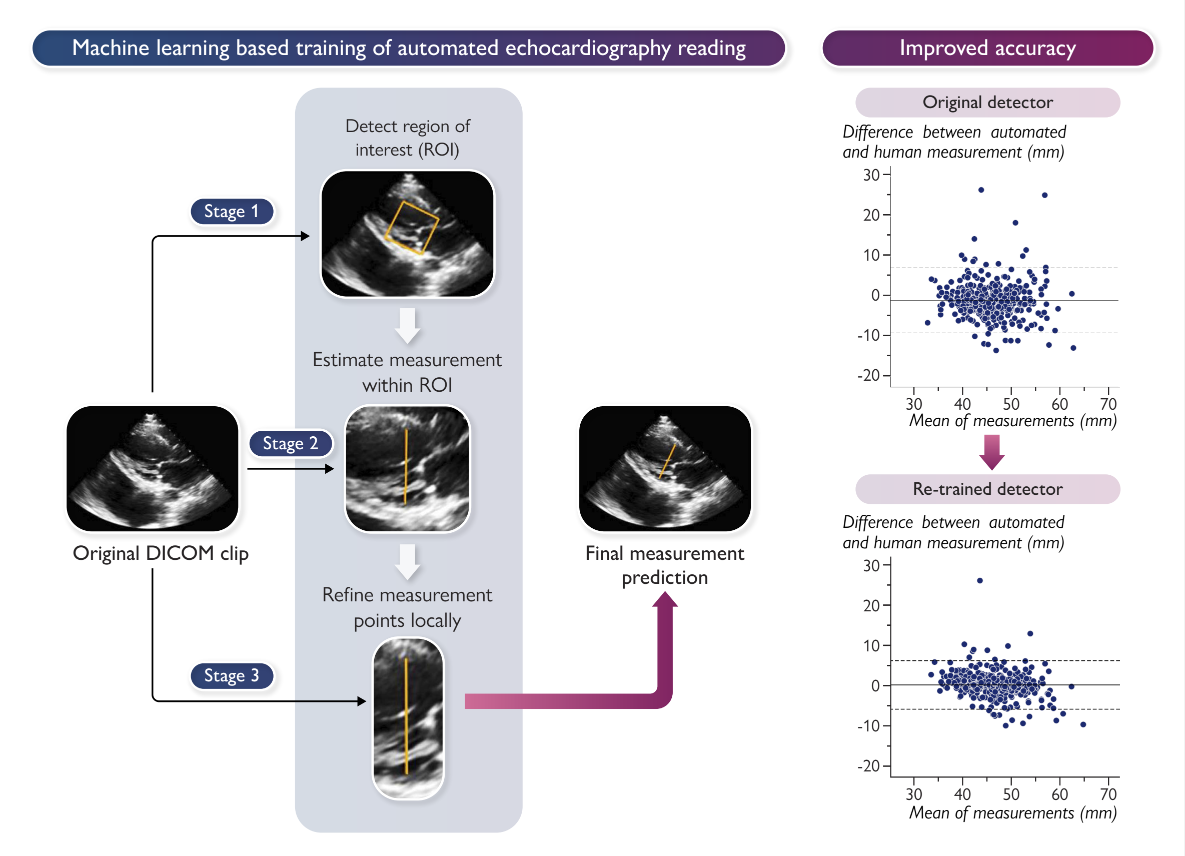
Conclusion

Population data–based ML in a federated ML set-up was feasible. The re-trained detector exhibited a much lower measurement variability than human readers. This gain in accuracy and precision strengthens the confidence in automated echocardiographic readings, which carries large potential for applications in various settings.

* Corresponding author. Tel: +49 931 201 46248, Email: Morbach_c@ukw.de

[†]These authors contributed equally to this work.

Graphical Abstract


Machine learning based training of automated echocardiography reading
Improved accuracy
Detect region of interest (ROI)
Stage 1
Estimate measurement within ROI
Stage 2
Original DICOM clip
Refine measurement points locally
Stage 3
Final measurement prediction
Original detector
Difference between automated and human measurement (mm)
30
20
10
0
-10
-20
30 40 50 60 70
Mean of measurements (mm)
Re-trained detector
Difference between automated and human measurement (mm)
30
20
10
0
-10
-20
30 40 50 60 70
Mean of measurements (mm)

Keywords Automated measurement • Machine learning • Measurement variability • Observer variability • Sample size • Population-based cohort

Introduction

Echocardiography is widely used to guide the diagnosis and management of cardiac diseases. Clinical decisions regarding pharmacotherapy or devices are frequently based on echocardiography-derived measurement values and/or serial assessment of cardiac structure and function, e.g. when targeting potentially cardiotoxic chemotherapy.[1,2] Another field of application of this imaging modality is clinical and epidemiologic research. There, the sample size required to reveal a significant between-group difference or temporal change of an echocardiography parameter critically depends on the respective measurement error. Minimizing measurement variability, hence, is of utmost importance for both patient care and clinical research.

Both image acquisition and reading generate measurement variability and thus impact the total difference between two measurements.[3–5] Acquisition variability can be reduced by the deployment of standardized sonographer training and regular credentialing, provision of state-of-the-art equipment, alignment of ultrasound machine system pre-sets, and standardization of imaging protocols.[6] Reading variability can be reduced by the implementation of central reading in an echocardiography core laboratory.[7,8] These approaches, however, are frequently cost- and/or time-intensive and therefore have not yet become routine clinical standard.

Machine-learning (ML)-based automated measurement of echocardiography images have emerged as a potential modality to reduce reading variability.[9–12] Only recently, Tromp *et al.*[13] showed that the disagreement between the measurements of a deep-learning-based automated workflow and a human measurement was lower than the disagreement among three core laboratory readers. The World Alliance of Societies of Echocardiography Normal Values (WASE) study recruited healthy individuals from 15 different countries to derive reference values from standardized echocardiograms.[14,15] Based on these echocardiograms and the respective core laboratory readings, Lang *et al.*[16] trained a detector by ML to automatically determine standard echocardiography parameters. Validation of the detector in a subset of WASE echocardiograms showed excellent agreement with the manual measurements of an expert reader.[16] Further, measurement variability was comparable with the human inter-observer variability between two expert readers.[16]

Given the excellent performance originating from close-to-perfect conditions found in highly selected cohorts of healthy participants, the next step of advancement necessitates training of detectors on the echocardiograms of the general population. Here, detectors need to master suboptimal acoustic windows and deviations from normal cardiac morphology and function. This requires access to large amounts of echocardiographic data and the respective phenotypic characterization of the imaged participants. As this process is frequently impeded by stringent data protection laws posing major barriers, alternative modalities such as federated learning have to be explored.

Federated learning is an ML technique that allows multiple training sites to collectively build, train, and refine an ML model without the need to access or share a central data pool. Each site has its own local data pool, and only the trained model weights are exchanged among them. As no sensitive patient data are shared outside the clinical network, federated learning offers a promising approach for collaborative ML developments among multiple (clinical) sites, considering critical aspects such as data privacy, data security, access rights, and access to heterogeneous data.

The present study aims to (i) implement a federated ML system in a clinical environment and (ii) apply the above-described detector to the echocardiograms of a large and well-characterized population-based cohort with the goal of further improving measurement accuracy by federated artificial intelligence–supported re-training.

Methods

Study population

The population-based Characteristics and Course of Heart Failure Stages A–B and Determinants of Progression (STAAB) Cohort Study included individuals without self-reported heart failure from the general population of Würzburg, Germany, aged 30–79 years and stratified for age (10:27:27:27:10 for the respective decades) and sex (1:1) between December 2013 and October 2017. The detailed study design and methodology have been published.[17,18] We implemented a rigid and regular quality control process for all study-related procedures.[17,18] The STAAB cohort study protocol and procedures complied with the Declaration of Helsinki and received positive votes from the Ethics Committee of the Medical Faculty (vote #98/13) as well as from the data protection officer of the University of Würzburg. All participants provided written informed consent prior to any study examination.

Echocardiography

All participants underwent transthoracic echocardiography (Vivid S6 or Vivid E95; GE Healthcare, Horten, Norway) performed by dedicated trained personnel, which was internally certified and quality-controlled on a regular basis.[3] Cine loops were recorded based on three R-R intervals, labelled only with the participant's study identification number (pseudonymized) and stored in raw data format. The pre-specified assessment protocol included a parasternal long-axis view for the measurement of the left ventricular outflow tract diameter (LVOT), the left ventricular (LV) end-diastolic (LVDd) and end-systolic (LVDs) diameters, the end-diastolic thickness of the interventricular septum (IVSd), and the LV posterior wall (LVPWd), as well as an M-mode recording of the basal LV with the ultrasound beam perpendicular to the IVS. We further recorded apical four- and two-chamber views and performed tissue-Doppler pulsed-wave Doppler measurements of the septal and lateral mitral annulus to assess the respective velocities septal *e′* and lateral *e′*. Early (*E*) and late (*A*) diastolic mitral inflow velocities were assessed by using pulsed-wave Doppler with the acquisition window positioned at the mitral leaflet tips.[18–20] Along the scanning procedure, the sonographer immediately performed pre-specified measurements [human on-site measurement; i.e. LVDd, LVDs, IVSd, LVPWd, septal *e′*, lateral *e′*, *E*, *A*, as well as LV end-diastolic and end-systolic volumes to calculate the LV ejection fraction (LVEF)] and congregated them in a result sheet for each study participant.

The Automatisierte Vermessung der Echokardiographie project

Automatisierte Vermessung der Echokardiographie (AVE) was performed as co-operation project of the University Hospital and University of Würzburg and TOMTEC Imaging Systems GmbH, Unterschleißheim, Germany. The study protocol was approved by the Ethics Committee of the Medical Faculty Würzburg in the format of an amendment to the STAAB study protocol. All study procedures were in concordance with the German data protection law as confirmed by the data protection officer of the University of Würzburg. All echocardiography images remained within the University Hospital premises at any time. The stored echocardiography images of all STAAB participants were exported in DICOM format and imported into the clinical image analysis software (TOMTECArena®, TOMTEC Imaging Systems GmbH) of the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the Comprehensive Heart Failure Center (CHFC), Würzburg, Germany (*Figure 1*). Trained and internally certified personnel performed measurements on the complete STAAB population according to a pre-specified protocol, including LVDd, LVDs, IVSd, LVPWd, septal *e′*, lateral *e′*, *E*, and *A*. Thus, derived measurements served as a *human referent* in statistical analyses on measurement differences. In order to quantify the performance of the original detector and to train and validate a potentially improved detector, the STAAB population was divided into three distinct subgroups (training pool, validation Pool 1, and validation Pool 2) stratified for sex and age decades using a random algorithm applying a 4:1:1 ratio. The respective group size allowed for a larger training pool and two equally sized validation pools. In addition, images of the first 250 participants from validation Pool 1 were measured repetitively by observers of different experiences, who were blinded to the results of the other observers (*repeat human measurements*). Validation Pool 2 has not been used in the study yet and is part of the ongoing utilization phase of the AVE project.

Federated machine learning

At no point during the AVE project were patient data or DICOM studies shared with the external partner TOMTEC Imaging Systems or processed outside of the hospital network. The generation of ground truth data and the computations for the ML training cycles took place strictly on computers within the hospital network. Only non-personal model weights and training states were exchanged with the software engineers for analysis and refinement of the algorithms (*Figure 1*). A centralized server-based federated ML mechanism was developed for this project. Each node had exclusive access to its local data pool and only the current solver state, and a model snapshot was exchanged with the server (*Figure 1*). Because of the sensitivity of patient data, internet access was highly restricted on all nodes.

The server's main task is to take the training state from one node and pass it to the next one. It considers the pool size on each node and decides when the training is completed. Local epoch counts and learning rates are adapted with respect to the pool sizes. An additional worker process is launched on each node, which is responsible for observing the local training (states) and synchronizing the training data with the centralized server (*Figure 1*). In contrast to conventional federated learning approaches, the averaging of model weights after each training round was omitted for our project. Instead, the model was passed from one node to the next (including the solver state). If the model/solver exchange happens very frequently, the training conditions become similar to a traditional non-federated set-up. This approach is time-consuming, as nodes do not work in parallel but must wait for each other. Also, the binary files for solver state and model weights must be exchanged frequently. One data exchange includes between 10 and 30 Megabytes and happens every 20–60 s. This is equivalent to ∼1000 learning iterations on one node before the next model exchange. There is a trade-off between network transfer times and the smallest possible learning increment. Because training time efficiency was not considered critical for this project, equivalence to a result that would be expected using a common pool on a central site was a matter of priority for our approach.

Re-training of measurement detectors

All detector models were trained from the ground up and no pre-trained model weights were used. Re-training of the detectors was based on the measurements of the human referent. The de-identified measurements values and coordinates were extracted from the clinical image analysis software by using a specifically developed database extraction tool (TOMTEC Imaging Systems GmbH). The clinical image analysis software provides a specialized detector for each labelled measurement; for example,

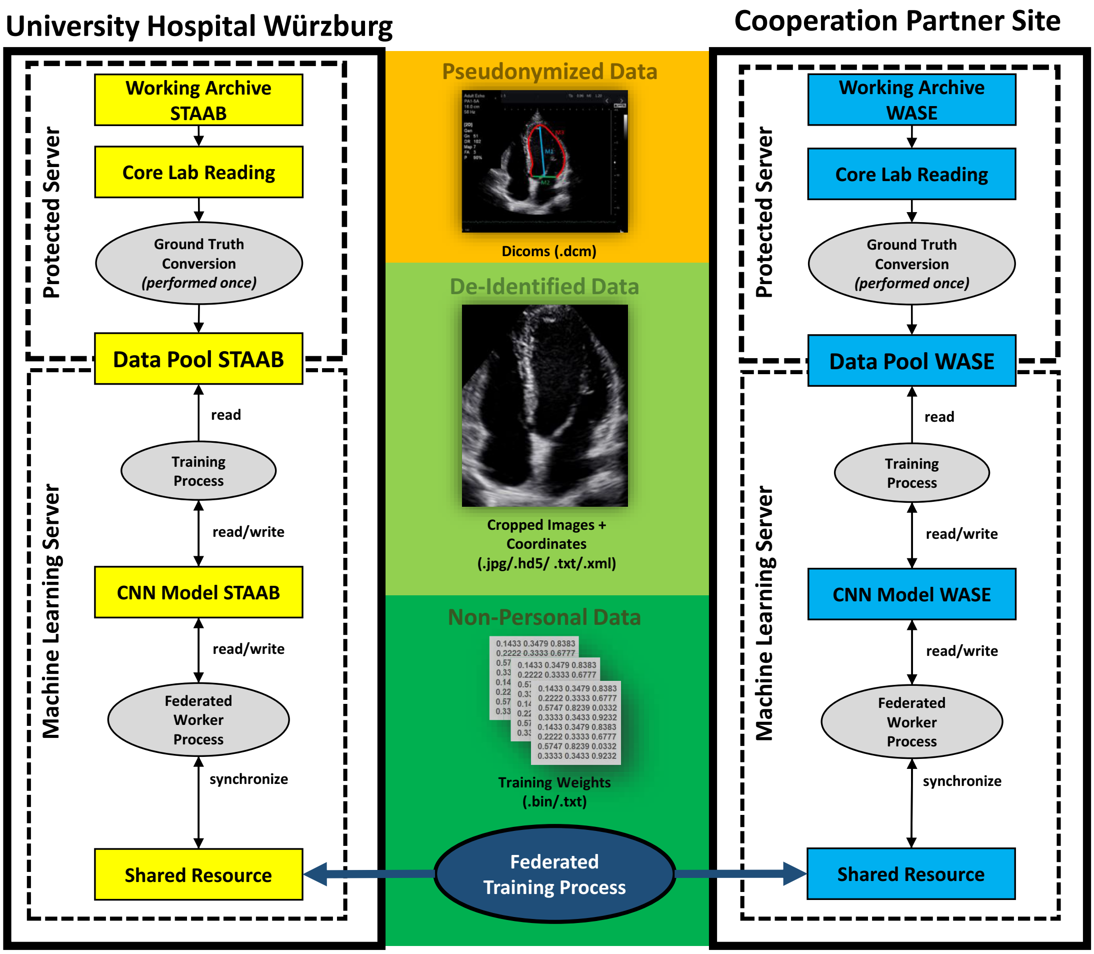


Figure 1 Federated machine-learning set-up. The pseudonymized echocardiography images of the participants of the population-based *Characteristics and Course of Heart Failure Stages A–B and Determinants of Progression* (STAAB) Cohort Study were imported in DICOM format into the TomtecArena® (TOMTEC Imaging Systems GmbH) of the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the Comprehensive Heart Failure Center, University Hospital Würzburg, Germany. Trained and internally certified personnel (human referent) performed measurements according to a pre-specified protocol. The de-identified values and coordinates of measurements performed by the human referent were extracted as ground truth from the TomtecArena® using a specifically developed database extraction tool (TOMTEC Imaging Systems GmbH). The ground truth conversion and the computations for the machine-learning training process took place on computers within the hospital network. The same was true for the Cooperation partner site using the echocardiography images from the World Alliance of Societies of Echocardiography Normal Values (WASE) study. For the centralized server-based federated machine learning, the respective convolutional neural network models had exclusive access to their local data pools within their own virtual private networks. Only the current solver status and a model snapshot were exchanged with the federated server that had access to both networks.

there are separate line segment detectors for LVOT, LVPWd, IVSd, e′, etc. which had been trained previously. Each measurement has a geometry type; for example, a Doppler velocity measurement is represented by a 2d point, a distance measurement by two 2d points, and a tracked endocardial contour by a list of splines. For each geometry type, there is a separate detector class consisting of a cascade of multiple convolutional neural networks (CNNs) + optional trackers. Within the CNN cascade, coarse features such as the bounding box are detected at the beginning and the position of individual points is refined towards the end. In the re-training of the detectors, both grid search and manual steps were used for obtaining cardiac ultrasound data. The training pool was used for a 10-fold cross validation (9:1 split) and hyperparameter optimization.

Each detector consists of several cascaded detection stages with individual CNNs. In earlier detection stages, coarse features like oriented regions of interests are trained, whereas in later stages, the focus is on finer details like point positions. The output from each stage is used as an input by the next one. This way complexity is reduced, and the detection task is distributed across specialized subdetectors (*Graphical Abstract*).

The image pre-processing pipeline has a key role in generating the actual input for the neural networks. We identified the frames that had been selected for the generation of the ground truth and sampled the input frames equidistantly from the corresponding RR cycle. For 2D measurements, from the original B-mode clip, an interval of frames between two R-wave events is extracted. In the case of a Doppler clip, one complete R-R cycle is cropped from the tissue region. The selected images are then processed by using a detector-specific pipeline, which, for example, includes contrast normalization, image orientation normalization, signal wrapping and scaling (Doppler), down sampling, ultrasound sector mask extraction, etc. The processed frames are then merged into a multi-channel image and forwarded to the CNN inference engine.

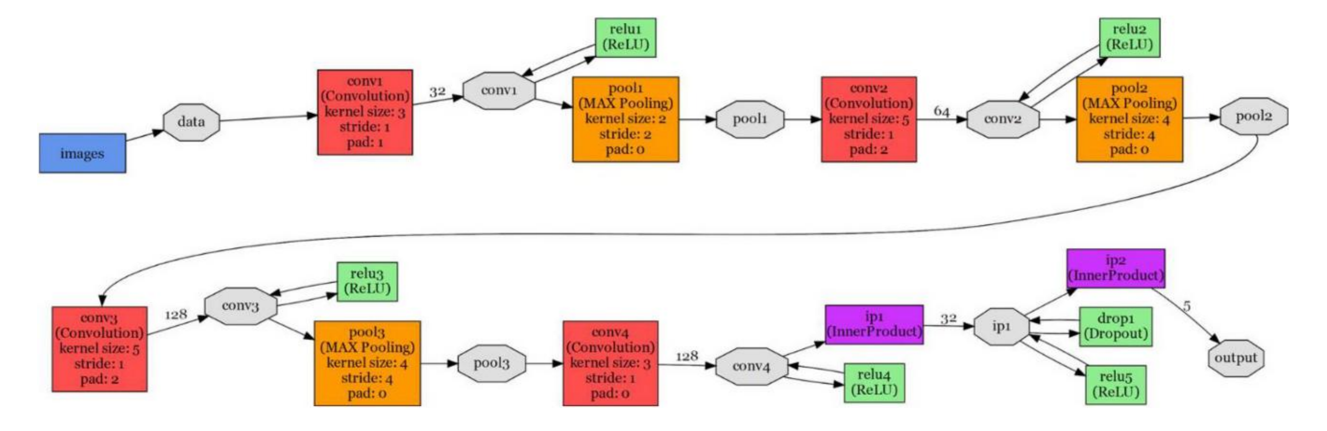


Figure 2 Example of one convolutional neural network as applied for each echocardiography parameter. One convolutional neural network contains convolutional rectified linear units and fully connected layers. Its output is of low dimensionality, e.g. a vector of five floating point values representing a bounding box and its orientation. Only the combination of multiple convolutional neural networks leads to the desired precision. conv, convolution; MAX, maximal; ip, inner product.

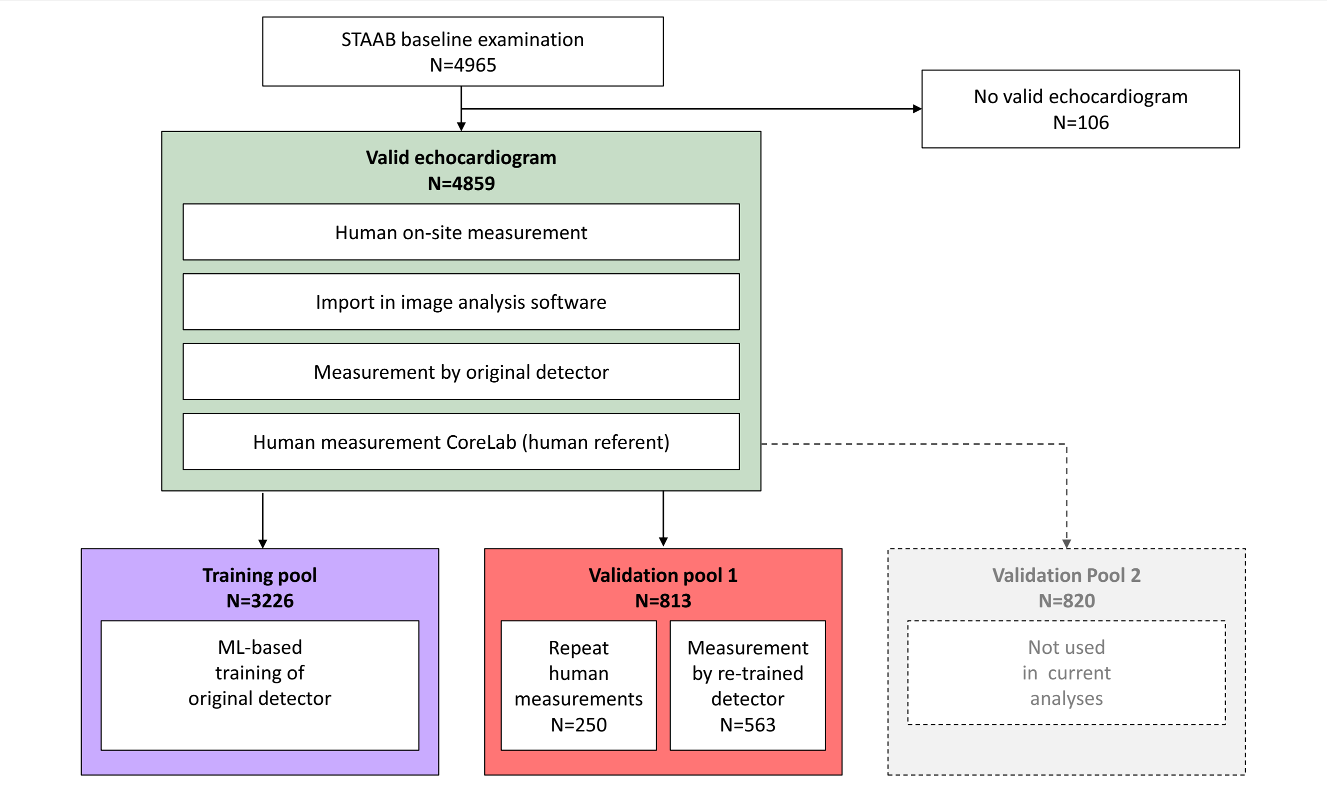


Figure 3 Study flow. Of the 4965 participants of the population-based Characteristics and Course of Heart Failure Stages A–B and Determinants of Progression (STAAB) Cohort Study, 4859 had valid echocardiograms and entered the Automatisierte Vermessung der Echokardiographie project. All echocardiograms had been read at the time point of scanning (human on-site measurement). At project start, the stored pseudonymized echocardiography images of the STAAB participants were imported into the TomtecArena®. Subsequently, all scans were read by the original detector. Further, trained and internally certified personnel of the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the Comprehensive Heart Failure Center, Würzburg, Germany, performed measurements in all echocardiograms, serving as the human referent in further analyses. The STAAB population was then divided into three distinct subgroups using a random algorithm applying a 4:1:1 ratio. The respective group size allowed for a larger training pool (*n* = 3226), and two equally sized validation pools. The training pool served for machine-learning-based training of the original detector. Images of the first 250 STAAB participants allocated to validation Pool 1 were measured repetitively by observers of different experience, blinded to the results of the other observers (repeat human measurements). The remaining images of validation Pool 1 served for measurements of the re-trained detector. The validation Pool 2 was not used in the current analysis.

Table 1 Characteristics of STAAB participants with valid echocardiograms entering Automatisierte Vermessung der Echokardiographie

	Total sample ($n = 4859$)	Training pool ($n = 3226$)	Validation Pool 1 ($n = 813$)	Validation Pool 2 ($n = 820$)
Female sex	2561 (52.7)	1700 (52.7)	429 (52.8)	432 (52.7)
Age (years)	54.9 (11.8)	54.9 (11.8)	55.0 (11.8)	54.9 (11.9)
BSA [m^2]	1.9 (0.2)	1.9 (0.2)	1.9 (0.2)	1.9 (0.2)
Hypertension	2246 (46.2)	1473 (45.7)	411 (50.6)	362 (44.1)
Diabetes mellitus	451 (9.3)	299 (9.3)	72 (8.9)	80 (9.8)
Coronary heart disease	191 (3.9)	122 (3.8)	34 (4.2)	35 (4.3)
Heart rate [min^{-1}]	68 (11)	68 (11)	67 (10)	68 (10)
Systolic BP [mmHg]	131 (18)	131 (18)	132 (18)	130 (18)
Diastolic BP [mmHg]	78 (12)	78 (11)	79 (15)	78 (13)
Echocardiography[a]				
LVDd [mm]	48.5 (5.1)	48.5 (5.1)	48.4 (5.3)	48.6 (4.9)
LVDs [mm]	32.6 (7.0)	32.7(7.4)	32.5 (5.9)	32.5 (5.8)
IVSd [mm]	8.8 (1.6)	8.7 (1.6)	8.8 (1.6)	8.8 (1.5)
LVPWd [mm]	7.7 (1.5)	7.7 (1.5)	7.8 (1.5)	7.7 (1.4)
LVOT diameter [mm][b]	21.6 (1.9)	21.6 (1.9)	21.8 (2.1)	21.7 (1.8)
E [cm/s]	70 (20)	70 (20)	70 (20)	70 (30)
A [cm/s]	60 (20)	60 (20)	60 (20)	60 (30)
e′ lateral [cm/s]	10.9 (3.2)	10.9 (3.2)	10.9 (3.3)	10.8 (3.2)
e′ septal [cm s^{-1}]	8.5 (2.5)	8.5 (2.5)	8.4 (2.5)	8.5 (2.5)
LVEF [%]	59.8 (4.8)	59.8 (4.9)	59.9 (4.9)	59.9 (4.6)
LVVd [mL]	101 (27)	101 (27)	100 (28)	103 (27)

Data are count (per cent) or mean (SD).

A, late mitral inflow velocity; BP, blood pressure; BSA, body surface area; *E*, early mitral inflow velocity; *e′*, mitral annular early diastolic velocity; IVSd, end-diastolic thickness of the interventricular septum; LVDd, left ventricular end-diastolic diameter; LVDs, left ventricular end-systolic diameter; LVEF, left ventricular ejection fraction; LVPWd, end-diastolic thickness of left ventricular posterior wall; LVOT, left ventricular outflow tract diameter; LVVd, left ventricular end-diastolic volume.

[a] Human on-site measurement.

[b] Available in $n = 2313$ participants.

The CNN architecture has not been modified for the federated learning set-up. The only adaption was an appropriate learning rate reduction in order to prevent the fine-tuned model weights from different nodes from overwriting each other during the final solver state exchanges. An Adam solver was selected as optimizer, reducing the initial learning rate by half every 30–90 epochs with a total epoch count of 100–300. All networks were designed with a focus on fast object detection for cardiologic ultrasound images. A single CNN follows a simplistic approach by using a few blocks of convolutional, Rectified Linear Unit and fully connected layers. Their output is also of low dimensionality, e.g. a vector of five floating point values representing a bounding box and its orientation. Only the combination of multiple CNNs leads to the desired precision. Depending on the specific detector and detection stage, the parameters differ for neuron count, kernel size, padding, etc. *Figure 2* shows the layout of a typical neural network applied in this study.

The 'caffe' deep-learning framework (https://caffe.berkeleyvision.org/) was used for all training activities on a Linux machine with a NVIDIA GeForce RTX 3070 TI.

Data analysis

Statistical analysis was performed using SPSS (Version 26; SPSS Inc., Chicago, IL, USA). Descriptives were summarized as frequencies (per cent) and mean (standard deviation), as appropriate. To assess measurement variability, we performed Bland–Altman analyses calculating the mean differences between the human referent and the respective measurements of the original detector (Scenario A) as well as of the re-trained detector (Scenario B). To quantify changes in measurement variability purported by the re-trained detector, we also determined the absolute differences between Scenario A vs. Scenario B. Accordingly, a negative value for measurement differences would favour Scenario B, indicative of an improvement of the detector by ML-based training. We further assessed, how often the measurement difference was smaller in Scenario B. Correlations were calculated using Pearson's correlation coefficient. Groups were compared by parametric or non-parametric tests, as appropriate. We calculated percentiles for the absolute difference between measurements. All tests were performed two-sided, and *P*-values <0.05 were considered statistically significant.

Results

For the AVE study, development of a federated ML system was feasible providing training conditions similar to a traditional non-federated set-up. This approach ensured that no personal data left the hospital network and that the measurement detectors could be re-trained within the clinical environment.

From the total STAAB sample,[18] which comprised $n = 4965$ participants [52% women, mean age 55 (12) years], $n = 4859$ (98%) had valid echocardiograms and qualified for the AVE protocol (*Figure 3*). The random algorithm allocated $n = 3226$ participants to the training pool, $n = 813$ to the validation Pool 1, and $n = 820$ to the validation Pool 2. From validation Pool 1, the first consecutive $n = 250$ participants were evaluated from five different observers (repeat human measurements; *Figure 3*). Distribution of age, sex, and body dimensions as well as of cardiovascular risk factors, comorbidities, and cardiac structure and function were comparable among the three pools (*Table 1*).

We first applied the original detector to the imported images. Comparing these automated measurements with the human referent

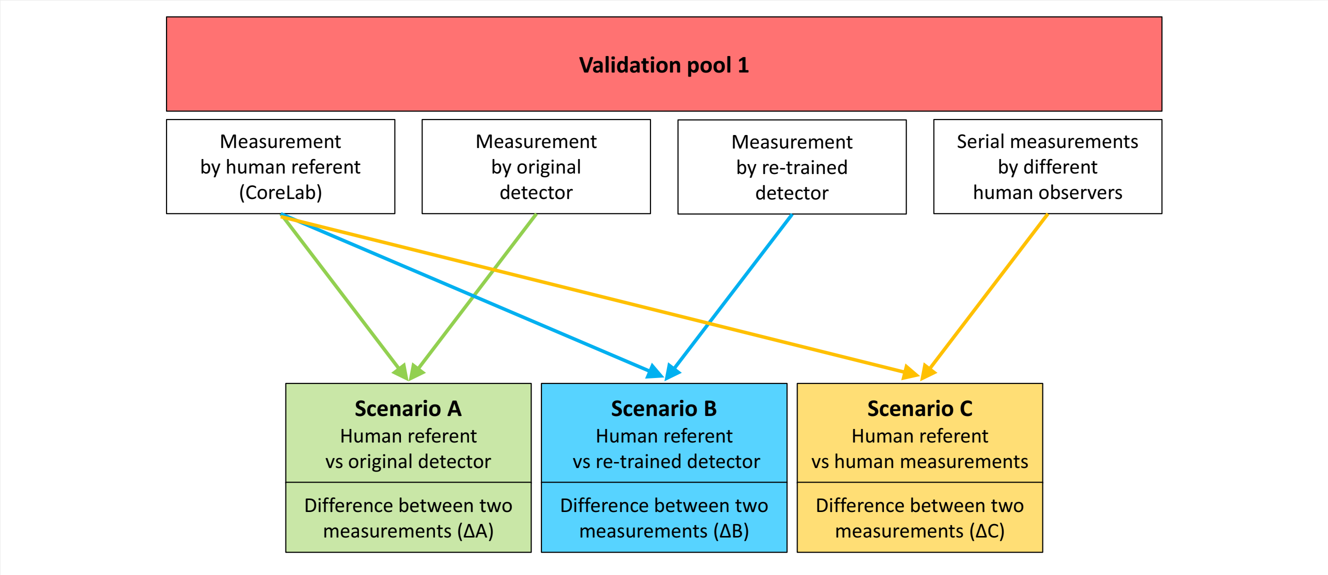

Validation pool 1
Measurement by human referent (CoreLab)
Measurement by original detector
Measurement by re-trained detector
Serial measurements by different human observers
Scenario A
Human referent vs original detector
Difference between two measurements (ΔA)
Scenario B
Human referent vs re-trained detector
Difference between two measurements (ΔB)
Scenario C
Human referent vs human measurements
Difference between two measurements (ΔC)

Figure 4 Scheme of measurement scenarios. The images of the 813 STAAB participants allocated to validation Pool 1 were used for the assessment of measurement accuracy. All images had been measured by trained and internally certified personnel of the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the Comprehensive Heart Failure Center, Würzburg, Germany, serving here as human referent. Scenario A compares the measurements of the original detector against the human referent, Scenario B compares the measurements of the re-trained detector against the human referent, and Scenario C compares the measurements of different human observers against the human referent.

Table 2 Measurement variability of the original detector (A) and the trained detector (B) in comparison with the human referent, respectively

	No. of measurements	Original detector vs. human referent (Scenario A) Mean difference between measurements (95% CI)	Re-trained detector vs. human referent (Scenario B) Mean difference between measurements (95% CI)	Absolute difference between measurement differences (Scenarios A vs. B) Mean (95% CI)	Smaller difference observed in Scenario B *n* (% of measurements)
LVDd [mm]	428	−1.1 (−1.5, −0.7)	0.2 (−5.9, 6.2)	−0.9 (−1.2, −0.7)**	281 (66)**
LVDs [mm]	388	2.9 (2.3, 3.4)	1.0 (0.6, 1.4)	−1.4 (−1.8, −1.0)**	242 (62)**
IVSd [mm]	427	0.0 (−0.1, 0.2)	−0.1 (−0.3, 0.0)	−0.1 (−0.2, 0.0)	223 (52)
LVPWd [mm]	427	1.2 (1.1, 1.3)	0.7 (0.6, 0.8)	−0.2 (−0.3, −0.1)**	248 (58)*
LVOT diameter [mm]	440	0.2 (−0.0, 0.4)	0.1 (−0.1, 0.2)	−0.3 (−0.5, −0.2)**	264 (60)**
E [cm/s]	494	3.7 (3.3, 4.0)	0.1 (−0.3, 0.4)	−1.5 (−1.8, −1.20)**	341 (69)**
A [cm/s]	489	5.8 (5.2, 6.3)	0.1 (−0.4, 0.6)	−3.3 (−3.7, −2.9)**	389 (80)**
e′ lateral [cm/s^{1}]	467	1.0 (0.9, 1.1)	0.0 (−0.1, 0.1)	−0.6 (−0.7, −0.5)**	380 (81)**
e′ septal [cm/s^{-1}]	471	0.5 (0.4, 0.6)	0.0 (−0.0, 0.1)	−0.2 (−0.2,−0.1)**	322 (68)**

The original detector performed well regarding determination of IVSd, for example, and there was no significant difference between the automated measurement and the human referent. The re-trained detector was able to achieve even smaller measurement differences in 52% of cases. On the other hand, re-training of the original detector for e′ lateral, for example, improved measurement accuracy to a large extent: there, the re-trained detector was superior in 81% of cases. Abbreviations as in Table 1.

*$P < 0.01$.

**$P < 0.001$.

(Scenario A; Figure 4), we observed a significant measurement difference for LVDd, LVDs, LVPWd, *E*-wave velocity, *A*-wave velocity, e′ lateral velocity, and e′ septal velocity, respectively. No significant differences emerged for IVS and LVOT diameter (Table 2 and Figure 5).

When compared with the original detector (Scenario A), the re-trained detector (*B*) arrived at significantly smaller mean differences in all but one parameter (IVS) with respect to the human referent (Scenario B; Figure 4 and Table 2). Specifically, all differences between

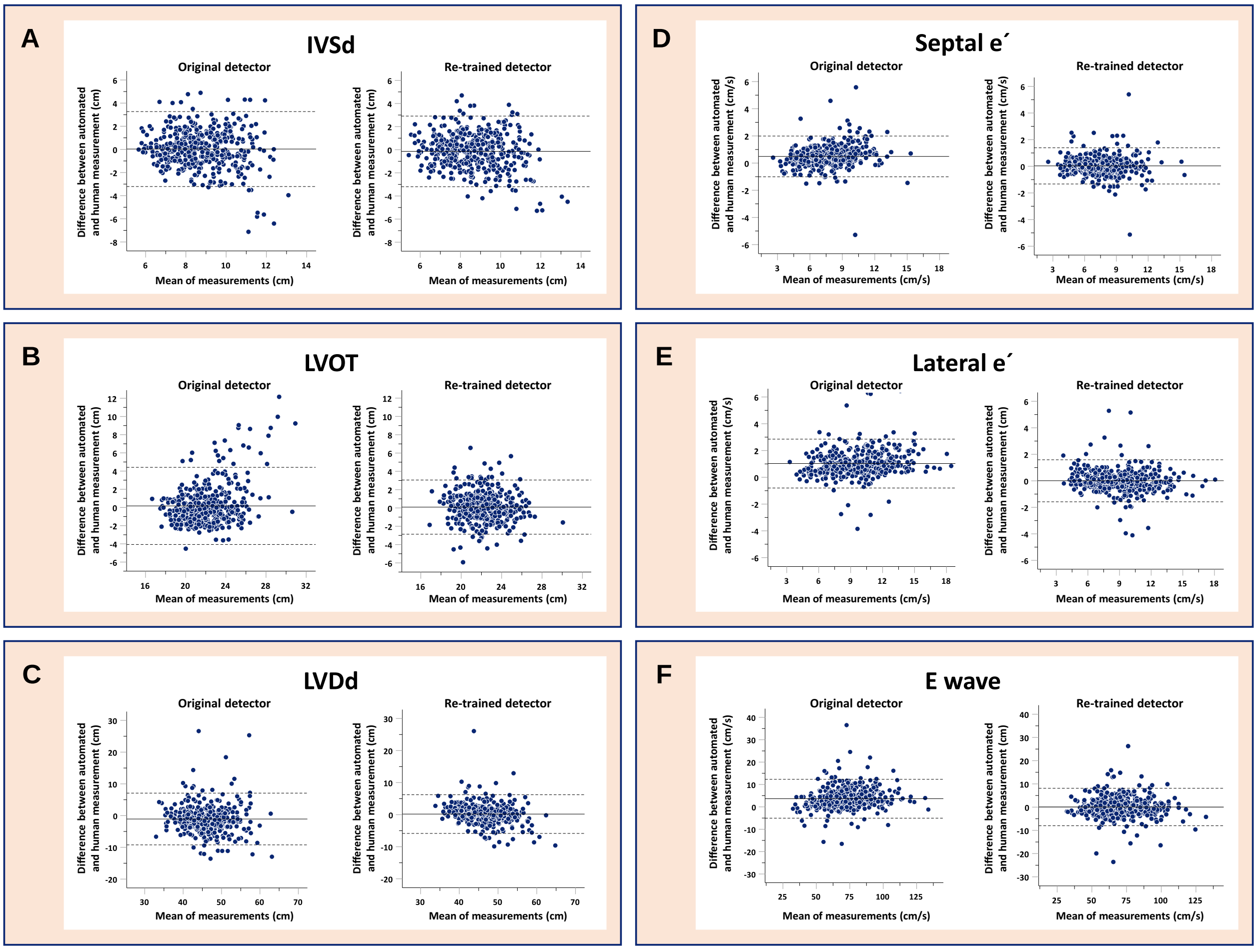


Figure 5 Bland–Altman plots displaying differences between measurements performed by the original and the re-trained detector, respectively, in comparison with the human referent (Scenario A vs. Scenario B) regarding (*A*) IVSd, (*B*) LVOT, (*C*) LVEDd, (*D*) septal e´, (*E*) lateral e´, and (*F*) E wave velocity. LoA, limits of agreement; LVDd, left ventricular end-diastolic diameter; IVSd, end-diastolic thickness of interventricular septum; LVOT, left ventricular outflow tract diameter; *E*, early mitral inflow velocity; *e*′, mitral annular diastolic velocity.

the re-trained detector and the human referent became non-significant, with the exception of LVDd and LVPWd, respectively. Further, the amount of absolute measurement differences was more likely to be smaller in Scenario B than in Scenario A ($\Delta B < \Delta A$) regarding all parameters except for IVS (*Table 2*).

Table 3 displays the 50th and the 95th percentiles of the absolute differences between measurements in different scenarios, respectively. The 50th and 95th percentiles of the absolute difference between the automated measurement performed by the re-trained detector and the human referent (Scenario B) for LVDd were 1.3 and 6.3 mm, respectively. This means that in 50% of cases, the absolute difference between the automated and the human reference measurement was ≤1.3 mm, and in 95% of cases ≤6.3 mm. To put these results into perspective, we also evaluated the absolute differences between measurements of different human observers and the human referent (Scenario C). For LVDd, the 50th and the 95th percentiles were 2.5 and 8.6 mm, respectively (*Table 3* and *Figure 6*). The measurement difference was significantly more often smaller in Scenario B when compared with Scenario C ($\Delta B < \Delta C$). The same was true for IVSd, LVOT diameter, *E*-wave, *A*-wave, *e*′ lateral, and *e*′ septal, respectively (*Table 3* and *Figure 6*).

Discussion

We implemented a federated ML system into the clinical environment providing training conditions similar to a traditional non-federated setup but ensuring that no personal data left the hospital network. This novel approach enables the privacy-compliant handling of sensitive patient data and a location-independent refinement of ML algorithms. Application of an automated detector to the standardized transthoracic echocardiograms of the population-based STAAB cohort revealed small but significant and clinically meaningful measurement differences for all but two of the selected parameters, when compared with the human referent. Machine-learning-based re-training of the detector resulted in significantly smaller differences between automated and human referent measurements, and the differences between automated measurements and human referent lost significance in all but two parameters. Hence, population data–based machine-learning algorithm can improve automated echocardiographic quantification of selected parameters of cardiac structure and function. When compared with the measurements of the human referent, the absolute difference between the measurements was smaller for the re-trained detector than for a group of different human observers implying that

Table 3 Absolute differences between measurements performed by the trained detector (B) as well as performed by other human observers (C) in comparison with the human referent, respectively

	Re-trained detector vs. human referent (Scenario B)			Human observers vs. human referent (Scenario C)			*P* for smaller measurement difference in Scenario B
	No. of measurements	Absolute difference between measurements		No. of measurements	Absolute difference between measurements		
		50th percentile	95th percentile		50th percentile	95th percentile	
LVDd [mm]	428	1.3	6.3	1325	2.5	8.6	<0.001
IVSd [mm]	427	0.9	3.1	1326	1.2	3.8	<0.001
LVPWd [mm]	427	1.0	3.3	1322	1.0	3.5	ns
LVOT diameter [mm]	440	0.8	3.3	1345	1.3	4.3	<0.001
E [cm/s]	494	2.0	8.0	1441	4.5	13.4	<0.001
A [cm/s]	489	1.8	7.8	1441	4.3	14.2	<0.001
e′ lateral [cm/s]	467	0.3	1.4	1456	1.1	3.5	<0.001
e′ septal [cm/s]	471	0.3	1.4	1441	0.8	2.7	<0.001

Abbreviations as in *Table 1*.

the automated measurements of the re-trained detector might be at least inter-changeable with human measurements.

The reliability of measurements is the key in the clinical and scientific application of echocardiography. Smaller measurement differences allow for more accurate diagnoses and for early and valid detection of changes over time, which both are essential in clinical care and clinical science. Major efforts including training and certification of sonographers and readers, standardization of measurements, alignment of system pre-sets, and acquisition protocols have been made and resulted in a reduction of measurement variability.[3,4,6–8] With regard to linear measurements, a relative difference between the measurements of two observers of 7 ± 2% could be achieved for LV mass,[21] for example, in NORRE, a large multi-centre study aiming to derive echocardiography reference values, while data on reproducibility of Doppler-derived parameters are scarce. Previous own data from the STAAB quality assurance program showed an inter-reader variability of ≤1.1 mm for IVS, ≤ 3.2 mm for LVDd, ≤ 1.7 mm for LV posterior wall, as well as ≤1.3 cm/s for lateral *e*′, and ≤0.2 m/s for *E*-wave velocity in 95% of cases, respectively.[3] As such low measurement variability was achieved under optimized study conditions, measurement variability likely is higher in clinical routine.

New echocardiography machines provide automated measurements, which can be manually adopted to speed up the reading process and reduce measurement variability. Nevertheless, until now, the perceived low accuracy in non-optimal images prevents broader application of automated measurement by clinical users. Hence, there is need for its further optimization in echocardiography.

To overcome the challenge of data privacy, we implemented a federated ML system achieving training conditions equivalent to those that would be expected using a common pool on a central site. This technology allows various (clinical) sites to build and refine ML models together without exchanging sensitive patient data.

Based on the WASE study[14] echocardiography images, an ML detector was trained to measure structural and functional cardiac parameters.[16] In a subset of WASE echocardiograms not used for training, the detector showed high accuracy when compared with manual measurements of an expert reader as well as an inter-observer variability comparable with two expert readers but with substantial reduction in the time required for reading.[16] World Alliance of Societies of Echocardiography Normal Values was performed to derive echocardiography reference values from a multi-nation multi-ethnicity cohort of healthy individuals covering a wide age range and aiming for a 1:1 stratification for sex.[14,15] Hence, the sample was appropriate to derive widely applicable reference values, but at the same time was highly selected and thus not representative of the general population.

When compared with WASE participants,[15] STAAB participants were about 10 years older and had higher body surface area as well as higher systolic and diastolic blood pressure. Further, a substantial number of participants had cardiovascular risk factors like hypertension, diabetes, and coronary heart disease. Hence, STAAB participants showed mean measures of LV structure within the reference ranges proposed by WASE,[15] but revealed lower mean values for diastolic functional parameters when compared with WASE.[22] In summary, we expected the STAAB cohort with more frequent abnormal cardiac structure and function than WASE, challenging the original detector and constituting a good basis for further optimization of the detector by ML-based re-training. Prediction models, by definition, optimally fit to the data they are derived from. Application of the prediction model—here the automated detector—to new, unknown data usually results in a lower degree of agreement but reflects the clinical reality to which the prediction model should be applied. For this reason, we split our large population-based sample into a training pool and two distinct validation pools. Stratification according to a random allocation algorithm resulted in three subgroups of comparable age and sex distribution, body composition, and comorbidity burden as well as of comparable cardiac structure and function.

Application of the original detector to STAAB echocardiograms showed good performance regarding the selected measurement parameters in the range of previous reports,[9,13] but for most parameters, a significant and clinically relevant measurement difference remained when compared with the human referent. Indeed, these differences were larger than the measurement differences reported from the application of the detector to the WASE validation sample.[16]

Based on >14 000 echocardiograms from routine care, Zhang *et al.* used CNN models to train automated segmentation of cardiac chambers and determination of cardiac structure and function measurements. Comparison with the study report values revealed median absolute deviations of 15–17% for LV mass and LV end-diastolic

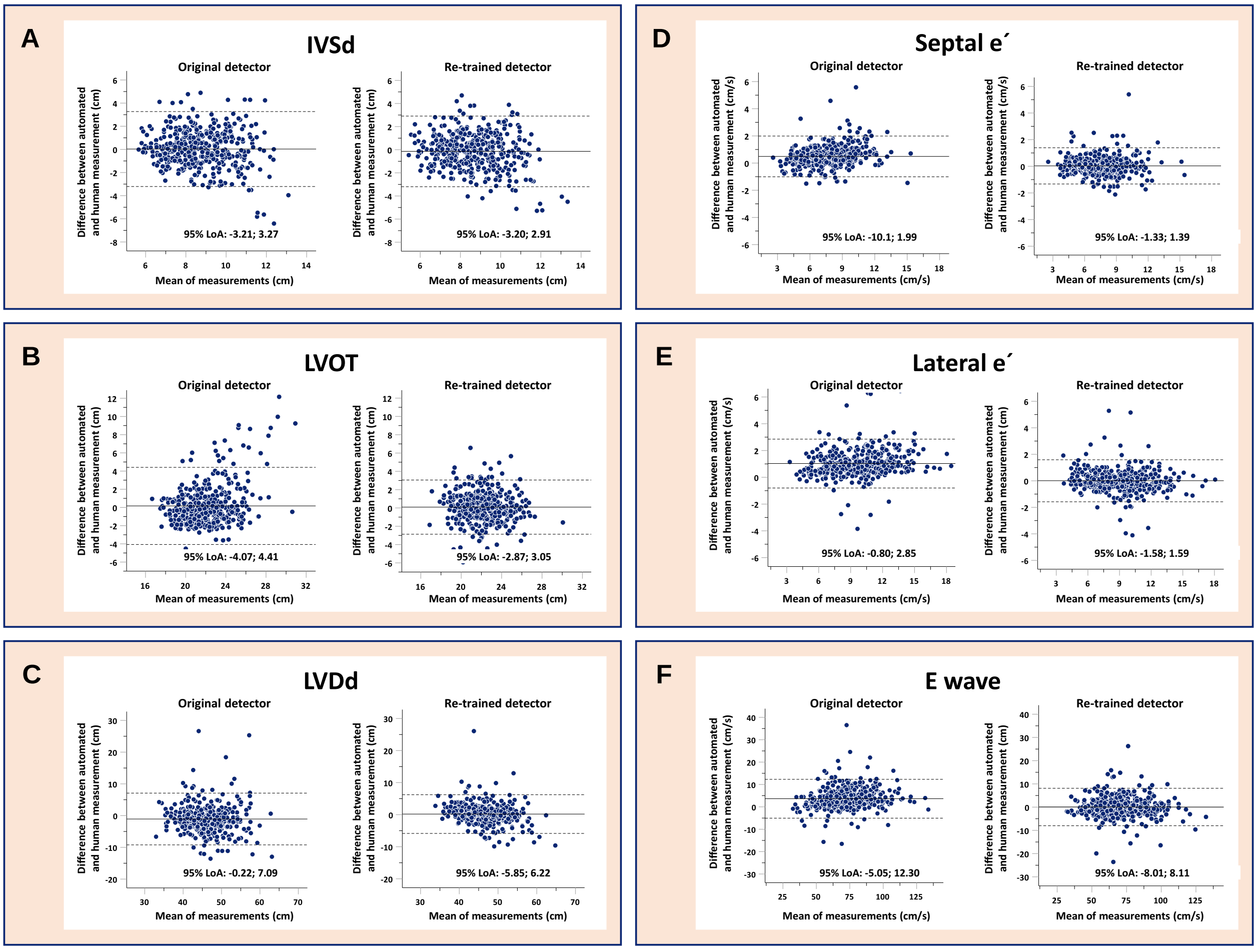


Figure 6 Percentiles of the absolute difference between the measurements performed by the re-trained detector and human observers, respectively, in comparison with the human referent (Scenario B vs. Scenario C) regarding (*A*) IVSd, (*B*) LVOT, (*C*) LVEDd, (*D*) septal e´, (*E*) lateral e´, and (*F*) E wave velocity. LVDd, left ventricular end-diastolic diameter; IVSd, end-diastolic thickness of interventricular septum; LVOT, left ventricular outflow tract diameter; *E*, early mitral inflow velocity; *e*´, mitral annular diastolic velocity.

volume as well as for left atrial volume. The respective comparison with commercial software-derived values showed a median deviation of 9.7 and 7.5% for LVEF and LV strain.[9] Although the median values were convincing, there was a substantial number of outliers with large deviation preventing unsupervised independent use in a clinical setting, so far.[9]

EchoNET-Dynamic[23] published a video-based artificial intelligence for beat-to-beat assessment of cardiac function. Their software predicted LVEF with a mean absolute error of 4.1% in their own data set while application to an external data set resulted in a mean absolute error of 6.0%. Prospective evaluation with repeated human measurements confirmed that the model had variance comparable with or less than that of human experts, which is in line with the general finding of our analyses, as well. The EchoNET-LVH[24] deep-learning algorithm measured LV wall thickness and diameter in patients with increased wall thickness and achieved performance measures in the range of experienced human readers. In our study, re-training of the original detector based on the echocardiography images of the STAAB training pool resulted in significantly lower differences between the automated measurements and the measurements performed by the human referent for all but one parameter and measurement differences lost statistical significance. Further, when compared with the measurements of different human observers, differences between the automated measurements and the human referent were significantly smaller in all but one parameter. These results suggest that the automated measurements might be interchangeable with human measurements when applied to the general population.

Interpreting our results, we have to consider several limitations. The set of echocardiography parameters on which the detector has been re-trained is incomplete. We first concentrated on linear and Doppler-derived measures. In the next step, we will extend the automated measurements to volumes and deformation parameters as well. Further, the STAAB cohort is a representative sample of the general population of Würzburg, a city with predominantly Caucasian inhabitants; hence, the performance of the re-trained detector might be different in other populations. Finally, a relevant proportion of STAAB participants exhibited cardiovascular risk factors, but only a small proportion had overt cardiac diseases. In a future step, the detector will need to be re-trained in patient populations to validly measure impaired cardiac structure and function, too. Nevertheless, our results are based on strong methodology using standardized echocardiograms of a population-based cohort for re-training of the detector, and the results were consistent throughout the selected echocardiography parameters.

Conclusions

The implemented federated ML set-up was feasible and population data–based ML improved the echocardiography detector with regard to the automated determination of the selected parameters. Performance measures suggest that the automated measurements might at least be interchangeable with human measurements when applied to the general population.

Lead author biography

Caroline Morbach currently serves as cardiologist and echocardiography specialist at the University Hospital Würzburg, Germany. She founded the Academic Core Lab Ultrasound-based Cardiovascular Imaging at the Comprehensive Heart Failure Center Würzburg and established internal training and certification algorithms to ensure high-quality standards in clinical and scientific echocardiography. She is a fellow of the European Association of Cardiovascular Imaging (EACVI); she is a certified specialist in transthoracic and transoesophageal echocardiography, both from the EACVI and from the German ultrasound association (DEGUM). She further also has long-term experience in clinical research and care for patients with advanced heart failure and patients with cardiac amyloidosis.

Clinical perspectives

It has been demonstrated that the use of an ML-based reading algorithm markedly enhanced workflow efficiency of echocardiographic interpretation and improved the inter-reader variability of common measurements.[16] We applied this original ML-based algorithm to >3000 echocardiography studies of a population-based cohort and re-trained the algorithm. The improved version yielded superior performance and exhibited a much lower measurement variability than human readers. This gain in accuracy and precision strengthens the confidence in automated echocardiographic readings, which carries large potential for applications in various settings, including, but not refined to (i) clinical studies: using the re-trained detector is expected to result in smaller sample sizes required to identify group differences or changes over time; (ii) routine setting: faster reading with higher accuracy is expected to reduce costs and improve the level of certainty when changes over time have to be judged (monitoring) or when treatment options depend on measurement thresholds (e.g. implantable defibrillator); (iii) point-of-care echocardiography performed by non-experts in the emergency care setting or in the general practitioner's office, for example, may aid the confirmation or exclusion of defined cardiac pathologies and thus accelerate patient triage.

This project proves that it is possible to train and improve automatic cardiac measurements on ultrasound images across two clinical research institutions without exchanging sensitive patient data. In times of increasing regulatory and privacy requirements, the federated learning technology offers the possibility to better evaluate systemic differences or discrepancies among larger groups of clinical observers and, through the more robust ML models, could provide more realistic standard values for cardiac measurements and correction factors for different observers and clinical sites.

Funding

This work was supported by the German Federal Ministry of Education and Research by grants to the Comprehensive Heart Failure Centre Würzburg (BMBF 01EO1004 and 01EO1504). The AVE project was a cooperation project between the University Hospital Würzburg and the University of Würzburg and TOMTEC Imaging Systems GmbH, Unterschleißheim, Germany. C.M., G.G., and M.S. received funding from the Bavarian Ministry of Economic Affairs, Regional Development and Energy, Germany (Nr. 07 03/686 68/203/19/28/20/29/21), respectively. C.M. receives financial support from the Interdisciplinary Center for Clinical Research IZKF Würzburg (advanced clinician-scientist programme; IZKF, Project No. Z-2/CSP-3) as well as from the German Research Foundation (DFG) within the Comprehensive Research Center 1525 'Cardio-immune interfaces' (453989101, Project C5).

Conflict of interest: C.M. reports research cooperation with the University of Würzburg and TOMTEC Imaging Systems GmbH funded by a research grant from the Bavarian Ministry of Economic Affairs, Regional Development and Energy, Germany. G.G. reports research cooperation with the University Hospital of Würzburg and TOMTEC Imaging Systems GmbH funded by a research grant from the Bavarian Ministry of Economic Affairs, Regional Development and Energy, Germany. M.S. is a full-time employee of TOMTEC imaging systems. He further reports research cooperation with the University and University Hospital of Würzburg funded by a research grant from the Bavarian Ministry of Economic Affairs, Regional Development and Energy, Germany. N.H., H.W., and D.S. are full-time employees of TOMTEC imaging systems. H.W. and S.S. are supported by the CHFC Würzburg and by the German Federal Ministry of Education and Research (BMBF). All other authors report no conflict of interest.

Data availability

The data underlying this article will be shared on reasonable request with the corresponding author.

Consent

The STAAB cohort study protocol and procedures complied with the Declaration of Helsinki and received positive votes from the Ethics Committee of the Medical Faculty (Vote #98/13) as well as from the data protection officer of the University of Würzburg. All participants provided written informed consent prior to any study examination.

References

1. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Bohm M, *et al.* 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* 2021;**42**:3599–3726.
2. Zamorano JL, Lancellotti P, Rodriguez Munoz D, Aboyans V, Asteggiano R, Galderisi M, *et al.* 2016 ESC position paper on cancer treatments and cardiovascular toxicity developed under the auspices of the ESC committee for practice guidelines: the task force for cancer treatments and cardiovascular toxicity of the European Society of Cardiology (ESC). *Eur Heart J* 2016;**37**:2768–2801.
3. Morbach C, Gelbrich G, Breunig M, Tiffe T, Wagner M, Heuschmann PU, *et al.* Impact of acquisition and interpretation on total inter-observer variability in echocardiography: results from the quality assurance program of the STAAB cohort study. *Int J Cardiovasc Imaging* 2018;**34**:1057–1065.
4. Thorstensen A, Dalen H, Amundsen BH, Aase SA, Stoylen A. Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study. *Eur J Echocardiogr* 2010;**11**:149–156.
5. Letnes JM, Eriksen-Volnes T, Nes B, Wisløff U, Salvesen Ø, Dalen H. Variability of echocardiographic measures of left ventricular diastolic function. The HUNT study. *Echocardiography* 2021;**38**:901–908.
6. Picard MH, Adams D, Bierig SM, Dent JM, Douglas PS, Gillam LD, *et al.* American Society of Echocardiography recommendations for quality echocardiography laboratory operations. *J Am Soc Echocardiogr* 2011;**24**:1–10.

7. Douglas PS, Waugh RA, Bloomfield G, Dunn G, Davis L, Hahn RT, *et al.* Implementation of echocardiography core laboratory best practices: a case study of the PARTNER I trial. *J Am Soc Echocardiogr* 2013;**26**:348–358.e3.
8. Kataoka A, Scherrer-Crosbie M, Senior R, Gosselin G, Phaneuf D, Guzman G, *et al.* The value of core lab stress echocardiography interpretations: observations from the ISCHEMIA trial. *Cardiovasc Ultrasound* 2015;**13**:47.
9. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, *et al.* Fully automated echocardiogram interpretation in clinical practice. *Circulation* 2018;**138**: 1623–1635.
10. Gandhi S, Mosleh W, Shen J, Chow CM. Automation, machine learning, and artificial intelligence in echocardiography: a brave new world. *Echocardiography* 2018;**35**: 1402–1418.
11. Knackstedt C, Bekkers SCAM, Schummers G, Schreckenberg M, Muraru D, Badano LP, *et al.* Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. *J Am Coll Cardiol* 2015;**66**: 1456–1466.
12. Nabi W, Bansal A, Xu B. Applications of artificial intelligence and machine learning approaches in echocardiography. *Echocardiography* 2021;**38**:982–992.
13. Tromp J, Bauer D, Claggett BL, Frost M, Iversen MB, Prasad N, *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat Commun* 2022;**13**:6776.
14. Asch FM, Banchs J, Price R, Rigolin V, Thomas JD, Weissman NJ, *et al.* Need for a global definition of normative echo values-rationale and design of the World Alliance of Societies of Echocardiography Normal Values study (WASE). *J Am Soc Echocardiogr* 2019;**32**:157–162.e2.
15. Asch FM, Miyoshi T, Addetia K, Citro R, Daimon M, Desale S, *et al.* Similarities and differences in left ventricular size and function among races and nationalities: results of the World Alliance Societies of Echocardiography Normal Values study. *J Am Soc Echocardiogr* 2019;**32**:1396–1406.e2.
16. Lang RM, Addetia K, Miyoshi T, Kebed K, Blitz A, Schreckenberg M, *et al.* Use of machine learning to improve echocardiographic image interpretation workflow: a disruptive paradigm change? *J Am Soc Echocardiogr* 2021;**34**:443–445.
17. Wagner M, Tiffe T, Morbach C, Gelbrich G, Störk S, Heuschmann PU, *et al.* Characteristics and Course of Heart Failure Stages A-B and Determinants of Progression—design and rationale of the STAAB cohort study. *Eur J Prev Cardiol* 2017;**24**:468–479.
18. Morbach C, Gelbrich G, Tiffe T, Eichner FA, Christa M, Mattern R, *et al.* Prevalence and determinants of the precursor stages of heart failure: results from the population-based STAAB cohort study. *Eur J Prev Cardiol* 2021;**28**:924–934.
19. Morbach C, Sahiti F, Tiffe T, Cejka V, Eichner FA, Gelbrich G, *et al.* Myocardial work—correlation patterns and reference values from the population-based STAAB cohort study. *PLoS One* 2020;**15**:e0239684.
20. Morbach C, Walter BN, Breunig M, Liu D, Tiffe T, Wagner M, *et al.* Speckle tracking derived reference values of myocardial deformation and impact of cardiovascular risk factors—results from the population-based STAAB cohort study. *PLoS One* 2019;**14**: e0221888.
21. Kou S, Caballero L, Dulgheru R, Voilliot D, De Sousa C, Kacharava G, *et al.* Echocardiographic reference ranges for normal cardiac chamber size: results from the NORRE study. *Eur Heart J Cardiovasc Imaging* 2014;**15**:680–690.
22. Miyoshi T, Addetia K, Citro R, Daimon M, Desale S, Fajardo PG, *et al.* Left ventricular diastolic function in healthy adult individuals: results of the World Alliance Societies of Echocardiography Normal Values study. *J Am Soc Echocardiogr* 2020;**33**:1223–1233.
23. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;**580**:252–256.
24. Duffy G, Cheng PP, Yuan N, He B, Kwan AC, Shun-Shin MJ, *et al.* High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol* 2022;**7**:386–395.