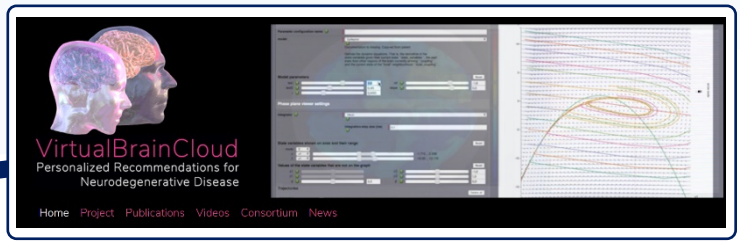
VirtualBrainCloud

Personalized Recommendations for Neurodegenerative Disease



www.VirtualBrainCloud-2020.eu

Public deliverable report

D3.17 “Publication of a study on deconfounding for inference”

Date	November 2021
Authors	Darya Chyzyk, Bertrand Thirion, Gael Varoquaux [Inria]
Dissemination level	**public**
Website	www.VirtualBrainCloud-2020.eu

This project has received funding from the **European Union´s Horizon 2020** research and innovation programme under **grant agreement No 826421**

Table of content

1. Introduction

Predictive models, using machine learning, are becoming a standard tool for scientific inference. In cognitive neuroscience, they can be used for decoding, to conclude on mental processes given observed brain activity. With the rise of large-scale brain-imaging cohorts, they can extract imaging biomarkers that predict across subjects phenotypes such as neuropsychiatric conditions or individual traits.

A crucial aspect of these biomarkers is their ability to predict the outcome of interest, ie generalize to new data. However, these predictions can be driven by confounding effects. Such effects affect both the brain-imaging data and the prediction target but are considered as irrelevant. For instance, subjects' in-scanner motion has been shown to severely affect the link between brain-imaging signals and their age: in-scanner motion varies with subjects' age and it creates systematic differences in brain signals. Given this confounding effect, MRI biomarkers of brain aging may be nothing more than expensive measurements of head motion. Other examples may be more subtle: brain imaging reflects age quite accurately, and age matters for diagnosing Alzheimer's disease, yet an important question is whether brain imaging yields an accurate diagnosis of Alzheimer disease beyond the mere effect of age.

More generally, the data at hand often capture effects not of direct interest to the investigation. In many situations, some confounds such as head motion cannot be fully avoided. To make matters worse, large cohorts developed in population imaging to answer epidemiological questions as UK biobank are observational data: there is no controlled intervention or balanced case-control group; rather individuals are recruited from diverse populations with various sampling or selection biases. To conclude on the practical use of biomarkers, it is important to control that their predictions are not fully driven by such unwanted effects. This requires measuring model predictive accuracy after controlling for nuisance variables. Confounding effects can also make it hard to interpret brain-behavior relationships revealed by predictive models, as confounds can mediate the observed association or be a latent common cause to observations.

In experimental settings, eg as in a small cohort, it can be suppressed by balancing the acquisition for confounds, or using randomized control trials. However, constraints in the data acquisition, eg recruitment of a large cohort, often imply that confounds are present in the data, and a suitable analysis is needed to avoid reaching erroneous conclusions. The statistical literature on controlling confounding variables is well developed for classic statistical analysis, such as statistical testing in a linear model at the heart of the standard mass-univariate brain mapping. However, these procedures need to be adapted to high-dimensional predictive-modeling settings, where the focus is to achieve high-prediction accuracy based on imaging data. Indeed, predictive models do not rely on the same parametric assumptions, namely linearity of effects and Gaussian noise. Often, a predictive analysis does not build on a generative model of the signal but on optimizing discrimination. In addition, predictive models draw their purpose and validity from out-of-sample prediction, rather than in-sample statistical testing. The question tackled here is thus whether one can assess the predictive accuracy of brain measurements free of unwanted confounds. It is not to identify treatment effects size nor to perform other types of causal inference.

In this work, we study statistical tools to control for confounding effects in predictive models. We consider that practitioners should primarily avoid or reduce the impact of confounds on their model, but this is not always feasible or hard to check, hence, we choose to put the emphasis on the unbiased evaluation of models even in the presence of confounds. A preliminary version of the work discussed here was presented at the PRNI conference. While the core method is the same, its presents limited insights on the theoretical underpinnings and practical value of the method proposed. Experiments on simulated data are absent and experiments on neuroimaging data are limited to just one data set. In

particular, statistical significance is not established thoroughly, and only one alternative approach is considered. In short, the conference publication provides limited insights on the method, while the current work provides a complete description and points to the code for reuse.
We first review how the classic deconfounding procedures can be used in predictive-modeling settings, i.e. together with cross-validation. We then expose a complementary approach that is not based on removing confounding effects, but rather testing whether a given predictive model – e.g. a biomarker– predicts well when these confounds are not present. For this we introduce the confound-isolating cross-validation method, that consists in sampling test sets in which the effect of interest is independent from the confounding effect. The benefits of this approach are that it is non-parametric and that it directly tests the quantity of interest in a predictive analysis. We then run an extensive empirical study on three population-imaging biomarker extraction problems as well as simulations. We draw practical recommendations to test predictive models in the presence of confounding effects.

2. Partners involved

Inria

3. Description of work performed

Here we study how to adapt methods that control for confounds in statistical analyses to predictive modeling settings. We review how to train predictors that are not driven by such spurious effects. We also show how to measure the unbiased predictive accuracy of these biomarkers, based on a confounded dataset. For this purpose, cross-validation must be modified to account for the nuisance effect. To guide understanding and practical recommendations, we apply various strategies to assess predictive models in the presence of confounds on simulated data and population brain imaging settings. Theoretical and empirical studies show that deconfounding should not be applied to the train and test data jointly: modeling the effect of confounds, on the train data only, should instead be decoupled from removing confounds. Cross-validation that isolates nuisance effects gives an additional piece of information: confound-free prediction accuracy.

4. Results

Confound-isolating cross-validation

Figure 1 displays the evolution of the association between confound and target during Confound-isolating cross-validation in the CamCan dataset, predicting Fluid Intelligence with Age as a confound. In the full dataset, comprising 608 subjects, the correlation between confound and target is $\rho = -0.67$. Iterating the algorithm to remove half of the subjects leads to $\rho = -0.17$.
The final test set contains 1/5 of the subjects and achieves $\rho = -0.07$, showing that it indeed cancels the dependency between aging and motion. The joint distribution between target and confound displayed in Figure 1 shows that the initial statistical dependency between these two variables vanishes after a few tens of iterations of the algorithm.

Results on CamCan and UKBB datasets

Figure 2 reports the mean absolute error for the different approaches to control for confounds. The figure also reports the p-value of predictive accuracy, from permutations. The first thing to note is that without controlling for confounding effects, all models lead to significant prediction. But are these driven by the confounds? Given that the various approaches measure predictions on different data, we compare how far these predictions are above chance, rather than their absolute value.

Deconfounding test and train sets jointly –removing the linear effect of the confounding variable on the full data– has little impact on the prediction performance on all datasets. On the other hand, out-of-sample deconfounding changes significantly prediction performance in a way that varies across tasks. Prediction accuracy of fluid intelligence on CamCan falls to chance level. Age prediction on CamCan is little impacted. However, Age prediction on UKBB gives results worse than chance, i.e. worse than a model that learns to predict age on data where this relationship has been shuffled by permutation.

Confound-isolating cross-validation also gives varying results on different datasets. For fluid-intelligence prediction on Cam-Can, it also gives results at chance level. For age prediction on CamCan, it does alter significantly prediction accuracy, and on UKBB, it leads to a slightly worse prediction, but still above chance. Finally, Prediction from confounds leads to chance-level or good prediction of the target depending on the dataset. In particular, it does better than chance for Fluid intelligence prediction.

These results show that in all these datasets, the confounds z are associated with both the data X and the target y. For fluid intelligence prediction on CamCan, all the prediction of y from X is mediated by z. However, for age prediction in CamCan, there exists within X some signal that is unrelated to z but predicts y. Age prediction in UKBB is a more subtle situation: X contains signals from z and y with shared variance, but there is enough signal beyond the effect of z to achieve a good prediction, as demonstrated by confound-isolating cross-validation, where the prediction cannot be driven by z. Yet, out-of-sample deconfounding removes the shared variance and hence creates predictions that are worse than chance.

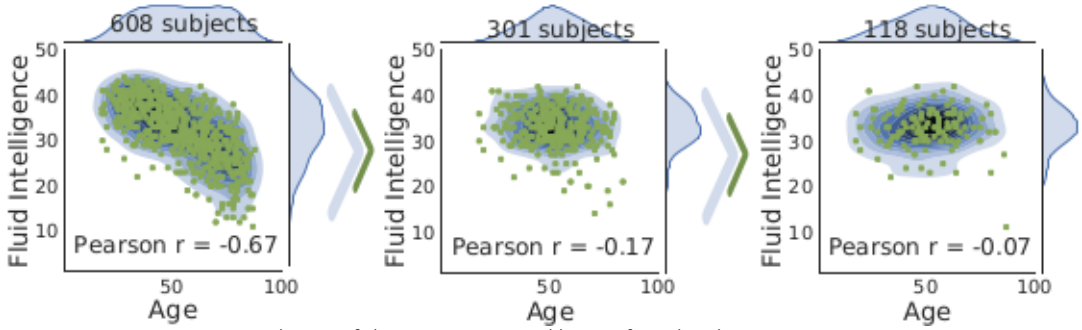


Figure 1: Evolution of the test set created by Confound-isolating cross-validation.

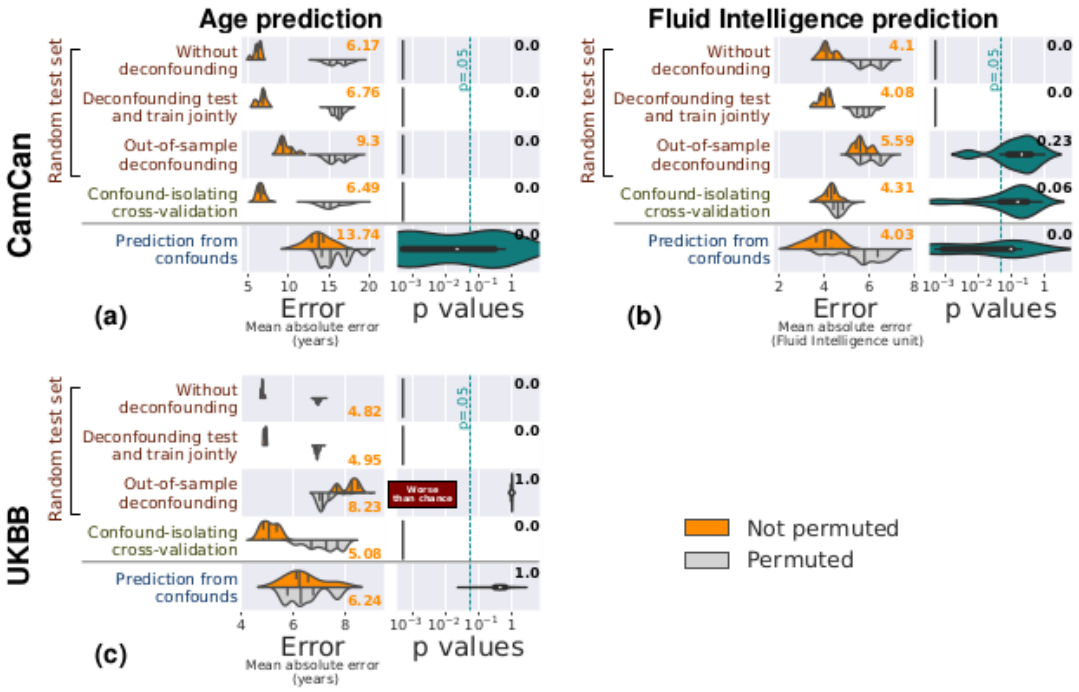


Figure 2: Comparisons on population-imaging data

5. Conclusion, next steps

Deconfounding strives to remove confounding effects from the data, after which successful prediction can be interpreted as a direct link from the remaining brain signals to the outcome of interest. However, in biomarkers settings, the primary focus may be on the quality of detection, rather than interpretation, for instance to improve diagnosis or prognosis. In such settings, an important question is: how much do the brain signals improve the prediction upon a simpler measure of the confounding effect? Answering this question calls for a cross-validation procedure isolating this confounding effect. The corresponding prediction accuracy can then safely be interpreted as not resulting in any way from the confounding effect.

6. References

Controlling a confound in predictive models with a test set minimizing its effect Darya Chyzhyk, Gaël Varoquaux, Bertrand Thirion, Michael Milham. *PRNI 2018 – 8th International Workshop on Pattern Recognition in Neuroimaging*, Jun 2018, Singapore, Singapore. pp.1-4

How to remove or control confounds in predictive models, with applications to brain biomarkers, Darya Chyzhyk, Gaël Varoquaux, Michael Milham, Bertrand Thirion. GigaScience in Press.