Pitfalls related to computer-aided diagnosis system learned from multiple databases

Hugo Touvron, Sylvain Faisan, Florian Tilquin, Vincent Noblet

PITFALLS RELATED TO COMPUTER-AIDED DIAGNOSIS SYSTEM LEARNED FROM MULTIPLE DATABASES

Hugo Touvron, Sylvain Faisan, Florian Tilquin, Vincent Noblet

ICube UMR 7357, Strasbourg University, CNRS, FMTS, Strasbourg, France

ABSTRACT

The growing availability of large neuroimaging databases offers exceptional opportunities to train more and more efficient machine learning algorithms. Nevertheless, these databases may be prone to several sources of variability (e.g., age, gender, acquisition parameters,...). These nuisance variables can hamper the performance of a classification method and can even lead to misinterpret its behavior. We focus in this paper on how to account for data coming from different databases. First, we present experiments on simulated data that illustrate how interactions with other confounds such as age can be problematic for the adjustment of data from multiple databases. Then, we compare three strategies to adjust data and evaluate them in the real scenario of a Computer-Aided Diagnosis (CAD) system that discriminates healthy from Alzheimer's Disease (AD) subjects based on volumetric characteristics derived from structural MRI.

***Index Terms*—** Computer-aided diagnosis, nuisance variables, classification and regression

1. INTRODUCTION

The growing availability of large neuroimaging databases offers a unique opportunity to collect considerable amount of data to train artificial intelligence algorithms. A side effect is that it inevitably leads to the introduction of new biases related especially to heterogeneous protocol guidelines in terms of subject enrollment and acquisition parameters [1, 2]. In this paper, we investigate several strategies to account for data coming from different databases. This is exemplified through the design of a CAD system that discriminates healthy from Alzheimer's Disease (AD) subjects based on volumetric characteristics derived from structural MRI.

A *nuisance variable* is an external factor which is not of interest for the study, but which causes an increase in variability within groups. In our case, nuisance variables can be database, age, gender and Estimated Total Intracranial Volume (ETIV). A particular subcategory of nuisance variables are the *confounding variables* (or *confounds*) which also have a direct effect on the target variable (*i.e.*, the diagnosis for a CAD system). For instance, gender is a confounding variable for AD diagnosis since almost two-thirds of the individuals diagnosed with AD are women [3]. We also define the concept of *artificial confound*, which is a nuisance variable that should theoretically not have an effect on the target variable, but which in practice has one because it has been sampled non uniformly. For instance, the database should not theoretically have an impact on AD diagnosis, but since the ratio of AD *vs* healthy subjects can be very different according to the database, an artificial link is created. Note that such artificial confounds are as far as possible avoided in prospective study by collecting data stratified into groups that have the same distribution for the confounds. Nonetheless, it may frequently occur when merging several databases that may have heterogeneous recruitment schemes in terms of age distribution, gender ratio, or pathology prevalence.

Particular attention should be paid to both *natural* and *artificial* confounds since they can introduce a positive bias into the performance of a classifier and can lead to misinterpret its behavior. Several papers already addressed how to handle confounds such as age and gender in neuroimaging study [4, 5, 6]. Here, we investigate more specifically how to account for data coming from multiple databases and the interplay that occurs with other confounds such as age and gender. First, we present experiments on simulated data that illustrate how interactions with other confounds such as age can make the handling of data coming from multiple databases an intricate problem. Then, we compare three strategies to adjust data and evaluate them in the real scenario of a CAD system dedicated to Alzheimer's disease.

2. INTERACTIONS BETWEEN DATABASE EFFECT AND A CONTINUOUS CONFOUND

2.1. Data simulation and adjustment procedure

For simplicity, we consider a single feature x that is influenced by both age and database effect. We denote by x_a the age of x ($-1 \leq x_a \leq 1$). x is assumed to come from two databases (*DB1* and *DB2*) that are modeled by two binary variables x_{c_1} and x_{c_2} ($x^j_{c_i} = 1$ if subject j belongs to *DBi*, 0 otherwise). The generative model of feature x is expressed as $x = f(x_a) + x_{c_1} - x_{c_2} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.01^2)$. The terms x_{c_i} represent a bias in the measure and $f(x_a)$ represents the influence of age.

We consider two hypotheses for modeling the effect of age: a linear function $f(x_a) = -x_a$, and a nonlinear one $f(x_a) = 4 - (x_a + 1)^2$. We also investigate two scenarios concerning the distribution of ages within each database: either Similar Database (*SD*) age distribution (x_a is drawn for both databases and each subject following a uniform distribution $\mathcal{U}$(-1,1)) or Dissimilar Database (*DD*) age distribution (x_a is drawn following $\mathcal{U}$(-1,1) for the first database *DB1* and following $\mathcal{U}$(-0.2,0.2) for the second one *DB2*). In order to highlight only the potential biases induced by model misspecification without being impacted by estimation inaccuracies in the regression parameters, we consider a large number of subjects (10^6 for each database).

These experiments aim at simulating a given population (e.g., healthy subjects) acquired in two different studies and recruited according to different age sampling schemes. Under these conditions, the distributions of the observational measures from each database are different. The adjustment procedure should ideally make the two distributions similar, so that all data can be merged to efficiently learn a CAD system.

We evaluate here the most commonly used way to suppress the effect of nuisance variables, namely to regress them out from the features of interest through a multilinear model. The confound effects are modeled using the generalized linear model with a design matrix composed of three columns: the first one corresponds to the age, and the two others encode the database belonging. The parameter vector $\boldsymbol{\beta} = [\beta_a, \beta_1, \beta_2]^T$ is estimated with a least square fit using half of the data. Then, the effects of the confounds are removed for the rest of the data. The adjusted value denoted x_c is given by: $x_c = x - \beta_a x_a - \beta_1 x_{c_1} - \beta_2 x_{c_2}$. The distribution of x_c is plotted in Fig. 1 for each database under four different conditions.

2.2. Results

In the case of linear influence of age, the fitted model corresponds exactly to the data generating process: the nuisance variable effect can perfectly be regressed out. Consequently, the distributions of the adjusted data x_c of the two databases are identical and reflect the *true* population variability (*i.e.*, $x_c \sim \mathcal{N}(0, 0.01^2)$) (Fig. 1(a)). The same behavior is observed for both *SD* and *DD* scenarios.

When considering the nonlinear influence of age, the fitted model is misspecified and the effect of age cannot be regressed out properly. For the *SD* case (Fig. 1(b)), the distributions of x_c for each database are no more reflecting the true population variability (the variability is higher), but the adjusted data of the two databases have still similar distributions. Consequently, although the age effect is not regressed out properly, the database effect is successfully corrected. This conclusion is no more true for the *DD* case (Fig. 1(c)), which leads to different distributions of the adjusted data for each database.

For the *SD* case, since the two databases share the same distribution of ages, the distribution of the non-adjusted features x of *DB1* is just a translated version of the distribution of *DB2*. This translation (*i.e.*, the database effect) can be perfectly captured by the regression model even if the shape of the distributions (*i.e.*, the age effect) is not modeled properly. For the *DD* case (Fig. 1(c)), the distribution of ages varies across the databases: the distribution of the non-adjusted features x of *DB1* is no more just a translated version of the distribution of *DB2*. Consequently, the effects of database and age cannot be properly disentangled to explain the shape variation of the distributions. It may be surprising to notice on Fig. 1(c) that the distribution of *DB2* seems to reflect the *true* population variability (see Fig. 1(a)). Indeed, the nonlinear function that models the effect of age is almost linear on the age interval $[-0.2, 0.2]$, thus explaining why it has been well regressed out for *DB2*.

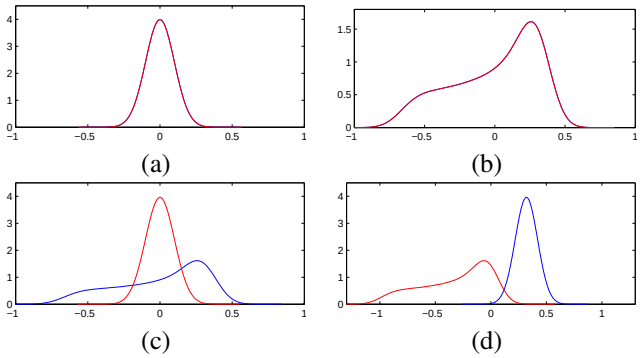

Fig. 1. Distributions of adjusted data x_c for *DB1* (blue) and *DB2* (red) (see section 2.1 for details).

Note finally that, when the effect of age cannot be modeled properly, the worst case appears when the learned model is applied on a population whose distribution of ages significantly differs from the distribution of the training population. In Fig. 1(d), we use the model learned in the last experiment (Fig. 1(c): nonlinear model of age, *DD* case) and we apply it to a population generated with the same model except that x_a is now drawn following $\mathcal{U}$(-0.2,0.2) for *DB1* and following $\mathcal{U}$(-1,1) for *DB2* (distributions have been switched between databases for the testing phase). In addition to the shape variation, a shift between the distributions of the two databases is also observed in that case.

To conclude, these experiments reveal that the database effect cannot be properly regressed out if the effect of another continuous confound, whose distribution varies across databases, is not properly modeled.

3. AD VS CONTROLS DISCRIMINATION

We focus here on a CAD system learned on multiple databases that discriminates healthy from AD subjects based on volumetric characteristics derived from structural MRI. Three

	Adni1.5T	Adni3T	OASIS_CS	OASIS_LGT	IXI1.5T	IXI3T	total
n (MA)	102/97	14/24	15/26	32/23	0	0	163/170
age (MA)	75.6 ± 7.6	74.3 ±8.3	76.5 ± 7.3	75.4 ± 6.9			75.5(7.6)
n (controls)	120/111	20/36	30/88	23/54	70/117	33/46	296/452
age (controls)	76.0 ± 4.9	75.2 ± 4.8	71.8±11.5	75.7 ± 8.1	63.8±7.8	62.0±6.5	70.7 ± 9.5

Table 1. Demographic data. Number of subjects (n) is given as males/females. Age is given as mean ± standard deviation.

	S1	*S2*	*S3*
age	0.38 ± 0.02 / 0.41 ± 0.01	0.95 ± 0.04 / 0.59 ± 0.07	1.00 ± 0.04 / 0.82 ± 0.08
ETIV	0.11 ± 0.01 / 0.14 ± 0.01	0.96 ± 0.04 / 0.90 ± 0.03	0.97 ± 0.04 / 0.96 ± 0.05
gender	0.50 ± 0.02 / 0.74 ± 0.03	0.49 ± 0.01 / 0.62 ± 0.02	0.50 ± 0.01 / 0.62 ± 0.03
database	0.69 ± 0.06 / 0.67 ± 0.04	0.34 ± 0.05 / 0.46 ± 0.04	0.17 ± 0.03 / 0.34 ± 0.04
diag-std	0.89 ± 0.02 / 0.89 ± 0.03	0.89 ± 0.02 / 0.88 ± 0.03	0.90 ± 0.02 / 0.87 ± 0.02
diag-biased	0.71 ± 0.05 / 0.69 ± 0.05	0.84 ± 0.04 / 0.83 ± 0.04	0.84 ± 0.04 / 0.84 ± 0.05

Table 2. Prediction of the confounds and of the diagnosis for the three strategies. Each result is given as mean FUV or BAC ± standard deviation. The first score is obtained with a linear method (LSVM or LASSO) and the second with a nonlinear one (GBDT or RF).

public databases are considered: OASIS [7, 8], ADNI [9] and IXI (http://brain-development.org/ixi-dataset/). The IXI and ADNI datasets have been split into two parts, one with 1.5T MRI, and the other one with 3T MRI. We extracted 116 volumetric characteristics from structural MRI using Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). As it can be observed in Tab1, age distribution, gender ratio and pathology prevalence vary across databases.

The objective is to design a CAD system that does not exploit any prior knowledge about age or gender (natural confounds) in order to evaluate the potential for discrimination related to brain morphology. We would like not to account for global shape difference of the head, which induces unwilling variability. To this end, the Estimated Total Intracranial Volume (ETIV) will be considered as a nuisance variable. Finally, we also want the CAD system not to be biased by any database effect (artificial confound).

To achieve these specifications, we compare three strategies to adjust the data from confounds:

- *S1*: no data adjustment (baseline method),
- *S2*: effects of database, age, gender and ETIV are regressed out from each feature independently using the generalized linear model as described in Sec. 2.1,
- *S3*: data are adjusted using *S2* and then each feature is linearly transformed so that its distribution for each database has similar median and Median Absolute Deviation (MAD) .

We propose the *S3* strategy in order to compensate for potential residual variation in position and shape between the distributions of adjusted data (see experiments on simulated data, Fig. 1(c-d)). Since we use a regularized linear classifier, each feature has finally been scaled to be zero-mean and unit variance for the three strategies.

To compare the three strategies, we first investigate if all the effects are correctly removed from the data. This is done by evaluating the ability to predict each of the confound from the adjusted data (Sec. 3.1). Then, we evaluate the impact on a CAD system dedicated to AD *vs* controls discrimination (Sec. 3.2).

3.1. Confound predictability

In this section, only healthy subjects are considered. They are split into three datasets: 40% for computing the adjustment parameters (regression set), 40% for the learning step relative to the confound prediction (learning set), and 20% for testing (testing set). For the learning step, categorical confounds (gender and database) are predicted with a linear (linear support vector machine, LSVM) and a nonlinear (gradient boosted decision trees, GBDT) classifiers while accounting for imbalanced data. Continuous data (ETIV and age) are predicted with a linear (LASSO method) and nonlinear (random forests, RF) regressors. Cross-validation is used to estimate the hyperparameters of the different learning methods.

The testing set is finally used to assess the accuracy of the prediction, namely the balanced accuracy (BAC) for classification and the fraction of unexplained variance (FUV) for regression. To improve the reliability of the assessment, the whole procedure is repeated 10 times by splitting randomly the controls into the three datasets. Results are reported in the four first rows of Tab. 2. A good adjustment strategy should lead to a low BAC and to a high FUV.

Without data adjustment (*S1*), all the confounds can be well predicted using both linear and nonlinear methods (except the gender which can only be predicted with the nonlinear method). Strategies *S2* and *S3* are relatively efficient because it becomes much more difficult to predict the confounds. Indeed, except for database prediction with *S2*, linear approaches do not allow anymore to predict confounding variables properly: the classification (resp. regression) performance is similar to a random classifier (resp. a regression method that always predicts the mean). The failure of *S2* to remove the database effect is related to the phenomenon described on simulated data (Sec. 2.2, Fig. 1(c-d)). We can observe that the proposed *S3* strategy can successfully overcome this limitation. However, in all cases except for the ETIV, nonlinear approaches are still able to predict confounds with above-chance accuracy, thus pointing out the need for further more sophisticated strategies to adjust data.

3.2. AD *vs* controls CAD system

AD and healthy subjects are split into three datasets as follows: 40% of the controls in the regression set (only controls are used for computing the regression parameters to adjust data [6]), 40% of the controls and of 80% of the AD subjects

in the learning step, and the rest of the data in the testing set. The same classifiers as before (LSVM and GBDT) are trained from the learning set. As previously, mean BAC is computed over 10 realizations. Results are presented in the second last row (*diag-std*) of Tab. 2.

Surprisingly, similar results are obtained whatever the data adjustment strategy and the classifier used. Our intuition is that there is a positive bias in the results obtained without data adjustment (*S1*). Since some confounds are strongly correlated with the diagnosis and can be predicted with above chance accuracy, this may help to improve diagnosis accuracy. Considering only age, gender, ETIV and database enables us to predict diagnosis with a BAC of $56\% \pm 4$ (LSVM) and $63\% \pm 2$ (GBDT). Classifiers learned with *S2* and *S3* strategies are expected to capture more relevant information related to brain morphology only, since confounds are less well predicted. Age seems to play a prominent role in the positive bias induced by *S1*. *S1* tends to classify old subject as AD since AD subjects are on average significantly older than controls. To fault this approach, we build a testing set composed of "young" AD subjects and "old" controls. Results are given in the last row (*diag-biased*) of Tab. 2.

The performance obtained with the *S1* strategy is largely degraded, thus confirming that the information of age (and probably also of the other confounds) is indirectly exploited to drive the decision process. Conversely, the performances obtained with *S2* and *S3* strategies are altered in a much lesser extent. This illustrates that classifiers with apparently similar performance do not necessarily all rely on relevant information. This phenomenon is related to the ambiguity problem [4]: different sources of information can help to predict the target variable. The ambiguity problem makes it difficult to decode what information is used by a classifier to make the decision. Among all the methods that present similar performance in Tab. 2, the most reliable is likely the one based on *S3* data adjustment strategy with the linear classifier. Indeed, since confounds cannot be predicted with above chance accuracy when using linear classification/regression methods, the CAD linear classifier cannot exploit (and consequently cannot be biased by) this information to predict the diagnosis.

4. CONCLUSION

In this paper, we demonstrate that database effect cannot be properly regressed out if the effect of another confound, whose distribution varies across databases, is not properly modeled. We propose a simple strategy that compensates for the residual variation in position and shape that can appear between the distributions of data adjusted with the generalized linear model. The benefit of this strategy has been highlighted in the context of a CAD system discriminating AD *vs* healthy subjects. However, the fact that confounds can still be predicted from adjusted data suggests that there is still some room for improvement in the adjustment procedure. The risk of processing data corrupted by several confounding variables is that the adjusted data may still permit the prediction of confounds with above chance accuracy. In the context of a CAD system, confounds that are correlated with the diagnosis may be responsible for ambiguity. To assess the reliability of a CAD system, we suggest the following guidelines: *(i)* test if the confounds are correlated with the target, *(ii)* test if the adjusted data still allow a good prediction of the confounds, *(iii)* test if the classifier can be misled with new testing data that have not the same distributions of confounds than those of the training set.

5. REFERENCES

[1] J.-P. Fortin, N. Cullen, and Y.I. Sheline et al., "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104 – 120, 2018.

[2] Ch. Wachinger, B. Gutiérrez-Becker, and A. Rieckmann, "Detect, quantify, and incorporate dataset bias: A neuroimaging analysis on 12, 207 individuals," *CoRR*, vol. abs/1804.10764, 2018.

[3] M.M. Mielke, P. Vemuri, and W.A. Rocca, "Clinical epidemiology of Alzheimers disease: assessing sex and gender differences," *Clinical Epidemiology*, vol. 6, pp. 37 – 48, 2014.

[4] L. Snoek, S. Miletić, and S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *bioRxiv*, 2018.

[5] A. Rao, J. M. Monteiro, and J. Mourao-Miranda, "Predictive modelling using neuroimaging data in the presence of confounds," *NeuroImage*, vol. 150, pp. 23 – 49, 2017.

[6] J. Dukart, M.L. Schroeter, K. Mueller, and The Alzheimer's Disease Neuroimaging Initiative, "Age correction in dementia–matching to a healthy brain," *PloS one*, vol. 6, no. 7, 2011.

[7] D.S Marcus, T.H. Wang, and J. Parker et al., "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[8] D.S. Marcus, A.F. Fotenos, and J.G Csernansky et al., "Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults," *Journal of Cognitive Neuroscience*, vol. 22, no. 12, pp. 2677–2684, 2010.

[9] C.R. Jack, M.A. Bernstein, and N.C. Fox et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.