Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal

Nicola K. Dinsdale[a], Mark Jenkinson[a,b,c,*], Ana I. L. Namburete[d,*]

[a] *Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, UK*
[b] *Australian Institute for Machine Learning (AIML), School of Computer Science, University of Adelaide, Adelaide, Australia*
[c] *South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, Australia*
[d] *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK*

Abstract

Increasingly large MRI neuroimaging datasets are becoming available, including many highly multi-site multi-scanner datasets. Combining the data from the different scanners is vital for increased statistical power; however, this leads to an increase in variance due to nonbiological factors such as the differences in acquisition protocols and hardware, which can mask signals of interest.

We propose a deep learning based training scheme, inspired by domain adaptation techniques, which uses an iterative update approach to aim to create scanner-invariant features while simultaneously maintaining performance on the main task of interest, thus reducing the influence of scanner on network predictions. We demonstrate the framework for regression, classification and segmentation tasks and two different network architectures.

We show that not only can the framework harmonise many-site datasets but it can also adapt to many data scenarios, including biased datasets and limited training labels. Finally, we show that the framework can be extended for the removal of other known confounds in addition to scanner. The overall framework is therefore flexible and should be applicable to a wide range of neuroimaging studies.

Keywords: Harmonization, MRI, Joint Domain Adaptation

1. Introduction

The ability to combine datasets in neuroimaging between scanners and protocols is vital to achieve higher statistical power and is especially important when studying neurological conditions where limited data is available. Although some large scale neuroimaging projects now exist, such as the UK Biobank [40], most studies remain small and many of the larger studies are multi-site, such as ABIDE [11] and ADNI [24]. Pooling data across scanners and

*Indicates Equal Contribution

sites leads to an undesirable increase in non-biological variance, even when attempts have been made to harmonise acquisition protocols and use identical phantoms across imaging sites [50]. Multiple studies have confirmed this variation caused by scanner and acquisition differences including scanner manufacturer [22, 43], scanner upgrade [22], scanner drift [42], scanner strength [22], and gradient nonlinearities [26]. The ability to identify the presence of this variability has been further confirmed by several works applying the *'Name the Dataset'* game [45] to neuroimaging datasets [20, 48].

The removal of this scanner-induced variance is therefore vital for many neuroimaging studies. Many existing methods are based on ComBat [25], an empirical Bayes method originally developed to remove 'batch effects' in genetics, which has been applied for harmonising values derived from structural [13, 35, 48], diffusion [14] and functional MRI [50], successfully removing non-biological variance while preserving biological associations. ComBat uses multivariate linear mixed effects regression to account for both the biological variables and the scanner, to allow the modelling of the imaging features. For each site, a site-specific scaling factor δ is calculated, yielding a model that adjusts for additive and multiplicative effects. Furthermore, ComBat uses empirical Bayes to learn the model parameters, which assumes that model parameters across features are drawn from the same distribution; this improves the estimation of the parameters where small sample sizes are available. ComBat has been further developed to include: a term explicitly to model for variables of interest to preserve after harmonisation [48], model covariances [8], a generalized additive model (GAM) into the model, extending it to include nonlinear variations [35], and longitudinal studies [3]. ComBat, however, is limited to the harmonisation of image-derived values and defined associations. Random forests have also been used to approach the problem similarly, with the approach demonstrated for harmonising structural ROIs [19]. Other models also approach the harmonisation task through finding scaling factors for the raw image values, for instance MICA [49], which estimates nonlinear transformations of the voxel intensity values.

Harmonisation has been explored more for diffusion MRI than other modalities [5, 6, 7, 14, 23, 32, 30, 31, 33, 39] with many methods using spherical harmonics to harmonise the data [30, 31, 32, 33, 39] limiting their applicability to other modalities. Moyer et al [33] adopt variational autoencoders to produce scanner-invariant representations of the data. These feature representations can then be used to reconstruct the input images so that they are minimally informative of their original collection site. This work shows that using deep learning techniques to create feature representations that are invariant to scanner presents a strong candidate for MRI data harmonisation.

In addition to the variational autoencoders used in [33], other generative models have been used to harmonise MRI, largely based on deep learning models including U-Net [36] based models and cycleGAN [10, 53] based models [52]. These are limited by needing either paired or 'travelling heads' data for training for each site, which is expensive and infeasible to acquire in large numbers, but hard to evaluate without them [33].

The key measure of success for harmonisation methods is to be discriminative for the biological variable of interest whilst being indistinguishable with respect to the scanner used

to acquire the data. Following the framework introduced by Ben-David et al [4], domain adaptation routinely concerns the scenario where we have a source domain with a large set of labels and a target domain with either no labels or a low number of labels. For this work, we consider only the case where the task to be performed is identical across domains. We would expect that these two domains would be related but not identical: for instance, the source and target domains could indicate different scanners or different acquisition protocols. The difference between the two domains means that a neural network trained on the source data would have a reduced performance on the target domain [47]; this degree of difference is known as *domain shift* [44]. The smaller the degree of the domain shift, the more likely the domain adaptation is to be successful. Domain adaptation techniques therefore attempt to find a feature space that performs a given main task whilst being invariant to the domain of the data. This is achieved by learning features using the source domain labels, but these features are transformed to a representation that should generalise to the target domain by simultaneously adapting the features to have the same distribution for both datasets. Therefore, domain adaptation should be applicable to the task of MRI data harmonisation, creating features that are indiscriminate with respect to scanner but discriminate with respect to the task of interest.

There have been many methods proposed for domain adaptation, the simplest of which use a divergence-based approach that tries to minimise some divergence criterion between feature distributions for the source and target data distributions, which should provide a domain-invariant representation. This relies on a feature representation existing, where the classifier is able to perform approximately equally on both tasks. Frequently used divergence measures include maximum mean discrepancy [37] that compares the means of two samples to determine if they are from the same distribution, and correlation alignment [41] that tries to align second-order statistics of the two distributions using a linear transformation. These methods are limited by the assumption that all scanner information can be removed by satisfying a simply definable condition, which is unlikely when dealing with a highly non-linear system such as MRI.

The similarity between domains is generalised further by using adversarial methods, where the features for each domain are forced to be identical for the family of possible non-linear functions that can be produced by a given convolutional neural network. Adversarial domain adaptation was formalised with *Domain Adversarial Neural Networks* (DANNs) [18]. Based on the premise that for effective domain transfer to be achieved, predictions must be based on features that cannot discriminate between domains, they propose a generic framework that jointly optimises the underlying features as well as two discriminative classifiers that operate on these features. The first is a label predictor that predicts the class labels and the second is a domain classifier that aims to predict the source of the data. The overall aim of the network is then to minimise the loss of the label predictor and to maximise the loss of the domain predictor, such that the learnt features are discriminative for the main task but are completely unable to discriminate between domains. They show that this behaviour can be achieved for feedforward neural networks simply through the addition of a domain predictor formed of standard convolutional layers and a gradient reversal layer [17]. The gradient reversal layer allows the maximisation of the loss of the domain predictor whilst

minimising the loss on the main task. Placed between the feature extractor and domain classifier, the gradient reversal layer acts as an identity function in the forward step, and during the backward pass, it multiplies the gradient function by $-\lambda$, where λ is a hyperparameter set empirically for a given experimental setup [18]. The extension of domain adaptation to N source domains – key to enable the harmonisation of more than two scanners – was formalised in [4] and demonstrated for adversarial domain adaptation in [52].

Another method for adversarial domain adaptation was introduced in [46] where, rather than using a gradient reversal layer to update the domain predictor in opposition to the task, they used an iterative training scheme. By alternating between learning the best domain classifier for a given feature representation, and then minimising a confusion loss which aims to force the domain predictions to be closer to a uniform distribution, they obtain a domain classifier that is maximally confused [46]. Compared to DANN-style unlearning networks, it is better at ensuring a domain classifier that is equally uninformative across the domains [2] because the confusion loss tries to force all softmax output values to be equal, which is vital for successful data harmonisation, especially as we extend to a larger number of source scanners. This work was then applied in [2] to face recognition, not for domain adaptation, but for the explicit removal of identified sources of bias in their data, showing that the framework can be extended to 'unlearn' multiple factors simultaneously.

In this work we will show that the adversarial framework proposed in [46] can be adapted for use in harmonisation. By considering harmonisation to be a multiple source, joint domain adaptation task, we will show that we can produce shared feature representations that are invariant to the acquisition scanner while still completing the main task of interest across scanners and acquisition protocols with negligible performance compromise. Further, we will consider a range of likely data scenarios, such as the effect of having limited amounts of training data and the effect of having different distributions of data for different scanners, and show that the training framework is simply adapted to deal with these additional challenges. Finally, we will show that the framework can be used to remove other confounds in addition to harmonising for scanner. Whilst we demonstrate the network on age prediction and tissue segmentation, the framework is applicable to all feedforward networks and tasks and so should be applicable to the harmonisation of data across a wide variety of neuroimaging studies.

2. Methods

Consider the case where we have data for S_n subjects, where $n \in \{1..N\}$ indicates the scanner on which the data was acquired. For each scanner, we have MRI scans $\boldsymbol{X}$ which are the input for our network training. The network is formed of three parts: a *feature extractor* that takes the input images $\boldsymbol{X}$ and transforms these to a feature representation $\boldsymbol{\Theta}_{repr}$, a *label predictor* that takes the feature representation as input and outputs the predicted main task labels $\boldsymbol{y}_p$, and the *domain predictor* that takes the same feature representation and outputs the predicted domain $\boldsymbol{d}$ that corresponds to the acquisition scanner.

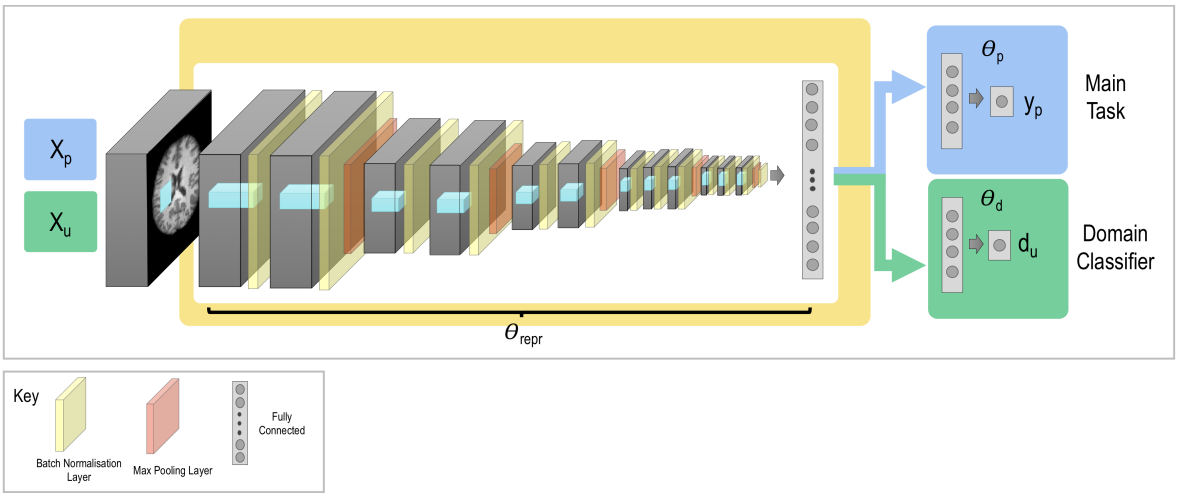


Figure 1: General network architecture. The network is formed of three sections: the feature extractor with parameters $\boldsymbol{\Theta}_{repr}$, the label predictor with parameters $\boldsymbol{\Theta}_p$, and the domain predictor with parameters $\boldsymbol{\Theta}_d$. $\boldsymbol{X}_p$ represents the input data used to train the main task with labels $\boldsymbol{y}_p$, and $\boldsymbol{X}_d$ represents the input data used to train the steps used to unlearn scanner with labels $\boldsymbol{d}$.

Three loss functions are used in the training of the network. The first loss function is for the primary task and is conditioned on each domain such that the training is not driven by the largest dataset:

$$L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) = \sum_{n=1}^{N} \frac{1}{S_n} \sum_{j=1}^{S_n} L_n(\boldsymbol{y}_{j,n}, \hat{\boldsymbol{y}}_{j,n}) \tag{1}$$

where N is the number of domains and S_n is the number of subjects from the domain n such that $\boldsymbol{y}_{j,n}$ is the true task label for the jth subject from the nth domain. $\boldsymbol{X}_p$ represents the subset of the the total dataset $\boldsymbol{X}$ used for the main task, which in all cases will be the full set of $\boldsymbol{X}$ for which we have main task labels $\boldsymbol{y}_p$ available.

The domain information is then unlearned using two loss functions in combination. The first is the domain loss that assesses how much domain information remains in $\boldsymbol{\Theta}_{repr}$ and simply takes the form of the categorical cross-entropy:

$$L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) = -\sum_{n=1}^{N} \mathbb{1}[d = n] log(p_n) \tag{2}$$

where p_n are the softmax outputs of the domain classifier. The softmax outputs are also used in the confusion loss, which removes domain information by penalising deviations from a uniform distribution:

$$L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) = -\sum_{n=1}^{N} \frac{1}{N} log(p_n) \tag{3}$$

Therefore, the overall method minimises the total loss function:

$$\begin{aligned} L(\boldsymbol{X}_p, \boldsymbol{X}_u, \boldsymbol{y}_p, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = & L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ & + \alpha L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ & + \beta L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{4}$$

where α and β represent the weights of the relative contributions of the different loss functions. Equations (2) and (3) cannot be optimised in a single step because they act in direct opposition to each other. Therefore, we have to update the loss functions iteratively and the total method results in three forward and backward passes to be evaluated per batch. Each loss function is used to update a subset of the parameters, with the rest being frozen. In the first stage, the parameters in the feature extractor and label predictor are updated and those in the domain predictor frozen; in the second stage, parameters in the feature extractor and label predictor are frozen and only those in the domain predictor are updated; in the final stage, the parameters in the label predictor and the domain predictor are frozen and those in the feature extractor are updated. This freezing of weights allows the network to converge on a feature representation that is invariant to the acquisition scanner.

As shown by the loss functions, different sets of data can be used to evaluate the different loss functions. This not only means that we can unlearn scanner information for data points for which we do not have main task labels but also means that we can unlearn scanner information using curated (or 'matched') subsets of data points when potentially problematic biases exist in the data as a whole. These scenarios are explored in the experiments to follow.

2.1. Age Prediction Task

We first consider the task of brain age prediction [9, 15, 16] as an example task to demonstrate the framework: $\boldsymbol{X}$ takes the form of the T1-weighted MRI images and $\boldsymbol{y}$ the true biological age values.

For these experiments, an architecture similar to the VGG-16 [38] network was used. A batch size of 32 was used throughout, with each batch constrained to have at least one example from each dataset, increasing the stability during training: to achieve this, the smaller datasets were over sampled. The main task loss (Eq. (1)) took the form of the Mean Square Error (MSE) for each scanner. All experiments were completed on a V100 GPU and were implemented in Python (3.6) and Pytorch (1.0.1) [34].

2.2. Datasets

We used three datasets for these experiments: UK Biobank [40] (Siemens Skyra 3T) preprocessed using the UK Biobank Pipeline [1] (5508 training, 1377 testing); healthy subjects from the OASIS dataset [29] (Siemens Telsa Vision 1.5T) at multiple longitudinal time points, split into training and testing sets at the subject levels (813 training, 217 testing), and healthy subjects from the Whitehall II study [12] (Siemens Magnetom Verio 3T) (452 training, 51 testing).

The input images for all the three datasets were resized to $128 \times 128 \times 128$ and then every fourth slice was retained, leaving 32 slices in the z direction (axial slices). This was done

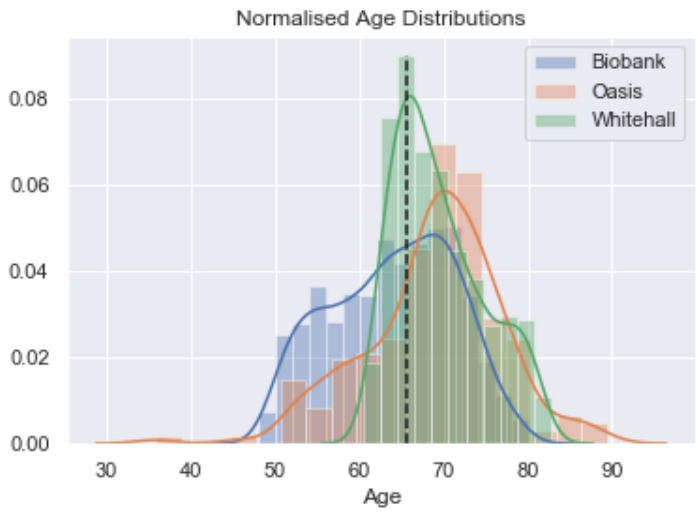


Figure 2: Normalised age distributions for the three datasets: Biobank, OASIS and Whitehall. The dashed line indicates the mean age of the three datasets.

so as to maximise coverage across the whole brain whilst minimising the redundancy and allowing us to use a larger batch size and number of filters. The inputs were also normalised to have zero mean and unit standard deviation. It can be seen from Fig. 2 that the age distributions for the three datasets are broadly the same and there is not a high degree of correlation between age and acquisition scanner.

2.2.1. Basic Fully-Supervised Learning

We first consider the simplest training scenario where we have main task training labels available for all datasets being used and that all the datasets have similar distributions for the main task label. This means that there is not a large degree of correlation between the age and the scanner and therefore, we should be able to remove scanner information from the feature representation without removing information that is discriminative for the age prediction task.

In this scenario, all three loss functions can be evaluated on a single combined dataset $\boldsymbol{X}$ where we have labels $\boldsymbol{y}$ for the full set and know the acquisition scanner for all data points $\boldsymbol{d}$, meaning that the overall method minimises the total loss function:

$$\begin{aligned} L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = & L_p(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ & + \alpha L_d(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ & + \beta L_{conf}(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{5}$$

To allow comparison, we also train the network using normal training - training only the feature extractor and label predictor - on the different possible combinations of datasets, and compare to all combinations of datasets. As a form of ablation study, we also compare to standard training with the main task loss function evaluated as in Eq.(1). These comparisons

are vital because we do not want the harmonisation process to significantly degrade the performance on the main task. We do not compare to existing harmonisation methods such as ComBat because these methods approach harmonisation differently to our method and so are not directly comparable.

2.2.2. Biased Domains

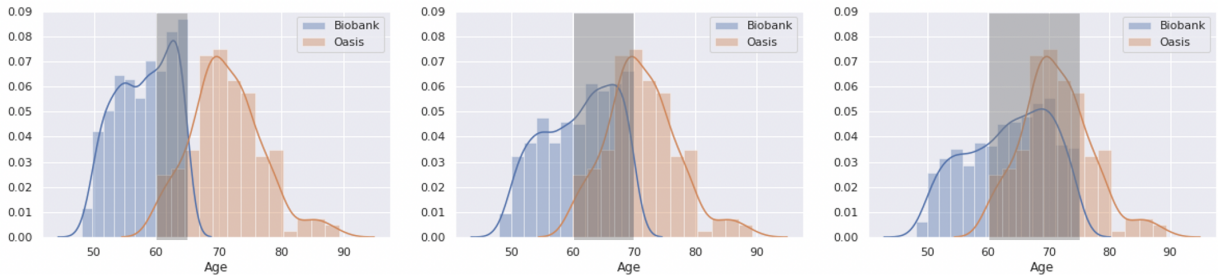


Figure 3: The three biased datasets used in this experiment with 5 years, 10 years and 15 years overlap. Only the shaded overlap region is used for unlearning scanner information and all the data points are used to evaluate the main loss function.

When the distribution of the data for the main task is similar across all data points, unlearning can simply be completed on all of the data points. However, where there exists a large difference between the two domains, such that the main task label is highly indicative of the scanner, it is likely that the unlearning process will also remove information that is important for the main task. This latter scenario could, for instance, be where the age distributions for the two studies are only slightly overlapping or where nearly all subjects with a given condition are collected on one of the scanners.

To reduce this problem, we utilise the flexibility of the training framework, and, whilst evaluating the main task on the whole dataset, we perform the scanner unlearning (Eqs. (2) and (3)) on only a subset of the data. For the case of age prediction, we perform unlearning only on the overlapping section of the data distributions. If we were to consider the case where we had data from both subjects and healthy controls and, for instance, most of the subjects had been scanned on one of the two scanners, we could perform unlearning on only the healthy controls rather than the whole dataset. As we do not need main task labels for the data used for unlearning, unlearning could be performed on a separate dataset so long as the scanner and protocol remained identical.

We sample the Biobank and OASIS datasets so as to create three degrees of overlapping datasets: 5 years, 10 years and 15 years (Fig 3) and test on the same test sets as before, spanning the whole age range. We compare three training options: i) training normally on the full combination of the biased datasets, ii) unlearning on the whole distribution, and iii) unlearning only on the overlapping section of the distributions.

2.2.3. Extension to ABIDE data

Having demonstrated the network on three scanners, we now demonstrate the network on a multi-site dataset, ABIDE [11]. We split the data into 90% training and 10% testing across

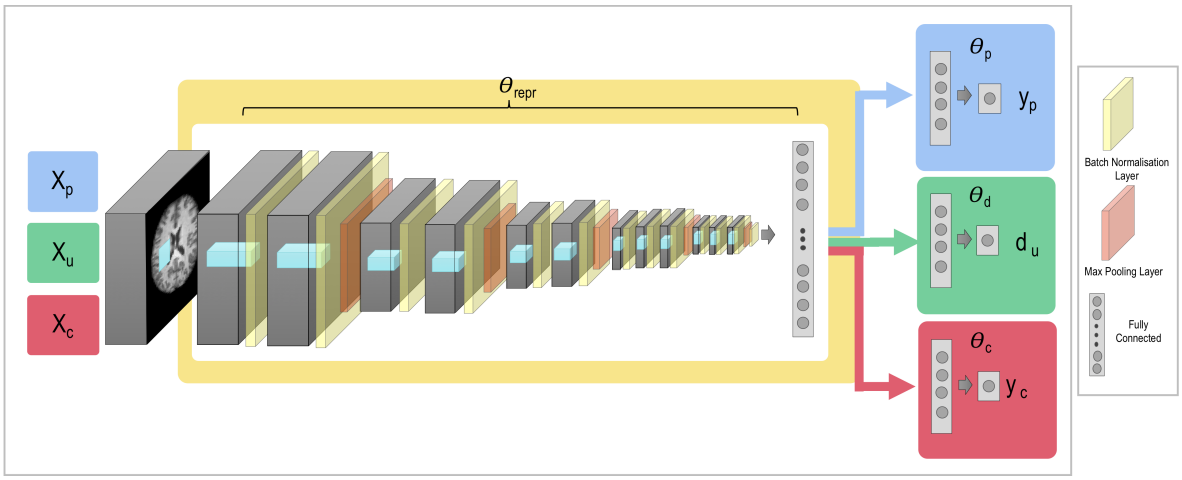


Figure 4: The general network architecture can be adapted to allow us to also remove other confounds from the data. In addition to the datasets used for the main task and the scanner unlearning, we add a data pair where $\boldsymbol{X}_c$ are the input images used for deconfounding and $\boldsymbol{y}_c$ are the confound labels.

each scanner and process as described in Section 2.2.1 so that we have data from 16 scanners, each with between 27 and 165 scans per site, for training. This is representative of many neuroimaging studies, with relatively low numbers of scans available for each site. Following the same fully supervised framework, as demonstrated in Section 2.2.1, we compare training on the combination of data from all 16 sites, to the unlearning approach.

A batch size of 32 was used in training, with the batch constrained to have at least one example from each scanner. To achieve this, the smaller datasets were over sampled such that the number of batches was limited by the largest dataset. This was performed because it was found that having an example of each dataset in each batch led to higher stability during training. The over sampled data points were not augmented so that we could be sure improvement was not due to the augmentation, but, in practice, this would be a sensible step.

2.2.4. Removal of other categorical confounds

In addition to unlearning scanner information to harmonise the data, we can also adapt the framework to explicitly remove other confounds. As shown in Fig 4, an additional data pair is added to training and an additional two loss functions are added to the overall loss function per confound. For each confound we wish to remove, we add a *confound classification loss*, which, like the domain classification loss, identifies how much information relating to the confound remains in the feature space, and a *confound confusion loss*, which aims to modify the feature space so as to remove the confound information by penalising deviations from a uniform distribution for the softmax predictions for the given task.

Therefore, the overall method will minimise the loss:

$$
\begin{aligned}
L(\boldsymbol{X}_p, \boldsymbol{X}_u, \boldsymbol{X}_c, \boldsymbol{y}_p, \boldsymbol{d}_u, \boldsymbol{y}_c, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_c) = {} & L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\
& + \alpha L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\
& + \beta L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \\
& + \sum_{j}^{J} (\gamma_j L_{c_j}(\boldsymbol{X}_{c_j}, \boldsymbol{y}_{c_j}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_{c_j}) + \phi_j L_{Cconf_j}(\boldsymbol{X}_{c_j}, \boldsymbol{y}_{c_j}, \boldsymbol{\Theta}_{c_j}; \boldsymbol{\Theta}_{repr}))
\end{aligned} \tag{6}
$$

where we consider J different confounds we wish to remove, γ_j is the weighting for the classification loss L_{c_j} for the jth confound, and similarly ϕ_j is the weighting of the confound confusion loss L_{Cconf_j} for the jth confound. We demonstrate this with sex as the confound and so the classification loss can simply be the binary cross-entropy loss. If the confound to be unlearned took the form of a continuous variable, such as age, then following [2] it would need to be split into discrete bins.

We consider three different scenarios for removing sex as a confound while completing the age prediction task. We first consider the simplest case, where sex is approximately equally distributed across age and scanner. In this case we can simply evaluate all the loss functions for the whole set of input images – assuming all labels are available for input images for all tasks.

We also explore the scenarios of the confound being correlated with a) the scanner and b) the main task. For correlation with scanner, we create datasets where 80% of the subjects in the OASIS dataset are male and 80% of the subjects in the Biobank dataset are female. For correlation with age, we consider the case where 80% of the subjects under the median age across the two datasets (65) are female and 80% of the subjects over the median age are male. For both scenarios, we still test on the full test set, and we compare normal training, unlearning sex on the whole training set, and unlearning sex on a subset curated to be balanced with respect to the generated bias. In the last case, the loss functions for the main task and unlearning of scanner are evaluated across the whole datasets and only the equations controlling unlearning the sex confound are evaluated for the subsets.

2.2.5. Removal of Continuous Confounds

While the approach for the removal of categorical confounds, such as age, is clear from the section above, the extension to continuous variables is less clear. If we reframe our network such that sex prediction is now the main task, we can consider the process of removing the age information.

The continuous age labels could be approximately converted into categorical labels by binning the data into single-year bins spanning across the age range. This, however, would not encode the fact that a prediction of 65 for a true label of 66 encodes more true age information than a prediction of 20. We therefore convert the true age labels into a softmax label around the true age, normally distributed as a $\mathcal{N}(\mu, \sigma^2)$ where μ was the true age label and σ was set to 10 empirically, allowing us to maintain relative information between bins. The value of 10 was chosen as, when used in normal training, this minimised the mean

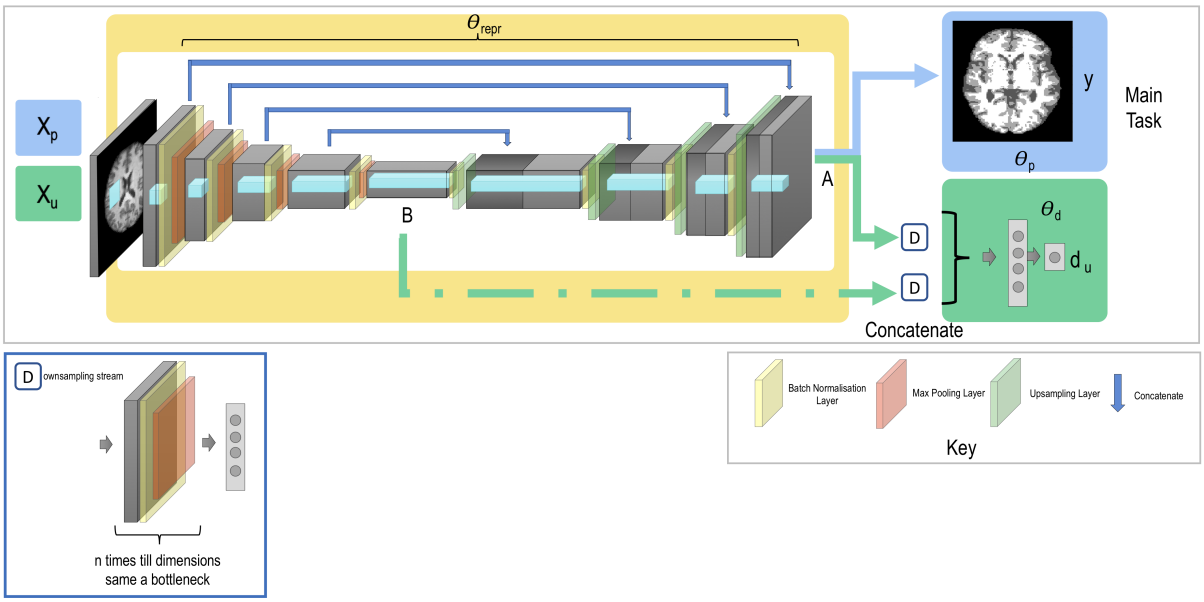


Figure 5: Network architecture used for unlearning with segmentation. $\boldsymbol{X}_p$ represents the input images used to evaluate the primary task with $\boldsymbol{y}$ being the main task label segmentations. $\boldsymbol{X}_u$ are the input images used for unlearning scanner information with domain labels d_u. The domain discriminator for unlearning can be attached from A, B or the two in combination. If it is attached from A and B together, the first fully connected layers are concatenated together to produce a single feature representation.

absolute error (MAE) on the test set. The loss function for the main task then becomes the Kullback-Leibler (KL) divergence between the the true softmax distribution and the predicted softmax outputs. Unlearning can then be achieved as before, with the aim still being to make all the softmax outputs equal, so that no information about the age of the subject remains.

2.3. Segmentation Task

Having demonstrated the unlearning process on the age prediction task, we now consider the task of segmentation. We specifically consider the case where our network architecture takes the form of a U-Net [36], the most popular network for biomedical image segmentation, where the skip-connections, which form a crucial part of the architecture, have the possibility of increasing the complexity of unlearning. Again, the input images $\boldsymbol{X}$ are T1-weighted input images but the labels take the form of tissue segmentations (grey matter/white matter/CSF) produced using FSL FAST [51] as a proxy to manual segmentations, which were converted to one-hot labels for the segmentation. We consider examples from the UK Biobank [40] (2095 training, 937 testing) and healthy subjects from the OASIS dataset [29] (813 training, 217 testing). As before, these were resized to $128 \times 128 \times 128$ voxels and the intensity values normalised, but then they were split into 2D slices so that we trained a 2D network. Multi-class Dice Loss was used as the primary task loss function. All experiments were completed on a V100 GPU and were implemented in Python (3.6) and Pytorch (1.0.1).

2.3.1. Basic Fully Supervised Learning

As can be seen in Fig. 5, the general structure is identical to that used for age prediction, apart from the location of the domain predictor. In the case of the age prediction task, the unlearning was performed at the first fully connected layer, chosen as it was both a reduced representation of the data and because it was at the end of the network so there were few weights afterwards for the network to re-learn how to encompass any remaining scanner information were it not fully removed by the unlearning process. In the case of the U-Net architecture, the most compressed representation of the information is at the bottleneck (B) but there are both several upsampling layers and skip-connections after this point. Therefore, it is likely that, were unlearning only to be completed at the bottleneck, all scanner information would be re-learned by the segmentation layers and so might still influence the output segmentations.

We therefore consider unlearning from both the end of the network before the label predictor (A), the bottleneck (B) and the combination of the two, formed by concatenating the first fully-connected layers to form a single domain prediction. We also compare to standard training as a benchmark. We utilise the simplest training regime, where we have segmentation labels available for all the data from all scanners and that we have identical segmentation tasks. Therefore the overall loss function takes the same form as equation (5) and all of the loss functions are evaluated on all of the data points.

2.3.2. Semi-Supervised Learning

Finally, we consider the scenario where we have very few or no labels available for one of the scanners – this is a very likely scenario for segmentation where manual labels are time consuming and difficult to obtain. We therefore assume access to one fully labelled dataset (UK Biobank) - and another dataset for which we do not have many labels. While the unlabelled data points cannot be used to evaluate the main task, they can be used for scanner unlearning.

No changes to the architecture are required; rather, we simply evaluate the main task for those data points for which we have main task labels and use all data points for unlearning such that the overall method minimises:

$$\begin{aligned} L(\boldsymbol{X}, \boldsymbol{y}_p, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = & L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ & + \alpha L_d(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ & + \beta L_{conf}(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{7}$$

where $\boldsymbol{X}_p$ is the subset of $\boldsymbol{X}$ for which we have main task labels $\boldsymbol{y}_p$ available and the full dataset $\boldsymbol{X}$ is used in unlearning scanner information.

We explore the effect of increasing numbers of data points on the final segmentation, comparing to normal training on the combination of the full UK Biobank dataset and OASIS with available segmentations.

3. Results

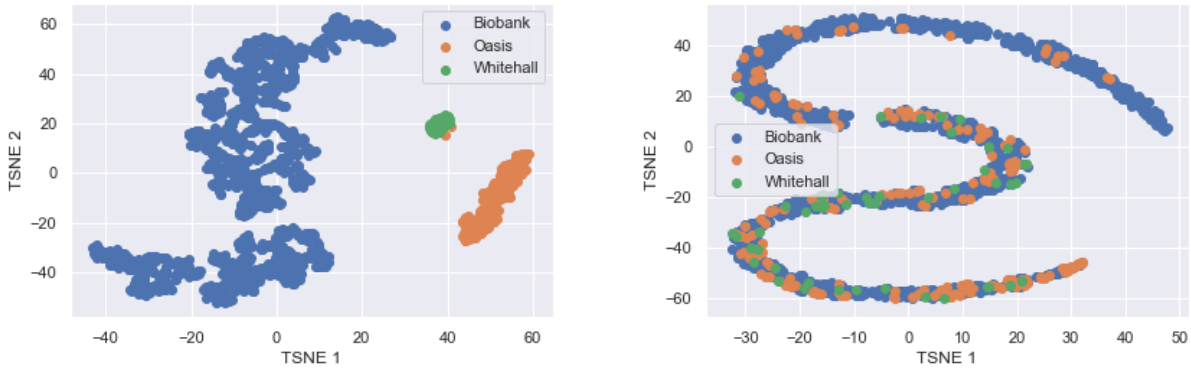
3.1. Age Prediction Task

3.1.1. Basic Fully-Supervised Learning

	Training Data B	Training Data O	Training Data W	Biobank MAE	OASIS MAE	Whitehall MAE	Scanner Classification Accuracy %
Normal Training							
1.	✓	×	×	3.25 ± 2.36	16.50 ± 6.77	13.81 ± 5.42	-
2.	×	✓	×	5.61 ± 3.52	4.27 ± 3.79	6.73 ± 4.82	-
3.	×	×	✓	5.61 ± 3.65	5.22 ± 4.83	3.15 ± 2.81	-
4.	✓	✓	×	3.30 ± 2.50	4.00 ± 2.78	4.71 ± 3.42	98 (50)
5.	✓	×	✓	3.31 ± 2.49	4.45 ± 3.53	3.05 ± 2.84	100 (50)
6.	×	✓	✓	5.71 ± 3.59	4.05 ± 3.71	3.21 ± 2.94	100 (50)
7.	✓	✓	✓	3.24 ± 2.47	4.19 ± 3.50	2.89 ± 2.70	96 (33)
Normal Training - Eq (1)							
9.	✓	✓	×	3.45 ± 2.63	3.99 ± 2.85	4.56 ± 3.38	100 (50)
10.	✓	×	✓	3.42 ± 2.60	4.19 ± 4.01	2.99 ± 2.69	100 (50)
11.	×	✓	✓	4.56 ± 3.05	4.08 ± 3.73	3.18 ± 2.93	100 (50)
12.	✓	✓	✓	3.55 ± 2.68	3.90 ± 3.53	2.62 ± 2.65	98 (33)
Unlearning							
13.	✓	✓	×	3.41 ± 2.04	3.79 ± 2.99	4.60 ± 3.47	48 (50)
14.	✓	×	✓	3.41 ± 2.58	4.07 ± 4.12	2.81 ± 2.57	52 (50)
15.	×	✓	✓	3.38 ± 2.64	3.91 ± 3.53	2.82 ± 2.65	50 (50)
16.	✓	✓	✓	3.38 ± 2.64	3.90 ± 3.53	2.56 ± 2.47	34 (33)

Table 1: Results comparing unlearning to training the network in different combinations on the datasets. Mean absolute error is reported in years. Scanner accuracy is the accuracy achieved by a domain predictor given the fixed feature representation at convergence, evaluating only for the datasets the network was trained on. Number in brackets indicates random chance. B = Biobank, O = OASIS, W = Whitehall. p values can be found in the supplementary material.

The results from training with all three datasets separately and the different combinations can be seen in Table 1. The scanner classification accuracy is achieved by taking the frozen feature representation $\mathbf{\Theta}_{repr}$ from training, and training a classifier based on this. For the unlearning method, this is the domain classifier that was used during training but fine tuned to convergence. For normal training, an identical domain predictor is trained using the frozen feature representation as input until convergence. Therefore, the closer the value is to random chance, the less informative the feature representation is about scanner, thus meaning the influence of scanner on the prediction is reduced. The results from training on all combinations of datasets both with standard training, and standard training using our loss function for the main task conditioned on each scanner, are shown for comparison.



(a) (b)

Figure 6: a) T-SNE plot of the fully connected layer in $\mathbf{\Theta}_{repr}$ before unlearning. It can be seen that the domains can be almost entirely separated, except for two data points grouped incorrectly, showing that data from each scanner has its own distinct distribution. b) T-SNE plot of the fully connected layer in $\mathbf{\Theta}_{repr}$ after unlearning. It can be seen that, through the unlearning, the distributions become entirely jointly embedded.

In Fig. 6, a T-SNE plot [28] can be seen, allowing visualisation of the frozen feature representation $\mathbf{\Theta}_{repr}$. It can be seen that, before unlearning, the features produced are entirely separable by scanner, but after unlearning the scanner features become jointly embedded and so the feature embedding is not informative of scanner.

The T-SNE demonstrates that without unlearning, not only are there identifiable scanner effects but that they also affect the age predictions. On the other hand, it can be seen that we are able to remove scanner information using our unlearning technique, such that the data points for all three scanners share the same embedding. This is confirmed by the scanner classification accuracy being almost random chance after unlearning has been completed.

It can also be seen that unlearning does not decrease substantially the performance on the main task. Graphs showing the different loss functions with training can be found in the supplementary material. If we consider the case of training on all three datasets, we actually achieve an overall improvement in performance using unlearning across the three datasets (comparing lines 7 and 16 in Tab. 1) with the performance on the OASIS and Whitehall datasets substantially improving. The performance on the UK Biobank dataset decreases, but this is as would be expected because the UK Biobank dataset has a much larger size and so, with normal training, the network is driven by this dataset whereas, with unlearning, all of the datasets drive the performance. Comparing to line 12, where the same loss function was used for the main task, evaluated separately for each dataset, but unlearning was not performed, we can see that there was an improved performance when the unlearning was added and so the unlearning itself provides actual improvement on the main task. It can also be seen that the scanner classification accuracy is almost perfect when using normal training (96%), or the loss function evaluated on the datasets separately (98%), and is random chance after unlearning (34%), clearly demonstrating that the unlearning process removes scanner information from the feature space.

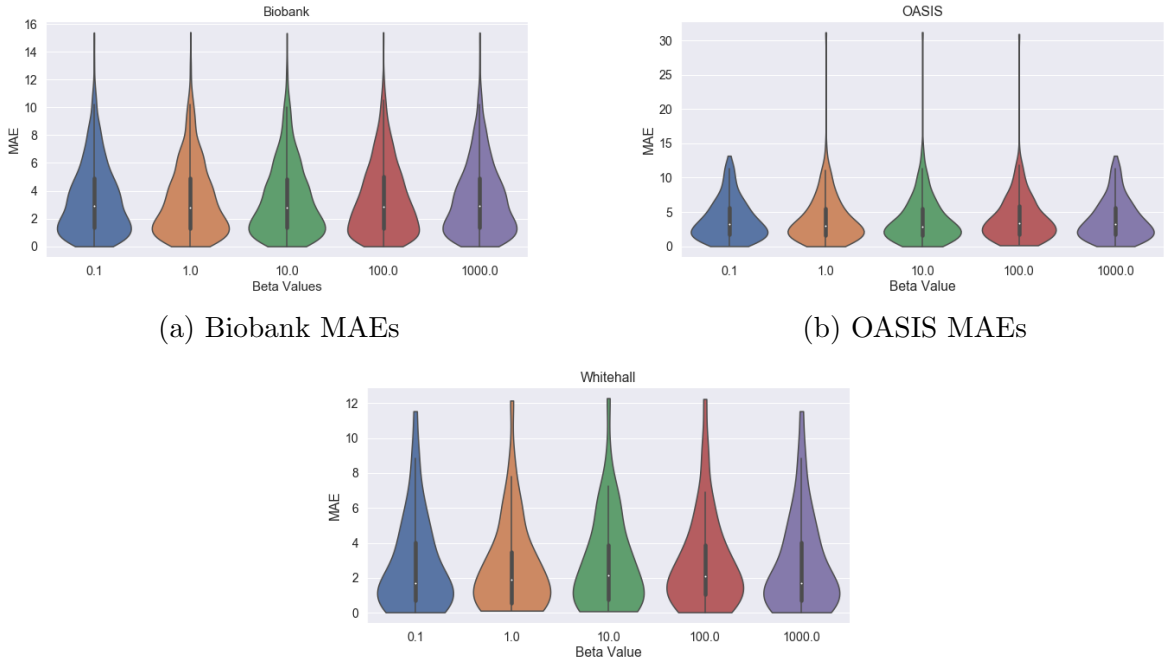
Training Method	Testing Dataset (MAEs)			
	Biobank	OASIS	Whitehall	Scanner Classification Accuracy
Full Dataset				
Normal Training	3.24 ± 2.47	4.19 ± 3.50	2.89 ± 2.70	100 (33)
Unlearning	3.38 ± 2.64	3.90 ± 3.53	2.56 ± 2.47	34 (33)
Balanced Dataset				
Normal Training	5.31 ± 3.82	5.80 ± 3.93	3.50 ± 3.09	98 (33)
Unlearning	4.11 ± 3.23	4.10 ± 3.68	3.11 ± 2.99	34 (33)

Table 2: Comparison of the effect of training with the full datasets with imbalanced numbers from each scanner versus training with balanced numbers of subjects from each scanner (452 per scanner, randomly selected compared to 5508 Biobank, 813 OASIS and 452 Whitehall), comparing both normal training and unlearning.

It can also be seen that the unlearning process creates a feature space that generalises better to the third unseen dataset. If, for instance, we compare lines 5 and 14 we can see that not only is there an improved performance on the OASIS and Whitehall datasets used for training, but there is also a significant improvement in the performance on the UK Biobank dataset, which was unseen by the network. This, therefore, suggests that the unlearning process not only creates features that are invariant to the scanners that were present in the training set, but also that these features are more generalisable and so can be more effectively applied to other scanners. It can again be seen there is an improvement between normal training with our loss function evaluated on each scanner separately and the performance with unlearning (lines 10 and 14), showing that the increase in generalisability of the features is due to the unlearning process. The same pattern can be seen for training on the other possible pairs of datasets.

The generalisability of the features between datasets and the reduction of the scanner classification accuracy together demonstrate that the unlearning process successfully harmonises the differences between scanners and protocols.

The effect of several of the design decisions was also explored. First, we considered if it were better to use the full imbalanced datasets to maintain the number of data points or if this hindered the unlearning process. Therefore, we randomly downsampled the Biobank and OASIS datasets to have the same number of examples as the Whitehall dataset (452 subjects). Testing was evaluated on the same test dataset as in the previous experiments. The results can be seen in Table 2. For both normal training and unlearning, there was a large decrease in performance using the balanced datasets, but the decrease was less pronounced when using unlearning than when using normal training. Using balanced datasets made no difference to the ability to classify scanner: with normal training the classifier was still able fully to distinguish between scanners; with unlearning, the classification was almost random chance. Therefore, any advantage gained from having balanced datasets is outweighed by the reduction in performance from having reduced numbers of training examples and so we



(a) Biobank MAEs

(b) OASIS MAEs

(c) Whitehall MAEs

Figure 7: MAEs for each dataset with increasing values of the constant β used in training.

are better off training on the whole dataset. Given that the method is able to remove nearly all information with significant imbalance in this case (Biobank 5508 training data points compared to Whitehall's 452) it is likely that unlearning would be sufficient in most cases.

For all experiments, the value of α remained set to 1 and no change was required. The value of β was varied between experiments, controlling the training stability and the rate of convergence: higher β values provide more stability for the domain unlearning but the decrease rate of convergence for the main task. It should not have any effect on the final age prediction values achieved, given that for all values we are able to reach convergence, and so we varied the value of β between 0.1 and 1000 and compared the MAEs for the three datasets. It can be seen that the value of β has no effect on the achieved MAEs across these values from Fig. 7. Therefore, the segmentation performance is robust to the choice of β and the value chosen can be selected to maximise the stability of training without impacting on the final prediction values. Stability could also be controlled by using different learning rates for each stage, however this was found to be harder to tune in practice.

Finally, we explored the choice of the batch size used. For training, the largest batch that could fit into memory was used; here the effect of smaller batches is explored. Figure 8 shows the validation loss during training for batch sizes of 3, 8 and 32 (other batch sizes are not shown for clarity) and the pattern was the same across the different loss functions. It can be seen that the training was more stable with the smaller batch sizes. However, it can be seen from Fig. 9 that the overall MAEs achieved across the datasets were better with

larger batch sizes. This is most likely because larger batch sizes force the network to unlearn scanner information more thoroughly than, for instance, with a batch size of 3, where given our constraint on the batches, there can only be one example per scanner. Given the result from varying the value of β above, we should select the batch size so as to maximise the performance on the main task and vary the β value to control the stability of training.

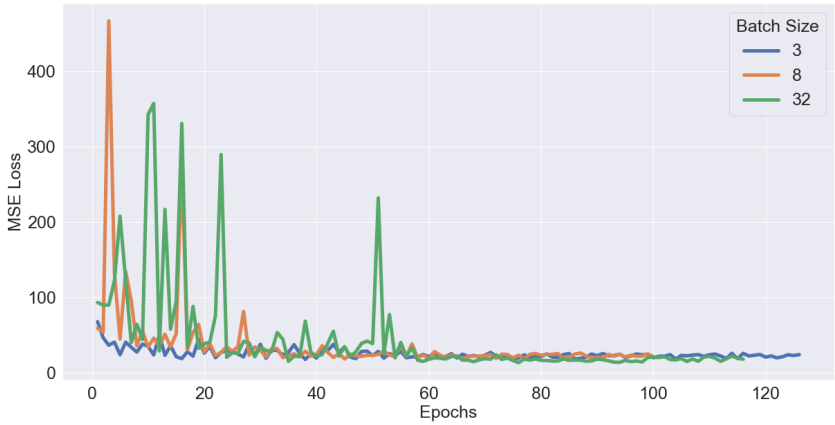
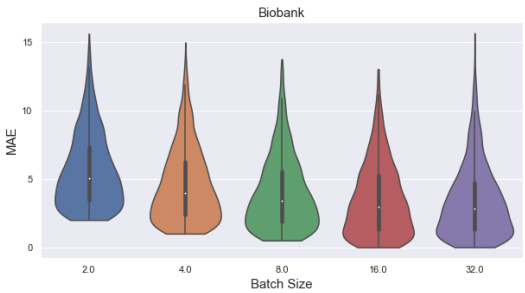
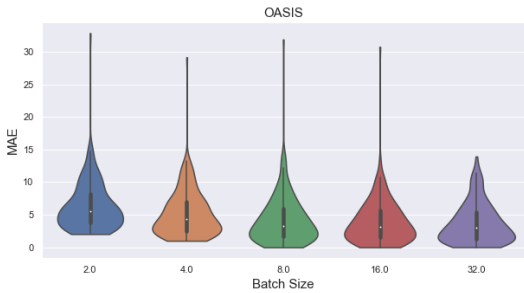


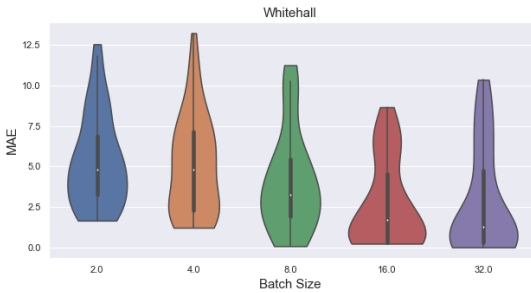
Figure 8: Validation loss with epochs for batch sizes of 3, 8 and 32.



(a) Biobank MAEs



(b) OASIS MAEs



(c) Whitehall MAEs

Figure 9: MAEs for each dataset with increasing batch sizes used during training.

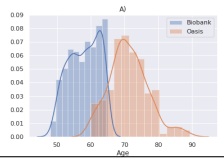
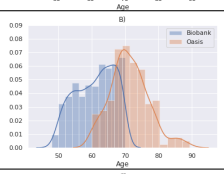
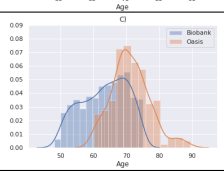
	Method	Biobank MAE	OASIS MAE	Scanner Classification Accuracy (%)
	**5 Years Overlap**			
	1. Standard	16.5 ± 5.94	15.5 ± 6.95	100 (50)
	2. Naïve Unlearning	6.11 ± 3.99	4.44 ± 4.20	58 (50)
	3. Unlearning on Overlap	5.49 ± 3.67	4.37 ± 4.05	53 (50)
	**10 Years Overlap**			
	4. Standard	9.66 ± 5.83	13.6 ± 6.58	100 (50)
	5. Naïve Unlearning	4.20 ± 2.90	4.29 ± 4.01	56 (50)
	6. Unlearning on Overlap	3.93 ± 2.81	4.04 ± 3.86	52 (50)
	**15 Years Overlap**			
	7. Standard	8.91 ± 5.31	10.4 ± 5.55	100 (50)
	8. Naïve Unlearning	3.82 ± 2.84	4.39 ± 4.07	57(50)
	9. Unlearning on Overlap	3.75 ± 2.78	3.99 ± 3.52	50 (50)

Table 3: MAE results for Biobank and OASIS data from training with datasets with varying degrees of overlap as shown in the first column. Standard refers to training normally and naïve unlearning refers to unlearning on the whole of the datasets. Scanner accuracy is calculated by training a domain classifier on the fixed feature representation. Random chance is given in brackets.

3.1.2. Biased Datasets

We assessed the performance of the network training with biased datasets curated from subsets of the Biobank and OASIS datasets. We considered the cases of a 5-year, 10-year and 15-year overlaps where the smaller the overlap, the harder the task. The trained networks are tested on the same test sets used above, with subjects across the whole age range included.

The results can be seen in Table 3 where we compare normal training, naïve unlearning on the whole dataset, and unlearning on only the overlapping age range. It can be seen that for all three degrees of overlap, standard training leads to much larger errors than unlearning and that unlearning on only the overlap gives a lower error than unlearning on all data points. Plots of the MAEs with age can be seen in Fig. 10 for the 10-year overlap case. It can be seen that the overall lowest MAEs are achieved across the age range when unlearning is performed only on the overlapping subjects.

As expected, it can be seen that the normal training regime produces large errors, especially outside of the range of the Biobank training data, as the learnt weights are very largely driven by the much larger size of the Biobank training data. With naïve unlearning, the network is not able to correct for both scanners and the results for the OASIS data are poor, whereas by unlearning on just the overlapping subjects, the error is reduced on both of the datasets during testing. The only area where we see a reduction in performance is the lower end of the OASIS dataset, possibly because when the network was being driven by the Biobank data, the network generalised well to the OASIS subjects from the same range. Naïve unlearning also performs slightly less well at removing scanner information

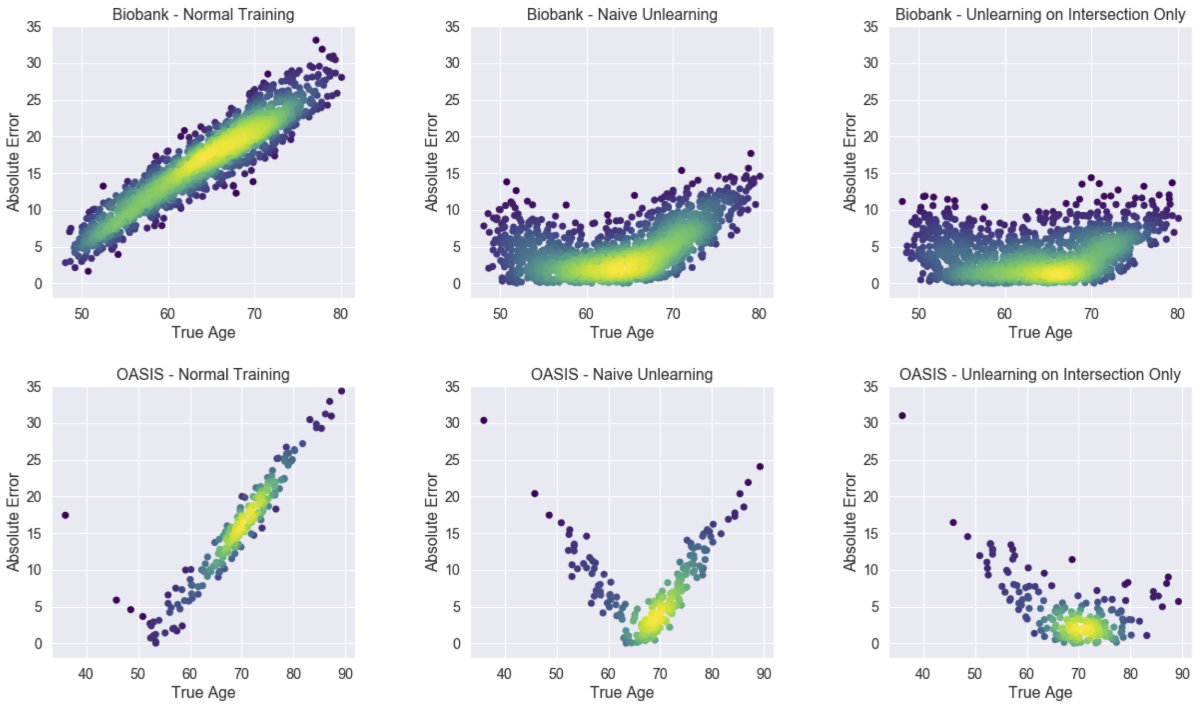


Figure 10: Density plots showing the absolute errors for the three different training regimes: standard training, naïve unlearning and unlearning only on the overlap data for 10-year overlap. It can be seen that unlearning only on the overlap dataset leads to substantially lower losses across both datasets.

on the testing data, probably indicating that the features removed also encode some age information and so generalise less well across the whole age range.

3.1.3. Extension to ABIDE data

Figure 11 shows the results of applying the unlearning process to the ABIDE data, comparing normal training to unlearning. It can be seen that unlearning improves the MAE across all sites apart from MPG, which contains significantly older subjects compared to the other sites. The scanner classification accuracy before unlearning was 55.9% and after unlearning was reduced to 6.42%, where random chance was 6.25%. Therefore, we can see that scanner information is present in the feature embeddings, even when harmonised protocols are used.

These results therefore show that the framework can be applied to many sites with no changes needed. It also shows that the framework is applicable to lower numbers of subjects, with some of the sites having as few as 27 subjects for training. We thus foresee that the framework should be applicable to many studies.

3.1.4. Removal of other categorical confounds

The effect of removing sex information as an additional confound in addition to harmonising for scanner was investigated. In Fig. 12 it can be seen that there is no significant effect on the MAE results when removing sex information in addition to scanner informa-

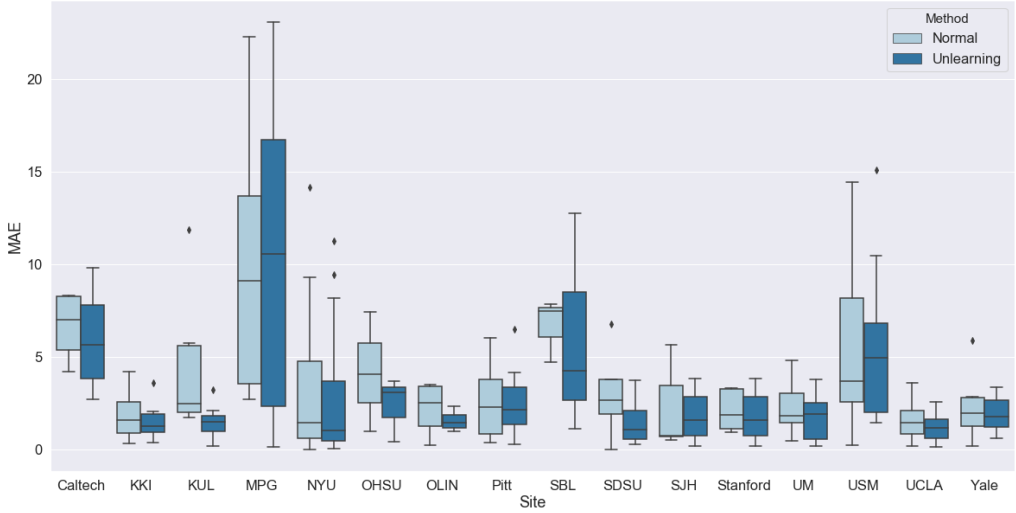


Figure 11: MAE results broken down by site for the ABIDE data, comparing normal training to unlearning.

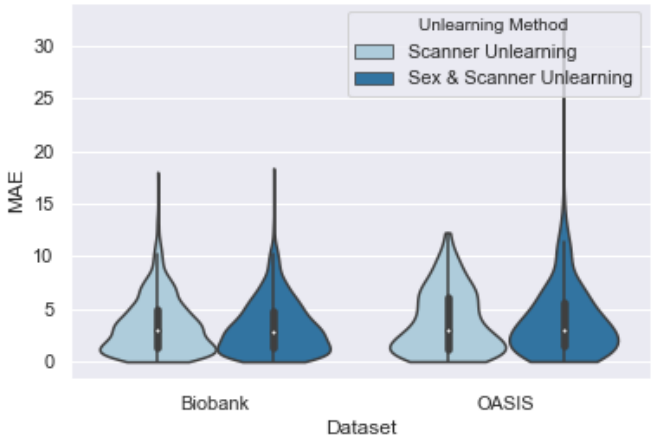


Figure 12: Results comparing unlearning scanner only to unlearning scanner and sex information.

tion. Unlearning sex information had no substantive effect on the ability to remove scanner information with the scanner classification accuracy being 48% and 49%, respectively. The sex classification accuracy was 96% before unlearning and 54% after unlearning. Therefore, we can remove multiple pieces of information simultaneously and, so long as the information we wish to unlearn does not correlate with the main task, we can do so without significantly reducing the performance on the main task.

We then considered the case where sex information was correlated with the acquisition

Training Method	Biobank MAE	OASIS MAE	Scanner Classification Accuracy	Sex Classification Accuracy
Normal Training	3.44 ± 2.56	3.98 ± 3.33	100 (50)	98 (50)
Unlearning All	3.41 ± 2.60	4.05 ± 3.66	52 (50)	57(50)
Unlearning Balanced	3.43 ± 2.60	4.04 ± 3.60	49 (50)	56 (50)

Table 4: MAE results for unlearning scanner and sex when sex is highly correlated with scanner.

Training Method	Biobank MAE	OASIS MAE	Scanner Classification Accuracy	Sex Classification Accuracy %
Normal Training	3.45 ± 2.60	4.85 ± 3.73	100 (50)	95 (50)
Unlearning All	5.25 ± 3.70	5.62 ± 4.07	100 (50)	65 (50)
Unlearning Balanced	3.53 ± 2.61	4.53 ± 3.52	52 (50)	54 (50)

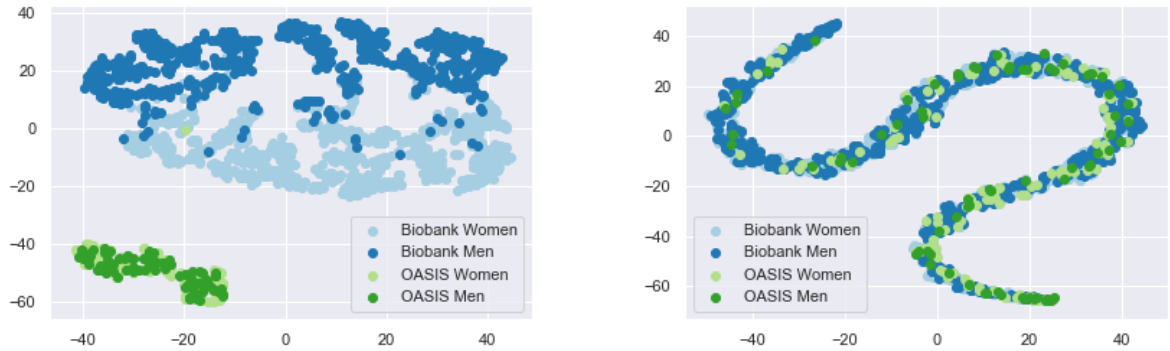
Table 5: MAE results for unlearning scanner and sex when sex is highly correlated with age.

scanner given that 80% of the subjects in the OASIS dataset are male and 80% of the subjects in the Biobank dataset are female. Table 4 shows the comparison of normal training on these datasets compared to unlearning sex on all data points and unlearning sex on a subset with balanced numbers of each sex for each scanner. The full testing set was still used.

It can be seen that there is little difference between unlearning on all, or just a balanced subset, compared to the normal training baseline. It can also be seen that neither method affects the network's ability to unlearn scanner or sex information and so there is no need to change the unlearning training from the standard case.

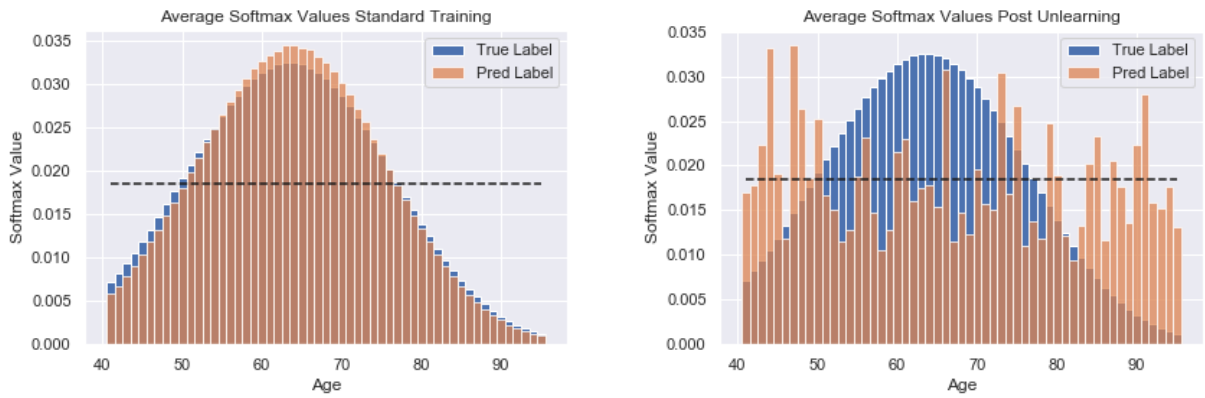
Finally, we considered the case where sex – the confound we wish to unlearn – is highly correlated with age – the main task – such that 80% of the subjects below the median age are female and 80% of the subjects over the median age are male. Again, normal training was compared to unlearning on the whole dataset and to unlearning on a curated subset, with uniform distributions of sex with age for each scanner. Testing was again performed on the full testing datasets.

It was found that it was not possible to unlearn on the whole distribution as it caused the training to almost (after approximately 5 epochs) become unstable almost immediately, with the training and validation loss exploding - before we were able to unlearn sex and scanner information. The results from unlearning scanner can be seen in Table 5 where it can be seen that by unlearning only on a subset of data points with equal sample numbers we can almost entirely unlearn both scanner and sex information. Figure 13 shows a T-SNE [28] of the feature embeddings where it can be seen that, before unlearning, the data was largely separable by scanner and sex, and that, after unlearning, these are indistinguishable.



(a) (b)

Figure 13: a) T-SNE plot of the fully connected layer in $\mathbf{\Theta}_{repr}$ from training without unlearning, having trained on a dataset where sex correlates highly with age. It can be seen that the scanners can still be almost entirely separated, except for two data points grouped incorrectly (located within the cluster of light blue data points), showing that data from each scanner has its own distinct distribution and that the data points can also be entirely split by sex for the Biobank data. b) T-SNE plot of the fully connected layer in $\mathbf{\Theta}_{repr}$ after unlearning. It can be seen that, through the unlearning, the distributions become entirely jointly embedded in terms of both scanner and sex.



(a) (b)

Figure 14: Softmax labels for the age prediction task, averaged across all examples for the Biobank dataset showing both the true values converted into Gaussian distributions and the predicted values. The dashed line corresponds to the value if all the softmax values were equal. a) Predictions with standard learning b) Predictions after unlearning. It can be seen that after unlearning the softmax values are much closer to the dashed line and there is no trend towards the true maximum.

3.1.5. Removal of Continuous Confounds

For the removal of age as a confound, we used sex prediction as the main task. We achieved an average of 96.3% on the sex prediction task before unlearning and 95.9% after unlearning. As before, we were able to unlearn scanner information, reducing the scanner classification accuracy from 100% to 53%. Figure 14 a) shows the the averaged softmax

labels from the age prediction task with normal training, where it can be seen that there is a large degree of agreement between the true labels and the predicted labels, showing that we are able to learn age using the continuous labels and KL divergence as the loss function. We achieved MAE values of 3.26 ± 2.47 for Biobank and 4.09 ± 3.46 for OASIS with normal training in this manner.

Figure 14 b) shows the softmax labels after unlearning. It can be seen that the predicted labels no longer follow the same distribution as the true labels and that they are distributed around the random chance value indicated by the dotted line. It can also be seen that there is no trend towards the true age value, indicating that a large amount of the age information has been removed. After unlearning, the MAEs for the age prediction task were 17.54 ± 7.65 for Biobank and 22.86 ± 9.40 for OASIS. The average KL divergence has increased from 0.0022 with standard learning to 0.251 after unlearning. This therefore demonstrates that, by adapting the labels, we can also use the same training regime to remove continuous information.

3.2. Segmentation Task

3.2.1. Supervised Learning

Location of Domain Predictor	Biobank Dice	OASIS Dice	Scanner Classification Accuracy
Final Convolution (A)	0.910 ± 0.023	0.916 ± 0.021	51 (50)
Bottleneck (B)	0.871 ± 0.046	0.882 ± 0.030	100 (50)
Both (A+B)	0.903 ± 0.025	0.912 ± 0.021	55 (50)

Table 6: Dice scores comparing different locations for attaching the domain predictor in the network (as indicated in Fig. 5) A) at the final convolutional layer, B) at the bottleneck, and A+B) the combination of the two locations. The scanner classification accuracy was the accuracy achieved by a separate domain predictor using the fixed feature representation at the final convolutional layer. Random chance is given in brackets.

The results from comparing the location of the domain predictor can be seen in Table 6. We compare the Dice scores and consider the scanner classification accuracy. It can be seen that the best results for both Dice and the scanner classification accuracy were achieved with the domain predictor simply connected at the final convolution. Training with the domain predictor attached to the bottleneck was also much less stable. Given this result, the location of the domain predictor is fixed to the final convolutional layer for the rest of the experiments that follow.

The results for training on the combination of datasets with normal training and unlearning can be seen in Table 7, averaged across tissue type, and in Fig. 15 by tissue type for each scanner. It can be seen that the the unlearning process does not reduce the Dice score achieved across tissue types and that we are able to almost entirely unlearn scanner information.

Training Data	Biobank Dice	OASIS Dice	Scanner Classification Accuracy (%)
Biobank Only	0.910 ± 0.022	0.836 ± 0.043	-
OASIS Only	0.874 ± 0.032	0.917 ± 0.020	-
Both (Normal Training)	0.906 ± 0.024	0.915 ± 0.020	100 (50)
Both (Unlearning)	0.910 ± 0.023	0.916 ± 0.021	51 (50)

Table 7: Dice scores comparing unlearning to training the network on different combinations of the datasets, averaged across the tissues types. Scanner accuracy is the accuracy achieved by a domain predictor given the fixed feature representation, with random chance given in brackets.

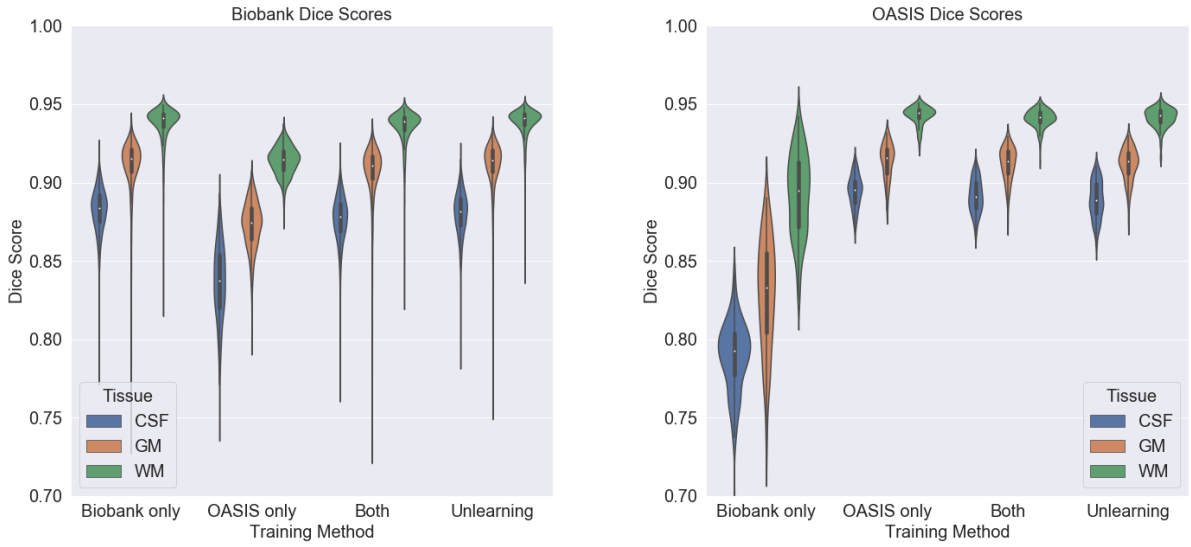
3.2.2. Semi-Supervised Segmentation

To allow us to explore the effect of training the network with low numbers of labelled training points for one of the scanners, we trained the network with both normal learning and with unlearning for different numbers of OASIS data points with labels. Only the examples with labels were used for evaluating the main task but all of the data points were still used for unlearning scanner as these loss functions do not require main task labels to be evaluated. The results can be seen in Fig. 16: for all cases, the scanner classification accuracy for normal training was 100% and between 50% and 55% for unlearning. It can be seen that unlearning gives a large improvement in the segmentations with low numbers of data points, not only in terms of the mean value but also the consistency of the segmentations, including when the network is unsupervised with regards to the OASIS dataset and so has no training labels. Even with large numbers of training examples, there is never a disadvantage to using unlearning.

4. Discussion

In this work, we have shown how a domain adaptation technique can be adapted to harmonise MRI data for a given task using CNNs, by creating a feature space that is invariant to the acquisition scanner. Removal of scanner information from the feature space prevents the information driving the prediction. The cost of the harmonisation is an approximately two fold increase in training time (from 0.88 seconds per batch to 1.53 seconds) but there is no increase in prediction time, and so, once the model is trained, there is no disadvantage to having unlearned scanner information. The duration of a batch increase is not the three times that might be expected, with the increase from one-forward and backward pass to three forward and backward passes per epoch, because the second two loss functions only update subsets of the parameters, reducing the number of calculations that must be completed.

We have demonstrated our technique for both regression and segmentation and so it should be applicable to nearly all medical imaging problems. The applicability to a wide range of networks, even those with skip connections such as the U-Net, provides a high level of flexibility. It does, however, only harmonise the features for a given task and therefore a separate network must be trained for each task. Consequently, an area for future research would be to harmonise across both scanners and tasks. As a comparison, a DANN-style



(a) (b)

Figure 15: Dice scores for the two datasets for each method broken down by tissue type. CSF = Cerebral Spinal Fluid, WM = White Matter, GM = Grey Matter

network [18], with the same architecture as our network but an additional gradient reversal layer inserted for the domain classifier branch, was also trained, but it was found that when training with only two datasets the domain predictor was unstable and predicted the domains with 7.2% accuracy (results can be seen in the supplementary material). This was a local minimum but clearly meant that the scanner information was not successfully removed. This instability in adversarial tasks posed as a minmax optimisation has also been experienced for other applications across the literature [27, 21]. By contrast, the confusion loss used in our iterative approach forces the logits towards random chance and we experience much more stability in the domain unlearning branch. We have not been able to compare to existing harmonisation methods, such as ComBat, because the different approaches are not easily compared, especially for the segmentation task, and focus on different stages of the pipeline.

The ablation study on the age prediction task shows that not only are we able to unlearn the scanner information but that there is an increase in performance that is not just due to the main task loss function. This indicates that the learned features generalise better across datasets, which is confirmed by the performance on the third dataset when only training on two. This corroborates the scanner classification accuracy and T-SNE plots, showing that scanner information has largely been removed and that the features that remain encode age information that transfers across datasets. This result is also robust to the choice of the weighting factors and so the unlearning is not sensitive to the hyper-parameters.

The results also show that we can use the unlearning scheme even when there is a strong relationship between the main task label and the acquisition scanner. This could be

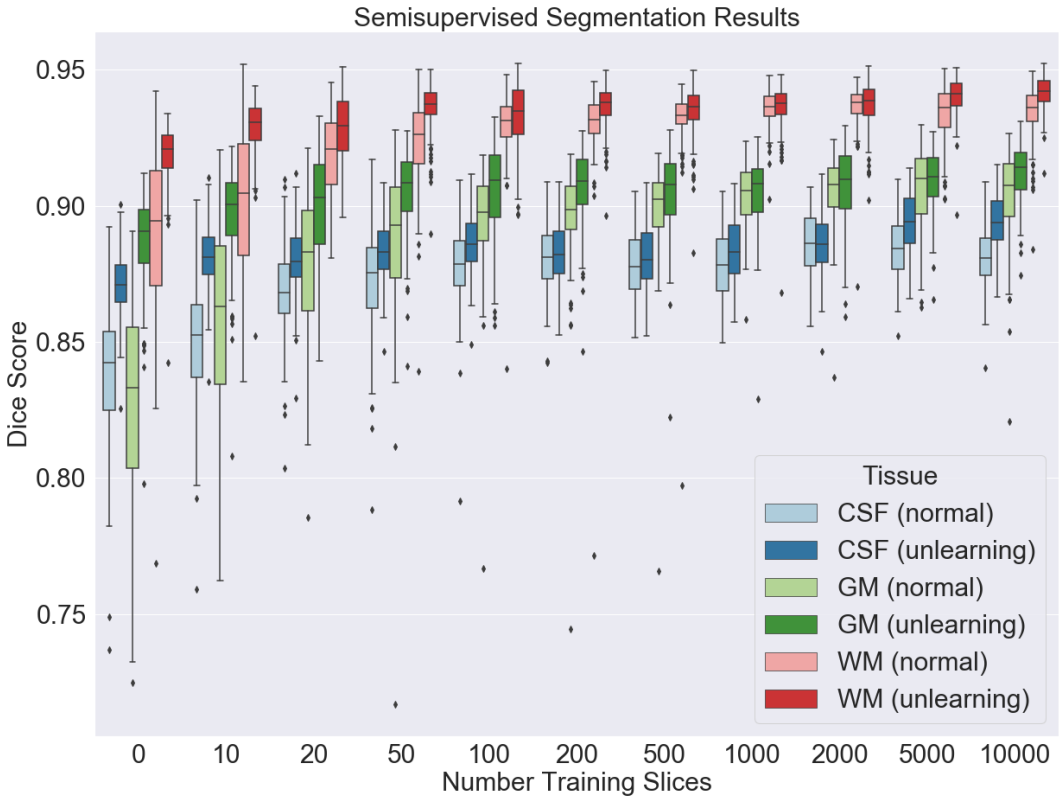


Figure 16: Dice scores for the three different tissue types for the OASIS data with increasing numbers of OASIS training slices, comparing both normal training and unlearning with the full Biobank dataset used throughout. For clarity, the x axis is not plotted to scale.

especially useful when combining data between studies with different designs and with very different numbers of subjects from each group. As we can perform unlearning on a different dataset to the main task, we have the flexibility to apply the method to a range of scenarios.

At the limit where there exists no overlap between the datasets' distributions, unlearning scanner information would be highly likely to remove information relating to the main task. This would then only be able to be solved by having an additional dataset for each scanner, acquired with the same protocol, and so represents a potential limitation of the method. The extension to the ABIDE dataset, however, shows the applicability of the method to many small datasets, and therefore this may be solvable using a set of small datasets which together span the whole distribution.

We have also shown that the approach can be extended to allow us to remove additional confounds, and have demonstrated a way that this could also be extended to allow us to remove continuous confounds such as age. For each confound to be removed, we require two additional loss functions and two additional forward and backward passes. Therefore, the

training time will increase with each additional confound, presenting a potential limitation. Were many confounds to be removed, we might also need to increase the number of times the passes for the main task are performed to prevent the performance on the main task suffering too much degradation due to the feature space being optimised for multiple tasks. We have, however, also shown how confounds can be removed even when they correlate with the main task. By carefully selecting a subset of the data with which to unlearn the scanner, we can remove the confounds, including, for instance, where sex is correlated with age, which is a case where both the main task and confound to be removed are associated with structural changes.

The approach can also be applied when no labels are available for one or more of the domains. We have demonstrated for the segmentation task that we are able to use this technique effectively when we only have labels for one domain and we would expect that this should extend to multiple unlabelled datasets.

Therefore, the presented method should be applicable for the harmonisation of data for many tasks and data scenarios. The framework is flexible and can be applied to any feedforward network. The ability to remove scanner bias influencing CNN predictions should enable the combination of small datasets and the exploration of problems for which currently there are no single-scanner datasets of adequate size to consider.

5. Data and Code Availability Statement

The data used in these experiments are available on application to the relevant studies. The code used is available at www.github.com/nkdinsdale/Unlearning_for_MRI_harmonisation and weights from training are available on request through emailing the corresponding author.

6. Acknowledgements

ND is supported by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. MJ is supported by the National Institute for Health Research (NIHR), Oxford Biomedical Research Centre (BRC), and this research was funded by the Wellcome Trust [215573/Z/19/Z]. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [203139/Z/16/Z]. AN is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme.

This research has been conducted in part using the UK Biobank Resource under Application Number 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. Analysis was carried out on the clusters at the Oxford Biomedical Research Computing (BMRC) facility and FMRIB (part of the Wellcome Centre for Integrative Neuroimaging). BMRC is a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute, supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

The computational aspects of this research were supported by the Wellcome Trust Core Award [Grant Number 203141/Z/16/Z] and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

7. References

[1] Alfaro-Almagro, F., Jenkinson, M., Bangerter, N., Andersson, J., Griffanti, L., Douaud, G., Sotiropoulos, S., Jbabdi, S., Hernandez Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D., Zhang, H., Dragonu, I., Matthews, P., Smith, S.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage **166**, 130385 (04 2017). https://doi.org/10.1016/j.neuroimage.2017.10.034

[2] Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. Workshop on Bias Estimation in Face Analytics, ECCV (09 2018)

[3] Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., : Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. bioRxiv (2020). https://doi.org/10.1101/868810

[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning **79**, 151–175 (2010)

[5] Blumberg, S., Palombo, M., Khoo, C., Tax, C., Tanno, R., Alexander, D.: Multi-stage prediction networks for data harmonization. Medical Image Computing and Computer Assisted Intervention - Conference Proceedings pp. 411–419 (10 2019)

[6] Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., Rathi, Y.: Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters. NeuroImage **184**, 180–200 (January 2019). https://doi.org/10.1016/j.neuroimage.2018.08.073

[7] Cetin Karayumak, S., Kubicki, M., Rathi, Y.: Harmonizing diffusion mri data across magnetic field strengths. Medical Image Computing and Computer Assisted Intervention - Conference Proceedings pp. 116–124 (09 2018)

[8] Chen, A.A., Beer, J.C., Tustison, N.J., Cook, P.A., Shinohara, R.T., Shou, H., : Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. bioRxiv (2020). https://doi.org/10.1101/858415

[9] Cole, J., Poudel, R., Tsagkrasoulis, D., Caan, M., Steves, C., Spector, T., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage **163** (12 2016). https://doi.org/10.1016/j.neuroimage.2017.07.059

[10] Dewey, B., Zhao, C., Reinhold, J., Carass, A., Fitzgerald, K., Sotirchos, E., Saidha, S., Oh, J., Pham, D., Calabresi, P., Zijl, P., Prince, J.: Deepharmony: A deep learning approach to contrast harmonization across scanner changes. Magnetic Resonance Imaging **64** (07 2019)

[11] Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F., Alaerts, K., Anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Birgit, E.W., Fair, D., Gallagher, L., Kennedy, D., Keown, C., Keysers, C., Milham, M.: The autism brain imaging data exchange: Towards large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19** (06 2013). https://doi.org/10.1038/mp.2013.78

[12] Filippini, N., Zsoldos, E., Haapakoski, R., Sexton, C., Mahmood, A., Allan, C., Topiwala, A., Valkanova, V., Brunner, E., Shipley, M., Auerbach, E., Moeller, S., Urbil, K., Xu, J., Yacoub, E., Andersson, J., Bijsterbosch, J., Clare, S., Griffanti, L., Ebmeier, K.: Study protocol: The whitehall ii imaging sub-study. BMC psychiatry **14**, 159 (05 2014). https://doi.org/10.1186/1471-244X-14-159

[13] Fortin, J.P., Cullen, N., Sheline, Y., Taylor, W., Aselcioglu, I., Cook, P., Adams, P., Cooper, C., Fava, M., McGrath, P., Mcinnis, M., Phillips, M., Trivedi, M., Weissman, M., Shinohara, R.: Harmonization of cortical thickness measurements across scanners and sites. NeuroImage **167** (11 2017). https://doi.org/10.1016/j.neuroimage.2017.11.024

[14] Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M., Ruparel, K., Roalf, D., Satterthwaite, T., Gur, R., Gur, R., Schultz, R., Verma, R., Shinohara, R.: Harmonization of multi-site diffusion tensor imaging data. NeuroImage **161** (08 2017). https://doi.org/10.1016/j.neuroimage.2017.08.047

[15] Franke, K., Gaser, C.: Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? Frontiers in Neurology **10** (08 2019). https://doi.org/10.3389/fneur.2019.00789

[16] Franke, K., Ziegler, G., Klöppel, S., Gaser, C.: Estimating the age of healthy subjects from t1-weighted

mri scans using kernel methods: Exploring the influence of various parameters. NeuroImage **50**, 883–92 (04 2010). https://doi.org/10.1016/j.neuroimage.2010.01.005

[17] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. International Conference on Machine Learning (09 2014)

[18] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (05 2015)

[19] Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., Mcdonald, C., Tordesillas-Gutiérrez, D., Cannon, D., Mothersill, D., Hernaus, D., Morris, D., Setién-Suero, E., Donohoe, G., Frisoni, G., Tronchin, G., Mechelli, A.: Neuroharmony: A new tool for harmonizing volumetric mri data from unseen scanners. NeuroImage p. 117127 (07 2020). https://doi.org/10.1016/j.neuroimage.2020.117127

[20] Glocker, B., Robinson, R., Coelho de Castro, D., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects (10 2019)

[21] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans (2017)

[22] Han, X., Jovicich, J., Salat, D., Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, R., Rosas, H., Makris, N., Dale, A., Dickerson, B., Fischl, B.: Reliability of mri-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. NeuroImage **32**, 180–94 (09 2006). https://doi.org/10.1016/j.neuroimage.2006.02.051

[23] Huynh, K.M., Chen, G., Wu, Y., Shen, D., Yap, P.T.: Multi-site harmonization of diffusion mri data via method of moments. IEEE transactions on medical imaging **38**(7), 1599 – 1609 (July 2019). https://doi.org/10.1109/TMI.2019.2895020

[24] Jack, C., Bernstein, M., Fox, N., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P., Whitwell, J., Ward, C., Dale, A., Felmlee, J., Gunter, J., Hill, D., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., Weiner, M.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of magnetic resonance imaging : JMRI **27**, 685–91 (05 2008). https://doi.org/10.1002/jmri.21049

[25] Johnson, W., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics **8**, 118–27 (01 2007)

[26] Jovicich, J., Czanner, S., Greve, D., Haley, E., Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., Macfall, J., Fischl, B., Dale, A.: Reliability in multi-site structural mri studies: Effects of gradient non-linearity correction on phantom and human data. NeuroImage **30**, 436–43 (05 2006). https://doi.org/10.1016/j.neuroimage.2005.09.046

[27] Lezama, J.: Overcoming the disentanglement vs reconstruction trade-off via jacobian supervision. International Conference on Learning Representations (2019)

[28] van der Maaten, L., Hinton, G.: Viualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (11 2008)

[29] Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**, 1498–507 (10 2007). https://doi.org/10.1162/jocn.2007.19.9.1498

[30] Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Kubicki, M., Westin, C.F., Shenton, M., Rathi, Y.: Inter-site and inter-scanner diffusion mri data harmonization. NeuroImage **135** (04 2016). https://doi.org/10.1016/j.neuroimage.2016.04.041

[31] Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C., Morey, R., Flashman, L., George, M., Mcallister, T., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R., Coleman, M., Kubicki, M., Rathi, Y.: Multi-site harmonization of diffusion mri data in a registration framework. Brain Imaging and Behavior **12** (02 2017). https://doi.org/10.1007/s11682-016-9670-y

[32] Mirzaalian, H., Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C.F., Shenton, M., Rathi, Y.: Harmonizing diffusion mri data across multiple sites and scanners. Medical Image Computing and Computer Assisted Intervention - Conference Proceedings **9349**, 12–19 (10 2015)

[33] Moyer, D., Ver Steeg, G., Tax, C., Thompson, P.: Scanner invariant representations for diffusion mri harmonization. Magnetic Resonance in Medicine (04 2020). https://doi.org/10.1002/mrm.28243

[34] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32 (NIPS 2019) (2019)

[35] Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I., Satterthwaite, T., Fan, Y., Launer, L., Masters, C., Maruff, P., Zhuo, C., Valzke, H., Johnson, S., Fripp, J., Koutsouleris, N., Wolf, D., Davatzikos, C.: Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage **208**, 116450 (12 2019). https://doi.org/10.1016/j.neuroimage.2019.116450

[36] Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015)

[37] Rozantsev, A., Salzmann, M., Fua, P.: Beyond sharing weights for deep domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (03 2016). https://doi.org/10.1109/TPAMI.2018.2814042

[38] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

[39] St-Jean, S., Viergever, M., Leemans, A.: Harmonization of diffusion mri datasets with adaptive dictionary learning. bioRXiv (10 2019). https://doi.org/10.1101/824706

[40] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age **12**, e1001779 (03 2015)

[41] Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. European Conference on Computer Vision (07 2016)

[42] Takao, H., Hayashi, N., Ohtomo, K.: Effect of scanner in longitudinal studies of brain volume changes. Journal of magnetic resonance imaging : JMRI **34**, 438–44 (08 2011). https://doi.org/10.1002/jmri.22636

[43] Takao, H., Hayashi, N., Ohtomo, K.: Effects of study design in multi-scanner voxel-based morphometry studies. NeuroImage **84** (08 2013). https://doi.org/10.1016/j.neuroimage.2013.08.046

[44] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. 27th International Conference on Artificial Neural Networks (2018)

[45] Torralba, A., Efros, A.: Unbiased look at dataset bias. pp. 1521 – 1528 (07 2011). https://doi.org/10.1109/CVPR.2011.5995347

[46] Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. International Conference on Computer Vision pp. 4068–4076 (12 2015)

[47] Wachinger, C., Reuter, M.: Domain adaptation for alzheimer's disease diagnostics. NeuroImage **139** (06 2016). https://doi.org/10.1016/j.neuroimage.2016.05.053

[48] Wachinger, C., Rieckmann, A., Polsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. ArXiv (02 2020)

[49] Wrobel, J., Martin, M., Shinohara, R., Goldsmith, J.: Intensity warping for multisite mri harmonization. bioRXiv (06 2019). https://doi.org/10.1101/679357

[50] Yu, M., Linn, K., Cook, P., Phillips, M., McInnis, M., Fava, M., Trivedi, M., Weissman, M., Shinohara, R., Sheline, Y.: Statistical harmonization corrects site effects in functional connectivity measurements from multisite fmri data. Human Brain Mapping **39** (07 2018). https://doi.org/10.1002/hbm.24241

[51] Zhang, Y., Brady, M., Smith, S.: Segmentation of brain mr images through a hidden markov random

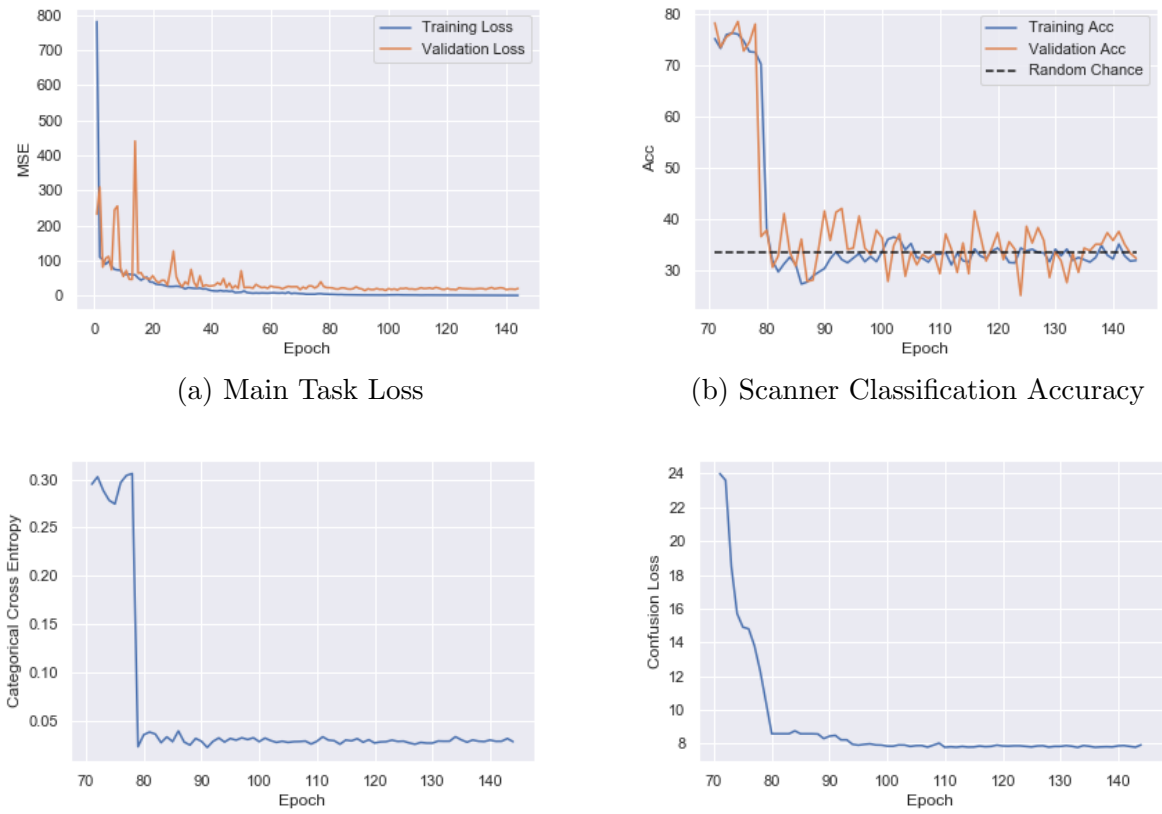
field model and the expectation maximization algorithm. IEEE transactions on medical imaging **20**, 45–57 (02 2001). https://doi.org/10.1109/42.906424

[52] Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Li, G.: Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks. Medical Image Computing and Computer Assisted Intervention - Conference Proceedings pp. 475–483 (10 2019)

[53] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2242–2251 (2017)

8. Supplementary Materials

8.1. Training Graphs



(a) Main Task Loss

(b) Scanner Classification Accuracy

(c) Domain Loss - Training

(d) Confusion Loss - Training

Figure 17: Graphs for the different loss functions from training. Only the main task loss is used during pretraining and so all other losses begin at 70 epochs when convergence on the main task is achieved.

Example training graphs from the age prediction task with three datasets. Only the main task loss (MSE for age prediction) is used during pretraining and so the other loss functions begin at epoch 70 when unlearning started. Unlearning began when convergence was achieved on the main task (patience = 15 epochs). It can be seen that beginning unlearning does not reduce the performance on the main task. It can also be seen that all the loss functions are stable throughout training.

8.2. DANN Results

A DANN-style domain adaptation approach was also tried due to its simpler implementation. As confirmed across the literature, it was unstable to train, limiting its performance. The results from training can be see in Table 8 where it can be seen that the scanner classification accuracy was much lower than random chance, indicating that scanner information remained in the feature representation. It can also be seen that the performance on the main task was reduced, with the domain adaptation leading to lower performance compared to normal training.

Training Data			Biobank MAE	OASIS MAE	Scanner Classification Accuracy %
	B	O			
Normal Training					
1.	✓	×	3.25 ± 2.36	16.50 ± 6.77	-
2.	×	✓	5.61 ± 3.52	4.27 ± 3.79	-
3.	✓	✓	3.30 ± 2.50	4.00 ± 2.78	98 (50)
DANN					
4.	✓	✓	3.35 ± 5.53	4.21 ± 6.02	7.2 (50)

Table 8: Comparing normal training on the different combinations of datasets to using DANN to remove scanner information.

8.3. p Values

Paired t-tests have been completed between experiments and the uncorrected p-values are reported separately for each dataset. For each table in the results section, the rows compared are indicated in the first column of the table.

8.3.1. Age Prediction - Basic Fully-Supervised Learning

Rows Compared	Biobank	OASIS	Whitehall
4) - 9)	0.038	0.567	0.621
5) - 10)	0.094	1.63×10^{-14}	0.001
6) - 11)	1.69×10^{-14}	0.001	1.21×10^{-10}
7) - 12)	0.002	0.042	4.61×10^{-7}
4) - 13)	0.223	0.923	0.032
5) - 14)	0.099	9.42×10^{-15}	8.62×10^{-20}
6) - 15)	4.21×10^{-23}	0.012	0.008
7) - 16)	0.145	3.25×10^{-12}	1.96×10^{-22}
9) - 13)	0.416	0.004	0.070
10) - 14)	0.893	0.034	0.041
11) - 15)	3.96×10^{-24}	1.30×10^{-4}	3.61×10^{-8}
12) - 16)	0.011	0.666	0.098

1) - 16) Biobank Only: 0.103
2) - 16) OASIS Only: 5.62×10^{-31}
3) - 16) Whitehall Only: 9.71×10^{-71}

8.3.2. Age Prediction - Balanced Datasets

Rows Compared	Biobank	OASIS	Whitehall
1) - 3)	4.65×10^{-11}	2.34×10^{-18}	2.98×10^{-10}
2) - 4)	5.67×10^{-4}	1.98×10^{-3}	7.45×10^{-6}

8.3.3. Age Prediction - Biased Datasets

Rows Compared	Biobank	OASIS
1) - 2)	5.68×10^{-89}	8.62×10^{-89}
1) - 3)	2.98×10^{-92}	5.78×10^{-93}
2) - 3)	5.63×10^{-4}	0.098
4) - 5)	3.69×10^{-65}	7.14×10^{-75}
4) - 6)	3.48×10^{-72}	9.01×10^{-83}
5) - 6)	9.72×10^{-4}	0.032
7) - 8)	6.91×10^{-42}	4.78×10^{-61}
7) - 9)	9.01×10^{-51}	1.00×10^{-63}

8) - 9)	0.045	0.029

8.3.4. Age Prediction - Sex Correlated with Scanner

Rows Compared	Biobank	OASIS
1) - 2)	0.456	0.892
1) - 3)	0.416	0.493

8.3.5. Age Prediction - Sex Correlated with Age

Rows Compared	Biobank	OASIS
1) - 2)	0.001	0.009
1) - 3)	0.752	0.042

8.3.6. Segmentation - Location of Domain Predictor

Rows Compared	Biobank	OASIS
1) - 2)	9.72×10^{-4}	3.00×10^{-3}
1) - 3)	0.003	0.091

8.3.7. Segmentation - Method Comparison

Rows Compared	Biobank	OASIS
3) - 4)	0.645	0.986

1) - 4) Biobank Only: 0.833
2) - 4) OASIS Only: 0.654