NeuroImage 56 (2011) 517–524

Contents lists available at ScienceDirect

NeuroImage

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

Common component classification: What can we learn from machine learning?

Ariana Anderson [a,b], Jennifer S. Labus [a,c,d], Eduardo P. Vianna [a,c,d], Emeran A. Mayer [a,c,d], Mark S. Cohen [a,b,*]

[a] Department of Psychiatry and Behavioral Sciences, David Geffen School of Medicine at UCLA, USA
[b] Center for Cognitive Neuroscience, David Geffen School of Medicine at UCLA, USA
[c] Center for Neurobiology of Stress, David Geffen School of Medicine at UCLA, USA
[d] Brain Research Institute, David Geffen School of Medicine at UCLA, USA

ARTICLE INFO

Article history:
Received 9 December 2009
Revised 10 May 2010
Accepted 25 May 2010
Available online 25 June 2010

Keywords:
Classification
Discrimination
fMRI
Bias
Machine learning
Independent components analysis
Cross-validation
Irritable bowel

ABSTRACT

Machine learning methods have been applied to classifying fMRI scans by studying locations in the brain that exhibit temporal intensity variation between groups, frequently reporting classification accuracy of 90% or better. Although empirical results are quite favorable, one might doubt the ability of classification methods to withstand changes in task ordering and the reproducibility of activation patterns over runs, and question how much of the classification machines' power is due to artifactual noise versus genuine neurological signal. To examine the true strength and power of machine learning classifiers we create and then deconstruct a classifier to examine its sensitivity to physiological noise, task reordering, and across-scan classification ability. The models are trained and tested both within and across runs to assess stability and reproducibility across conditions. We demonstrate the use of independent components analysis for both feature extraction and artifact removal and show that removal of such artifacts can reduce predictive accuracy even when data has been cleaned in the preprocessing stages. We demonstrate how mistakes in the feature selection process can cause the cross-validation error seen in publication to be a biased estimate of the testing error seen in practice and measure this bias by purposefully making flawed models. We discuss other ways to introduce bias and the statistical assumptions lying behind the data and model themselves. Finally we discuss the complications in drawing inference from the smaller sample sizes typically seen in fMRI studies, the effects of small or unbalanced samples on the Type 1 and Type 2 error rates, and how publication bias can give a false confidence of the power of such methods. Collectively this work identifies challenges specific to fMRI classification and methods affecting the stability of models.

Introduction

Machine learning classification applied to fMRI data have shown strong potential to diagnose cognitive disorders and identify behavioral states (Fan et al., 2006; Zhang and Samaras, 2005; Ford et al., 2003), but drawing inference to the general population from small-sample studies can be difficult. The assumptions of reproducibility of reactions over different fMRI runs may not be realistic (Lange et al., 1999; McKeown et al., 2003), and factors such as small sample sizes, feature selection methods, and sampling variation may cause the cross-validation results one sees in publication to be a biased estimate of the testing accuracy one realizes in practice. Even when some care is taken to exclude obvious artifacts the resulting classifiers may be difficult to interpret, as they typically are formed without prior functional hypotheses. To illustrate these methodological susceptibilities we present and then deconstruct a classifier to test the true power of machine learning. In Anderson et al. (2010) the *spectral classification* method was presented which allows classification among fMRI scans that have not been aligned spatially using the temporal correlations among the independent components. From this, there arises the question of which components temporal activity differs enough between groups to power the classifier. To identify the discriminative component relationships we present a method called **Common Component Classification** that facilitates *post-hoc* identification of the components powering the classifier. Multi-session temporal concatenation (MSTC), a procedure based on independent components analysis, extracts common spatial maps across subjects as well as component-specific time series for each subject (Smith et al., 2004). Classification is performed by characterizing correlations between pairs of components, revealing which components behaved differently between patients and controls.

Our classifier is tested on data from irritable bowel syndrome (IBS) patients and healthy controls (HC) undergoing a gastrointestinal stress task. IBS is a common functional pain disorder associated with chronic abdominal pain, discomfort, and associated altered bowel habits (Drossman, 2006; Mayer et al., December, 2006). When applying our machine-learning classifier to fMRI scans acquired

* Corresponding author. UCLA Center for Cognitive Neuroscience, 760 Westwood Plaza, Suite 17-369, Los Angeles, CA 90095, USA.
E-mail address: mscohen@ucla.edu (M.S. Cohen).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.05.065

during controlled rectal distension in IBS patients and HCs, these methods identified which participants were IBS or HCs and exposed entire networks differing between groups corresponding to identifiable neurological phenomena.

We next deconstruct this classifier by training and testing it within and across two runs to assess its sensitivity to permutation of the stimulus set as well as the reproducibility of stimulus effects across runs. We show how models can be made biased by mistakes made in the feature selection, parameter choice and cross-validation stages and measure the magnitude of this error. We further assess the strength of group-ICA methods by extracting components within and across runs, and evaluate the effectiveness of ICA-based methods to identify and remove artifacts. The classifier is also evaluated on data that has been cleaned of physiological noise to evaluate how much of the classification ability is attributable to scan artifacts such as motion versus true neurological signal. We examine the impact of motion artifacts on the classifier and the ability to remove it without also removing signal. Finally, we examine the statistical assumptions underlying machine learning classifiers, discussing the reproducibility of stimulus effects across runs, how bias can skew the predictive accuracy of the model and how the small sample sizes typical in fMRI affect our Type 1 and Type 2 errors and limit the ability to draw inference from findings of such machine learning studies. From our exercise of creating and deconstructing a classifier we seek collectively to identify what is being learned from machine learning.

Materials and methods: Common component classification

Data characteristics

Functional MRI data sensitive to blood oxygen content were recorded from 13 female IBS subjects and 11 HC, each scanned multiple times in a single day, in a block designed protocol that included anticipated and delivered mild and moderate rectal distention.

Experimental design

Full details of the experimental design are presented in Berman et al. (2008). In brief, four to six 10 min stimulus sets or fMRI runs containing 16 inflation trials each were administered to HC and IBS. The first trial in each stimulus was a 45 mmHg pressure followed by five additional 45 mmHg inflations, five 25 mmHg inflations, and five trials at a baseline pressure of 5 mmHg (sham inflation) in pseudo-random order. Each trial or block comprised 18 s before balloon inflation, followed by 15 s of inflation at the designated pressure and 3 s for deflation and rating (36 s). A visual cue preceded the inflations by 2 to 5 s and was removed at the end of the inflation period. This signaled the end of a trial, at which point subjects rated the intensity of the stimulus on a simple three-point scale. There was a brief rest (<1 min) between fMRI runs. Data for all patients were not available for each fMRI run; we utilized the scans from the runs most patients completed, runs 1 and 4. These runs are referred to Run 1 and Run 2 respectively within this manuscript.

fMRI acquisition

Brain images were acquired on a 1.5 T MRI scanner (Siemens Sonata; Siemens, Erlangen, Germany). First, a sagittal scout was used to position the head. Then, functional T2*-weighted gradient-recalled echo-planar images with blood oxygen level-dependent (BOLD) contrast (repetition time, 2970 ms; echo time, 42 ms; flip angle, 80; slice thickness, 4 mm with a 1 mm interslice interval; matrix, 64×64; in-plane resolution, 3.12 mm^2) were acquired during each 10 min stimulus set and constituted an fMRI run. After discarding the first two volumes of each 10 min run, 25 axial slices were recorded for each of 200 functional whole-brain volumes and saved to disk for off-line analysis.

fMRI data preprocessing

The data were motion corrected using MCFLIRT, spatially smoothed using an 8 mm Gaussian kernel, high-pass filtered and spatially aligned using the registration tool FLIRT (Smith et al., 2004).

Methods

In this section we present our classifier, *common component classification*, which will identify the group membership of a subject based on their specific temporal response pattern on group-wide independent components. Because the stimulus set is reordered across runs and artifacts are highly subject-specific, any features used for classification need to generalize across both subjects and runs. We wish to identify interactions between components that may differ enough between groups to power a classifier. To identify these relationships we use summary statistics based on the within subject cross-correlation of the components temporal activity. The procedure is outlined as follows:

Common component classification procedure

- Component Extraction: For all scans extract common spatial maps and individual timecourses reflecting the subject-specific response patterns using Multi-Session Temporal Concatenation (MSTC)
- Feature Creation and Selection: Create a feature matrix for every subject describing the temporal relationship among the components within a subject. Perform feature selection by identifying component pairs within a subject that exhibit a different relationship between groups.
- Subject Classification: Train a random forests classifier using selected features and evaluate accuracy.

This procedure is tested over different sets of data, where every set has a unique combination of runs in which the components are extracted and the model is trained and tested.

Component extraction

To extract common spatial maps we use MSTC in which the data from different subjects are concatenated into a single data matrix and ICA is performed on this aggregate matrix (Calhoun et al., 2001b; Mckeown et al., 1998). We use components to explain the scan activity rather than the task-relatedness of the component timecourses under the hypothesis that the components form a sparse basis set that efficiently capture activity. MSTC is recommended where one is looking for common spatial patterns but cannot assume that the associated temporal response is consistent between sessions/subjects (Calhoun et al., 2001a). In this case the stimulus presentation is pseudo-randomized between subjects and runs so that subjects with different presentation orders will show engagement of specific spatial components at different points in time. Because of this, we hold the spatial maps constant across subjects and extract unique time series from common spatial maps, where scans from N subjects having spatial dimension S and time length T are concatenated into a matrix $\mathbf{X}$ having rows $N \times T$ and columns S. This can be expressed as a linear combination of $\tau < T$ common spatial components and their corresponding timecourses:

$$X_{ts}^{n} = \sum_{\mu=1}^{\tau} M_{t\mu}^{n} C_{\mu s}$$

where X_{ts}^{n} represents the raw scan intensity at timepoint $t < T$ within subject block n and spatial location $s \leq S$, $M_{t\mu}^{n}$ is the temporal intensity of component μ at time t within subject n, and $C_{\mu s}$ is the spatial intensity for component μ at spatial location s that is common across subjects. The element $M_{t\mu}^{n}$ measures the multiplicative weighting of the voxels and can be converted to subject-specific timeseries by

partitioning the timeseries for each of the n subjects. The number of components is determined using the Laplace approximation to the model order. MSTC was performed in FSL (Smith et al., 2004), and the resulting time series were then analyzed using the statistical software package R (R Development Core Team, 2006).

Feature creation and selection

We wish to classify our patients using the *relationship* between subject-specific component time series. The stimulus presentation differed across runs and the strongest artifact of motion we expect to vary highly across subjects. To counteract this, we allow the timeseries extracted within each run to be unique for every subject, and summarize the temporal cohesiveness within each subject using the correlation. This connectivity has been shown useful in analyzing psychiatric disorders such as schizophrenia (Jafri et al., 2007; Anderson et al., 2010), and will be calculated using the cross-correlation function (CCF), a linear measure of the similarity between two time series separated by a time lag. The CCF between time series within a subject eliminates the problem of the stimulus being presented at different times (Jafri et al., 2007). The lag allows for differences in reaction time to the stimuli between subjects. The maximal correlation that two time series may take over a range of lags is an indicator of the amount of information shared between them.

The CCF is calculated within a subject between all $\binom{\tau}{2}$ pairs of components.

$$\mathrm{CCF}\left(M_\alpha, M_\beta, l\right) = \frac{E\left[\left(m_{\alpha,t+l}-\overline{M_\alpha}\right)\left(m_{\beta,t}-\overline{M_\beta}\right)\right]}{\sqrt{E\left[\left(m_{\alpha,t}-\overline{M_\alpha}\right)^2\right]E\left[\left(m_{\beta,t}-\overline{M_\beta}\right)2\right]}}$$

where l is the time lag separating the two time series, $\overline{M_\alpha}$ is the mean of the entire time series, and E is the expected value operator. For conceptual interpretability this is converted to a distance measure, where a smaller distance between two components indicates that they are more correlated, with

$$d(\alpha,\beta) = \frac{1}{max_{l\in\mathrm{lags}}\left[\left|\mathrm{CCF}\left(M_\alpha, M_\beta, l\right)\right|\right]} - 1$$

where two components behaving similarly would yield a smaller $d(\alpha,\beta)$. For the Session 1 data in Model A there are $\binom{33}{2} = 528$ possible component comparisons. We reduce this dimensionality again by selecting 40 pairs with the largest signal-to-noise (SNR) ratio between groups to use as features, where the SNR is a measure of how different the correlations are between groups. For each feature pair (α,β):

$$\mathrm{SNR}(\alpha,\beta) = \frac{\left|\overline{d}(\alpha,\beta)_{\mathrm{IBS}} - \overline{d}(\alpha,\beta)_{\mathrm{Ctrl}}\right|}{\sigma_d}$$

where σ_d is the pooled standard deviation of the set $\{d(\alpha,\beta)_{IBS} \cup d(\alpha,\beta)_{Ctrl}\}$ and $\overline{d}$ is the average distance for the feature pair (α,β) for either the patients or controls. We perform a prior feature selection step because our classifier, random forests, faces difficulty when given many irrelevant features to sort among (Gashler et al., 2008). The threshold of 40 was chosen by visually inspecting the distribution of the sorted SNRs; the steepest drop off in the SNR occurred after roughly 40 nominees. Although classification accuracy would likely increase if this parameter would be optimized within the model, doing so would introduce bias in our estimate of the testing error. This extraction gives a feature matrix Φ of dimension $N \times 40$ to be used for classification.

Subject classification

Using the selected features, we trained and tested a random forests classifier (Breiman, 2001), selected because of its resiliency to overtraining in problems with limited numbers of observations. Random forests create many classification trees by resampling from both the observations and features at each node and subsequently making decision rules to minimize the misclassification rate of the sampled data within each tree. Decision trees are constructed and combined to create a "forest" that decides an observation's class by voting over the decisions made by each tree. This classifier has the added benefit of producing decision trees that indicate how the classifier actually operates, instead of a *black box* tool such as SVM where the actual decision boundaries are hyperplanes in a high-dimensional space. Because of the hierarchical structure of this model, a conclusion reached could be "the linkage between ICs A–D is important only after accounting for linkages A–B and C–D." The tree is then tested on observations that were not selected in the initial sampling to give the "out-of-bag" error, which is usually an unbiased estimate of the testing error. We estimate the classification testing accuracy using the out-of-bag error from the random forests classifier, where the accuracy is an estimate of how well the classifier would do if categorizing a scan from a previously unseen subject.

Because of the usage of *all* subjects to select the features used for classification, the out-of-bag error will become a *biased* estimate of the testing accuracy when the same run of data are used to test and train models (Simon et al., 2003; Kriegeskorte et al., 2008). We will calculate the amount of this bias within our faulty model to show that a biased model can lead to flawed inference. Because of this bias, we cannot generalize our predictive results outside of our model. We interpret the predictive accuracy instead as a metric to compare the models created within this study and will use it to evaluate the effects of stimulus changes, across-run predictability and artifact removal. Finally, we will measure the mistake in the out-of-bag error by running leave-one-out cross-validation outside the entire procedure in Models A^* and B^*, comparing it to the biased procedure in Models A and B. These results are shown in Table 1.

The "importance" of the features within the classifier is obtained to measure their contribution to discrimination. The features with the greatest importance correspond to component pairs that exhibit enough difference between the groups to classify upon. This is derived by permuting each feature randomly within the data matrix, and calculating how much the prediction accuracy of the out-of-bag data changes on the permuted data matrix. The normalized difference between the accuracy of the permuted and non-permuted data matrix indicate how essential that predictor was to classification compared to other features, where a higher importance indicates a feature more

Table 1
Predictive accuracy.

Model	Component extraction run	Model training run	Model testing run	Overall testing accuracy	Control testing accuracy	Patient testing accuracy	Artifact removal methods
A	1	1	1	79.17%	72.73%	84.62%	FSL
A*	1	1	1	58.33%	45.45%	69.23%	FSL
B	2	2	2	91.67%	100%	84.62%	FSL
B*	2	2	2	66.67%	36.36%	92.30%	FSL
C	1,2	1	1	70.83%	54.55%	84.2%	FSL
D	1,2	2	2	70.83%	72.73%	69.23%	FSL
E	1,2	1	2	62.5%	36.36%	84.62%	FSL
F	1,2	2	1	62.5%	45.45%	76.92%	FSL

The accuracy obtained in "Overall Testing Accuracy" is computed by extracting group-wide components in "Component Extraction Run", training a model based on temporal activity of these components from data in "Model Training Run", and evaluating the model using the temporal activity of the group-wide components from data in "Model Testing Run". In models* with cross-validation is performed outside all feature selection procedures. FSL artifact removal includes McFlirt for motion correction, Spatial Smoothing, High-Pass Filtering as described in the text.

essential for the classifier. If two features are correlated and the stronger predictor is permuted then the importance would be larger than if the weak predictor would be permuted, but a strong predictor with no covariates would have an even greater importance because it had no substitute. These "most important" features in Models G and H will be shown in Fig. 1.

We applied these methods to six different models from two runs, Run 1 and Run 2 (which were runs 1 and 4 in the original data).

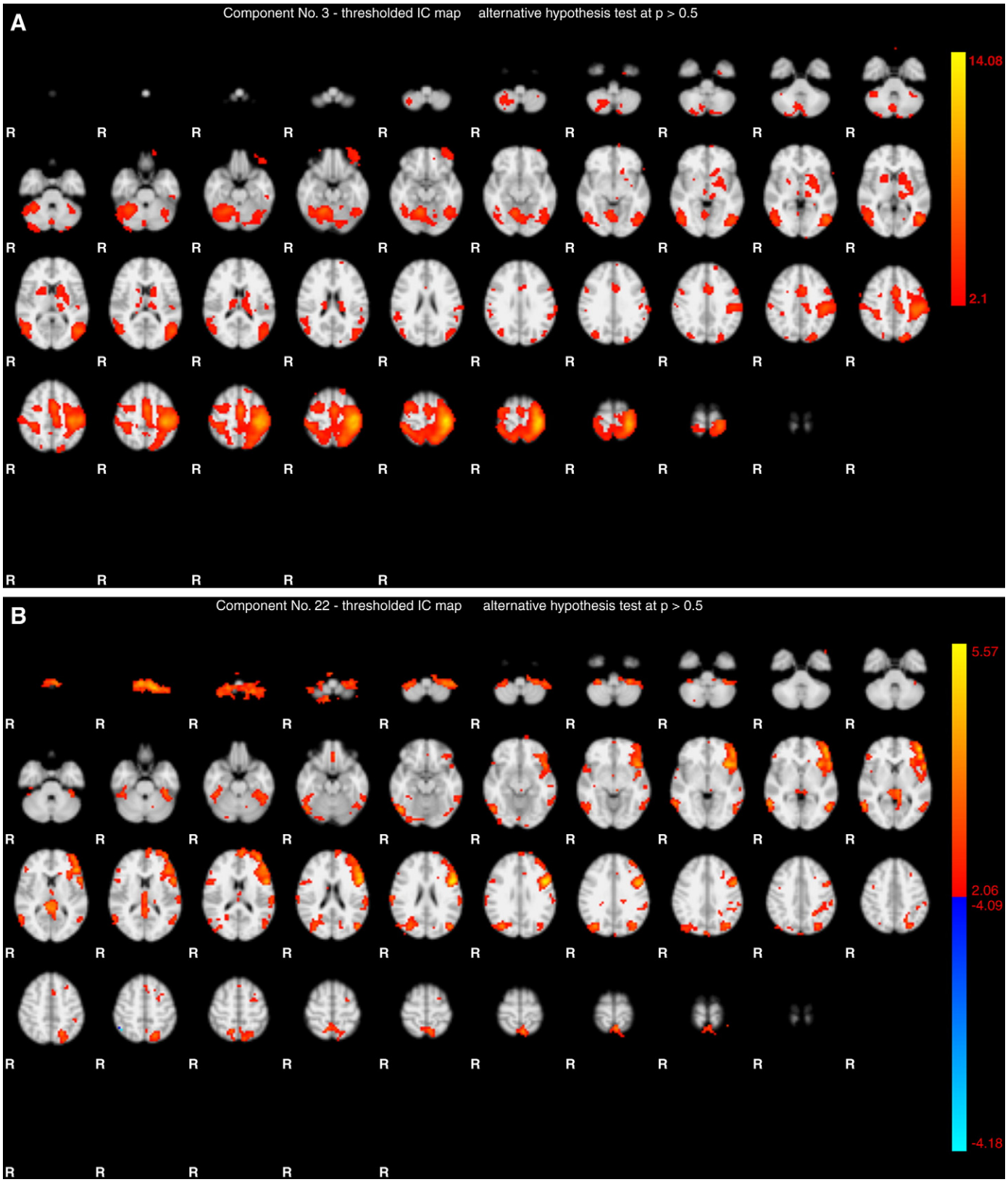

Fig. 1. Spatial map from the group-wide independent components analysis depicting the most important feature, component pair A and B, discriminating between patients and healthy control subjects during gastrointestinal stress task.

These models were created by altering the run in which the components were extracted and the models were trained and tested. Stimulus presentation order differed between subjects and between runs. Model A and Model B were component extracted, trained and tested all within a single run. Models C–F were component extracted, trained and tested on different runs to evaluate the stability of component extracted methods across runs as well as to test across-scan predictive accuracy. For Model E and Model F, components were extracted across both runs simultaneously to evaluate how components in one session predicted behavior in another. The average accuracy of these two models is an indicator of how well a model would perform when predicting the outcome of another session, when the patient groups are identical to those on which the model was trained. The results are shown in Table 1.

To investigate classifier dependence on physiological noise we applied the **Common Component Classification** method after removing artifacts. This additional cleaning is in addition to the standard artifact removal steps taken during the preprocessing stages. In Model G and H blatant noise group-wide components have been identified manually by inspecting each of the group wide components extracted from Run 2. We specifically searched for the following patterns in each component:

- A. Residual movement: pattern on the rim of the brain, forming a halo or semi halo.
- B. Pattern in ventral areas with known signal dropout, such as orbitofrontal cortex or temporal pole that does not follow anatomical boundaries.
- C. Physiological noise: pattern within cerebrospinal fluid areas, and/or pattern in areas adjacent to the great cerebral vein.

The components considered associated with artifacts are removed from consideration in the feature extraction stage and only components likely associated to BOLD response are used for model training and testing in Model G and Model H. If a component showed evidence of the aforementioned noise patterns noise, it was not removed. Because motion frequently occurred with the presentation of the stimulus, there were artifacts removed that also appeared to contain some task-related signal. These results are also presented in Table 2.

In Model I we used an automated ICA classifier (Tohka et al., 2008) to remove ICA components that reflected residual movement, physiological noise, and signal dropout. The classifier was trained with a subset of 15 runs, and was used to automatically classify and remove noise related ICA components on the individual runs. After the data were cleaned within subject, group-ICA was performed on the cleaned scans as described above in Methods. These results are also presented in Table 2.

Table 2
Predictive accuracy over artifact removed data.

Model	Component extraction run	Model training run	Model testing run	Overall testing accuracy	Control testing accuracy	Patient testing accuracy	Artifact removal methods
G	1	1	1	41.67%	27.27%	53.85%	FSL, Manual
H	2	2	2	79.17%	63.64%	92.31%	FSL, Manual
I	2	2	2	75%	75%	75%	FSL, Tohka

The decrease in testing accuracy with artifact removal indicates that some of the classification ability is due to systematic noise in the scans. "FSL": artifact removal using standard routines in software package. "Tohka": automated ICA classifier (within subject) and removal of artifactual components. "Manual": remove group-wide ICA components associated with artifacts.

Results

Using the procedure outlined above, we test our methods on six different models in which the component extraction run, model training run, and model testing run are altered as described in Methods. This allows us to test the reproducibility of effects across runs and the effectiveness of ICA methods for both signal extraction and artifact removal. Results of training and testing models within and across runs are seen in Models A–D in Table 1, showing that classification within a run is easier than classification across a run in which the stimulus has been reordered. For example, although our predictive accuracy was estimated at 79.2% and 91.67% for Models A and B, the predictive accuracy decreased once components were extracted across runs that contained different orders of stimulus presentation (70.83%.) When group-wide components were extracted for Models E and F, the predictive accuracy decreased further. The bias in the classifier is exhibited in the differences between Model A* and A and Model B*and B. These results collectively are shown in Table 1. The disparity between the biased and unbiased feature selection methods falsely inflates our testing accuracy by over 20% and makes the model effectively useless for inference.

The results of testing classification on data manually cleaned of artifacts are shown in Model G and Model H in Table 2. In Run 1, 66% of the components are identified as containing noise and subsequently removed, whereas 44% are identified as containing noise in Run 2. Noise is defined as any component that has one of the patterns described in the Methods section. The decrease in accuracy indicates that some of the classification power is due to systematic "noise" in the data and there is more noise present in the first session. The group-wide components most useful to classification for H correspond to the 3rd and 22nd components in the group-wide ICA. These are shown in Fig. 1.

Classification results using the data that had been artifact removed within subject (Tohka et al., 2008) are shown in Model I in Table 2. It was initially observed that due to the nature of the task (a warning followed by an uncomfortable visceral stimulation), some group-wide components contained residual motion artifacts. This residual movement had the same frequency, and co-occurred at the individual level with what was thought to be task-related BOLD response. After classification, the ICA components that were the best features of disease contained residual movement even though obvious motion artifacts were removed during the data cleaning process. Approximately 44% of these group-wide ICs extracted from the data cleaned previously using the Tohka method showed slight evidence of artifacts.

Discussion

Artifacts and classification

Three methods of artifact removal were performed on the Models using combinations of the FSL preprocessing routines, manual group-ICA removal and the methods of Tohka et al. (2008) where components associated with artifacts were removed *within* each subject prior to the group-ICA methods. The standard FSL routines were applied to Model B, resulting in classification of over 91%. Group-wide ICs still revealed components associated with residual movement, physiological noise, and signal dropout, so these were removed in Model H and the accuracy dropped to 79%. In Model I data were cleaned using the Tohka approach which led to a predictive accuracy of 75%, but the most predictive relationships were between components associated with motion, and roughly 44% of the group-wide ICs *still* showed evidence of artifacts. This points to the difficulty as a whole to remove motion from data. Although artifacts were removed within subject, this process was incomplete and left enough motion, physiological noise and signal dropout to be identified later during the group component extraction.

Motion is known to be a strong artifact in fMRI, and ICA methods are able to extract motion components (McKeown et al., 2003). However, these results show that artifact removal using ICA-based methods may be too strict in cases where the stimulus presentation is uncomfortable, as it induces motion concurrent with genuine neurological activation. The timeseries corresponding to the task activation may be confounded with the timeseries corresponding to motion, leading to spatial maps containing both movement and activation.

Even with a component containing purely motion there may exist signal; although these artifacts are considered "noise", this "noise" does in fact contain signal for classification when one group moves more than the other during scans. Such physiological artifacts might be predictive of the patient's diagnosis and increase classification accuracy, yet offer little neurophysiological information to the scientist. Using these components is at the discretion of the researcher and depends on the objectives of the classifier. If the objective is mainly to diagnose, then all useful signal should be used to build predictive models. If instead the objective is to find BOLD response activity specific to the IBS patients, then artifacts such as movement might offer little in the way of interpretation.

Reproducibility of stimulus effects across runs

As expected, classification within a run was easier than classification across runs. Predictive models across runs (Models E, F) resulted in a lower testing accuracy (Strother et al., 2002; Poline et al., 2006). A possible reason for the lower accuracy may be due to order effects that were present in the task design, where reactions to stimuli changed over repetitions and reduced the reproducibility of the classifier. Even with the reordered stimulus, the models still had some ability to predict across runs indicating as a whole stability of ICA methods.

The differences between A and B may be due to higher levels of noise present in Run 1 as well as effects of the run ordering, as there may be similar reactions to the stimulus between patients and controls during early scans. Curiously, although Models E and A are trained on the same runs and are both unbiased, Model E performs better at the more difficult task of classifying *across* runs. This could be due either to the component extraction being more stable across both Runs because of high noise levels in Run 1 or because of the easier separability of patients in Run 2.

Sources of bias in the predictive accuracy

The cross-validation accuracy can become biased when using all available data to select optimal features, and then running cross-validation on models produced using subsets of this data. This common mistake is present in some of the flawed models presented here where each subject had $\frac{1}{24}$ of the influence in choosing which "features" were most useful. Although the models were created without using all the subjects, the *a priori* usage of all subjects for the feature selection introduces bias into the final model (Demirci et al., 2008). The bias within the models inflated the estimate of the accuracy by roughly 20%(Models A, A* and Models B, B*). Any bias in the cross-validation error makes generalization of the results dubious, as a flawed model can only lead to flawed inference.

We investigate the strength of this bias further by observing how the selected features change using cross-validation. There were 24 total cross-validation runs (using leave-one-out) where 40 features were selected at each iteration to be features. We tally the frequency at which each feature is selected over these 24 runs. Only 77 (14.14%) of the 528 possible features were ever selected. Of the features selected at all, 50% of those were selected at least 9 times and 25% chosen all 24 times. This indicates some stability in the feature-selection process, where a biased and unbiased model would likely share half of the same selected features.

In addition to models being sensitive to task ordering, feature selection, and artifactual noise, it is often not appreciated that the predictive accuracy also can be influenced by sample sizes and composition (Demirci et al., 2008). These are instances of bias coming from the data, and not from errors in the methods themselves. This often occurs because of the non-randomness in the recruitment of subjects, leading to a patient sample that exhibits more homogeneity than the patient population. Cross-validation of an "unbiased" model based on that sample would still produce a biased estimate of the testing error in this situation because it would assume the incoming data were similar to the data used to create the model.

The composition of the sample can also bias the predictive accuracy; a disproportionate fraction of cases from one category makes it very easy to build an "accurate" classifier just by chance (Demirci et al., 2008; Pereira et al., 2009). For example, if a set of training scans contained 90% Syndrome X patients, a classifier could be 90% accurate just by assigning to every case a "Syndrome X" label. When comparing models then across different data sets, the overall accuracy is not an appropriate measure of which classifier is strongest. Rather, the accuracy of that classifier relative to *chance* for that dataset is a stronger indicator of the ability.

The predictive accuracy can also become biased when the ratio of patients/controls differs between the sample and the population. Using our classifier that assigns to every subject the label of "Syndrome X", we would correctly identify 90% of the people in our sample. If the *testing population* however had 90% healthy normal patients, the *actual* testing accuracy of this same classifier would drop to 10%. This inferential error is illustrated in this manuscript; in practice the classifier created in Model E in Results section would have an accuracy rate much closer to 36.36% than to 84.62% on a set of random women, since we would expect a disproportionate fraction of the population is *not* to be labeled as IBS. In circumstances where the proportion of patients in which the model is trained does not match the proportion of patients one expects to see in practice, the estimated testing accuracy of the entire model cannot be assumed equivalent to the accuracy one will see in practice, but instead must be estimated and adjusted for using the error rates within each subject group (Wood et al., 2007). In addition, most classification methods can be informed with prior probabilities to adjust for this bias within the actual model.

In addition to randomness within the data, randomness can also exist within the methods. This randomness becomes greater when the models are constructed on smaller samples, as is done here. It can be introduced in the cross-validation estimate because of sampling variation in splitting the data into training and testing sets. For example, in split-half cross-validation the data are divided into two parts, and two models are trained on one part of the data and tested on the other. The testing accuracy is averaged across both models to give the estimate of the cross-validation error. This averaging minimizes the variation of the estimate with respect to the initial partitioning of the data. However with a total of n observations, there were $\frac{1}{2}\binom{n}{\frac{n}{2}}$ ways to divide the data into two sets to create the cross-validation error. Exacerbated by small sample sizes, these $\frac{1}{2}\binom{n}{\frac{n}{2}}$ partitions may vary in their estimates of the cross-validation error. This is not an issue with all methods of estimating the cross-validation error. Leave-one-out cross validation is not subject to this issue of partitioning because there is only one way to partition the data such that only one observation is excluded each time. Other less-affected cross-validation approaches include repeated random sub-sampling validation and random forests, both which create new partitions with each iteration (Bouckaert and Frank, 2004), but they are hindered in that they may not sample all observations since they are sampling with replacement. As sample-sizes become larger, this variability will decrease because the partitions will also become larger. Thus, this risk is greatest for small-sample studies.

Some models, such as random forests, use resampling both of the data *and* the features. Because of this, even with identically sampled data, different learning rules can be created for each "tree" leading to slightly different forests. Even though results are averaged over entire forests of 500 trees, small samples still result in larger variations of results. Running these models repeatedly will lead to slightly different estimates of the testing error. In circumstances when the model is run repeatedly, one must be careful not to "cherry pick" the best run as being representative of the model as a whole, since this is undoubtedly an unrealistic estimate in practice.

Errors and inference

In standard published literature a Type 1 error of $\alpha = .05$ is usually assigned, and *p*-values less than this threshold lead to a rejection of the null hypothesis. This parameter choice implies that with an average of 20 studies, a false positive will be experienced. Publication bias is a phenomenon in which a disproportionate number of studies reject the null hypothesis. For example in one study involving experimental psychology journals, a total of 93.5% of studies rejected their chosen null hypothesis, even when the Type 1 error was set at $\alpha = .05$ (Sterling et al., 1995). Usually only "statistically significant" results are published, so when a new study rejects the null hypothesis they will have very little opposition in the literature stating the contrary even though the result is a false positive. This problem may also exist in the literature of fMRI classification, as every paper published shows their selected patient group to be classifiable with their choice of methods. One possible remedy to validate the methods is to test them against multiple sets of data, to guard against the classifier working merely because of a particular parameter choice or combination of routines. Another remedy is to work on patient groups known to be classifiable in the literature.

This paper discriminated between IBS and normal controls, a patient group which has not been previously used for classification in the literature. However, studies have demonstrated discrimination between schizophrenia patients and normal controls using the cross-correlations of ICs (Sakoglu et al., 2009; Anderson et al., 2010), so it is conceivable that the methods presented here would prove more effective discriminating on a mental illness patient group. However, because of the small sample size presented here, it would quite brash for us to declare that we have created a model capable of diagnosing IBS. The predictive accuracy claimed in fMRI studies is likely overly optimistic but will become more realistic when studies are able to train and test their methods on larger data sets.

In pairwise comparisons of component activity, the small sample sizes frequently used in fMRI classification can result in an inflated Type 2 (β) error where we fail to recognize genuine between-group differences (Cohen, 1992). In relation to this study, a Type 2 error would be committed by assuming that a component relationship between groups is identical, when it is actually different. At a confidence level of $\alpha = .05$, 128 subjects total are needed to test the difference between group means (medium effect size) with a power level of $1 - \beta = .8$ (Cohen, 1992). This study only has 24 subjects, giving us a much lower power and higher Type 2 error. For classification analysis, confidence intervals can be a more stable way to illustrate accuracy rates over the iterations of cross-validation (Brown et al., 1999).

In the context of fMRI classification studies, with enough data changes and alteration of methods even random algorithms on identical patient groups can eventually "classify" between identical groups. In addition, we demonstrated here that artifacts can aid in classification ability. Even if our models are capable of capturing genuine signal within a run, this signal may in fact change across runs and subjects. Because of this, even with solid methodology, the cross-validation error may still be an overly optimistic and biased estimate of the in-practice testing error, and it can be difficult to draw predictive inference outside of the sample used to create and test it. To increase confidence in the power of fMRI classification, methods should be tested on multiple runs practicing "statistical pluralism", interchanging different classification machines (support vector machines, random forests, etc.) and feature extraction methods when possible (Lange et al., 1999). If the proposed methods are valid, then reproducibility should be seen.

Conclusion

This analysis identified how the coupling of components was a measure capable of discrimination between patients and controls. We have demonstrated that group-wide component extraction methods such as MSTC can extract and identify sparsely coded basis functions (independent components) useful for classification within runs, but that ICA methods in general can be sensitive to session and task ordering as well as to systematic "noise" across sessions. When possible, components should be extracted over as large a pool of subjects as possible. ICA-based artifact removal may be less useful when using an uncomfortable stimulus since activation and motion are induced simultaneously, leading to confounding of the onset of true signal and motion. The elimination of artifacts through preprocessing can lead to better neurophysiological interpretations of the underlying disorders, but often result in less accurate classification.

The changes in classifier performance that followed manual artifact removal emphasize another important issue in machine methods. Namely, that the decision boundaries are not guaranteed to reflect important functional processes, shown here by the power of artifacts for classification. We believe that a more principled selection of classifier dimension, such as the independent spatial components used here, is likely to lead to more physiologically *interpretable* results that are better able to inform neurological research goals; the loss of classification accuracy that results from the removal of physiologically implausible dimensions is more than made up for by the potential for discovery.

While the present classification methods are somewhat unconventional in that they utilize independent components as dimensions for classification, we argue that the general findings of the limitations of machine learning applied to fMRI are broader in scope because of the nature of fMRI data. Because fMRI data are so costly to obtain, sample sizes are small causing difficulties in extracting true signal as well as difficulties in establishing error rates and drawing inference. fMRI data suffer from low SNR, so physiological artifacts can produce stronger signal than even genuine neurological activity. Preprocessing (motion correction, slice timing correction) can reduce physiological noise, but these steps alone do not completely eliminate systematic artifacts from contributing to classification accuracy. With smaller samples comes the risk of greater statistical errors.

The reported accuracy in fMRI classification studies may not be realized in practice partially due to improper statistical methodology and inference (Simon et al., 2003). The bias also can be introduced by having small samples that do not capture accurately the testing population or by having skewed proportions of patients in the model construction phase that would not be seen in practice, as well as the sensitivity to scan conditions. Models may become very accurate simply by chance in circumstances where the patient/control ratios are skewed; therefore models should not be compared across different studies without first adjusting for different subject ratios within each training dataset. Even then comparison is hazy because the predictive accuracy within a study can easily be inflated by running the models multiple times, as there exists variation in partitioning the data and also in constructing models.

IBS is currently an ill-defined syndrome with patients presenting a high variety of symptoms (Tillisch and Chang, 2005). Furthermore, high percentage of IBS patients has comorbidity with anxiety disorders leading to several authors to propose the disease as a disruption of the gut–brain axis (Mayer et al., 2006). Thus, our

classification results might be due to the heterogeneity of the disease and patient population, in addition to being a product of the high motion levels associated with the task paradigm.

So what, then, do we ultimately learn from machine learning? Although fMRI classification methods can produce quite impressive results, one must exercise caution before attributing high accuracy rates to the mathematical machinery behind them. Testing the data across runs can help assess the true strength of these models, as accurate subject classification based on chance or systematic physiological noise (i.e. motion) is always possible. With the small sample sizes typical in many fMRI classification studies, the results obtained must be scrutinized carefully for statistical inference errors. Even with powerful machine-learning tools at our disposable, a model is ultimately limited by the data used to create it. Machine learning gives as much insight into the shortcomings of our methods and data as it does the potential of them, and without properly critiquing our own work we are ultimately left with more questions raised than answered.

Acknowledgments

National Institutes of Health Grants DA023422 (A.A), DA026109 (M.S.C., A.A.), R24DK48351 (E.A.M.), R24 AT002681 (E.A.M.), and DK071626 (J.L). The authors gratefully acknowledge Steve Berman, Brandall Suyenobu, and Jean Stains for their invaluable efforts in the acquisition of the data, and Pamela Douglas and Alan Yuille for helpful conversations.

References

Anderson, A., Dinov, I.D., Sherin, J.E., Quintana, J., Yuille, A., Cohen, M.S., 2010. Classification of spatially unaligned fMRI scans. NeuroImage 49, 2501–2519.

Berman, S.M., Naliboff, B.D., Suyenobu, B., Labus, J.S., Stains, J., Ohning, G., Kilpatrick, L., Bueller, J.A., Ruby, K., Jarcho, J., Mayer, E.A., 2008. Reduced brainstem inhibition during anticipated pelvic visceral pain correlates with enhanced brain response to the visceral stimulus in women with irritable bowel syndrome. J. Neurosci. 28 (2), 349–359 URL http://www.jneurosci.org/cgi/content/abstract/28/2/349.

Bouckaert, R.R., Frank, E., 2004. Evaluating the replicability of significance tests for comparing learning algorithms. In PAKDD. Springer, In. pp. 3–12.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brown, L.D., Cai, T.T., Dasgupta, A., 1999. Interval estimation for a binomial proportion. Statistical Science 16, 101–133.

Calhoun, V., Adali, T., Pearlson, G., Pekar, J., 2001a. Group ICA of functional MRI data: separability, stationarity, and inference.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001a. A method for making group inferences from functional MRI data using independent component analysis November Hum Brain Mapp 14 (3), 140–151 URL http://view.ncbi.nlm.nih.gov/pubmed/11559959.

Cohen, J., 1992. A power primer. Psychological Bulletin 112, 155–159.

Demirci, O., Clark, V.P., Magnotta, V.A., Lauriello, J., Kiehl, K.A., Pearlson, G.D., Calhoun, V.D., 2008. A review of challenges in the use of fMRI for disease classification/characterization and a projection pursuit application from a multi-site fMRI schizophrenia study. Brain Imaging and Behavior 2, 207–226 URL http://www.springerlink.com/content/870x362571283 0l0.

Drossman, D.A., 2006. The functional gastrointestinal disorders and the Rome III process. Gastroenterology 130 (5), 1377–1390 URL http://www.sciencedirect.com/science/article/B6WFX-4JW7 C37-D/2/537d0372c9053457d2636ba6c881af1f.

Fan, Y., Shen, D., Davatzikos, C., 2006. Detecting cognitive states from fMRI images by machine learning and multivariate classification, 89.

Ford, J., Farid, H., Makedon, F., Flashman, L.A., Mcallister, W., Megalooikonomou, V., Saykin, A.J., 2003. Patient classification of fMRI activation maps. Proc. of the 6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'03, 58–65.

Gashler, M., Giraud-Carrier, C., Martinez, T., 2008. Decision tree ensemble: small heterogeneous is better than large homogeneous. Machine Learning and Applications. ICMLA '08. Seventh International Conference on, pp. 900–905. URL http://dx.doi.org/10.1109/ICMLA.2008.154.

Jafri, M.J.J., Pearlson, G.D.D., Stevens, M., Calhoun, V.D.D., 2007. A method for functional network connectivity among spatially independent resting-state components in schizophrenia November Neuroimage URL http://dx.doi.org/10.1016/j.neuroimage.2007.11.001.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2008. Circular inference in neuroscience: the dangers of double dipping 5 J. Vis. 8 (6), 88-88 URL http://journalofvision.org/8/6/88/.

Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.A., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis September NeuroImage 10 (3), 282–303 URL http://dx.doi.org/10.1006/nimg.1999.0472.

Mayer, E.A., Naliboff, B.D., Craig, A.D., December 2006. Neuroimaging of the brain–gut axis: from basic understanding to treatment of functional GI disorders. Gastroenterology 131, 1925–1942.

Mckeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., Sejnowski, T., 1998. Analysis of fMRI data by blind separation into independent spatial components. URL citeseer.ist.psu.edu/mckeown98analysis.html

McKeown, M.J., Hansen, L.K., Sejnowsk, T.J., 2003. Independent component analysis of functional MRI: what is signal and what is noise? Current Opinion in Neurobiology 13 (5), 620–629.

MarchPereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1 Suppl) URL http://dx.doi.org/10.1016/j.neuroimage.2008.11.007.

Poline, J., Strother, S., Dehaene-Lambertz, G., Egan, G., Lancaster, J., 2006. Motivation and synthesis of the FIAC experiment: the reproducibility of fMRI results across expert analyses. Human Brain Mapping 27, 351–359 organization for Human Brain Mapping 2009 Annual Meeting.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0. URL http://www.R-project.org.

Sakoglu, U., Michael, A., Calhoun, V., 2009. Classification of schizophrenia patients vs. healthy controls with dynamic functional network connectivity. NeuroImage 47 (Supplement 1), S57-S57 organization for Human Brain Mapping 2009 Annual Meeting. URL http://www.sciencedirect.com/science/article/B6WNP-4X3P HYG-7P/2/c7ec35d57bdba23c296512add3d34cab.

Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M., 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl. Cancer Inst. 95 (1), 14–18 URL http://jnci.oxfordjournals.org.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansenberg, H., Bannister, P.R., Luca, M.D., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N.D., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, 208–219.

Sterling, T.D., Rosenbaum, W.L., Weinkam, J.J., 1995. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician 49 (1), 108–112 URL http://www.jstor.org/stable/2684823.

Strother, S.C., Anderson, J., Hansen, L.K., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. URL http://www2.imm.dtu.dk/pubdb/p.php?1445

Tillisch, K., Chang, L., 2005. Diagnosis and treatment of irritable bowel syndrome: state of the art July Current Gastroenterology Reports 4, 249–256 URL http://www.springerlink.com/content/u85637006256m382.

Tohka, J., Foerde, K., Aron, A.R., Tom, S.M., Toga, A.W., Poldrack, R.A., 2008. Automatic independent component labeling for artifact removal in fMRI February NeuroImage 39 (3), 1227–1245 URL http://dx.doi.org/10.1016/j.neuroimage.2007.10.013.

Wood, I.A., Visscher, P.M., Mengersen, K.L., 2007. Classification based upon gene expression data: bias and precision of error rates. Bioinformatics 23 (11), 1363–1370.

Zhang, L., Samaras, D., 2005. Machine learning for clinical diagnosis from functional magnetic resonance imaging. IEEE Conference on Computer Vision and Pattern Recognition 1211–1217 CVPR.