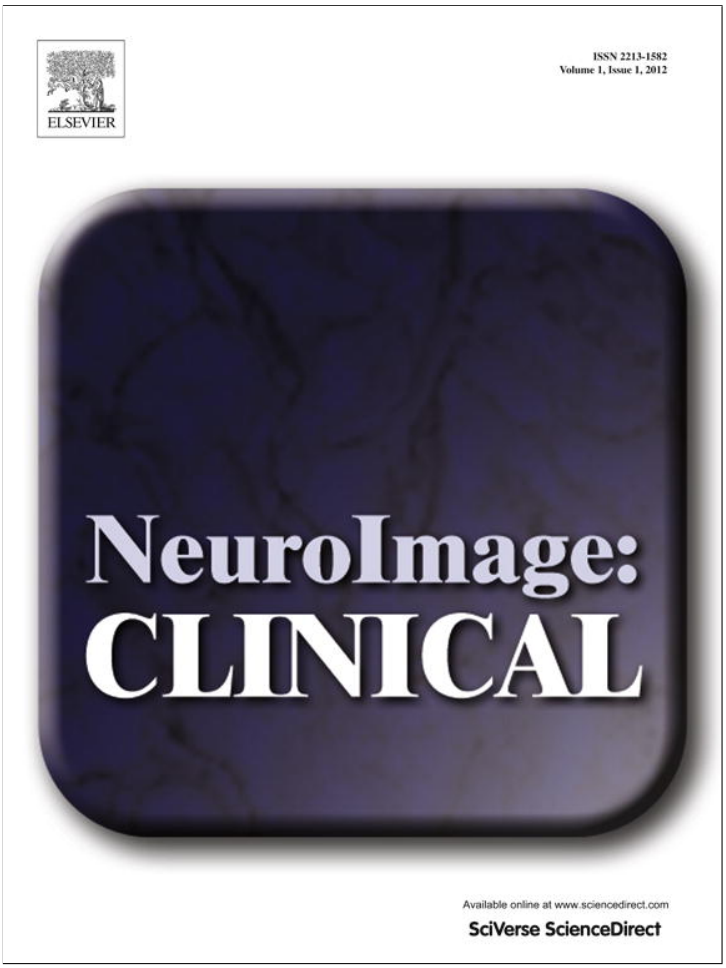

NeuroImage: Clinical 4 (2014) 687–694

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions

Quentin Noirhomme[a,b,*], Damien Lesenfants[a,b], Francisco Gomez[c], Andrea Soddu[d], Jessica Schrouff[a,e], Gaëtan Garraux[a], André Luxen[a], Christophe Phillips[a,f,1], Steven Laureys[a,b,1]

[a]Cyclotron Research Centre, University of Liège, Liège, Belgium
[b]Coma Science Group, Neurology Department, University Hospital of Liège, Liège, Belgium
[c]Complexus Group, Computer Science Department, Universidad Central de Colombia, Bogotá, Colombia
[d]Department of Physics & Astronomy, Brain and Mind Institute, University of Western Ontario, London, ON, Canada
[e]Laboratory of Behavioral and Cognitive Neurology, Stanford University, Palo Alto, USA
[f]Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form 8 April 2014
Accepted 8 April 2014

Keywords:
classification
cross-validation
binomial
permutation test

ABSTRACT

Multivariate classification is used in neuroimaging studies to infer brain activation or in medical applications to infer diagnosis. Their results are often assessed through either a binomial or a permutation test. Here, we simulated classification results of generated random data to assess the influence of the cross-validation scheme on the significance of results. Distributions built from classification of random data with cross-validation did not follow the binomial distribution. The binomial test is therefore not adapted. On the contrary, the permutation test was unaffected by the cross-validation scheme. The influence of the cross-validation was further illustrated on real-data from a brain–computer interface experiment in patients with disorders of consciousness and from an fMRI study on patients with Parkinson disease. Three out of 16 patients with disorders of consciousness had significant accuracy on binomial testing, but only one showed significant accuracy using permutation testing. In the fMRI experiment, the mental imagery of gait could discriminate significantly between idiopathic Parkinson's disease patients and healthy subjects according to the permutation test but not according to the binomial test. Hence, binomial testing could lead to biased estimation of significance and false positive or negative results. In our view, permutation testing is thus recommended for clinical application of classification with cross-validation.

1. Introduction

In the last few years, there has been a growing interest in the statistical assessment of classification results in biomedical applications. Machine learning approaches are now increasingly used to study brain function (Etzel et al., 2009; Pereira et al., 2009; Lemm et al., 2011) and have been proposed as a diagnostic and prognostic tool for patients (e.g., in the field of severe brain injury see (Phillips et al., 2011; Galanaud et al., 2012; Luyt et al., 2012; Lule et al., 2013) or Parkinson disease (Focke et al., 2011; Orru et al., 2012; Schrouff et al., 2012; Garraux et al., 2013; Schrouff et al., 2013)). Such classification machines have also been designed for many other applications such as analyzing DNA microarray and predicting tumor subtype and clinical outcome (Golub et al., 1999; Simon et al., 2003). Limitations and controversies of these approaches have been recently highlighted in a study using brain–computer interfaces (BCIs) to unravel signs of consciousness in patients with disorders of consciousness (Cruse et al., 2011; Goldfine et al., 2013). A statistically significant classification accuracy is one where we can reject the null hypothesis that there is no information about task, patient diagnosis or outcome in the data from which it is being predicted. In a two-class problem with an equivalent number of elements in each class, e.g., disease vs. no-disease, the theoretical chance level, which is valid in the case of an infinite number of trials, is 50%. In practice, we only have a limited number of trials, which can be in the order of 10, due to patient fatigue. If a specific set of features can classify the data with for example 58% accuracy, the question is whether this accuracy is trustworthy. To tackle this issue, several approaches have been proposed in the literature.

A frequently used method is based on the binomial distribution (Müller-Putz et al., 2008; Pereira et al., 2009; Billinger et al., 2013). With a limited number of trials, the results of a classifier are seen as the results of tossing a coin, an unfair coin, which can be modeled as a Bernoulli trial with probability $p = 50\%$ of success. The probability of achieving k successes out of N independent trials is given by the

[1] Both authors contributed equally.
* Corresponding author.
E-mail address: quentin.noirhomme@ulg.ac.be (Q. Noirhomme).

http://dx.doi.org/10.1016/j.nicl.2014.04.004

binomial distribution. Knowing the distribution and a given *p*-value, we can compute a lower bound for any classification accuracy. If the lower bound is higher than the chance level, we can reject the hypothesis that the accuracy was obtained by chance. Here, we are only interested on the accuracies higher than the chance level. We are not interested in the chance of coincidental deviations below the expected 0.50 because we would not pretend our features contain information in that case. Another approach is based on the Pearson chi-square coefficient (Kubler and Birbaumer, 2008). However, for small number of trials, as it is often the case in the neuroimaging and electrophysiology literature, this approach is not reliable (Pereira et al., 2009) and matches the binomial test for higher number of trials (Howell, 2012).

Alternatively, permutation test based methods (Good, 2005) have been employed (Mukherjee et al., 2003; Etzel et al., 2009; Pereira et al., 2009; Schrouff et al., 2013b). A permutation test is a non-parametric test that has also been proposed as a substitute to the Student *t*-test in functional neuroimaging (Nichols and Holmes, 2002) and electrophysiology (Maris and Oostenveld, 2007) experiments. A permutation test estimates the distribution of the null hypothesis from the data. Assuming that there is no class information in the data, the labels are randomly permuted and the accuracy computed with the new labels. As the new labels are random, the new accuracy estimate is expected to reflect the chance distribution. The permutation is repeated hundreds to thousands of times. Then, the *p*-value is given by the fraction of the sample that is larger than or equal to the accuracy actually observed when using the correct labels.

To estimate classification accuracy, ideally, the original data are split into two independent, complementary subsets: a training set (which is used to train the classifier and to define all parameters) and a testing set (which is used to validate the results). In practice, with small datasets, a cross-validation (CV) scheme is often used. The process of splitting the data into two is repeated several times using different partitions. The results obtained from all partitions are then averaged (Lemm et al., 2011). The classification accuracy can then be tested. Following common practice (Pereira et al., 2009; Pereira et al., 2011), the accuracy estimate obtained through a CV could be treated as if it came from a single classifier. In that case, the binomial test sees all accuracies as independent.

In the following, we will show on simulated and real data that the CV scheme has an effect on the calculation of the chance level and that this influence is accounted for by the permutation test but not by the binomial test. We will first present results from simulated data illustrating the influence of the CV scheme. Next, we will exemplify how this may influence the "diagnosis" of patients with disorders of consciousness on real data from a previous EEG-based brain–computer interface (BCI) study (Lule et al., 2013). We will then further illustrate the influence with an fMRI study on activation patterns in Parkinson's disease (Cremers et al., 2012;Schrouff et al., 2012, 2013a). Finally, we will discuss some hypotheses underlying the observed differences between classification testing methods. Our simulations make a simplifying assumption, e.g. type of features, and our example from real data does not cover all possible data source and classification approaches, but the issues presented here are quite general and apply to studies employing a cross-validation scheme to estimate the accuracy of the data.

2. Material and methods

2.1. Simulated data

To test the validity of the binomial and permutation tests to assess classification accuracy, we generated random datasets for a two-class problem. We simulated three cases. First, we tested several scenarios with low number of features and trials. Second, we tested the influence of the number of repetitions of the CV scheme. Third, we tested scenarios with high number of features and low number of trials as often the case in the neuroimaging literature. The generation of the random data and the classifiers used built-in MATLAB (The MathWorks, Natick, MA, USA) functions (*rand, randperm, classify*)[1] and libsvm functions (Chang and Lin, 2011). Datasets were generated with 10,000 simulations. Each simulation included two sets with an equal number of trials. Trial number was 100, 50 or 30. Trials of the 100 trial set (respectively 50 and 30 trial sets) had 40 features (respectively 20 and 10). Features and labels were randomly assigned 0 and 1 (*rand* function thresholded at .5). We tested four CV schemes. In an ideal CV scheme, all possible partitions of the data should be tested. This is the case for the leave-one-out (LOO) CV but in practice for classical N-fold CV schemes it is computationally intractable. Nevertheless, repeating the N-fold CV several times with different partitions is recommended and can reduce the variance of the estimator (Efron and Tibshirani, 1997; Etzel et al., 2009; Lemm et al., 2011). The CV schemes were LOO, 10-fold, 5-fold and 2-fold CVs. The first three are the most used and recommended in the literature (e.g., Lemm et al., 2011). The 2-fold CV is an extreme case at the opposite of the LOO CV. A linear discriminant analysis and a support vector machine (Burges, 1998) with linear kernel classified the data.

To compute the binomial lower bound, the binomial distribution is often approximated by a normal distribution; for example to compute the Wald interval or adjusted Wald interval (Kohavi, 1995; Martin and Hirschberg, 1996; Berrar et al., 2006; Billinger et al., 2013). However, the approximation of the binomial distribution by the normal distribution is only valid whenever the number of trials N and the accuracy p satisfy the following equation: $N \times p \times (1-p) > 5$ (Berrar et al., 2006). In the absence of problem specific knowledge, the best choice for estimation of the bound is derived from Jeffreys' Beta distribution (Martin and Hirschberg, 1996; Berrar et al., 2006). This approximation is adequate for $10 \leq N$ (Martin and Hirschberg, 1996). The binomial lower bound (λ) was computed using Jeffreys' Beta distribution (Berrar et al., 2006) as follows:

$$\lambda \approx \left\{ a + \frac{2(N-2m)z\sqrt{0.5}}{2N(N+3)} \right\} - z\sqrt{\frac{a(1-a)}{N+2.5}}$$

where N is the number of trials, m is the number of successful trials, a is the estimated accuracy and z is the z-score (1.65 for one sided test with $p < .05$ (resp. 2.33 for $p < .01$)).

The permutation test (Good, 2005) was based on 999 permutations plus the original accuracy (Ojala and Garriga, 2010). Only accuracies higher than 0.5 were assessed using permutation testing. We did not compute permutation test for accuracies smaller or equal than 0.50 because we would not pretend that our classifications contain information in that case. The permutation test consisted of randomly exchanging the label and classifying the data with the CV scheme. The *p*-value was calculated as the sum of all values of the permutation distribution equal or higher than the results of the original data divided by the number of permutations.

In a first experiment, 12 datasets were built, three for each of the four CV schemes with 100, 50 or 30 trials, and with 10,000 simulations each. Every simulation involved two subsets with an equal number of trials and features. First, the classification accuracy of the trials from the first subset obtained with linear discriminant analysis was assessed with a chosen CV scheme (Fig. 1A). The distribution of accuracies obtained from all simulations was called: CV distribution. Second, to build an empirical binomial distribution, all trials from the first subset were used to train a classification algorithm which was applied to the second, independent, subset (Fig. 1B). A third distribution, the CV-independent distribution, was built by applying a mixed CV scheme where the N-1 training folds came from the first subset

[1] The MATLAB code can be found at https://github.com/CyclotronResearchCentre/BinomPermTest .

and the test fold came from the second subset (Fig. 1C). At each step of the CV, the classifier trained on N-1 folds from the first subset was applied on a fold from the second subset. Differences between computed distributions and binomial distribution were assessed with a chi-square goodness-of-fit test (Howell, 2012). Results were considered significant at $p < .05$ with a Bonferroni correction for multiple comparison. In a second experiment, we further tested the influence of the number of repetition of the CV scheme on the binomial test. Datasets with 10,000 simulations, each containing 100 trials with 40 features, were generated as explained above. The CV schemes were tested without repetition and with 5, 10 and 20 repetitions to test the influence of the number of repetitions. A linear discriminant analysis classified the data. In a third experiment, we tested the influence of the number of features. To evaluate the binomial test, datasets with 10,000 simulations, each containing 100 trials, were generated as explained above. We tested the classification accuracy with 40, 100, 400, 1000 and 4000 features. These configurations with more features than trials are often the case in neuroimaging studies. To better accommodate the increasing number of features, a support vector machine with linear kernel classified the data. Classification accuracy was estimated with LOO CV. To evaluate the permutation test, we generated datasets with 1000 simulations, each containing 100 trials. Classification accuracy was estimated with a support vector machine and a LOO CV. The difference in number of simulations is due to the time of the permutation test. "1000 simulations" with the permutation test mean fifty million classifications. Each simulation generates 100 classifications with the LOO CV. On average, half of the simulations have a classification accuracy above 0.5 which are tested for significance with a permutation test (500 simulations $\times$ (1 + 999 permutations) $\times$ 100 classifications with the LOO CV). The other half are not tested for significance (500 simulations $\times$ 100 classifications). On the contrary, "10,000 simulations" with the binomial test mean only 1 million classifications (10,000 simulations $\times$ 100 classifications with the LOO CV).

2.2. Brain–computer interface diagnostic application

In a recent study (Lule et al., 2013), we used a stepwise linear discriminant analysis (LDA) to classify data from an EEG-based brain–computer interface (BCI) experiment with severely brain-damaged patients who had survived a coma. The experiment aimed at correctly diagnosing non-responding patients by determining if they were able to respond to command using a motor-independent BCI method. Response to command differentiates patients in a minimally conscious state from patients in a vegetative state/unresponsive wakefulness syndrome (Laureys and Schiff, 2012). We studied 16 severely brain damaged patients who had survived a coma. Thirteen were diagnosed with minimally conscious state (aged 42 $\pm$ 21 years, 9 males, 5 of traumatic etiology, mean time postinjury 70 $\pm$ 109 months) and three patients were in a vegetative state/unresponsive wakefulness syndrome (aged 61 $\pm$ 17 years, 2 males, 2 with anoxic etiology, time postinjury 10 $\pm$ 15 months). An auditory P3 four-choice speller paradigm was used (Sellers and Donchin, 2006; Furdea et al., 2009). Patients were presented with four stimuli ("yes", "no", "stop", "go") in a random sequence. Each trial encompassed 15 presentations of four words (60 words in total). The order of presentation was pseudo-randomized (sound duration: 400 ms; inter-stimulus interval: 600 ms, a trial lasting about 1 min). The participants' task was to count the number of times a target, either "yes" or "no", was presented. Stimulus presentation and data collection were controlled by the BCI2000 software[2] (Schalk et al., 2004). The EEG was recorded using an Ag/AgCl electrode cap with 16 channels (F3, Fz, F4, T7, T8, C3, Cz, C4, Cp3, Cp4, P3, Pz, P4, PO7, PO8, and Oz) based on the international 10–20 system (Sharbrough et al., 1991). Each channel was referenced to the right and grounded to the left mastoid. The recordings were divided in a training session and a question session. The training session lasted 4 trials, and participants were instructed to concentrate on either the "yes" or the "no" word. During the question session, participants had to respond to 10 questions with known answers using the BCI. Amplitude values from particular channel locations and time samples were classified with a stepwise linear discriminant analysis method (Farwell and Donchin, 1988; Donchin et al., 2000; Krusienski et al., 2006). Offline, all data were pooled together and a LOO scheme was used to determine the classification accuracy of each participant. From the 16 patients, 3 patients obtained an accuracy above chance level following the binomial test (accuracy equal or above 50% for a theoretical chance level at 25% and 14 trials). Two patients obtained an accuracy of 50% (7/14 questions) and one reached 57% (8/14 questions). These 3 patients were in a minimally conscious state. Here, we reassessed the previously published data with a permutation test (999 permutations) with a LOO CV and a 2-fold CV with 10 repetitions. We used a 2-fold CV scheme as it was one of the only possible partition of 14 trials and quite different from the LOO CV. The labels of the data were randomly exchanged within each trial. Results were considered significant at $p < .05$.

2.3. Discriminant BOLD activation patterns in Parkinson's disease

Recently, we used BOLD fMRI to study the brain activation pattern underlying mental imagery of walking in idiopathic Parkinson's disease as compared with healthy controls (Cremers et al., 2012; Schrouff et al., 2012; Schrouff et al., 2013). Behavioral and brain imaging data acquisition and processing have been described in Cremers et al. (2012). In brief, participants enrolled in this study were 14 patients (8 males; aged 65.1 $\pm$ 9.5 years) diagnosed with idiopathic Parkinson's disease (Hughes et al., 1992) with different degrees of severity of gait disturbances and 15 controls matched for age (63.8 $\pm$ 8.1 years) and gender (7 males). Before fMRI, all participants were trained to walk comfortably and then briskly on a 25 m path and to mentally rehearse themselves walking on the path. Brain activity changes were recorded using BOLD fMRI during three main experimental conditions: mental imagery of standing (STAND), walking at a comfortable pace (COMF) and walking briskly (BRISK). Eight trials of each condition (12 for BRISK to account for shorter trial duration) were randomly presented within and between subjects. The COMF and BRISK conditions were self-paced, subjects indicating when they had completed each trial by a key press, while each trial of the STAND condition was constrained by the duration of the previous COMF trial. fMRI data preprocessing and first-level univariate analyses were performed using SPM8[3] as previously reported (Cremers et al., 2012). Three images per subject were generated from these first-level fMRI analyses representing BOLD signal changes associated with STAND, COMF and BRISK conditions, respectively.

We aimed to assess whether the multivariate analysis of these images using binary SVM (Burges, 1998) as implemented in PRoNTo[4] could be used to accurately discriminate patients from controls. A leave-one-subject per group out CV was performed to compute model performance, its significance being assessed by a permutation testing using 1000 permutations. Either all voxels within the brain served as features (140,305 voxels), or only voxels from the areas involved in gait (both in healthy subjects and in patients), as described in Table 1 of Maillet et al. (2012) ("motor mask", 45,825 voxels). The between group classification was based on either individual task (e.g., STAND in controls vs. STAND in patients) or a combination of task (e.g.,

[2] http://www.bci2000.org/ .

[3] http://www.fil.ion.ucl.ac.uk/spm .

[4] http://www.mlnl.cs.ucl.ac.uk/pronto .

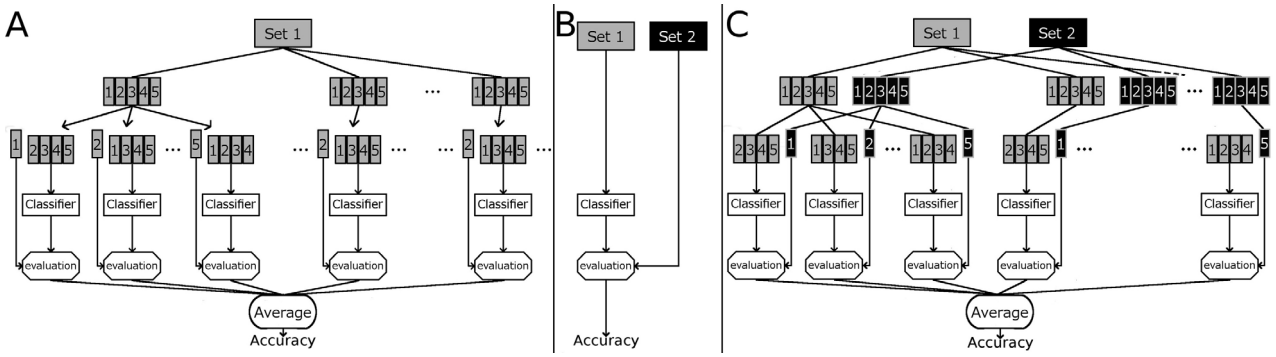

Fig. 1. For each simulation, three distributions of accuracies were computed. The CV distribution (A) was computed through the estimation of accuracy with a CV scheme. Here a 5-fold CV with repetition is used as an example. The empirical binomial distribution (B) was computed by training the classification algorithm on the first subset and testing on the second, independent, subset. In the CV-independent distribution (C), the classification algorithm was trained on N-1 fold from the first subset and the accuracy estimated on one fold from the second subset.

BRISK + COMF in controls vs. BRISK + COMF in patients). Here, we reassessed the previously published data with a binomial test. Results were considered significant at $p < .05$.

3. Results

3.1. Simulated data

Results of the first experiment on evaluating binomial and permutation tests with a two-class problem and low number of features and trials are shown in Fig. 2 for the 10-fold CV with 100 trials and 40 features, and Table 1 for the LOO, 10-fold, 5-fold and 2-fold CVs with 100, 50 and 30 trials. The binomial lower bound is 59% accuracy for 100 trials with significance level $p < .05$ (62% accuracy at $p < .01$) independently of the CV scheme. For the simulations with 50 and 30 trials, the lower bound at $p < .05$ was, respectively, 62% and 65% (67% and 71% at $p < .01$). All computed distribution differed significantly of the binomial distribution (chi-square goodness-of-fit test, $p < .05$). LOO CV produced the widest distribution. More than 8% of the accuracy values from random data were above the binomial accuracy lower bound at $p < .05$, and 3% at $p < .01$. 10 × 10-fold CV also produced a wider distribution than the binomial distribution. The 10 × 5-fold CV distribution was closest to the binomial distribution. The 10 × 2-fold CV produced a distribution narrower than the binomial distribution with 0–1% of the random data above the binomial accuracy lower bound at $p < .05$ and 0% above the lower bound at $p < .01$. For the permutation test, the percentage of p-values below .05 and .01 was less than 5% and 1% respectively for all CV schemes. For all datasets, the empirical distribution matched the binomial distribution. The CV-independent distribution matched the binomial distribution with the LOO scheme. For all other schemes, the CV-independent distribution differed significantly from the binomial distribution (Fig. 3).

In the second experiment, the distributions built from the 4 CV schemes without repetition were wider than the binomial distribution (Fig. 4), with the LOO CV showing the most deviation. Repeating the CV narrowed the cumulative distribution function (CDF) of the 10-fold (Fig. 5), 5-fold and 2-fold CVs resulting in a mixed effect. The number of repetition had an influence up to 10 repetitions, increasing the number of repetitions to 20 changed only slightly the distribution. In the third experiment, the distributions estimated with LOO CV and 100 trials narrowed with the increased number of features. The binomial test evolved from being not enough conservative to being too conservative (Table 2). For the permutation test, the percentage of p-values below .05 and .01 was less than 5% and 1% respectively for all number of features (Table 3).

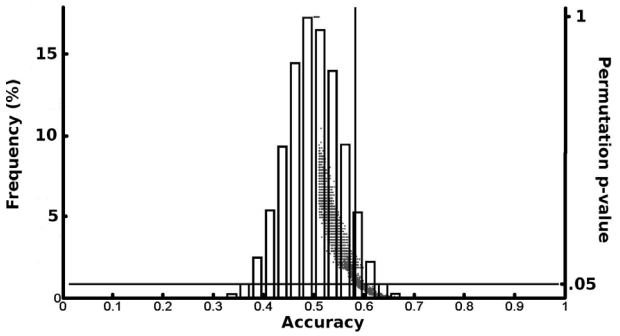

Fig. 2. Histogram of the distribution of the classification accuracy (bars; left axis) and p-values from the permutation test (for accuracy > .5; dots; right axis) for 10,000 simulations with 100 trials, 40 features, 10 × 10-fold cross-validation. The vertical thick line illustrates the binomial test lower bound and the horizontal thick line shows the permutation test accuracy level at $p < .05$.

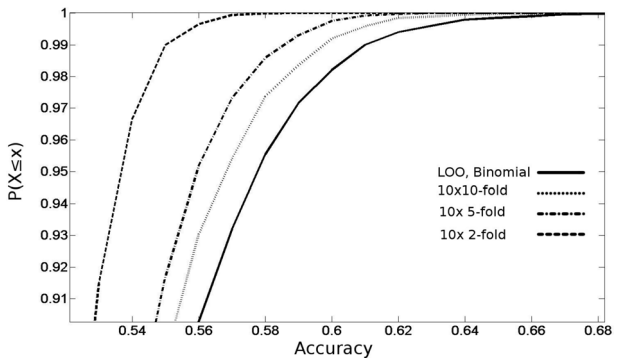

Fig. 3. Cumulative distribution functions (CDFs) of classification accuracy values obtained using a classifier trained on N-1 fold of one subset and applied on a fold from an independent subset. Classification accuracy values obtained from 10,000 simulations with 100 trials and 40 features. The leave-one-out independent CDF overlaps with the binomial CDF. Note that the 10-, 5- and 2-fold independent CVs show a narrower distribution.

3.2. Brain–computer interface diagnostic application

As presented in the original paper, three patients had an accuracy of 50%, 50% and 57% with the LOO CV. These three accuracies are above the binomial lower bound (above or equal to 7/14 compared to a theoretical chance level at 25%). Their permutation p-values were .06, .08 and .03 respectively. When reanalyzing the three patients' data with the 2-fold CV, they obtained an accuracy of 6%, 31% and 39%.

Table 1
Percentage of the 10,000 simulations with results thresholded for significance at $p < .05$ and $p < .01$ for the binomial and permutation tests. The simulations included either 100, 50 or 30 trials with, respectively, 40, 20 or 10 random features and randomly assigned binary labels. Lower bound thresholds for binomial test were computed using Jeffreys' priors. Permutation tests used 999 permutations. Cross validation schemes included leave-one-out (LOO), 10-fold, 5-fold and 2-fold cross validations. Folding and computing classification was repeated 10 times with different folds.

CV scheme	# of trials	Binomial		Permutation	
		$p < .05$	$p < .01$	$p < .05$	$p < .01$
LOO	100	8%	3%	4%	1%
	50	10%	3%	4%	1%
	30	9%	3%	4%	1%
10 × 10-fold	100	7%	2%	5%	1%
	50	7%	2%	5%	1%
	30	7%	2%	5%	1%
10 × 5-fold	100	5%	1%	5%	1%
	50	4%	1%	5%	1%
	30	5%	1%	5%	1%
10 × 2-fold	100	0%	0%	5%	1%
	50	1%	0%	5%	1%
	30	1%	0%	5%	1%

Table 2
Percentage of the 10,000 simulations with results thresholded for significance at $p < .05$ and $p < .01$ for the binomial tests. The simulations included 100 trials with random features and randomly assigned binary labels. Lower bound thresholds for binomial test were computed using Jeffreys' priors. Classification accuracy was estimated with a support vector machine with linear kernel and a leave-one-out cross validation.

# of features	$p < .05$	$p < .01$
40	13%	7%
100	7%	3%
400	8%	3%
1000	6%	2%
4000	5%	2%
10,000	2%	1%

Table 3
Percentage of the 1000 simulations with results thresholded for significance at $p < .05$ and $p < .01$ for permutation tests. The simulations included 100 trials with random features and randomly assigned binary labels. Permutation tests used 999 permutations. Classification accuracy was estimated with a support vector machine with linear kernel and a leave-one-out cross validation.

# of features	$p < .05$	$p < .01$
40	4%	1%
100	4%	1%
400	5%	1%
1000	4%	1%
4000	4%	1%
10,000	4%	1%

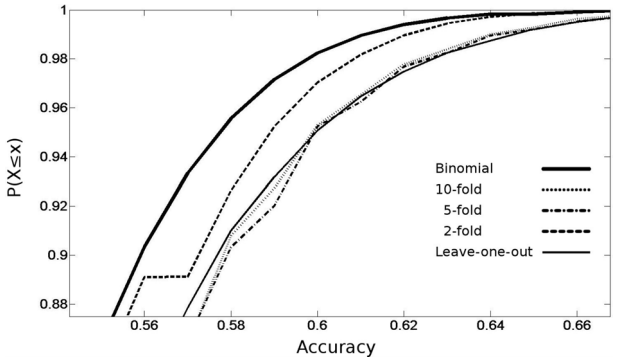

Fig. 4. Cumulative distribution functions for the binomial, leave-one-out, 10-fold, 5-fold and 2-fold cross-validations for 10,000 simulations of 100 trials with 40 features without repetition.

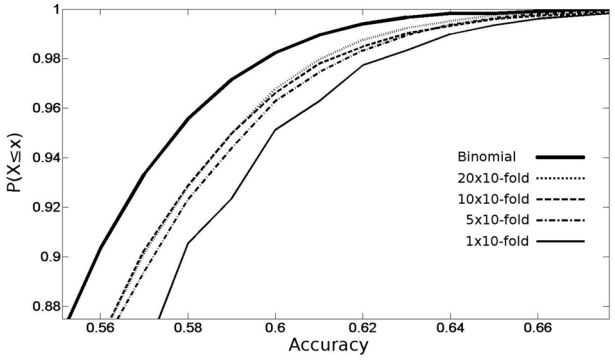

Fig. 5. Cumulative distribution functions for the binomial and the 10-fold cross-validated data with 1, 5, 10 and 20 repetitions for 10,000 simulations of 100 trials with 40 features.

All accuracies were below the binomial lower bound but the permutation p-values were .94, .17 and .046, respectively. The histograms of permuted accuracy for the patient with highest accuracy for the LOO and 2-fold CV are reported in Fig. 6. Both histograms peak at 0.25 which is the theoretical chance level. The use of a 2-fold CV narrowed the histogram. The binomial significant level at $p < .05$ (50%) was too wide for the LOO CV (8% of the data above the limit) and too narrow for the 10 × 2-fold CV (less than a 1% of the data above the limit) as compared to the accuracies obtained by permutation testing.

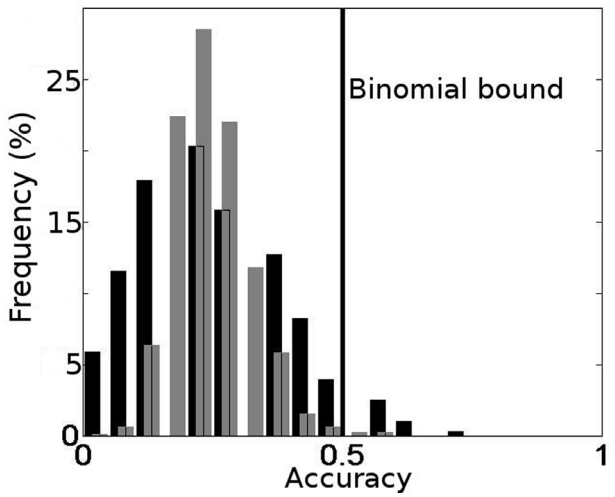

Fig. 6. Clinical data obtained from a patient with significant diagnostic accuracy using a BCI. Histogram of accuracies obtained with permutation testing for the leave-one-out cross-validation (black) and the 2-fold cross-validation (gray). The vertical line shows the binomial lower bound (50%) for significant accuracy at $p = .05$.

3.3. Discriminant BOLD activation patterns in Parkinson's disease

Overall, using all brain voxels led to a poor discrimination between idiopathic Parkinson's disease patient and controls. The binomial lower bound for 29 trials with equal probability of both classes is 66%. The estimated balanced accuracies with the different configurations of features were all below the binomial lower bound (Table 4). However, using the permutation test, one combination of features (BRISK + CONF) with accuracy reaching 62% was significant. The normalized weights from the classifier had a good overlap with the results from the univariate analysis (Schrouff et al., 2013). Slightly better results were obtained while decreasing the number of features with the motor mask, as shown by a higher balanced accuracy for the BRISK–COMF combination, as well as for the BRISK condition both significant at $p < 0.05$ with the permutation and binomial tests.

4. Discussion

Our results on artificially generated random data and real clinical data illustrate that the CV scheme has an influence on the statistical significance of obtained classification accuracies. This influence seems to bias results from binomial testing. The permutation test took the cross-validation scheme into account and was therefore not biased. We hypothesize that the observed differences between CV distribution and binomial distribution are due to counterbalancing factors. A first factor is the decreased independence among trials, a key assumption of the binomial testing, in CV scheme. The influence of this factor is well illustrated in the extreme case of the LOO scheme or using CV without repetition. A second factor is that the repetition of the CV scheme virtually increases the number of test examples. This is illustrated through the change in the CV-independent distributions. The number of repetitions and the CV-scheme both influence the size of the test set. In turn, the size of the test set influences the significance of the test, as a random classifier is less likely to maintain the same level of accuracy on an extended test set. This has been previously shown for the permutation test (Mukherjee et al., 2003) and is reproduced here using a real dataset. The 2-fold CV with 10 repetitions had a narrower distribution than the LOO CV (Fig. 6); therefore smaller accuracy could be significant. The reported simulated data here also illustrate this effect for the distribution of classification accuracies. A third factor is the number features. Increasing the number of features narrowed the CV distribution in our third simulation study. This effect was also illustrated in the Parkinson disease dataset where, despite the use of a LOO CV, the permutation distribution was narrower than the binomial distribution. High number of features makes the classifier more prone to generalization problem. The classifier has more chance to pick features that correlate well with training data but not with test data. The final accuracy is therefore less likely to be high. A feature selection method to reduce dimensionality (Lemm et al., 2011), a priori knowledge, or a regularization method may help reducing over-fitting. In our fMRI dataset, physiological a priori information helped reducing the features set and improved the classification. The feature selection or regularization method must be included in the CV loop and may also influence the CV distribution. Another factor which may influence the distribution of classified accuracies is the classifier. We show that the distributions build with LOO CV and with LDA and SVM classifiers yielded slightly different results for simulated data with 100 trials with 40 features.

It is important to stress, that the results and conclusions presented here were obtained on small dataset but with number of trials often found in neuroimaging or brain–computer interface studies. These results are not in line with current common practice (Pereira et al., 2009; Pereira and Botvinick, 2011) which treats the accuracy obtained through cross-validation as if it came from an independent dataset, and then test it in exactly the same way. One more point to take into consideration with small dataset is the stability of the classifier. The independence of accuracies obtained through cross-validation holds, as long as the classifier is stable under the perturbation induced by deleting one of the folds from the data in a cross-validation scheme (Kohavi, 1995). A classifier is stable for a given dataset and set of perturbations if it makes the same prediction with the perturbed datasets. This is most probably not the case for small datasets. How large should be a dataset to prevent these issues should be the subject of further studies.

Using a permutation test is more demanding than binomial testing, as the classification must be repeated hundreds of times. The number of permutations has an influence on the shape of the distribution. However, the p-value can be monitored to limit the number of permutations, computing all permutations only for a value around or below the level of significance and stopping the test much earlier for the others (Mukherjee et al., 2003; Ojala and Garriga, 2010). In the case of the two real datasets presented here with linear discriminant analysis and support vector machine classifiers, the permutation test took only a few seconds. With other classifier, e.g. Gaussian Processes, the computation time may be much longer. If the permutation test has to be applied independently on all voxels of an image, this could take a considerable time (thousands of voxels times a few seconds). Furthermore, it has been mentioned that a large number of permutations may be required to get p-values in a range that would survive multiple comparison correction (Pereira and Botvinick, 2011). Building a unique distribution for all voxels (Nichols and Holmes, 2002) or cluster based permutation test may circumvent that problem (Maris and Oostenveld, 2007).

In the data from the BCI dataset (Lule et al., 2013), one patient had significant accuracy with the permutation test. In 'clinical' settings, with a predefined and validated threshold of accuracy this would mean that the patient demonstrated command following, an important landmark for a diagnosis related to consciousness. In a scientific 'study', where the aim is to validate the approach, which is the case in the original and the present papers, we would protect ourselves against false claims, i.e., stating that the patient followed the command when he did not. If we test 20 patients with a threshold based on a p-value $< .05$, just by chance one patient may have positive results. In the original study, 16 patients were included. We therefore corrected for multiple comparisons via the false-discovery rate

Table 4
Balanced accuracy for the idiopathic Parkinson's disease patient vs. control classification for each combination of the three tasks. Significant results with the permutation test are displayed with an *. No result was significant with the binomial test.

Condition	Balanced accuracy (%)	
	Whole brain	Motor mask
STAND	14	35
COMF	58	62
BRISK	59	66*
STAND + BRISK	37	36
STAND + CONF	36	40
COMF + BRISK	62*	66*
STAND + COMF + BRISK	43	48

(Benjamini and Hochberg, 1995), significance at $p < .05$ and no patients survived the corrected threshold (Goldfine et al., 2013). The final threshold for clinical application should not depend on the number of patients tested as this number would permanently increase changing the threshold continuously (Cruse et al., Brain Injury). This threshold should be based on the accuracies obtained on an extended cohort of patients and healthy controls and balance the sensitivity and specificity of the method. This threshold should depend of the number of trials and the obtained accuracy. The quality of the data may also be taken into account but must be checked previously to any classification. The threshold may be adapted if the test is repeated or joined to results from other tests.

Here, we tested only a limited number of validation schemes. We have not tested bootstrapping (Efron and Tibshirani, 1997) or Monte-Carlo CV (Picard and Cook, 1984). These approaches should be tested in future studies even if the latter has most probably the same properties as the k-folds CV with repetitions. Furthermore, our results do not extend to the validation of an independent dataset which is still the gold standard for validating classification accuracy and recommended whenever possible; unfortunately this is not practical in the two diagnostic cases presented here: brain–computer interface applied to the detection of consciousness and the mental imagery of gait in idiopathic Parkinson's disease patients. With an independent validation set, the binomial test is perfectly valid. Eventually, a small test set may be tested multiple times with classifiers trained on slightly different subsets of the training set. The repetition of testing should virtually increase the size of the test set as illustrated by our CV-independent distribution. Regarding the selection of a CV scheme, the first priority should be to decrease the variance and the bias of the estimated classification accuracies. For a good compromise, the use of 10-fold or 5-fold CVs is often recommended (Lemm et al., 2011).

Here, we tested only a limited number of parameters (number of trials and features) and presented results for two classifiers. However, we believe that one example is enough to demonstrate that the distribution of accuracies obtained by classifying random data with a CV scheme does not follow a binomial distribution.

To conclude, the CV scheme has an influence on the distribution of classification accuracies. This influence biases the binomial testing. Therefore, a permutation test is recommended, especially when dealing with small sample sizes and non-independent CV schemes, as often is the case in clinical datasets.

Acknowledgments

The authors thank the anonymous reviewers for their helpful comments, which improved the quality of this paper. This study was funded by FEDER structural fund RADIOMED-930549; Fonds Léon Frédéricq; James S. McDonnell Foundation; Concerted Research Action (ARC 06/11-340); Public Utility Foundation "Université Européenne du Travail" and "Fondazione Europea di Ricerca Biomedica". S.L. is Research Director and C.P. is Research Associate at the Fonds de la Recherche Scientifique (FRS). The funding sources are not liable for any use that may be made of the information contained therein. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary materials

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.nicl.2014.04.004.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Berrar, D., Bradbury, I., et al. 2006. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics (Oxford, England) 22 (10), 1245–50. http://dx.doi.org/10.1093/bioinformatics/btl066, 16500931.

Billinger, M., Daly, I., et al. 2013. Is it significant? in: Allison, B.Z., Dunne, S., Leeb, R., Millan Jdel, R., Nijholt, A. (Eds.), Guidelines for Reporting BCI Performance. Towards Practical Brain–Computer Interfaces. Berlin, Heidelberg: Springer, pp. 333–54.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–67. http://dx.doi.org/10.1023/A:1009715923555.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 27:21–:=:7.

Cremers, J., D'Ostilio, K., et al. 2012. Brain activation pattern related to gait disturbances in Parkinson's disease. Movement Disorders: Official Journal of the Movement Disorder Society 27 (12), 1498–505. http://dx.doi.org/10.1002/mds.25139, 23008169.

Cruse, D., Chennu, S., et al. 2011. Bedside detection of awareness in the vegetative state: a cohort study. Lancet 378 (9809), 2088–94. http://dx.doi.org/10.1016/S0140-6736(11)61224-5, 22078855.

Donchin, E., Spencer, K.M., et al. 2000. The mental prosthesis: assessing the speed of a P300-based brain–computer interface. IEEE Transactions on Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society 8 (2), 174–9. http://dx.doi.org/10.1109/86.847808, 10896179.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the .632 + bootstrap method. Journal of the American Statistical Association 92 (438), 548–60. http://dx.doi.org/10.1080/01621459.1997.1047400710.2307/2965703.

Etzel, J.A., Gazzola, V., et al. 2009. An introduction to anatomical ROI-based fMRI classification analysis. Brain Research 1282, 114–25. http://dx.doi.org/10.1016/j.brainres.2009.05.090, 19505449.

Farwell, L.A., Donchin, E., 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology 70 (6), 510–23. http://dx.doi.org/10.1016/0013-4694(88)90149-6, 2461285.

Focke, N.K., Helms, G., et al. 2011. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic Parkinson syndrome and healthy controls. Human Brain Mapping 32 (11), 1905–15. http://dx.doi.org/10.1002/hbm.21161, 21246668.

Furdea, A., Halder, S., et al. 2009. An auditory oddball (P300) spelling system for brain–computer interfaces. Psychophysiology 46 (3), 617–25. http://dx.doi.org/10.1111/j.1469-8986.2008.00783.x, 19170946.

Galanaud, D., Perlbarg, V., et al. 2012. Assessment of white matter injury and outcome in severe brain trauma: a prospective multicenter cohort. Anesthesiology 117 (6), 1300–10. http://dx.doi.org/10.1097/ALN.0b013e3182755558, 23135261.

Garraux, G., Phillips, C., et al. 2013. Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical parkinsonian syndromes. NeuroImage: Clinical 2, 883–93. http://dx.doi.org/10.1016/j.nicl.2013.06.004, 24179839.

Goldfine, A.M., Bardin, J.C., et al. 2013. Reanalysis of "bedside detection of awareness in the vegetative state: a cohort study". Lancet 381 (9863), 289–91. http://dx.doi.org/10.1016/S0140-6736(13)60125-7, 23351802.
Golub, T.R., Slonim, D.K., et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.) 286 (5439), 531–7. http://dx.doi.org/10.1126/science.286.5439.531, 10521349.
Good, P., 2005. Permutation, Parametric and Bootstrap Tests of Hypotheses. United States of America: Springer.
Howell, D.C., 2012. Statistical Methods for Psychology. Wadsworth.
Hughes, A.J., Daniel, S.E., et al. 1992. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery, and Psychiatry 55 (3), 181–4. http://dx.doi.org/10.1136/jnnp.55.3.181, 1564476.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence.
Krusienski, D.J., Sellers, E.W., et al. 2006. A comparison of classification techniques for the P300 Speller. Journal of Neural Engineering 3 (4), 299–305. http://dx.doi.org/10.1088/1741-2560/3/4/007, 17124334.
Kubler, A., Birbaumer, N., 2008. Brain–computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 119 (11), 2658–66. http://dx.doi.org/10.1016/j.clinph.2008.06.019, 18824406.
Laureys, S., Schiff, N.D., 2012. Coma and consciousness: paradigms (re)framed by neuroimaging. Neuroimage 61 (2), 478–91. http://dx.doi.org/10.1016/j.neuroimage.2011.12.041, 22227888.
Lemm, S., Blankertz, B., et al. 2011. Introduction to machine learning for brain imaging. Neuroimage 56 (2), 387–99. http://dx.doi.org/10.1016/j.neuroimage.2010.11.004, 21172442.
Lule, D., Noirhomme, Q., et al. 2013. Probing command following in patients with disorders of consciousness using a brain–computer interface. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 124 (1), 101–6. http://dx.doi.org/10.1016/j.clinph.2012.04.030, 22920562.
Luyt, C.E., Galanaud, D., et al. 2012. Diffusion tensor imaging to predict long-term outcome after cardiac arrest: a bicentric pilot study. Anesthesiology 117 (6), 1311–21. http://dx.doi.org/10.1097/ALN.0b013e318275148c, 23135257.
Maillet, A., Pollak, P., et al. 2012. Imaging gait disorders in parkinsonism: a review. Journal of Neurology, Neurosurgery, and Psychiatry 83 (10), 986–93. http://dx.doi.org/10.1136/jnnp-2012-302461, 22773859.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164 (1), 177–90. http://dx.doi.org/10.1016/j.jneumeth.2007.03.024, 17517438.
Martin, J.K., Hirschberg, D.S., 1996. Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests. CA.: Department of Information and Computer Science, University of California, Irvine.
Mukherjee, S., Golland, P., et al. 2003. Permutation Tests for Classification. Cambridge, MA: Massachusetts Institute of Technology, 22.
Müller-Putz, G.R., Scherer, R., et al. 2008. Better than random? A closer look on BCI results. International. Journal of Bioelectromagnetism. 10 (1), 52–5.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 15 (1), 1–25. http://dx.doi.org/10.1002/hbm.1058, 11747097.
Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. Journal of Machine Learning Research 11, 1833–63.
Orru, G., Pettersson-Yeo, W., et al. 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience and Biobehavioral Reviews 36 (4), 1140–52. http://dx.doi.org/10.1016/j.neubiorev.2012.01.004, 22305994.
Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. Neuroimage 56 (2), 476–96. http://dx.doi.org/10.1016/j.neuroimage.2010.05.026, 20488249.
Pereira, F., Detre, G., et al. 2011. Generating text from functional brain images. Frontiers in Human Neuroscience 5, 72, 21927602.
Pereira, F., Mitchell, T., et al. 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1 Suppl.), S199–S209. http://dx.doi.org/10.1016/j.neuroimage.2008.11.007, 19070668.
Phillips, C.L., Bruno, M.A., et al. 2011. "Relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. Neuroimage 56 (2), 797–808. http://dx.doi.org/10.1016/j.neuroimage.2010.05.083, 20570741.
Picard, R.R., Cook, R.D., 1984. Cross-validation of regression models. Journal of the American Statistical Association 79, 575–83. http://dx.doi.org/10.1080/01621459.1984.10478083.
Schalk, G., McFarland, D.J., et al. 2004. BCI2000: A general-purpose brain–computer interface (BCI) system. IEEE Transactions on Biomedical Engineering 51 (6), 1034–43. http://dx.doi.org/10.1109/TBME.2004.827072, 15188875.
Schrouff, J., Cremers, J., et al. 2012. Discriminant BOLD activation patterns during mental imagery in Parkinson's disease.Machine Learning and Interpretation in NeuroImaging workshop at NIPS. South Lake Tahoe, United States of America: NIPS, p. 8.
Schrouff, J., Cremers, J., et al. 2013. Localizing and comparing weight maps generated from linear kernel machine learning models.3rd Workshop on Pattern Recognition in NeuroImaging (PRNI 2013). Philadelphia, USA: IEEE Computer Society Conference Publishing Services, p. 4.
Schrouff, J., Rosa, M.J., et al. 2013. PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11 (3), 319–37. http://dx.doi.org/10.1007/s12021-013-9178-1, 23417655.
Sellers, E.W., Donchin, E., 2006. A P300-based brain-computer interface: initial tests by ALS patients. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 117 (3), 538–48. http://dx.doi.org/10.1016/j.clinph.2005.06.027, 16461003.
Sharbrough, F., Chatrian, G.-E., et al. 1991. American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society 8, 200–2. http://dx.doi.org/10.1097/00004691-199104000-00007, 2050819.
Simon, R., Radmacher, M.D., et al. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 95 (1), 14–18. http://dx.doi.org/10.1093/jnci/95.1.14, 12509396.
Cruse D., Gantner I., Soddu A., Owen A.M. Lies, damned lies, and diagnoses: Estimating the clinical utility of assessments of covert awareness in the Vegetative State. Brain Inj: inpress.