scientific reports

Check for updates

OPEN

Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease

Akis Linardos[1]✉, Kaisar Kushibar[1], Sean Walsh[2], Polyxeni Gkontra[1] & Karim Lekadir[1]

Deep learning models can enable accurate and efficient disease diagnosis, but have thus far been hampered by the data scarcity present in the medical world. Automated diagnosis studies have been constrained by underpowered single-center datasets, and although some results have shown promise, their generalizability to other institutions remains questionable as the data heterogeneity between institutions is not taken into account. By allowing models to be trained in a distributed manner that preserves patients' privacy, federated learning promises to alleviate these issues, by enabling diligent multi-center studies. We present the first simulated federated learning study on the modality of cardiovascular magnetic resonance and use four centers derived from subsets of the M&M and ACDC datasets, focusing on the diagnosis of hypertrophic cardiomyopathy. We adapt a 3D-CNN network pretrained on action recognition and explore two different ways of incorporating shape prior information to the model, and four different data augmentation set-ups, systematically analyzing their impact on the different collaborative learning choices. We show that despite the small size of data (180 subjects derived from four centers), the privacy preserving federated learning achieves promising results that are competitive with traditional centralized learning. We further find that federatively trained models exhibit increased robustness and are more sensitive to domain shift effects.

Diagnostic tools based on artificial intelligence models have shown promising results in a variety of single-center studies across multiple medical imaging domains[1], but their generalizability to unseen distributions remains understudied, and their application in clinical practice is still far from realized. As data remains segregated in different institutions, such studies have mostly focused on limited single-center datasets for their training and evaluation. Aside from the obvious issue of having a small sample size, evaluation in such a set-up is questionable, as no assumptions can be made on how this performance translates to unseen centers. For ML methods to generalize to unseen datasets, it is often assumed that newly seen data is independent and identically distributed (IID) to the one seen during training—i.e. each data point comes from the same probability distribution and is mutually independent to all others. For this reason, the data heterogeneity present in multi-center medical data poses a significant problem, as in all cases such data is non-IID—a direct result of the usage of different acquisition protocols, different scanners and varying demographics.

To overcome the core obstacle of data scarcity and to better understand the effects of data heterogeneity that is present in between different centers, institutions need to come together in collaboration. This has thus far been difficult, as institutions are inclined to keep a tight grip on their medical data due to privacy regulations (e.g. GDPR in the European Union and HIPAA in the United States). While an obvious approach for collaborators would be to share their data on a central server (CDS, Fig. 1A), this endangers patients' privacy by increasing the chance of data leakage. Distributed learning allows for AI models to be trained across multiple edge devices or centers, without data ever leaving its original place[2,3]. In 2017, Google proposed Federated learning[4] (FL, Fig. 1B) a framework that allows deep learning models to be distributed and trained on local data, aggregating only their parameters in a central server. The central server only ever sees a complex representation of the initial data, as learned by the local models, and those representations are combined by an algorithm called *FederatedAveraging* before being redistributed for subsequent training.

Privacy preserving federated learning systems for medical image analysis have been mainly explored in the context of segmentation for brain[5–8], prostate[9] and COVID-19 affected regions[10–12]. Segmentation is still an open problem and faces multiple challenges, especially at the data collection stage as expert manual annotation is time-consuming and exhibits inter- and intra-operator variability in the segmentation masks. However, diagnosis is

[1]Department of Mathematics and Computer Science, University of Barcelona, 08007 Barcelona, Spain. [2]Radiomics, 4000 Liège, Belgium. ✉email: linardos.akis@ub.edu

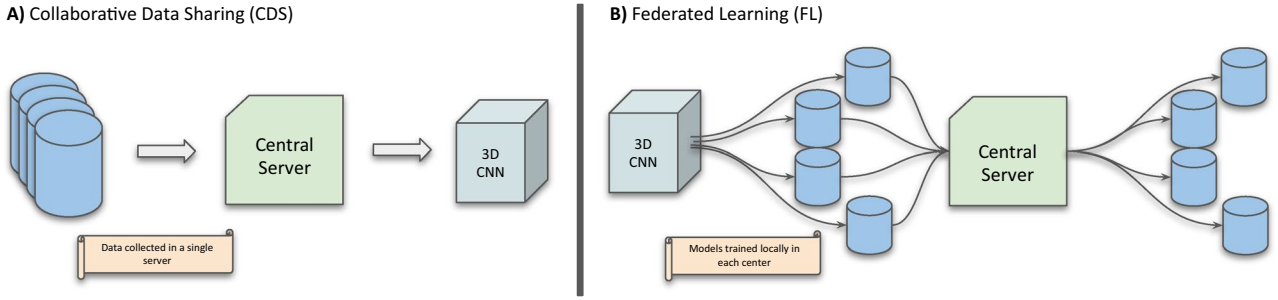

Figure 1. A diagram of the two collaborative learning frameworks. For FL only models are transferred, while in the CDS case, the data itself is transferred to the central server and all training happens there.

a less explored topic, and segmentation steps are often a prerequisite for diagnosis models to be accurate. The data scarcity problem is even more prominent in this case, as the ground truth information for the presence of disease is of a much more sensitive nature, and clinical registries are lacking. Label imbalance and demographic variabilities become more relevant as well, as different institutions usually contain different types of diseases, while in segmentation the expected ground truth is the same in all cases—i.e. a mask of the segmented parts. From a data science standpoint, segmentation allows for more flexible methods of data augmentation (even GAN-based generation), whereas, in diagnosis tasks, augmentation should be done with extreme care to avoid shifting the true value of corresponding labels. This often requires human-expert validation depending on the augmentation method, lest it risks adding noise. Segmentation also has the advantage of freely using 2D slices and patch-based approaches, which allows one to extract many training samples from a single scan, while in diagnosis, a single 3D volume (or sometimes series of longitudinal data) is used as a single training data point.

For these reasons, related research in domains other than segmentation has been more limited, with studies tangential to diagnosis popping up only as early as 2020 in breast density classification[13], and lung tumor survival prediction[14]. With the emergence of the COVID-19 pandemic, the urgent need for diagnostic tools circumvented common obstacles and allowed researchers to collaborate in federated learning diagnosis for the first time[15,16]. Very recently, an open-source framework that integrates FL along with the functionality of end-to-end encryption to protect against inversion attacks was developed, trained and tested on paediatric X-ray classification[17].

In this work, we focus on the diagnosis of cardiovascular disease (CVD) based on Cardiac MRI. The importance of furthering our understanding on the structure and function of the heart is highlighted by the prevalence of CVD in the population, which accounts for one third of annual deaths[18,19]. Cardiac MRI has been the go-to modality for this task, allowing the assessment and delineation of the three heart segments—i.e. the myocardium, and the left and right ventricle blood pools—to identify the presence of anomalies such as myocardic infarctions or cardiomyopathies. Based on this modality, and by leveraging hand crafted features, Machine Learning (ML) diagnostic tools have been developed with some success in single-center datasets[20,21].

To enable AI studies in CMR diagnosis, there have been public challenges for cardiac MRI segmentation and diagnosis such as "Automatic Cardiac Diagnosis Challenge" dataset (ACDC) and the "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation Challenge" dataset (M&M). Despite it's name, M&M now also includes diagnostic labels. We will be using subsets from these datasets in our study and provide more details on the data acquisition and annotation in "Dataset" section.

In terms of the Cardiac MRI modality, there has been a lot of literature on deep learning-based segmentation[22–30] with diagnosis typically being a follow-up step, leveraging the segmentation masks and utilizing models such as random forests[31,32], support vector machines[33] or a simple diagnostic rule[34]. These diagnosis models focus on two timepoints of the cardiac MRI per-patient: the phase of End-Diastole (ED) (maximum heart relaxation) and the phase of End-Systole (ES) (maximum heart contraction). Current such studies have emphasized on the aforementioned ACDC dataset[35], a single center dataset hosting 100 subjects and five labels (20 subjects each).

Khened et al.[31] reported promising results on multi-label diagnosis but evaluated on a limited hold-out test set of 10 samples. On the same task, Cetin et al.[33] used an SVM on top of radiomic features derived from manual segmentations, while Wolterink et al.[32] evaluated a random forest classifier instead, both using a cross validation scheme. Liu et al.[34] used an automated deep-learning based segmentation scheme, following up with a diagnostic rule. Despite the impressive classification performance reported in these studies, these models are both trained and evaluated solely on a single-center dataset (ACDC), and thus, no assumptions can be made regarding their generalization to unseen centers and larger datasets. To deploy such models in the real world, one has to assume that new subjects being tested are IID to those seen during training and evaluation. In the domain of medical imaging, where data is highly heterogeneous between centers, this assumption is far from true[36]. Furthermore, as these studies focus on a single center, no privacy preservation measures are studied, which are otherwise necessary for deployment of such models.

Despite the widespread interest in automated CMR diagnosis methods, multi-centric and distributed learning studies in the field are currently lacking. In this paper, we conduct our study with four centers, three of which were derived from the M&M dataset[37], and the fourth being a subset of ACDC[35]. We test the CDS and FL collaborative learning frameworks. As FL trains local models in each center, multi-label classification is a very challenging problem in the case of the M&M dataset where many labels have little to no overlap between the centers,

Center	Vendor	Spatial resolution (mm)	Slice thickness (mm^2)	NOR	HCM	Total
Vall d'Hebron	Philips	1.1516–1.2362	10.0	21	25	46
Sagrada Familia	Siemens	0.9765–1.6200	8.0–10.0	33	37	70
SantPau	Canon	0.7955–1.8228	10.0	14	10	24
ACDC	Siemens	1.3400–1.6800	5.0-10.0	20	20	40
Total				88	92	180

Table 1. Dataset description, including meta-data and the class distribution of Normal (NOR) and hyperthropic cardiomyopathy (HCM) for the selected subset from the M&M and ACDC datasets used in this study.

thus causing local models to overfit on different tasks. For this reason, we focus only on diagnosing hypertrophic cardiomyopathy (HCM)—i.e. a binary classification between normal (NOR) subjects and subjects suffering from HCM. HCM is the most common heritable cardiomyopathy, occurring in approximately as 0.29%—i.e. 1:344—of the adult population[38,39]. Contrary to previous work on cardiac MRI diagnosis, our main goal here is to test the feasibility of cardiac MRI diagnosis in between multiple centers and highlight the importance of evaluating both IID and non-IID performance (i.e. testing models on partitions of the centers seen during the training and also on unseen centers) in a principled manner.

Previous work in federated learning diagnosis on COVID-19[15,16] and paediatric X-ray classification[17] has focused on the development of state of the art federated learning frameworks—the latter one open-sourcing their pipeline which also integrates an encryption mechanism. In this study, we focus on the effects multi-center data has on this frameworks, conducting a systematic comparative analysis between the CDS and FL paradigms, testing a variety of data curation and augmentation techniques. By evaluating in two distinct cross-validation set-ups and repeating the experiments multiple times, we gain a robust estimate of both IID and non-IID performance, showcasing the gap between the two. Our model follows the notoriously hard to train 3D-CNN architecture[40], leveraging transfer learning by using an instance of the network that has been pretrained on action recognition. Concretely, our contributions can be summarized as follows:

- We present, to the best of our knowledge, the first federated learning study on CMR diagnosis and demonstrate that FL performance is comparable to CDS, while preserving patient privacy.
- We propose a technique of inducing different priors to the model by leveraging the ground truth masks, illustrating an effective way to constrain the solution space and improve performance for deep learning-based multi-center CMR diagnosis in both collaborative learning set-ups.
- We apply a diverse set of data augmentations to artificially increase the data size and study their effect in a principled way on the collaborative learning frameworks, repeating the experiments with different CNN weight initializations to gain an estimate of model robustness[41]. We also test a variation of the FL algorithm in this context, which assigns an equal vote to all centers in the training data and show that it is beneficial in some cases.
- Finally, by using two distinct repeated cross validation set-ups—one that uses a part of all centers as test set per fold and another that uses a whole center as test set per fold—we get an estimate of both on-site and out-of-site performance, showing that the two differ substantially and highlighting their importance for future diagnosis studies.
- To boost future research in the field, we make our code available for the research community.

Methods

Dataset. We base our experiments on a unique dataset derived from a combination of M&M (a multi-centric dataset gathered in a coordinated effort under the EuCanShare project[37]) and ACDC (a single center dataset presented as a challenge in 2018[35]). Both datasets are composed of T1-weighted Cardiac Cine MRI sequences. M&M consists of 6 centers and 4 labels corresponding to dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (RV) and subjects without cardiac disease (NOR)[37].

Because of severe label imbalance between centers, training a federated model on multi-label classification becomes a complex problem as local models overfit to a subset of the labels or even a single label (for example one center only has DCM cases). In our experiments, we consider the task of binary classification (HCM vs. NOR), using a subset of M&M in which the chosen labels are most balanced. Thus, we use 3 centers from the M&M dataset, namely Sagrada Familia, Vall d'Hebron and SantPau and complement them with a subset of the ACDC dataset as a 4th center. The final form of the dataset we used is outlined in Table 1.

For the M&M challenge, all patients signed the informed consent, the study protocol was approved by the Ethical Committee for Clinical Research for each institution involved, and it follows the ethical guidelines of the Declaration of Helsinki. The authors of the ACDC challenge have also received approval by their ethical committees to make it publicly available. As these are both public datasets, no direct approval from the ethical committee was necessary on our side.

Diagnosis rules. On the ACDC dataset, an HCM label was allocated if the patient's left ventricular cardiac mass exceeds 110 g/m^2 For the M&M centers, the diagnosis of HCM was allocated when left ventricular wall thick-

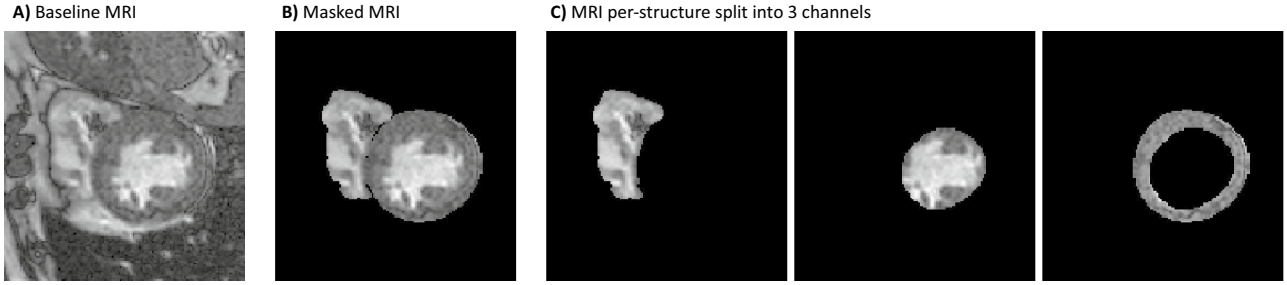

Figure 2. An example of the induced priors used in this study: (**A**) a baseline that is the 1-channel cardiac MRI, (**B**) the baseline multiplied by the segmentation mask, (**C**) the baseline split into three channels, one for each part of the heart (right ventricle blood pool, left ventricle blood pool, and myocardium).

ness exceeds 15 mm, given that this observation was otherwise unexplained by abnormal loading conditions (e.g., hypertension, valvular, congenital disease) or infiltrative cardiomyopathies.

Segmentation rules. ACDC's contours were manually drawn by experts on 3D volumes of the LV and RV cavities and the myocardium at the timepoints of interest–i.e. ED and ES. Annotation adhered to the following rules: LV and RV are completely covered. The papillary muscle is included into the cavity, and the contours follow the limit defined by the aortic valve. The RV cavity and the root of the pulmonary artery are clearly separated. RV is defined as the region on the right of heart with a significant contraction between ventricular diastole and systole. The data-providers also supply clear illustrations of the annotation rules in their work[35]. M&M segmentation[37] was built on top of ACDC's Standard Operating Procedure (SOP). The contours were corrected based on the consensus of two in-house annotators according to the following rules: a. LV and RV cavities must be completely covered, with papillary muscles included. b. No interpolation of the LV myocardium must be performed at the base. c. RV must have a larger surface in end-diastole compared to end-systole and avoid the pulmonary artery.

Data preprocessing. As a first step, we apply N4 bias field correction to remove non-uniformity of low frequencies inherent to MRI[42]. Then we resample the volumes to a common spacing of 1×1×1 mm^3 on the entirety of the data used.

The original spacing of the images vary, but all centers trim their original spacing, with the only exception of SantPau that in many cases interpolates its spatial resolution to reach a 1×1×1 mm^3 spacing. We crop the volumes using a 150×150×10 voxel window, centered on the center of the bounding box of non-zero values on their corresponding segmentation masks.

Induced priors. Leveraging the segmentation masks that are provided by clinicians as part of the M&M and ACDC datasets, we consider the following set-ups for our dataset: (1) a baseline that is the 1-channel MRI; (2) the MRI image multiplied by the segmentation mask (thus inducing a shape prior); and (3) the MRI image split into three channels, corresponding to left ventricle blood pool, right ventricle blood pool, and myocardium, inducing a strong prior of the heart structure (see Fig. 2). Notably, the initial layer of ResNet3D, was pretrained on action recognition RGB images and thus expects a 3-channel input. For this reason, the baseline and masked MRIs are copied three times before being fed to the model. Reinitializing the initial layer to a 1-channel input results in substantially worse performance, which, for brevity, we don't report in this paper. Two timepoints are extracted from the 3D volumes corresponding to the ED and ES phases. The two timepoints are fed as separate samples to the network so as to leverage more data; however, the two timepoints are always in the same set (train, validation, test).

Data harmonization. A main obstacle when learning from multi-center MRI data is the inter-site variability, which is present even when the same acquisition protocol is used[43,44]. To adapt the image intensities on the same scale, we apply Nyúl histogram matching[45] followed by a rescaling to [0, 1]. Concretely, we match the histogram of each image to the average histogram of all images from the training centers. As averaging is an aggregate query to the local data, it does not require access to individual subject data and thus preserves privacy. The final average histogram $H_{average}$ used for the histogram matching process is calculated as:

$$H_{average} = \sum_{k=1}^{K} \frac{N_k}{N} \sum_{n=1}^{N_k} H_n^k,$$

where K is the total number of centers, N_k the number of samples for center k, and N the total sample size of all centers.

In the majority of our experiments, we crop our inputs with segmentation masks, and thus only the values within the mask are used for the histogram matching as well. As we use two distinct evaluation set-ups, in some cases not all centers are calculated in the average. This is further clarified in "Evaluation Procedure" section where these set-ups are also explained.

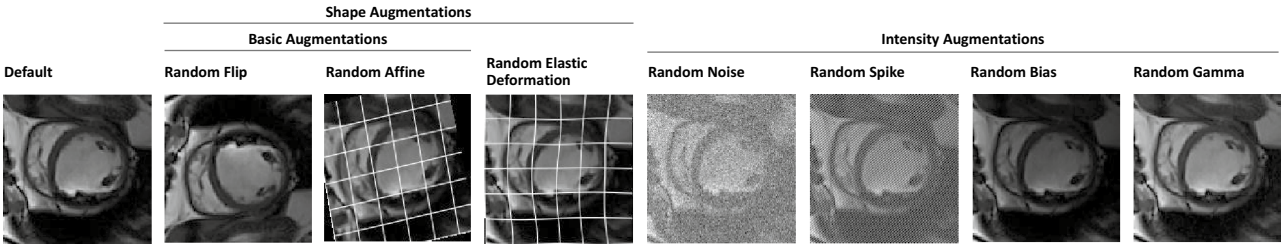

Figure 3. Illustration of augmentation techniques used in our analysis on a single CMR image slice.

Data augmentation. Due to the limited number of samples in the M&M dataset, we test several data augmentation techniques to artificially increase the size of the training set. These techniques also apply domain shift effects and make the representations learned more invariant to certain features, which can improve generalization to unseen datasets[46]. To test how this dataset shift affects our framework, we compare four different augmentation set-ups and repeat the experiments both for FL and CDS. These set-ups are as follows:

- No Augmentations.
- Basic Augmentations: affine transformations (rotation by varying degrees), horizontal and vertical flipping in axial view.
- Shape Augmentations: includes the basic augmentations plus elastic deformation.
- Shape and Intensity Augmentations: includes all aforementioned augmentations, plus random MRI spike artifacts, random MRI bias field artifact, noise sampled from a Gaussian with $\mu = 0$ and $\sigma \sim (0, 0.25)$, random gamma transformations (randomly changes contrast of an image by raising its values to a random power within a specified range).

During training on one of these set-ups (except the No Augmentations set-up), an augmentation is sampled from the aforementioned pools and is applied on the input data with a probability of 50%. That means that every image has a 50% chance of being augmented. These types of data augmentations are known to be effective in improving a model's generalization as they introduce domain shift effects. However, the chosen augmentations are also reflective of clinical reality as we avoid extreme cases of elastic deformation and consider the random noise augmentation as representative of bad acquisitions, which often escape quality control. All augmentations used and studied in this work were obtained using the TorchIO library[47] and are illustrated in Fig. 3.

3D-CNN model. 3D-CNNs have the potential to retrace the success story of 2D-CNNs; however, their immense size is cause of two major drawbacks—i.e. a high computational cost and the curse of dimensionality which causes them to overfit[40]. In medical imaging, where data is very scarce, 3D-CNNs have been used successfully by applying transfer learning schemes, utilizing data from entirely different domains[48]. In our case, we use the 3D-CNN ResNet18 model as defined by Tran et al.[49] and, instead of initializing the weights randomly, we load an instance of the model that has been pretrained on the action recognition dataset Kinetics-400[50]. This is beneficial because the early layers of the network tend to extract similar features (such as edges or blobs) irrespective of the domain that are beneficial to all imaging tasks[51]. To constrain the model from overfitting, we freeze the initial layers and train a newly initialized linear layer with a sigmoid activation function to the task of binary classification of HCM vs NOR. Concretely, out of the 33,166,785 parameters our network has in total, only the 512 parameters of the final linear layer are trained in this case. In our preliminary experiments, 256, 512, 1024 and 2048 channels on the linear layer were tested and 512 was found to perform best.

This, however, introduces a second problem: a strong bias on the model for a completely different domain and data modality. To find a proper trade-off between the issue of bias and overfitting, we experimented by fine tuning the pre-trained layers with varying learning rates (1e-4, 1e-5, 1e-6, 1e-7 and 0) while the newly initialized linear layer was always trained with a much higher learning rate of 0.01. Our preliminary experiments showed that for the pre-trained layers, a learning rate of 1e-5 performed best, and thus we used this configuration for all subsequent experiments. The models are trained for a maximum of 100 epochs, stopping early if validation performance stagnates for 10 epochs. In our experiments, the early stopping occurred on the 20–30th epoch.

Federated learning. We first initialize a global model and then distribute it across the four centers. It is fundamental that on every step, including the initialization the model being distributed is identical, otherwise the aggregation across models will result in a non-sense representation and training will not progress[4]. The models are trained for seven iterations—i.e. on seven batches of data—on each center and are aggregated after each epoch. To parse the entire data within the same number of iterations on each center, a different batch size is used.

After training locally, the models are aggregated using the FederatedAveraging algorithm[4]. Concretely, each model makes an update step in its respective center k using a learning rate η and the gradients g_k so that

$$w_{t+1}^k \longleftarrow w_t - \eta g_k, \forall k$$

Then, these weights are aggregated to the global model in a way that is proportionate to the sample size of each center,

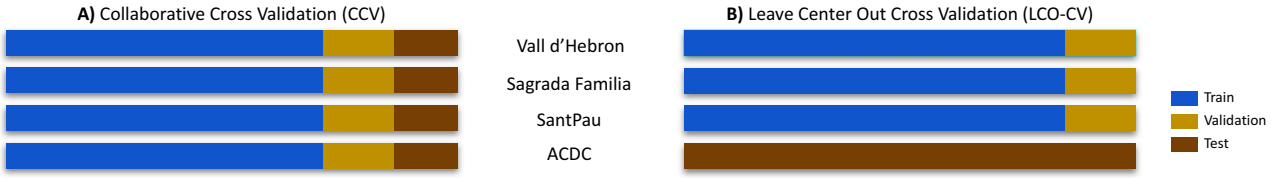

Figure 4. An illustration of a single iteration for: (**A**) Collaborative Cross Validation (CCV), where centers coordinate their splitting across 5 folds so that each center provides 20% of its data as test set and the rest as training and validation, (**B**) Leave Center Out Cross Validation (LCO-CV) which runs for as many iterations as there are centers in the dataset, each time using a different center as test set, and the rest as training and validation.

$$W_{t+1} \longleftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k,$$

where n is the total sample size and n_k the sample size of center k.

The reason that the FederatedAveraging algorithm weights the model parameters by a factor proportionate to the sample size of the center, is so that the better informed models outweigh the others[4]. However, medical centers are liable to domain shift effects due to different scanners, acquisition protocols, and different demographics from center to center. Given that data from different centers tends to be severely imbalanced (in our case Sagrada Familia has three times the size of SantPau), weighting the models by the sample is liable to add severe bias towards some cases over others. Thus we also tested a modified version of the FederatedAveraging algorithm where each model gets an equal weight during aggregation, which we will be refering to as FL-EV (as in Equal Voting). In this case the averaging equation becomes:

$$W_{t+1} \longleftarrow \frac{1}{K} \sum_{k=1}^{K} w_{t+1}^k.$$

Evaluation procedure. As we are dealing with data of small size, simple hold-out methods are inefficient to accurately represent the performance of our models[52]. Thus, all models are validated under a 5-fold cross validation scheme that utilizes the entire dataset. Concretely, we split the dataset into five folds, so that 20% is unseen during the training procedure and is used as a test set. From the remaining 80% we use 90% for training and 10% as validation. The purpose of the validation set is simply to find the early stop point.

In the past, federated learning performance has been evaluated under a collaborative cross validation (CCV) set-up[5]. In this set-up, cross validation is coordinated across centers so that each center splits its data into train, validation and test sets using the same percentage allocation. As a result, the final test set of each fold is an aggregate of different centers. However, when dealing with multi-center data, one has to take into account that a model should be deployable to centers unseen during the training and an estimate of out-of-site performance is required. For this reason, we are also estimating how our model generalizes to unseen centers under a different set-up that is Leave Center Out Cross Validation (LCO-CV). In this case, the data is split into as many folds as there are centers present (in our case 4 folds) and each fold uses a different center as testing set. This cross validation scheme has been used before in other distributed learning studies[53]. In both CCV and LCO-CV cases, FL and CDS are using the exact same split of the data so that all folds are comparable. A schematic of this evaluation procedure is presented in Fig. 4. For the CDS set-up, at each epoch a single model is trained and then evaluated. For the FL set-up, however, the evaluation process is more complex: (1) train the local copies of the model for one epoch; (2) send the local models to the server; (3) aggregate all local models to obtain an updated global model; (4) send copies of the updated global model to all centres; (5) evaluate the model on each centre locally. In this study, steps 2 and 4 are simulated within a single server; in the real world, however, these are the points where communication bottlenecks would occur due to possible bandwidth limitations[4]. The best model based on validation performance is used to predict the test set. After the entire cross validation procedure is complete, the Area Under the Receiver Operating Characteristic Curve score (AUC) is calculated based on the predictions derived from each respective test set.

In the case of CCV, to ensure fairness in comparison, the same folds are used for the FL and the CDS scheme. We repeat each experiment five times—i.e. using five different network weight initalizations. The reason for this is that a DL model's performance is liable to vary, as different initializations tend to converge to different optima[41]. By repeating each experiment five times and calculating the average AUC across repetitions, we obtain a more accurate estimate of the true performance and an estimate of the framework's robustness.

Regarding data harmonization, in the LCO-CV, the reference histogram used from matching does not include the test center, thus making the task harder when compared to CCV where all centers are used for the derived average. The histograms of each center before and after the standardization are outlined in Fig. 5. In this diagram, one example of an *LCO-CV* iteration is visualized where Vall d'Hebron is the test set. Vall d'Hebron in this case is matched to the histogram average as calculated by the train centers. Notably, the test set becomes more

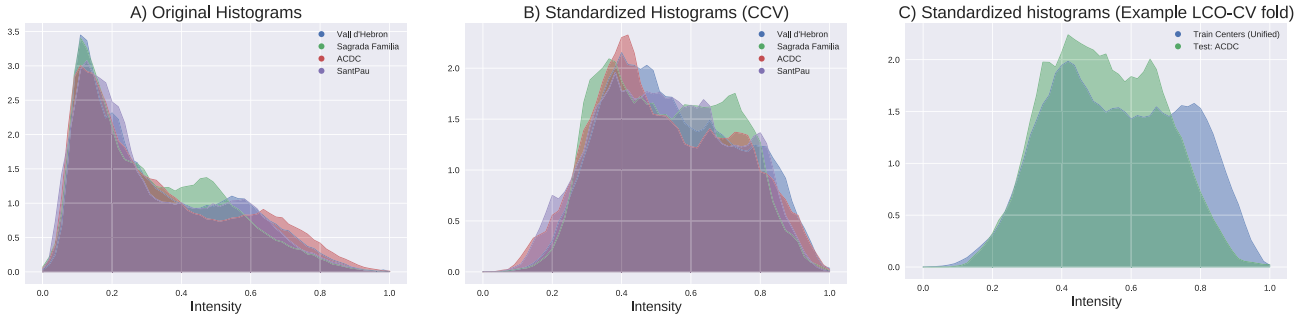

Figure 5. Histograms of different center intensities before and after histogram standardization. The two evaluation techniques imply different adaptation set-ups as for the *CCV* we are standardizing using an average from all of the centers, while for *LCO-CV* we consider one center entirely unseen.

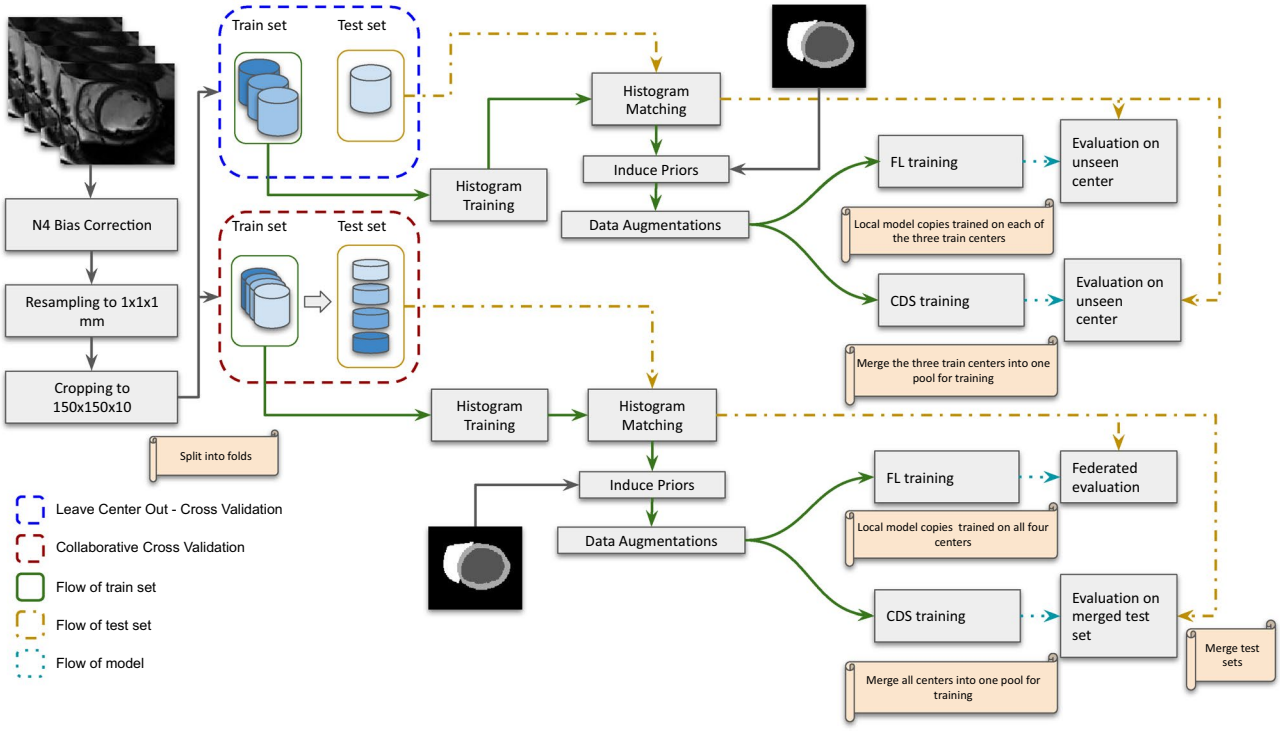

Figure 6. An overview of the pipeline used for the experiments in this study.

challenging because of this limitation, and in this example we can see this by the presence of a second peak on Vall d'Hebron's intensity distribution.

The overall pipeline is summarized in Fig. 6.

Results

In the initial experiments, we compare the baseline data to a curated version with induced shape priors as visualized in Fig. 2. The results are outlined in Table 2. For both collaborative learning frameworks, performance improves as we induce a shape prior (Masked MRI) and peaks when we use the per-structure split—a prior of the heart's structure. FL exhibits similar performance to CDS in all cases, however the standard deviations are much lower in the case of FL. Moreover, the difference in performances of CDS and FL decreases after constraining the solution space using the shape priors.

A systematic comparative study is conducted on the different augmentation set-ups using both the CCV and LCO-CV evaluation procedures. We find that CCV performance drops in all collaborative learning frameworks (Table 3 and Fig. 7) once augmentations are introduced.

In the case of LCO-CV, where the train and testing sets come from different distributions, CDS performance consistently improves from additional augmentations, reaching its highest when the most augmentations are applied (Fig. 7). FL also benefits from the basic augmentations of rotation and flipping, but its performance drops as shape and intensity augmentations are introduced, and CDS surpasses FL. Interestingly, models trained under an FL framework (either FL or FL-EV) exhibit incredibly consistent results across different initializations of the same set-ups, while CDS-framed models show a high amount of variance.

Curation type	Baseline MRI	Masked MRI	MRI per-structure split
CDS	0.727 ± 0.0298	0.810 ± 0.0120	0.856 ± 0.011
FL	0.747 ± 0.0005	0.826 ± 0.0005	0.861 ± 0.003

Table 2. AUC results for different curation types for Collaborative Data Sharing (CDS) and Federated Learning (FL). The reported numbers are calculated across five repeated experiments with different seeds.

Augmentations	Framework	Vall d'Hebron	Sagrada Familia	ACDC	SantPau	Total
None	CDS	0.898 ± 0.004	0.862 ± 0.008	0.797 ± 0.026	0.853 ± 0.005	0.856 ± 0.011
	FL	0.942 ± 0.002	0.896 ± 0.001	0.799 ± 0.003	0.867 ± 0.005	0.861 ± 0.003
	FL-EV	0.941 ± 0.002	0.874 ± 0.003	0.803 ± 0.002	0.875 ± 0.002	0.852 ± 0.002
Basic	CDS	0.861 ± 0.014	0.784 ± 0.012	0.778 ± 0.041	0.845 ± 0.016	0.809 ± 0.021
	FL	0.927 ± 0.002	0.872 ± 0.002	0.798 ± 0.001	0.807 ± 0.001	0.844 ± 0.002
	FL-EV	0.933 ± 0.001	0.859 ± 0.001	0.774 ± 0.001	0.818 ± 0.004	0.835 ± 0.002
Shape	CDS	0.897 ± 0.006	0.810 ± 0.023	0.770 ± 0.024	0.803 ± 0.017	0.827 ± 0.018
	FL	0.933 ± 0.002	0.871 ± 0.004	0.826 ± 0.001	0.901 ± 0.002	0.848 ± 0.002
	FL-EV	0.916 ± 0.001	0.852 ± 0.001	0.825 ± 0.001	0.883 ± 0.001	0.839±0.001
Shape and intensity	CDS	0.897 ± 0.009	0.821 ± 0.008	0.859 ± 0.028	0.833 ± 0.025	0.849 ± 0.018
	FL	0.905 ± 0.001	0.886 ± 0.001	0.785 ± 0.005	0.858 ± 0.003	0.839 ± 0.003
	FL-EV	0.917 ± 0.001	0.870 ± 0.002	0.800 ± 0.003	0.880 ± 0.002	0.842 ± 0.002

Table 3. Collaborative Cross Validation (CCV) performance evaluated using the AUC metric. Each number represents the average AUC, calculated across five repeated experiments with different seeds.

Augmentations	Framework	Vall d'Hebron	Sagrada Familia	ACDC	SantPau	Total
None	CDS	0.870 ± 0.020	0.809 ± 0.010	0.616 ± 0.064	0.784 ± 0.026	0.732 ± 0.008
	FL	0.897 ± 0.002	0.815 ± 0.001	0.599 ± 0.002	0.835 ± 0.001	0.746 ±0.001
	FL-EV	0.894 ± 0.001	0.816 ± 0.001	0.566 ± 0.002	0.837 ± 0.001	0.779 ± 0.004
Basic	CDS	0.916 ± 0.014	0.816 ± 0.016	0.654 ± 0.068	0.776 ± 0.019	0.759 ± 0.016
	FL	0.922 ± 0.002	0.855 ± 0.001	0.554 ± 0.002	0.818 ± 0.001	0.791 ± 0.001
	FL-EV	0.886 ± 0.007	0.868 ± 0.001	0.532 ± 0.003	0.810 ± 0.001	0.773 ± 0.005
Shape	CDS	0.900 ± 0.024	0.834 ± 0.028	0.641 ± 0.084	0.796 ± 0.013	0.764 ± 0.022
	FL	0.861 ± 0.002	0.879 ± 0.001	0.668 ± 0.003	0.845 ± 0.007	0.766 ± 0.001
	FL-EV	0.803 ± 0.003	0.869 ± 0.001	0.632 ± 0.008	0.821 ± 0.001	0.737 ± 0.004
Shape and intensity	CDS	0.901 ± 0.031	0.829 ± 0.013	0.743 ± 0.048	0.793 ± 0.018	0.776 ± 0.008
	FL	0.887 ± 0.001	0.807 ± 0.003	0.472 ± 0.003	0.840 ± 0.003	0.731 ± 0.009
	FL-EV	0.892 ± 0.001	0.854 ± 0.005	0.471 ± 0.002	0.844 ± 0.001	0.768 ± 0.003

Table 4. Leave Center Out (LCO-CV) performance evaluated using the AUC metric. Each number represents the average AUC calculated across five repeated experiments with different network initializations.

To analyze the different centers individually, it is more interesting to focus on the LCO-CV set-up (Table 4). In this case, the out-of-domain performance is evaluated as the testing center in each fold is entirely unseen. The most striking difference pertains to ACDC, where CDS outperforms the FL method in most cases, and by a margin of 0.3 on the AUC metric when intensity augmentations are introduced.

In terms of the FL-EV scheme, Sagrada Familia—the largest of all centers in the used cohort—seems to be the only center to benefit from the scheme by a substantial margin. This occurs when the maximum amount of augmentations are introduced to the data, where FL-EV outperforms FL by a margin of 0.05 on Sagrada Familia, reflected in the total result.

As HCM diagnosis is based mostly on the ED phase of the left ventricle, we also test the FL LCO-CV set-up using only the ED images. AUC performance is higher when no augmentations are introduced in this case at 0.78, dropping off to 0.77 with Basic augmentations down to 0.763 when shape and intensity augmentation are introduced. The results were again incredibly consistent between different initializations.

The highest score under the LCO-CV evaluation is achieved when both ED and ES are used in an FL set-up and Basic augmentations are present.

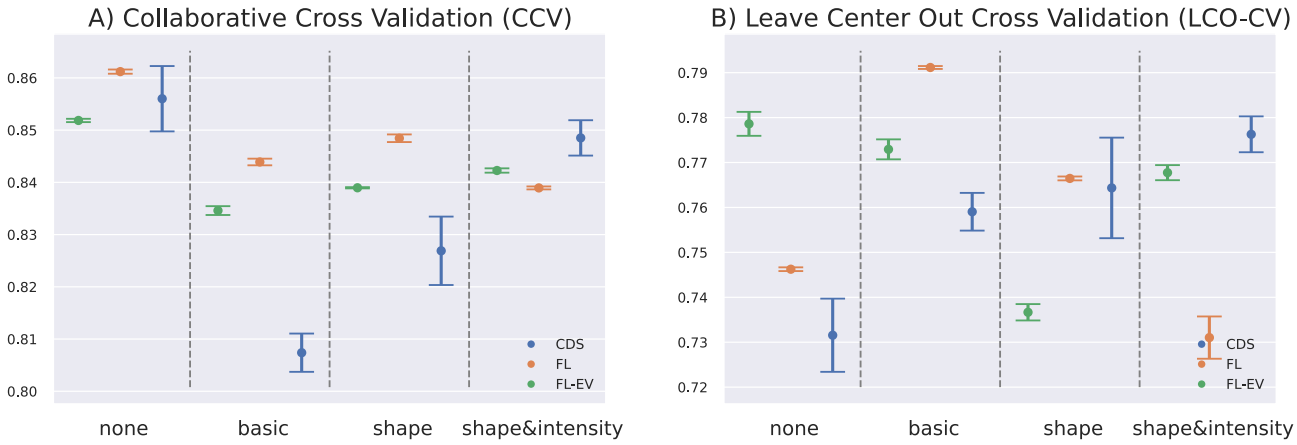

Figure 7. CDS, FL and FL-EV are tested given four data augmentation set-ups and for each set-up the experiment is repeated with 5 different network initializations, both for the CCV and LCO-CV set-ups. AUC is evaluated across all folds and we obtain 5 AUC metrics per set-up whose error bars are displayed here.

Discussion

Automatic CMR diagnosis has thus far focused on isolated centers, with training happening locally and evaluation limited to IID subsets of the data[31–34]. We have presented the first federated CMR diagnosis study and showcased two distinct evaluation set-ups to quantify both IID and non-IID performance. The gap in performance between the two evaluation frameworks was found to be critical, with LCO-CV consistently displaying worse results by a margin of approximately 0.1 on the AUC metric. This clearly suggests that for such models to be reliable and eventually adopted clinically, the LCO-CV set-up needs to be adopted in future research to account for non-IID performance. The necessity of this is exemplified by the presence of bias against ACDC on the shape and intensity set-up where FL exhibits an AUC performance of about 0.85 to 0.89 for the M&M centers but only 0.472 for ACDC (Table 4). Although lower performance is also seen in the CCV case for ACDC, the magnitude is not as critical (0.785 on AUC, Table 3). As M&M was part of a coordinated data collection within the EuCanShare project[37] acquisition protocols were bound to be more similar than that of ACDC which was collected in 2018 as part of a challenge[35]. We believe this to be a possible explanation for the exhibited bias against this center.

Our results reveal a couple of interesting behaviors on federated learning in this use case. Although FL has consistently been outperformed by CDS in medical imaging studies in the past, here FL outperforms CDS in the majority of our experiments. This result may come as surprising at first; however, FL has no inherent disadvantage over a CDS set-up. In the past, it has been shown that averaging the weights of copies of the same model in different stages of the training process boosts performance by approximating a lower point in the loss space[54,55]. We believe that, although in the case of FL the averaging happens at the same stage, with copies trained on different datasets, a similar effect occurs and becomes apparent in the use case of CMR, possibly due to the low amount of data. Furthermore, averaging across models seems to result in a stabilizing effect, and a higher performance of FL across different initializations of the same model, (exhibited by our repeated experiments over the same set-ups, Fig. 7).

Regarding the chosen augmentations, it's important to note that as CMR diagnosis largely relies on shape, deformation augmentations should be handled with care. As illustrated in Fig. 3, our deformations are of a low scale. Although this type of augmentation still gives us a boost in performance in the CDS case it's still less than the advantage we obtain from *Basic* augmentations. Furthermore, the FL set-ups both suffer from the addition of this augmentation compared to just using the *Basic* ones—even though they do better than no augmentation at all. In the case of using a single timepoint (ED) these domain shift effects are even greater, as FL performance drops even when *Basic* augmentations are introduced. This is indicative that even small deformations are causing domain shift effects that affect performance, to which federated set-ups are particularly sensitive.

This increased sensitivity to domain shift effects could stem from the way the Federated Averaging algorithm aggregates weights per round. Federated Averaging combines weights based on the respective data-size contribution, irrespective of possible domain shift effects. Interestingly, our FL-EV variant—which combines all datasets uniformly—does not see any benefit from augmentations and behaves very differently from the vanilla FL aggregation (Tables 3, 4). This variation hints that more sophisticated variants of the averaging algorithm accounting for domain shift effects will be interesting to explore.

With regards to this FL-EV variant, it's interesting to look into how the equal voting affects performance from a per-center standpoint. Sagrada Familia (Table 4) under shape and intensity augmentations benefits from the equal voting by a margin of 0.05 on the AUC metric. When Sagrada Familia is the test set, SantPau's vote increases while Vall d'Hebron decreases (based on their sample sizes). Thus, the explanation for the observed differences could be that Sagrada Familia's data distribution is much closer to SantPau's than Vall d'Hebron's and benefits from the effect on votes. We believe this warrants further exploration in the future, studying a spectrum of intermediate voting schemes that are not entirely equal nor as imbalanced as the per-center sample size.

We should also acknowledge that, while clinical reality mainly uses the ED timepoint for diagnosis, we focused the main bulk of our experiments on using both the ES and ED timepoints. This decision stemmed from our

need to tackle data scarcity, using more data both for more effective training and reliable evaluation. Our side experiment on just ED showed that our results are transferable, and FL translates well to this set-up with a decent performance when no augmentations are used.

This is one of the ways that we went about tackling the main limitations of this study—i.e. the small size of our data. While the impact of this limitation naturally puts in question the generalizability of our results, we conducted principled and diligent evaluations to minimize this negative effect. This includes our choice to emphasize on cross-validation schemes and repetitive experiments—i.e. conducting *five* experiments per configuration, each fully cross validated with 5 folds and different seeds. Furthermore, to ensure our evaluation was strong in spite of this limitation, we further evaluated our results from a collaborative standpoint (test sets from centers included in the training set) but also leave-center-out to account for out-of-site generalizability.

While we strongly believe that these evaluation principles should be ever-present in studies where such limitations are faced, they might not be enough. The solution then lies in the obvious direction: moving from simulated experiments to real-world federated studies, leveraging a high number of hospitals and institutes to overcome data scarcity.

Such a large-scale task, however, is anything but straightforward. Hospitals usually lack specialized staff and need to be closely guided on the technical requirements. This includes hardware powerful enough for deep learning models to train and a stable internet. The clinical requirements are even more challenging, requiring multiple expert annotators and consistency across sites. On that regard, crowd-sourcing portals that enable collaborative annotation across sites are a promising solution[56]. In addressing all of these requirements and tackling roadblocks along the way, the most important component will not so much be a matter of scientific mettle, but a skill in human communication and coordination.

Our focus is aimed at realizing such studies in the future.

Conclusions

In conclusion, through extensive analysis and experiments, we demonstrated that, even with a small sample size of 180 subjects derived from four centers, federated learning for Cardiac MRI diagnosis achieves promising performance that is comparable to collaborative data sharing. We highlighted the importance of a principled evaluation that accounts for both in and out of site performance and showed how models trained under a federated learning framework exhibit increased robustness and can be more sensitive to domain shift effects. As different centers seem to benefit in different ways from the interplay of augmentations and the collaborative learning frameworks, we believe that further research is required to delineate the underlying factors. In the future, bigger datasets with a wider diversity of centers should be used to systematically analyze and further verify these effects. Furthermore, our study was constrained to binary classification, but in the future semi-supervised federated learning methods will be needed to integrate labels that don't fully overlap between centers to enable multi-label classification. Also, as we used segmentation masks supplied by clinicians, the current pipeline still relies on expert support to provide these segmentations and could be replaced by an automated federated segmentation pipeline, appended as a first step to the current pipeline.

There is still a lot of fields in automated image-based diagnosis where federated learning has yet to make a presence, including prediction of prevalent diseases like retinopathy, diabetes, neurological disorders and cases of cancer in breast, liver or colon. Tangential fields like that of survival and treatment outcome prediction that are still underdeveloped due to the lack of data would also stand to benefit from the impact of federated learning and warrant similar exploration. We firmly believe that such studies will be a fundamental step to pave the way for multi-center studies going forward.

Data availibility

ACDC[35] was part of MICCAI 2017 Challenge and maintains its own website at: https://acdc.creatis.insa-lyon.fr. It can be made readily available after registration and/or contact with the authors. M&Ms was part of MICCAI 2020 challenge[37,57] and maintains its own website at https://www.ub.edu/mnms/. It can be made readily available by filling the form. The diagnostic labels are also available upon explicit request to the challenge organizers. The code used for all of the described experiments is made available open-source in the following link: https://github.com/Linardos/federated-HCM-diagnosis.

Received: 5 July 2021; Accepted: 2 February 2022
Published online: 03 March 2022

References

1. Zhou, S.K. *et al.* A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE* (2021).
2. Boyd, S., Parikh, N. & Chu, E. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* ((Now Publishers Inc, 2011).
3. Lambin, P. *et al.* Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. *Acta Oncol.* **54**, 1289–1300 (2015).
4. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282 (PMLR, 2017).
5. Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018).
6. Li, X. *et al.* Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Med. Image Anal.* **65**, 101765 (2020).
7. Li, W. *et al.* Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 133–141 (Springer, 2019).

8. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. Braintorrent. A peer-to-peer environment for decentralized federated learning (2019) arXiv preprint arXiv:1905.06731.
9. Sarma, K. V. *et al.* Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inf. Assoc.* **28**(6), 1259–1264 (2021).
10. Kumar, R. *et al.* Blockchain-federated-learning and deep learning models for covid-19 detection using CT imaging. arXiv preprint arXiv:2007.06537 (2020).
11. Yang, D. *et al.* Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan. *Med. Image Anal.* **70**, 101992 (2021).
12. Liu, B., Yan, B., Zhou, Y., Yang, Y. & Zhang, Y. Experiments of federated learning for covid-19 chest x-ray images. arXiv preprint arXiv:2007.05592 (2020).
13. Roth, H. R. *et al.* Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 181–191 (Springer, 2020).
14. Zerka, F. *et al.* Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (c-distrim). *IEEE Access* **8**, 183939–183951 (2020).
15. Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A. & Qadir, J. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. arXiv preprint arXiv:2101.07511 (2021).
16. Zhang, W. *et al.* Dynamic fusion-based federated learning for covid-19 detection. *IEEE Internet Things J.* **8**(21), 15884–15891 (2021).
17. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**(6), 1–12 (2021).
18. Ritchie, H. & Roser, M. Causes of death. *Our World in Data* (2018).
19. Wilkins, E. *et al.* European cardiovascular disease statistics 2017. European Heart. *Network* (2017).
20. Leiner, T. *et al.* Machine learning in cardiovascular magnetic resonance: Basic concepts and applications. *J. Cardiovasc. Magn. Reson.* **21**, 1–14 (2019).
21. Martin-Isla, C. *et al.* Image-based cardiac diagnosis with machine learning: A review. *Front. Cardiovascu. Med.* **7**, 1 (2020).
22. Zhang, N. *et al.* Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine mri. *Radiology* **291**, 606–617 (2019).
23. Luo, G., Sun, G., Wang, K., Dong, S. & Zhang, H. A novel left ventricular volumes prediction method based on deep learning network in cardiac mri. In *2016 Computing in Cardiology Conference (CinC)*, 89–92 (IEEE, 2016).
24. Isensee, F. *et al.* Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 120–129 (Springer, 2017).
25. Jang, Y., Hong, Y., Ha, S., Kim, S. & Chang, H.-J. Automatic segmentation of lV and RV in cardiac MRI. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 161–169 (Springer, 2017).
26. Zotti, C., Luo, Z., Humbert, O., Lalande, A. & Jodoin, P.-M. Gridnet with automatic shape prior registration for automatic MRI cardiac segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 73–81 (Springer, 2017).
27. Patravali, J., Jain, S. & Chilamkurthy, S. 2D-3D fully convolutional neural networks for cardiac MR segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 130–139 (Springer, 2017).
28. Baumgartner, C. F., Koch, L. M., Pollefeys, M. & Konukoglu, E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 111–119 (Springer, 2017).
29. Rohé, M.-M., Sermesant, M. & Pennec, X. Automatic multi-atlas segmentation of myocardium with svf-net. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 170–177 (Springer, 2017).
30. Yang, X., Bian, C., Yu, L., Ni, D. & Heng, P.-A. Class-balanced deep neural network for automatic ventricular structure segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 152–160 (Springer, 2017).
31. Khened, M., Alex, V. & Krishnamurthi, G. Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 140–151 (Springer, 2017).
32. Wolterink, J. M., Leiner, T., Viergever, M. A. & Išgum, I. Automatic segmentation and disease classification using cardiac cine mr images. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 101–110 (Springer, 2017).
33. Cetin, I. *et al.* A radiomics approach to computer-aided diagnosis with cardiac cine-mri. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 82–90 (Springer, 2017).
34. Liu, T., Tian, Y., Zhao, S., Huang, X. & Wang, Q. Residual convolutional neural network for cardiac image segmentation and heart disease diagnosis. *IEEE Access* **8**, 82153–82161 (2020).
35. Bernard, O. *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?. *IEEE Trans. Med. Imag.* **37**, 2514–2525 (2018).
36. Kwong, J. S. & Yu, C.-M. The need for multicentre cardiovascular clinical trials in Asia. *Nat. Rev. Cardiol.* **10**, 355 (2013).
37. Campello, V. M. *et al.* Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Trans. Med. Imag.* **40**, 3543–3554 (2021).
38. Marian, A. J. & Braunwald, E. Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res.* **121**, 749–770 (2017).
39. Geske, J. B., Ommen, S. R. & Gersh, B. J. Hypertrophic cardiomyopathy: Clinical update. *JACC: Heart Fail.* **6**, 364–375 (2018).
40. Hara, K., Kataoka, H. & Satoh, Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* 6546–6555 (2018).
41. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (2010).
42. Tustison, N. J. *et al.* N4itk: Improved n3 bias correction. *IEEE Trans. Med. Imag.* **29**, 1310–1320 (2010).
43. Nyholm, T. *et al.* Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer,-center and-sequence study. *Radiat. Oncol.* **8**, 1–12 (2013).
44. Mirzaalian, H. *et al.* Harmonizing diffusion MRI data across multiple sites and scanners. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 12–19 (Springer, 2015).
45. Nyúl, L. G., Udupa, J. K. & Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imag.* **19**, 143–150 (2000).
46. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 1–10 (2020).
47. Pérez-García, F., Sparks, R. & Ourselin, S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv:2003.04696 [cs, eess, stat] (2020). ArXiv:2003.04696.
48. Singh, S. P. *et al.* 3D deep learning on medical images: A review. *Sensors* **20**, 5097 (2020).
49. Tran, D. *et al.* A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* 6450–6459 (2018).
50. Kay, W. *et al.* The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017).

51. Kushibar, K. *et al.* Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Sci. Rep.* **9**, 1–15 (2019).
52. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).
53. Deist, T. M. *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: Eurocat. *Clin. Trans. Radiat. Oncol.* **4**, 24–31 (2017).
54. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A.G. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018).
55. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. & Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. arXiv preprint arXiv:1802.10026 (2018).
56. Diaz, O. *et al.* Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Phys. Med.* **83**, 25–37 (2021).
57. Campello, V.M. *et al.* Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge, https://doi.org/10.5281/zenodo.3886268 (2020).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 952103, EuCanImage.

Author contributions

A.L., K.K., P.G. and K.L. conceived the experiments, A.L. conducted the experiments and created the software used in this study, A.L., K.K., and K.L. analyzed the results. A.L., K.K., S.W. and K.L reviewed and edited the manuscript.

Competing interests

Sean Walsh declares the following financial interests/personal relationships which may be considered as potential competing interests: within and outside the submitted work, is the recipient of grants/sponsored research agreements in the areas of medical imaging, artificial intelligence, data science, applied to the clinical specialties of oncology and respiratory medicine. He holds a leadership position within Oncoradiomics SA, has shares in the company Oncoadiomics SA, and is co-inventor of submitted patents on behalf Oncoradiomics SA. The rest of the authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.L.