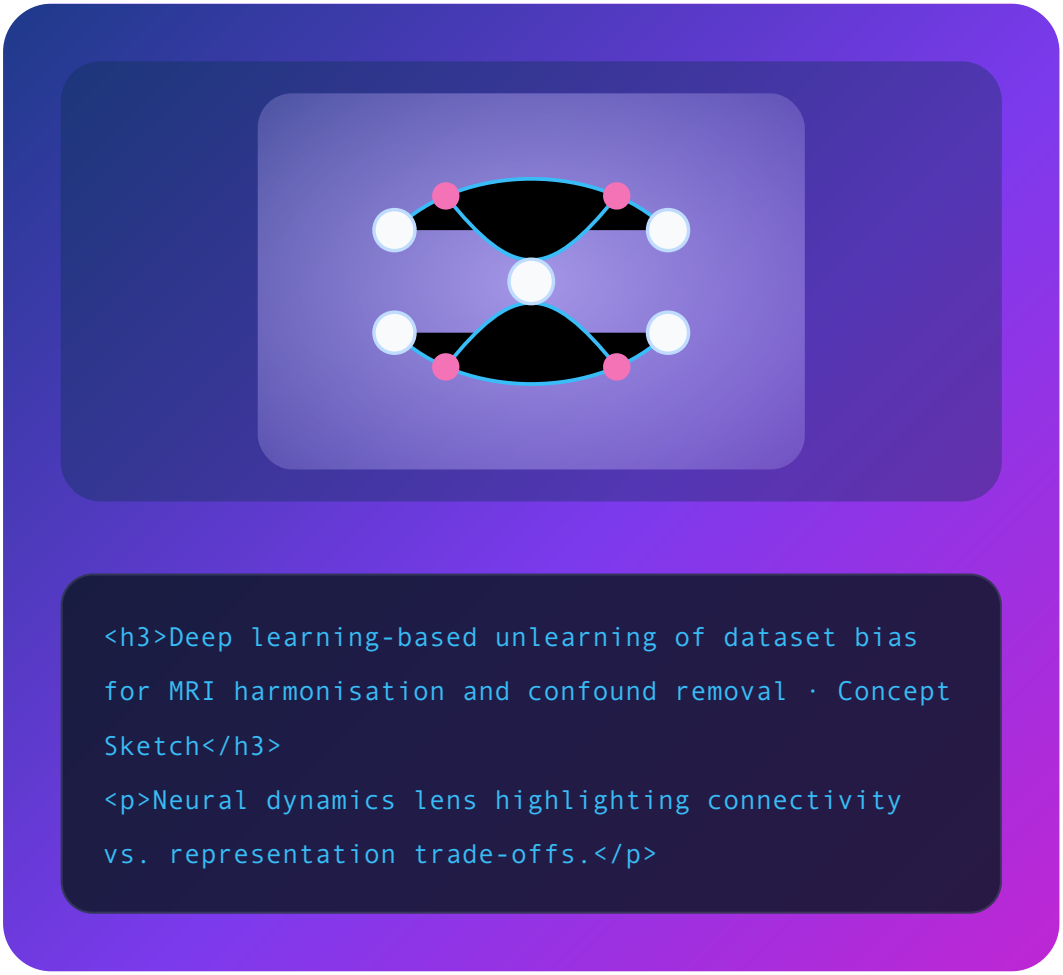

Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal

Authors: Nicola K. Dinsdale, Mark Jenkinson, Ana I.L. Namburete

Year: 2021

Venue: NeuroImage

1. Classification

- **Domain Category:**
  - Brain MRI harmonization and confound removal for multi-site neuroimaging studies.
- **FM Usage Type:**
  - Not a foundation model; proposes a **training framework for deep networks** that removes scanner and other confound information from learned representations while preserving task performance, making it highly relevant as a pre-processing and modeling strategy in FM-era pipelines.
- **Key Modalities:**
  - Structural and diffusion MRI (depending on experiment), with datasets spanning multiple scanners and acquisition protocols.

---

2. Executive Summary

This paper tackles the problem of **dataset bias in multi-site MRI**, where scanner and protocol differences introduce non-biological variability that can both confound analyses and hurt model generalization. The authors propose a **deep learning–based unlearning framework** that treats harmonization as a form of **joint domain adaptation**: networks are trained to perform a main task (e.g., age regression, tissue segmentation) while actively "unlearning" information about the scanner or other nuisance variables. The approach extends adversarial domain adaptation ideas by alternating between (1) training a domain classifier to predict scanner from latent features and (2) updating the feature extractor to **confuse** the domain classifier so that its predictions become maximally uncertain. Across experiments with multiple scanners, biased datasets, limited labels, and additional confounds, the method reduces scanner predictability while maintaining or improving main-task performance. The

framework is flexible, works with different architectures and tasks, and can be extended to remove non-scanner confounds such as sex or pathology labels when desired.

---

3. Problem Setup and Motivation

Scientific / practical problem

- Large multi-site neuroimaging datasets are essential for studying brain disorders and population variability, but combining data from different scanners introduces **non-biological variance** (scanner vendor, field strength, protocol differences).
- Traditional harmonization methods (e.g., ComBat on derived measures) partially correct these effects but often cannot operate directly in image or feature space and are less suited to modern deep learning workflows.
- Deep models trained naively on pooled multi-site data may inadvertently learn **scanner-specific shortcuts**, leading to biased predictions and poor generalization to new scanners or acquisition protocols.

Why this is hard

- **Scanner as a strong confound:** Scanner identity can be predicted very accurately from raw or preprocessed images, indicating strong domain shifts between sites.
- **Trade-off between harmonization and performance:** Removing all scanner information might also remove signal that is correlated with both scanner and the biological variable of interest, potentially harming the main task.
- **Multi-scanner, multi-task reality:** Real neuroimaging pipelines involve multiple scanners, heterogeneous labels, and sometimes missing annotations; harmonization must work in all of these regimes.

- **Confounds beyond scanner:** Other variables (e.g., sex, site, cohort, disease status in control groups) can also become entangled with the representations unless explicitly addressed.

---

4. Data and Modalities

While the paper covers several experimental setups, the overall data settings share common themes.

- **Datasets and settings (high level)**
  - Multi-site structural MRI datasets with images acquired on different scanners, possibly with different protocols and resolutions.
  - Experiments on tasks such as **age regression**, **tissue segmentation**, and other clinically relevant predictions, each involving subjects from multiple scanners.
  - Scenarios include: balanced vs biased scanner distributions, varying amounts of labeled data, and different numbers of scanners (domains).
- **Modalities**
  - Primarily **structural MRI**, but the framework is general and applicable to any modality where scanner/domain labels are available.
- **Preprocessing / representation**
  - Standard neuroimaging pipelines (intensity normalization, registration, possibly parcellation) produce images or feature maps fed into CNN-based architectures.
  - Input to the deep network is typically image patches or whole images, processed by a feature extractor followed by task-specific heads (e.g., regression head, segmentation decoder).

When precise dataset details are needed (e.g., number of subjects, exact scanner models), they should be taken from the full paper and supplementary material.

5. Model / Foundation Model

The core contribution is a **training framework** rather than a single fixed architecture. It augments standard CNNs with a domain classifier and tailored losses to create scanner-invariant yet task-relevant representations.

Model type

- Generic **feedforward CNN (or similar) backbone** for the main imaging task, augmented with:
  - A **label predictor** head for the primary task (e.g., age, segmentation).
  - A **domain classifier** head that predicts scanner identity or other confounds from the learned features.

Key components and innovations

Component	Description
Feature extractor	Shared base network that maps MRI images to latent representations used by both the main task and domain classifier.
Label predictor	Head trained with a standard task loss (e.g., regression or segmentation loss) to ensure good performance on the clinical or scientific objective.
Domain classifier	Head trained to predict scanner (or other domain labels) from the same features, making explicit how much scanner information is retained.

Component	Description
Domain loss	Categorical cross-entropy loss measuring how well the domain classifier can predict scanner, used to train the domain head given a fixed feature extractor.
Confusion loss	Loss that encourages the domain classifier's softmax outputs to be close to a **uniform distribution**, i.e., maximally uncertain about scanner, used to update the feature extractor while keeping the domain head fixed.
Alternating training scheme	Three-stage update in each batch: (1) optimize main task loss, (2) optimize domain classifier to best predict scanner, (3) optimize feature extractor to **confuse** the domain classifier using the confusion loss.

Training setup (conceptual)

- The total loss combines main task loss, domain loss, and confusion loss with weighting coefficients (,).
- Data subsets used for main-task training and for unlearning can differ (e.g., more unlabeled data for domain unlearning).
- The framework naturally extends from two domains to **multiple scanners**, and can incorporate additional confound labels (e.g., sex) as extra domain dimensions to unlearn.

6. Multimodal / Integration Aspects (If Applicable)

- The framework operates within a **single imaging modality** (MRI) but across multiple scanner domains.
- It does not perform multimodal integration in the sense of combining fundamentally different biological modalities; instead, it aligns feature distributions across acquisition conditions.
- This scanner-invariant representation can be a useful component inside broader multimodal or foundation-model pipelines that combine MRI with other data sources.

---

7. Experiments and Results

Tasks / benchmarks

- **Age regression:** Predicting subject age from MRI across multiple scanners, assessing the impact of unlearning on regression accuracy and scanner invariance.
- **Segmentation:** Tissue or structure segmentation tasks where labels are available on subsets of scanners.
- **Biased datasets and limited labels:** Experiments that intentionally skew scanner distributions or reduce labeled data to test robustness of the unlearning framework.
- **Additional confounds:** Extension to removing other confounds (e.g., site, sex) in addition to scanner.

Baselines

- Standard CNN-based models trained without any domain adaptation or unlearning (scanner information left intact).

- Classical harmonization methods (e.g., ComBat) applied to derived measures, where relevant, as conceptual comparators.
- Domain adaptation approaches such as **Domain Adversarial Neural Networks (DANNs)** relying on gradient reversal rather than confusion-based unlearning.

Key findings (trends)

- The unlearning framework substantially **reduces scanner predictability** from latent features (the domain classifier becomes near-chance when confusion loss is applied), indicating successful scanner invariance.
- Main-task performance (e.g., age prediction accuracy, segmentation quality) is **maintained or improved** compared to models trained without unlearning or with simpler domain adaptation schemes.
- The method adapts well to **biased datasets and low-label regimes**, avoiding models that overfit to the dominant scanner.
- Extending the framework to additional confounds shows that it can simultaneously reduce multiple unwanted biases while preserving the main-task signal.
- Compared to DANN-style gradient reversal, the iterative confusion-based scheme often yields more balanced and stable unlearning across scanners.

---

8. Strengths, Limitations, and Open Questions

Strengths

- Provides a clear, modular **training recipe** that can be plugged into many existing CNN architectures and tasks.
- Focuses directly on **feature-level scanner invariance**, which is closer to where deep models operate than purely statistical harmonization of derived measures.

- Demonstrates flexibility across multiple data scenarios (balanced/unbalanced, low labels, multiple scanners) and confound types.
- Bridges the literatures on domain adaptation and MRI harmonization, making it easier for practitioners to adopt robust techniques.

Limitations

- Requires explicit **domain labels** (e.g., scanner IDs) for the unlearning step; it does not handle unknown or latent domains.
- May not fully remove all scanner-related information, especially when scanner and biological variables are heavily entangled.
- If hyperparameters ((,), learning rates) are poorly tuned, the method could over- or under-correct, either leaving residual bias or harming main-task performance.
- Experiments, while diverse, focus on certain datasets and tasks; behavior on very large-scale heterogeneous cohorts or other modalities remains to be fully characterized.

Open questions / future directions

1. How robust is the unlearning framework when scaling to **dozens of scanners** and complex acquisition protocols?
2. Can similar ideas be incorporated into **large foundation models** for MRI, where scanner invariance needs to be preserved across pretraining and fine-tuning?
3. How should one choose or adapt the weights on domain vs confusion losses ((,)) in a principled, data-driven way?
4. Can unlearning be extended to **continuous confounds** (e.g., head motion, SNR, age) rather than discrete scanner categories?
5. How do we best evaluate whether important biological variation has been inadvertently removed along with nuisance variance?

9. Context and Broader Impact

- **Within MRI harmonization:** This work reframes harmonization as a **representation learning** and domain adaptation problem, moving beyond purely statistical corrections to directly controlling what deep networks remember about scanner and confounds.
- **Relation to foundation models:** The unlearning framework can be viewed as a building block for **scanner-robust feature extractors**, which is crucial when training or adapting brain FMs across many sites and cohorts.
- **Bias and fairness:** By making it possible to explicitly remove known confounds from representations, the method contributes tools for reducing certain kinds of dataset bias, though careful evaluation is needed to avoid discarding meaningful variation.
- **Practical impact:** The approach is easy to implement in modern deep learning frameworks and can be retrofitted into existing models, making it attractive for real-world neuroimaging pipelines.

---

10. Key Takeaways (Bullet Summary)

- **Problem:** Multi-site MRI datasets contain strong scanner- and protocol-induced biases that can distort analyses and hurt generalization.
- **Idea:** Treat harmonization as **joint domain adaptation**, training networks to perform main tasks while actively unlearning scanner and confound information from their latent representations.
- **Model / framework:** Standard CNN backbone plus a domain classifier and a confusion-based adversarial training loop that alternates between learning to predict scanner and learning to confuse that prediction.

- **Results:** The method reduces scanner predictability to near chance while keeping or improving performance on tasks like age regression and segmentation, even under biased or low-label conditions.
- **Impact:** Offers a flexible, architecture-agnostic recipe for building scanner-invariant MRI models, with clear relevance for large-scale neuroimaging analyses and the training of future foundation models that must work robustly across scanners and cohorts.

Generated via custom pipeline · 2025-11-26