

---

# Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects

---

Ben Glocker<sup>1</sup>, Robert Robinson<sup>1</sup>, Daniel C. Castro<sup>1</sup>, Qi Dou<sup>1</sup>, Ender Konukoglu<sup>2</sup>

<sup>1</sup> Biomedical Image Analysis Group, Imperial College London, UK

<sup>2</sup> Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

## Abstract

This is an empirical study to investigate the impact of scanner effects when using machine learning on multi-site neuroimaging data. We utilize structural T1-weighted brain MRI obtained from two different studies, Cam-CAN and UK Biobank. For the purpose of our investigation, we construct a dataset consisting of brain scans from 592 age- and sex-matched individuals, 296 subjects from each original study. Our results demonstrate that even after careful pre-processing with state-of-the-art neuroimaging pipelines a classifier can easily distinguish between the origin of the data with very high accuracy. Our analysis on the example application of sex classification suggests that current approaches to harmonize data are unable to remove scanner-specific bias leading to overly optimistic performance estimates and poor generalization. We conclude that multi-site data harmonization remains an open challenge and particular care needs to be taken when using such data with advanced machine learning methods for predictive modelling.

## 1 Motivation

Pooling data from different sites and previous studies is essential for analysis of large populations with sufficient statistical power (Smith and Nichols, 2018). However, due to differences in image acquisition, demographics, disease characteristics and other factors, naive combination of datasets for subsequent large-scale population analysis can be problematic. Here, we conduct a simple, empirical study to illustrate and highlight this problem in the context of machine learning. We are not suggesting a solution, but rather re-iterate that multi-center data harmonization is an open research challenge. For some recent attempts to tackle this problem, see for example (Fortin et al., 2017, 2018).

## 2 Data

We construct an age- and sex-matched dataset with T1-weighted brain MRI from  $n = 592$  individuals, where 296 subjects (146 females) are taken each from the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN)<sup>1</sup> (Shafto et al., 2014; Taylor et al., 2017) and UK Biobank imaging study (UKBB)<sup>2</sup> (Sudlow et al., 2015; Miller et al., 2016; Alfaro-Almagro et al., 2018). This is to simulate a somewhat ‘best case scenario’ for multi-site data where the age- and sex-matching intends to remove population bias. We note this is rarely possible in practice, and it is expected that current and previous analyses that pool data from different sites suffer from much larger site-specific biases.

**Cam-CAN:** All images were collected at a single site (Medical Research Council Cognition and Brain Sciences Unit (MRC-CBSU) in Cambridge, UK) using a 3T Siemens TIM Trio scanner with a 32-channel receive head coil. Imaging parameters are: 3D MPRAGE, TR=2250ms, TE=2.99ms, TI=900ms; FA=9 deg; FOV=256x240x192mm; 1mm isotropic; GRAPPA=2; TA=4mins 32s.

---

<sup>1</sup><http://www.cam-can.org/>

<sup>2</sup><http://www.ukbiobank.ac.uk/>

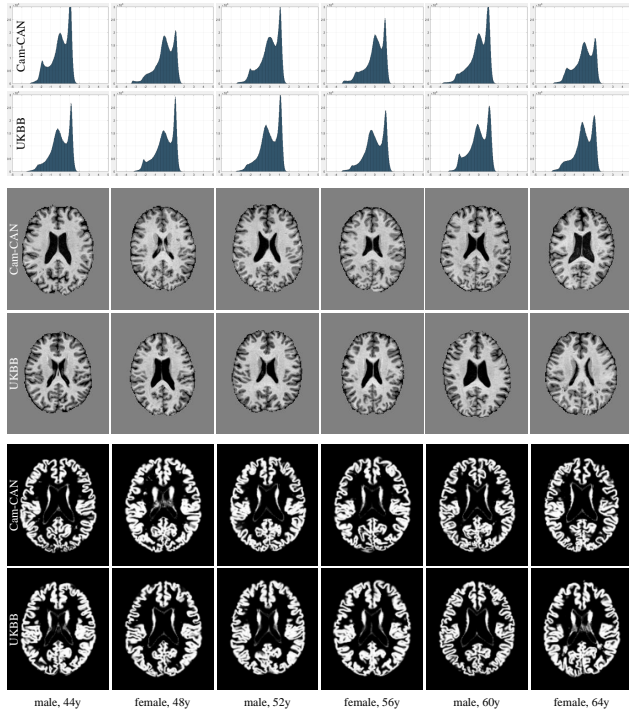


Figure 1: Example data for six age- and sex-matched subjects from the Cam-CAN and UKBB datasets after applying different pre-processing steps. Top two rows show the intensity histograms after skull-stripping, bias field correction, rigid registration to MNI, and whitening for intensity normalization. Rows three and four show the corresponding T1-weighted mid axial slices. Rows five and six show the spatially normalized graymatter maps obtained with SPM12. Site-specific differences are non-obvious from visual inspection.

**UK Biobank:** All images were collected at the UKBB imaging center using a 3T Siemens Skyra scanner with a 32-channel receive head coil. Imaging parameters are: 3D MPRAGE, R=2, TR=2000ms, TE=385ms, TI=880ms; FOV=208x256x256mm; 1mm isotropic; Duration 4mins 54s.

The acquisition protocols of the two studies are remarkably similar, and possibly much closer than typically found when pooling data from multiple sites. The subjects in both studies should be normal.

## 2.1 Pre-Processing Pipeline

We aimed at designing a common state-of-the-art pre-processing pipeline which in this or similar form is widely used in neuroimaging studies. In particular, we apply the following sequential steps: 1) Lossless image reorientation by swapping axes using the direction information from the NIfTI image header, such that all scans are in the same radiological orientation of left, posterior, superior; 2) Skull stripping with ROBEX v1.2<sup>3</sup> (Iglesias et al., 2011); 3) Bias field correction with N4ITK<sup>4</sup> (Tustison et al., 2010); 4) Intensity-based linear registration (rigid and affine) to MNI ICBM 152 2009a Nonlinear Symmetric<sup>5</sup> using an in-house registration tool with correlation coefficient as the similarity measure and downhill-simplex as the optimizer.

After these steps, we perform intensity normalization within brain regions with simple whitening (zero-mean/unit-variance). Voxels outside the brain are set to fixed value. Other techniques such as percentile matching and Nyul’s histogram standardization (Nyúl et al., 2000) led to similar subsequent observations. We also employ SPM12<sup>6</sup> (Friston et al., 2007; Ashburner, 2012) and FMRIB’s Automated Segmentation Tool (FAST) v4.0<sup>7</sup> (Zhang et al., 2001) to obtain brain tissue probability maps. SPM is run directly on the raw T1-weighted scans as it has its own pre-processing pipeline built-in including spation non-linear normalization to MNI space. FSL-FAST is run on our skull-stripped, bias field corrected and rigidly MNI aligned images.

<sup>3</sup><https://www.nitrc.org/projects/robex>

<sup>4</sup><https://itk.org>

<sup>5</sup><http://nist.mni.mcgill.ca/?p=904>

<sup>6</sup><http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

<sup>7</sup><https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST>

Table 1: Two-fold cross validation results for site classification. Reported are overall accuracy, average entropy, and average predictive probability. If the data were indistinguishable one would expect an accuracy of 50%, an entropy of 0.6931 (upper bound), and a probability of 0.5.

Stripped	Bias Field	Aligned	Intensities	Accuracy	Avg. Entropy	Avg. Prob.
✓	✓	rigid	whitening	96.96%	0.4039	0.8296
✓	✓	affine	whitening	98.82%	0.3876	0.8397
SPM12 – Segment				Accuracy	Avg. Entropy	Avg. Prob.
✗	✓	rigid	graymatter	80.24%	0.6363	0.6399
✗	✓	non-linear	graymatter	96.62%	0.5675	0.7234
FSL – FAST				Accuracy	Avg. Entropy	Avg. Prob.
✓	✓	rigid	graymatter	93.24%	0.4542	0.7968

Table 2: Two-fold cross validation results for sex classification under different data arrangements.

Data Arrangement	Aligned	Accuracy	Avg. Entropy	Avg. Prob.
Multi-site age/sex-matched	rigid	82.60%	0.5304	0.7388
Single-site (Cam-CAN)	rigid	81.42%	0.5592	0.7179
Single-site (UKBB)	rigid	84.46%	0.5049	0.7572
Cam-CAN females / UKBB males	rigid	94.59%	0.4036	0.8311
Cam-CAN 80/20% / UKBB 20/80%	rigid	85.87%	0.5038	0.7616
Cam-CAN train / UKBB test	rigid	81.42%	0.5617	0.7124
UKBB train / Cam-CAN test	rigid	78.04%	0.5284	0.7419
Multi-site age/sex-matched	affine	79.73%	0.6345	0.6389
Single-site (Cam-CAN)	affine	77.70%	0.6439	0.6269
Single-site (UKBB)	affine	81.08%	0.6393	0.6316
Cam-CAN females / UKBB males	affine	98.99%	0.4641	0.8013
Cam-CAN 80/20% / UKBB 20/80%	affine	84.78%	0.5713	0.7125
Cam-CAN train / UKBB test	affine	73.65%	0.6462	0.6245
UKBB train / Cam-CAN test	affine	62.16%	0.6075	0.6769

### 3 Experiments, Results & Conclusion

We conduct two image classification experiments to illustrate the impact of scanner effects which remain after careful pre-processing and are present even in image-derived tissue probability maps.

**Site classification:** We train random forest binary classifiers to distinguish between the origin of the imaging data. The classifiers are trained to distinguish between data from Cam-CAN and UKBB.

Results are summarized in Table 1. We make the following observations: i) classifiers are able to predict data origin with high accuracy; ii) scanner effects remain in derived tissue probability maps; iii) higher degrees of spatial normalization amplify scanner effects (possibly related to interpolation).

**Sex classification:** We consider a simple binary classification task of sex classification. We compare results of training random forest classifiers on single-site and multi-site data.

Results for sex classification are summarized in Table 2. We make the following observations: i) age/sex-matched multi-site data gives realistic estimates of accuracy (similar to single site); ii) sex imbalance in multi-site leads to overly optimistic accuracy; iii) training on one site and testing on the other shows drop of performance indicating poor generalization; iv) when discriminative features such as brain size are removed by affine registration, the drop in performance is more severe.

**Conclusions:** Scanner effects can be subtle yet significantly affect machine learning. Similar findings for multi-site neuroimaging data are reported in (Ferrari et al., 2018; Wachinger et al., 2019).

## Acknowledgements

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG). UK Biobank data has been accessed under Application Number 12579.

## References

- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L., and Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424.
- Ashburner, J. (2012). SPM: a history. *NeuroImage*, 62(2):791–800.
- Ferrari, E., Bosco, P., Spera, G., Fantacci, M. E., and Retico, A. (2018). Common pitfalls in machine learning applications to multi-center data: tests on the ABIDE i and ABIDE ii collections. In *Joint Annual Meeting ISMRM-ESMRMB*.
- Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.
- Fortin, J.-P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170.
- Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.
- Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634.
- Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*.
- Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150.
- Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T., et al. (2014). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14(1):204.
- Smith, S. M. and Nichols, T. E. (2018). Statistical challenges in “big data” human neuroimaging. *Neuron*, 97(2):263–268.
- 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., and Collins, R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3).
- Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., Tyler, L. K., Henson, R. N., et al. (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269.

- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.
- Wachinger, C., Becker, B. G., Rieckmann, A., and Pölsterl, S. (2019). Quantifying confounding bias in neuroimaging datasets with causal inference. *arXiv preprint arXiv:1907.04102*.
- Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57.