European Journal of Nuclear Medicine and Molecular Imaging (2019) 46:2370–2379
https://doi.org/10.1007/s00259-019-04400-w

ORIGINAL ARTICLE

Controls-based denoising, a new approach for medical image analysis, improves prediction of conversion to Alzheimer's disease with FDG-PET

Dominik Blum[1] · Inga Liepelt-Scarfone[2,3] · Daniela Berg[4] · Thomas Gasser[2,3] · Christian la Fougère[1] · Matthias Reimold[1] · for the Alzheimer's Disease Neuroimaging Initiative

Received: 11 February 2019 / Accepted: 11 June 2019 / Published online: 24 July 2019

Abstract

Objective The pattern expression score (PES), i.e., the degree to which a pathology-related pattern is present, is frequently used in FDG-brain-PET analysis and has been shown to be a powerful predictor of conversion to Alzheimer's disease (AD) in mild cognitive impairment (MCI). Since, inevitably, the PES is affected by non-pathological variability, our aim was to improve classification with the simple, yet novel approach to identify patterns of non-pathological variance in a separate control sample using principal component analysis and removing them from patient data (controls-based denoising, CODE) before calculating the PES.

Methods Multi-center FDG-PET from 220 MCI patients (64 non-converter, follow-up $\geq$ 4 years; 156 AD converter, time-to-conversion $\leq$ 4 years) were obtained from the ADNI database. Patterns of non-pathological variance were determined from 262 healthy controls. An AD pattern was calculated from AD patients and controls. We predicted AD conversion based on PES only and on PES combined with neuropsychological features and ApoE4 genotype. We compared classification performance achieved with and without CODE and with a standard machine learning approach (support vector machine).

Results Our model predicts that CODE improves the signal-to-noise ratio of AD-PES by a factor of 1.5. PES-based prediction of AD conversion improved from AUC 0.80 to 0.88 ($p = 0.001$, DeLong's method), sensitivity 69 to 83%, specificity 81% to 88% and Matthews correlation coefficient (MCC) 0.45 to 0.66. Best classification (0.93 AUC) was obtained when combining the denoised PES with clinical features.

Conclusions CODE, applied in its basic form, significantly improved prediction of conversion based on PES. The achieved classification performance was higher than with a standard machine learning algorithm, which was trained on patients, explainable by the fact that CODE used additional information (large sample of healthy controls). We conclude that the proposed, novel method is a powerful tool for improving medical image analysis that offers a wide spectrum of biomedical applications, even beyond image analysis.

Keywords Pattern expression score · Principal component analysis · Denoising · Physiological variance

Introduction

Already at preclinical stages of Alzheimer's disease (AD), typical metabolic alterations can be observed in cerebral gray matter and ^{18}F-fluorodesoxyglucose (FDG)-PET has been used to predict conversion from mild cognitive impairment (MCI) to AD, albeit with variable success [4, 5, 7, 8, 13, 16, 18, 22, 26]. Recently, the degree to which a single pathology-specific metabolic pattern is present

This article is part of the Topical Collection on Neurology

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-019-04400-w) contains supplementary material, which is available to authorized users.

✉ Dominik Blum
dominik.blum@med.uni-tuebingen.de

Extended author information available on the last page of the article.

(pattern expression score, PES) has been shown to be a strong predictor of conversion from MCI to AD, particularly when combined with clinical features [3].

Medical images contain different sources of non-pathological variance, that is, physiological variance and variance related to methodology (e.g., anatomical variability not compensated by spatial normalization and different scanners used in multicenter studies). Hence, inevitably, the pattern expression score not only reflects pathological alterations but is also affected by patterns of non-pathological variability present in the individual image data due to the fact that these patterns are likely to bear at least some resemblance to the pathology-related pattern.

Principal component analysis, PCA, [19] has been used extensively in statistical image analysis in order to decompose a set of images into uncorrelated components (principal components, PC) that are sorted with respect to the amount of variance they explain. While PCA is usually applied in order to identify pathology-specific patterns [16, 17, 23], we here propose the novel approach to calculate PCA in a separate control sample in order to detect patterns of non-pathological variability and removing them from patient data (*controls-based denoising*, CODE) before calculating the PES of a given pathology-related pattern.

The goal of this paper was to present and evaluate CODE using multi-center FDG-PET data from healthy controls and patients with MCI and AD from the ADNI database. To this end, we first derived the theoretical condition under which CODE improves PES-based image analysis and estimated the benefit of CODE (*net benefit index*) for an AD-related pattern. We then applied CODE to improve prediction of conversion from MCI to AD based on FDG-PET and, optionally, neuropsychological and genetic features. For comparison of classification performance, we also calculated a voxel-based standard machine learning approach.

In addition to the main manuscript, we provide a deeper analysis of theoretical aspects of CODE in online Supplementary Material.

Materials and methods

Mathematical notation

Scalars are denoted by lowercase italic letters. Image data are represented by row vectors, denoted by lowercase bold letters. The transpose of a row vector is given by a superscript T. When referring to all subjects $i = 1 \ldots N$, image data are given by a $(N \times D)$ - matrix, denoted by a bold uppercase letter $\mathbf{R}$. Coefficients of the principal component k are denoted by $\mathbf{c}_k$. Image data after denoising are denoted with a superscript d $\left(\mathbf{p}_i^{\mathbf{d}}\right)$.

Theory

Generally speaking, image data in patients can be decomposed into a pathology-related component and a residual component with non-pathological variability. Let $\mathbf{b}$ be a known pattern that represents pathology-related alterations, centered and normalized to unit length, here also denoted as pattern of interest. The image data of a single patient i can then be expressed as

$$\mathbf{p}_i = pss_i \mathbf{b} + \mathbf{r}_i \tag{1}$$

with pss_i (pathology-specific score) representing the degree to which the pattern of interest is expressed due to pathology. Of note, $\mathbf{b}$ and the residual image $\mathbf{r}_i$ are not orthogonal, corresponding to the fact that patterns of non-pathological variance present in patient data may bear some resemblance to $\mathbf{b}$. pss_i can be estimated from patient data by calculating the inner product between $\mathbf{p}_i$ and the pattern of interest $\mathbf{b}$, which is the *pattern expression score*

$$pes_i = \mathbf{p}_i \mathbf{b}^{\mathrm{T}} = (pss_i \mathbf{b} + \mathbf{r}_i)\mathbf{b}^{\mathrm{T}} = pss_i + \mathbf{r}_i \mathbf{b}^{\mathrm{T}} \tag{2}$$

with the estimation error

$$\varepsilon_i = \mathbf{r}_i \mathbf{b}^{\mathrm{T}}, \tag{3}$$

which is independent from pss.

Our aim is to minimize the error ε by removing patterns of physiological variance identified in a sample of controls using PCA. We calculated PC coefficients $\mathbf{c}_k$ using the subject-by-subject covariance matrix [2] so that $\mathbf{c}_k = (1/\sqrt{N\lambda_k})\mathbf{C}\mathbf{c}_k^{ss}$, with the $(N \times D)$ - matrix $\mathbf{C}$ containing the controls images, and λ_k and $\mathbf{c}_k^{ss}$ being the eigenvalues and eigenvectors of the subject–subject covariance matrix.

Before calculating PCA, we decided to center each voxel, in other words, to subtract the mean image of controls from each control subject. This had the effect that all PCs, including the first one, reflect patterns of variability across subjects, while, without centering, the first PC would correspond to the mean image of controls.

PCs determined in controls are removed from patient data after calculating the expression score of each $\mathbf{c}_k$ in each patient (i.e.. the inner product $\mathbf{p}_i \mathbf{c}_k^{\mathrm{T}}$) and using this expression score as a weighting factor for $\mathbf{c}_k$. Throughout the manuscript, removing $\mathbf{c}_k$ (with k referring to any subset of available PCs) from $\mathbf{p}_i$ will also be referred to as *denoising*

$$\mathbf{p}_i^{\mathbf{d}} = \mathbf{p}_i - \sum_k \mathbf{p}_i \mathbf{c}_k^{\mathrm{T}} \mathbf{c}_k. \tag{4}$$

Denoising is a linear operation, i.e.,

$$\mathbf{p}_i^{\mathbf{d}} = (pss_i \mathbf{b} + \mathbf{r}_i)^{\mathbf{d}} = pss_i \mathbf{b}^{\mathbf{d}} + \mathbf{r}_i^{\mathbf{d}} \tag{5}$$

with $\mathbf{b^d}$ being the denoised pattern of interest and $\mathbf{r^d}$ the denoised residuals, respectively:

$$\mathbf{b^d} = \mathbf{b} - \sum_k \mathbf{b}\mathbf{c}_k^{\mathrm{T}}\mathbf{c}_k \tag{6}$$

$$\mathbf{r}_i^{\mathbf{d}} = \mathbf{r}_i - \sum_k \mathbf{r}_i\mathbf{c}_k^{\mathrm{T}}\mathbf{c}_k. \tag{7}$$

We now calculate the pattern expression score from $\mathbf{p}_i^{\mathbf{d}}$ instead of $\mathbf{p}_i$:

$$pes_i^d = \mathbf{p}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}} = pss_i\mathbf{b^d}\mathbf{b}^{\mathrm{T}} + \mathbf{r}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}} \tag{8}$$

with pes_i^d now being an estimator for $pss_i\mathbf{b^d}\mathbf{b}^{\mathrm{T}}$, so that

$$\frac{pes_i^d}{\mathbf{b^d}\mathbf{b}^{\mathrm{T}}} = \frac{\mathbf{p}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}}}{\mathbf{b^d}\mathbf{b}^{\mathrm{T}}} = pss_i + \frac{\mathbf{r}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}}}{\mathbf{b^d}\mathbf{b}^{\mathrm{T}}} \tag{9}$$

with the new estimation error

$$\varepsilon_i^d = \frac{\mathbf{r}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}}}{\mathbf{b^d}\mathbf{b}^{\mathrm{T}}} \tag{10}$$

that is also independent from pss.

By comparing the estimation error before denoising, Eq. 3, and after denoising, Eq. 10, we can express the factor by which CODE reduces the error associated with estimating pss_i. We define this factor

$$\textit{net benefit index} = \frac{\sigma(\varepsilon)}{\sigma(\varepsilon^{\mathbf{d}})} = \frac{\sigma\left(\mathbf{R}\mathbf{b}^{\mathrm{T}}\right)}{\sigma\left(\mathbf{R^d}\mathbf{b}^{\mathrm{T}}\right)}\mathbf{b^d}\mathbf{b}^{\mathrm{T}}. \tag{11}$$

Of note, for calculation of the new pattern expression score pes_i^d, it is irrelevant if PCs are removed from the image data $\mathbf{p}_i$ or from the pattern of interest $\mathbf{b}$

$$\mathbf{p}_i^{\mathbf{d}}\mathbf{b}^{\mathrm{T}} = \mathbf{p}_i\mathbf{b}^{\mathbf{d}\mathrm{T}}. \tag{12}$$

For computational efficacy, we chose the latter option.

Subjects

From the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), we obtained multi-center FDG-PET scans from healthy subjects ($n = 262$, classified as healthy subjects at all visits), patients with AD ($n = 229$) and with MCI ($n = 220$, ADNI group *LMCI, Late MCI*). Of the MCI patients, 156 converted to AD within 4 years and 64 remained stable within the follow-up (minimum follow-up of 4 years). Subjects with missing clinical and neuropsychological scores were excluded. ADNI criteria for characterization of healthy subjects, MCI and AD patients can be found in [20]. Demographic and clinical information are summarized in Table 1.

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

Image pre-processing

We obtained image data at the ADNI pre-processing step *Co-reg, Avg, Standardized Image and Voxel Size*, i.e., images were already spatially oriented into a 160 × 160 × 96 voxel image grid with 1.5-mm cubic voxels. The image grid is oriented such that the anterior-posterior axis of the subject is parallel to the AC-PC line (no non-linear warping). We spatially normalized these images using SPM and the PET template (SPM12, Wellcome Department of Imaging Neuroscience, London, UK). We maintained the SPM default bounding box (79 × 95 × 78 voxels) and the template voxel size (2 mm, isotropic) and smoothed with a 10-mm Gaussian kernel. A whole-brain mask was created from the mean image of healthy controls by including all voxels with values greater than 0.2 of the maximum, resulting in 296,364 voxels in total. Images were intensity-scaled to achieve a mean voxel value of 1 in the whole-brain mask. All data analysis was done with MATLAB (MathWorks, Sherborn, MA, USA, Version 9.2).

Implementation of CODE

Calculation of PCs, denoising, and calculation of PES was implemented in MATLAB (MathWorks, Sherborn, MA, USA, Version 9.2). To reduce computational cost, we replaced MATLAB's built-in PCA with our own fast implementation as described above. The related MATLAB functions (fastpca.m and codenoise.m) are publicly available via MathWorks File Exchange.

Illustrating the effect of CODE on voxelwise variance

For illustration, we performed PCA on a randomly chosen subgroup of 200 out of 262 cognitively normal subjects. In the 62 remaining subjects, we calculated voxelwise variance before and after denoising three times using Eq. 4: First, we removed PCs 1 and 2, second, we removed PCs 1 to 10 and last, we removed PCs 1 to 45, which correspond to 80% of the variance of all PCs.

Estimating the benefit of CODE from healthy controls

Figure 1 gives an overview of the steps involved in estimating the net benefit index associated with our AD-related pattern, i.e., a pattern that had been calculated by

Table 1 Demographic and clinical data of ADNI subjects

	Sample size	Education (years)	Age at PET (years)	ADAS13[1]	ApoE4 (%)[2]	Gender (m/f %)
Healthy controls	262[3]	16.4 ± 2.8	74.4 ± 6.1	9.0 ± 4.2	75/23/2	50/50
AD	229	15.3 ± 2.9	75.0 ± 7.8	30.5 ± 8.2	34/48/18	59/41
MCI-stable[4] ('non-converter')	64	16.1 ± 2.7	73.5 ± 7.4	14.6 ± 5.5	63/31/6	61/39
MCI-AD converter[5]	156	15.8 ± 2.8	74.0 ± 7.2	21.9 ± 5.7	34/49/17	59/41

Education, age, and ADAS13 are reported as mean ± SD. MCI, mild cognitive impairment; AD, Alzheimer's disease

[1]Alzheimer's disease assessment scale 13

[2]Relative frequency of 0, 1, or 2 ApoE 4 alleles in percent

[3]Either 262 or 200 healthy controls for calculation of PCs, 262 for calculation of the AD pattern and 62 serving as residuals for calculation of the net benefit index

[4]No conversion to AD during follow-up (≥ 4 years)

[5]Conversion to AD within 4 years

means of a group difference (mean image from 229 AD patients minus mean image from 262 controls). Healthy controls were randomly subdivided into two groups: (1) 200 subjects were used to obtain 199 PCs used for denoising and (2) 62 subjects were used to represent the residuals R in Eq. 1. We repeated the analysis 100 times and calculated the mean and standard deviation of the net benefit index across all runs. In a prior version of the manuscript, to calculate the AD pattern, we used a smaller sample of AD subjects, which did not overlap with the two abovementioned groups. While a much smaller sample of AD and control subjects suffices to obtain an AD-related pattern, we decided to use all available subjects to avoid an overestimation of the net benefit of CODE due to an effect described in the Supplementary Material.

Applying CODE for conversion prediction

We applied CODE to FDG-PET from 220 MCI patients to improve prediction of conversion to AD. First, we calculated PCA in all 262 control subjects. We chose two different variants of CODE for evaluation: denoising with all available PCs (*full denoising*) and denoising with the number of PCs that correspond to 80% of the variance [10] in all PCs, which was the case for the first 45 PCs.

We classified converters and non-converters using i) PES only (ROC-analysis) and ii) PES in combination with a set of parameters recently recommended for predicting conversion to AD [11]. PES and parameters were combined using a linear support vector machine (SVM). Specifically, we used the following parameters: Alzheimer's Disease Assessment Scale (ADAS-Cog 11 and 13), ApoE4 genotype, Clinical Dementia Rating (CDR-SB), Functional Assessment Questionnaire (FAQ), Mini-Mental State Examination (MMSE) and Rey Auditory Verbal Learning Test (RAVLT immediate and RAVLT learning). For each clinical feature, we calculated its performance to classify converters from non-converters.

To evaluate the effect of CODE on physiological variability in the PES, we correlated PES with age before and after applying CODE and tested the difference as described in [24].

For comparison of classification performance, we classified converters and non-converters with a SVM based on

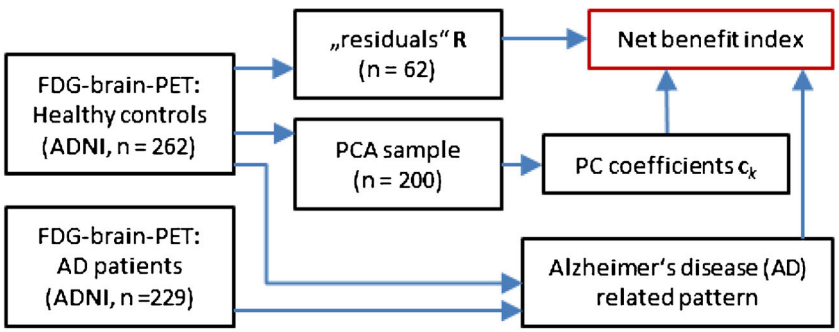

Fig. 1 Estimating the benefit of CODE (net benefit index)

voxel data (linear kernel, 5-fold cross-validation, CV). We chose AUC as the primary measure of classification performance and, using the fast implementation of the method of DeLong [6, 25], to test the improvement achieved by applying CODE. AUC has several advantages in comparison to the common evaluation criterion *accuracy* [9, 12], e.g., when sample sizes are unbalanced (64 non-converter vs. 156 converter), accuracy values can be considerably misleading. In addition to AUC, we report sensitivity and specificity (the cut-off was defined using the Youden's index, [27]). The disadvantage of using sensitivity and specificity is that each of these measures depends on the applied cutoff, so that both measures have to be interpreted together. Additionally, we report Matthews correlation coefficient (MCC), a metric suitable for imbalanced data and, together with AUC, chosen as the elective metrics for development and validation of predictive models [21].

Using voxel-wise SVM with training on MCI, we obtained the unexpected result of a classification performance that was considerably lower than with PES & CODE. We therefore modified voxel-based SVM analysis to more closely correspond to our approach. To this end, we trained the SVM with the same sample of healthy controls and AD subjects as in our approach, before applying the model to MCI subjects.

Results

Illustrating the effect of controls-based denoising

The effect of denoising on voxelwise variance is illustrated in Fig. 2. Before denoising (A), the highest variance was observed in the occipital cortex and at the border of gray matter (reflecting remaining spatial variability after normalization). Removing PCs 1 and 2 (B) already considerably reduced these edge effects and reduced the mean voxelwise variance to 85%. Removing PCs 1 to 10 (C) and 1 to 45 (D) reduced the variance to 66 and to 50%, respectively. The mask that was used for illustration was generated from the mean image across all controls (50% of the maximum voxel value).

Estimating the benefit of CODE from healthy controls

In 100 runs, we obtained a mean net benefit index of 1.5. The standard deviation was 0.2, reflecting the effect from random group assignments of healthy controls in each run.

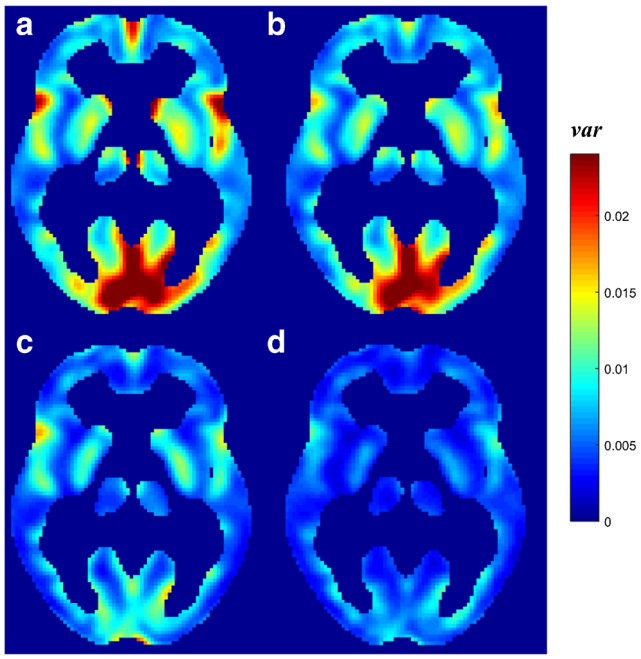

Fig. 2 Voxelwise variance in brain FDG-PET of cognitively normal subjects before (**a**) and after applying CODE (**b**: PC 1 and 2 removed; **c**: PCs 1 to 10 removed; **d**: PCs 1 to 45 removed). The mask that was used for illustration was generated from the mean image across all controls (50% of the maximum voxel value)

Applying CODE for conversion prediction

The effect of CODE on prediction of conversion to AD is shown in Table 2. In agreement with our model, CODE reduced the PES in converters by a factor that largely corresponds to $1/\mathbf{b}^{\mathbf{d}}\mathbf{b}^{\mathbf{T}}$, which is 10.34 ± 0.52 (suppl. Table 1). In the univariate analysis (ROC analysis of PES), removing the first 45 PCs (corresponding to 80% of the variance of all PCs) significantly improved AUC from 0.80 to 0.88, corresponding to an increase of sensitivity from 69 to 82%, specificity from 81 to 83% and MCC from 0.45 to 0.61. Additionally removing the remaining 217 PCs had a small effect on classification performance (AUC 0.88, sensitivity 83%, specificity 88% and MCC 0.66).

Combining PES with neuropsychological features and ApoE4 genotype (linear SVM, five-fold CV) increased the classification performance from AUC 0.80 to 0.88 (without CODE) and to 0.93 with CODE.

Classification performance (AUC) of the clinical features (when used as a single predictor without PET) were 0.83 (ADAS11), 0.83 (ADAS13), 0.65 (APOE4), 0.72 (CDRSB), 0.77 (FAQ), 0.71 (MMSE), 0.73 (RAVLT immediate), and 0.67 (RAVLT learning). Due to the high predictive value of ADAS, which was in the same order of magnitude as that

Table 2 Pattern expression score (PES) and classification performance before and after applying CODE

	Non-converter	Converter	Predicting conversion...									
	n = 64	n = 156	...with PES					...with PES & clinical features				
	PES (mean ± SD)	PES (mean ± SD)	AUC	SENS	SPEC	MCC	p[1]	AUC	SENS	SPEC	MCC	p[1]
Before CODE	7.93 ± 9.52	20.94 ± 11.58	0.80	69%	81%	0.45	—	0.88	90%	67%	0.62	—
After $CODE_{80}$	−0.28 ± 1.21	2.20 ± 1.95	0.88	82%	83%	0.61	0.002	0.91	92%	70%	0.70	n.s.
After $CODE_{100}$	0.92 ± 0.93	2.87 ± 1.51	0.88	83%	88%	0.66	0.001	0.93	94%	70%	0.71	n.s.

$CODE_{80}$, controls-based denoising using PC 1 to 45 (corresponding to 80% variance of all PCs); $CODE_{100}$, controls-based denoising using all 261 PCs. AUC, area under curve; SENS, sensitivity; SPEC, specificity; MCC, Matthew correlation coefficient, n.s.; not significant

[1]AUC with CODE as compared to AUC without CODE (method of Delong)

of PES, we calculated an explorative multiple regression analysis to demonstrate their independent predictive value. Using the two independent variables ADAS13 and PES (before and after CODE, respectively) to predict conversion, each of them showed to contribute with high significance (ADAS and PES, before CODE: $p < 10^{-7}$ and $p < 10^{-5}$; after CODE: $p < 10^{-4}$ and $p < 10^{-9}$).

CODE significantly reduced the contribution of age-related variance to the total variance of PES ($r = 0.27$ without CODE to 0.06 with CODE, $p = 0.01$).

Using voxel-wise SVM (Table 3) with training on MCI, we achieved a lower classification performance than with PES. In an attempt to modify SVM analysis so that it more closely corresponds to our approach, we used healthy controls and AD patients for training. This led to a higher classification performance (AUC 0.85, sensitivity 71%, specificity 86%, MCC 0.54) that was only marginally lower than that of our approach.

Discussion

In the present paper, we propose a novel approach (*controls-based denoising*, CODE) for improving image data analysis that is based on a pattern-expression score (PES), i.e., the degree to which a pathology-related pattern is present. The PES is a frequently used quantification approach in FDG-brain-PET and has been shown to be a powerful predictor of conversion from MCI to AD [3]. By applying CODE, we were able to significantly improve classification based on PES alone (AUC from 0.80 to 0.88, sensitivity from 69 to 83%, specificity from 81 to 88% and MCC from 0.45 to 0.66). The best classification performance (AUC 0.93) was achieved by combining the denoised PES with clinical features. In comparison to classification performances found in the literature [22], our results, achieved with CODE, are at the upper end of the range for sensitivity (56–100%) and specificity (24–100%).

Table 3 Predicting conversion to Alzheimer's disease with voxel-wise support vector machine (SVM)

Training sample	AUC	SENS	SPEC	MCC
MCI	0.76	85%	48%	0.39
Controls & AD	0.85	71%	86%	0.54

MCI, mild cognitive impairment; AUC, area under curve; SENS, sensitivity; SPEC, specificity

PES as basis for advanced analysis

It is now recognized that variability of classification performance in the literature can in large part be attributed to methodological aspects [15, 22] and the use of openly available databases such as ADNI was recommended for developing and testing automated quantification methods [15]. Of the available quantification methods, we chose the PES as basis for further improvements since it has been successfully applied to predict conversion [3] and due to the mathematical simplicity of its computation (once a disease specific pattern has been identified): it is calculated by an inner product which is a general metric that comprises traditional quantification methods (e.g., ROI analysis can be implemented by an inner product) as well as standard methods of machine learning (what we call *pattern of interest* corresponds to the normal of the hyperplane, e.g., calculated by SVM).

Further improvements of classification seem possible in two ways: first, the level of sophistication of image analysis may be increased by using a different pattern of interest than simply a group difference (here: group difference AD subjects versus controls) and, second, the PES can be combined with information from other sources, e.g., demographic, neuropsychological, and genetic features. CODE has shown to be a good solution with respect to these two aspects, as explained in the following two sections.

CODE versus SVM

CODE can be viewed as a method to modify a pattern of interest in order to improve classification performance (denoising image data is mathematically equivalent to denoising the pattern of interest). As mentioned above, the pattern of interest can be compared to the normal of the hyperplane calculated by SVM. Interestingly, our method achieved a higher classification performance than SVM when training was based on MCI subjects. While this may seem surprising, it can be explained by the fact that the number of MCI subjects was limited (64 non-converter, 156 converter) while information from controls was not used. This highlights a conceptual advantage of CODE: often controls are more easy to obtain than patients (using ADNI, we were able to include 262 control subjects in the analysis) and the same sample of controls may be used for multiple studies. A more fair comparison would therefore require SVM being trained on the same large sample of controls and AD subjects. Using a different group of subjects for training (healthy controls vs. AD) than for classification (converter vs. non-converter) is not a standard way to apply SVM, but lead to a classification performance that was higher than that from training based on MCI subjects. To statistically compare two methods with a good classification accuracy requires large samples. Our approach (PES & CODE) did perform slightly better than modified SVM, however the differences are not statistically significant.

Interestingly, the solution SVM found (i.e., the normal of the hyperplane) correlated more closely with the denoised pattern of interest (Pearson's r up to 0.51) than with the mere group difference (Pearson's $r = 0.31$), corresponding to the fact that machine learning faces the same challenge: to separate pathology-specific alterations from other sources of variability. A thorough comparison between our approach and machine learning is certainly beyond the scope of this paper, however, we would like to point out that in the present paper, we evaluated CODE in its basic form while optimizations seem possible as described below.

CODE and multivariate analysis

In agreement with the literature, we achieved a higher performance when including clinical information in our analysis. Combining parameters from imaging with clinical features may improve classification in two ways: first, clinical features may contain information not present in image data and therefore be of independent predictive value and, second, clinical information may, to some extent, explain unwanted variance in the image data.

Many different sources of variance are present in image data. An important example of physiological variance are age-related effects. The potential of CODE removing age-related variance could be demonstrated with the present data by significantly reducing the correlation between age and PES. In addition to physiological variance, there is a component of variance that is more closely related to methodology, e.g., different scanners used in multicenter studies or anatomical variability not compensated by spatial normalization (high voxelwise variance is often found at the border between gray and white matter). While this variance may in part be controlled for by external variables (e.g., the scanner used), the most important advantage of CODE is that it may account for variance that cannot—or only partially—be controlled for. This also includes biological age, which is only in part explained by chronological age.

Benefit versus cost of CODE

The net benefit index predicts the factor by which CODE improves the estimation of pss. It is conceivable that CODE does more harm than good, corresponding to a net benefit index < 1. Let us consider the extreme situation that the pattern of interest is identical to a linear combination of a subset of PCs. In this case, removing those PCs will completely remove he pathological effect from the data. Interestingly, as we have shown in the supplementary analyses, there is no straightforward way to tell if a particular PC improves or worsens the analysis: many PCs, when used in single-PC-denoising (i.e., only this PC is removed from patient data) decreased the net benefit index, but increased it when used in combination with other PCs. We therefore recommend calculating the net benefit index for each pattern of interest before applying CODE.

Outlook: Optimizing CODE

Our goal was to present and evaluate CODE in its basic form. Nevertheless, we would like to point out that several approaches for optimization of CODE are conceivable. While detailed evaluation of these approaches is certainly beyond the scope of this paper, we would like to shortly mention some approaches that yielded promising results in specific situations:

- Selecting a subgroup of PCs for denoising. In the present paper, for prediction of conversion, we focused on CODE using the first 45 PCs, which correspond to 80% of the variance of all PCs. However, more sophisticated methods for a priori selection of PCs are conceivable. As illustrated in the Supplementary Material, some PCs increased the net benefit index only in combination with other PCs. A priori identification of combinations of PCs that are particularly beneficial

(or particularly harmful) seem possible, e.g., by using the net benefit index.
- Calculating PCs from multimodal images. It is conceivable that inclusion of another modality (e.g., voxel-based morphometry) supports denoising, even when this modality has no additional predictive value and is not used for calculation of PES.
- Using voxelwise weights.
- Limiting the degree by which each PC is removed from patient data, e.g., by assuming that scores above 2 standard deviations likely reflect pathological alterations rather than physiological variance.

Limitations

We achieved a net benefit index of 1.5 for an AD-related pattern that has been calculated by means of a group difference. Since CODE also removes variance that is related to methodology, we would like to point out that this index in part reflects specifics of the database used (ADNI). It is conceivable that the net benefit index differs when applying CODE to FDG-PET from other databases (e.g., Australian Imaging Biomarkers & Lifestyle, AIBL) or to single center data.

In this paper, evaluation of predictive performance is limited by the number of non-converter (n = 64). While this number was sufficient to show that CODE significantly improved classification based on PES, a higher number of non-converter is required for a more detailed statistical comparison of methods, including comparison with various machine learning approaches.

In the present paper, we used SPM's spatial normalization that was based on PET only. One might suggest using DARTEL normalization, instead [1, 14]. While using a less accurate normalization approach may have theoretically increased the benefit from CODE as discussed above, we would like to point out that in clinical routine diagnostic, statistical analysis of FDG-PET is typically based on PET-only normalizations.

Applying CODE in a clinical setting

The suggested method (PES & CODE) seems suitable not only for analysis of multi-center data for scientific purpose, but also to support individual clinical diagnosis and prognosis. It is generally accepted that visual image interpretation (and particularly that of FDG brain PET) highly depends on the training of the reader and that the additional use of computer-generated diagnostic indices such as the one suggested here improves clinical image interpretation. However, for application of CODE, there is a number of caveats: in order to obtain reliable results, data preprocessing, particularly spatial normalization, has to be performed in exactly the same way as for calculation of the disease-related pattern and, particularly, the patterns of non-pathological variability (PCs). The demonstrated benefit of CODE is in part due to removing methodology-related variance present in multi-center data analysis so the benefit may be slightly smaller when using CODE in a monocentric setting. However, large portions of the variance (anatomical variance not compensated by spatial normalization, physiological variability) seem to be equally relevant for mono-center data analysis.

Conclusions

In the present paper, we propose a novel approach for data preprocessing prior to statistical analysis (controls-based denoising, CODE). In particular, we propose identifying principal components from a sample of healthy controls and removing them from patient data before calculating the degree to which a known pathology-related pattern is present. In patients with mild cognitive impairment, CODE has shown to considerably improve the analysis of FDG-PET with respect to predicting conversion to Alzheimer's disease. CODE seems to be promising for a wide spectrum of applications, even beyond image analysis, provided the pattern to be detected is already known to a sufficient degree of detail. CODE requires a rather large sample of controls, however controls may be recruited more easily than patients and the same sample can be used for multiple studies. While the suggested method performed well in comparison with a standard machine learning approach, we have shown that including information from healthy controls also improved prediction of conversion to AD by machine learning, demonstrating that information about patterns of variance that are not related to pathology are currently not sufficiently exploited in medical image analysis.

Acknowledgements We acknowledge the funding received from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 603646 (MultISyn).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.;

Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 603646 (MultISyn).

Compliance with Ethical Standards

Conflict of interest Dominik Blum declares that he has no conflict of interest. Inga Liepelt-Scarfone declares that she has no conflict of interest. Daniela Berg declares that she has no conflict of interest. Thomas Gasser declares that he has no conflict of interest. Christian la Fougère declares that he has no conflict of interest. Matthias Reimold declares that he has no conflict of interest.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage. 2007;38(1):95–113.
2. Bishop CM. Pattern recognition and machine learning (Information science and statistics). New York: Springer-Verlag; 2006.
3. Blazhenets G, Ma Y, Sorensen A, Rucker G, Schiller F, Eidelberg D, Frings L, Meyer PT. Principal component analysis of brain metabolism predicts development of Alzheimer's dementia. J. Nucl. Med. 2018.
4. Caroli A, Prestia A, Chen K, Ayutyanont N, Landau SM, Madison CM, Haense C, Herholz K, Nobili F, Reiman EM, Jagust WJ, Frisoni GB, Perani D, Pupi A, Holthoff V, Salmon E, Baron JC, Drzezga A, Perneczky R, Didic M, Guedj E, Van Berckel BN, Ossenkoppele R, Morbelli S. Summary metrics to assess Alzheimer disease-related hypometabolic pattern with 18F-FDG PET: head-to-head comparison. J Nucl Med. 2012;53(4):592–600.
5. De Carli F, Nobili F, Pagani M, Bauckneht M, Massa F, Grazzini M, Jonsson C, Peira E, Morbelli S, Arnaldi D. Accuracy and generalization capability of an automatic method for the detection of typical brain hypometabolism in prodromal Alzheimer disease. Eur J Nucl Med Mol Imaging. 2019;46(2):334–347.
6. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845.
7. Frings L, Hellwig S, Bormann T, Spehl TS, Buchert R, Meyer PT. Amyloid load but not regional glucose metabolism predicts conversion to Alzheimer's dementia in a memory clinic population. Eur J Nucl Med Mol Imaging. 2018;45(8):1442–1448.
8. Grimmer T, Wutz C, Alexopoulos P, Drzezga A, Forster S, Forstl H, Goldhardt O, Ortner M, Sorg C, Kurz A. Visual versus fully automated analyses of 18F-FDG and amyloid PET for prediction of dementia due to Alzheimer disease in mild cognitive impairment. J Nucl Med. 2016;57(2):204–207.
9. Huang J, Lu J, Ling LCX. Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In: 3rd IEEE international conference on data mining, ICDM 2003, pp 553–556. IEEE Computer Society; 2003.
10. Jolliffe I. Principal component analysis. Berlin: Springer Verlag; 1986.
11. Li K, Chan W, Doody RS, Quinn J, Luo S. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. J Alzheimers Dis. 2017;58(2):361–371.
12. Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In: Proceedings of 18th international conference on Artificial Intelligence (IJCAI-2003), pp 519–524; 2003.
13. Liu K, Chen K, Yao L, Guo X. Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the Cox model. Front Hum Neurosci. 2017;11:33.
14. Martino ME, de Villoria JG, Lacalle-Aurioles M, Olazaran J, Cruz I, Navarro E, Garcia-Vazquez V, Carreras JL, Desco M. Comparison of different methods of spatial normalization of FDG-PET brain images in the voxel-wise analysis of MCI patients and controls. Ann Nucl Med. 2013;27(7):600–609.
15. Morbelli S, Garibotto V, Van De Giessen E, Arbizu J, Chetelat G, Drezgza A, Hesse S, Lammertsma AA, Law I, Pappata' S, Payoux P, Pagani M. A Cochrane review on brain [18F]FDG PET in dementia: limitations and future perspectives. Eur J Nucl Med Mol Imaging. 2015;42(10):1487–1491.
16. Nobili F, Salmaso D, Morbelli S, Girtler N, Piccardo A, Brugnolo A, Dessi B, Larsson SA, Rodriguez G, Pagani M. Principal component analysis of FDG PET in amnestic MCI. Eur J Nucl Med Mol Imaging. 2008;35(12):2191–2202.
17. Pagani M, Giuliani A, Oberg J, Chincarini A, Morbelli S, Brugnolo A, Arnaldi D, Picco A, Bauckneht M, Buschiazzo A, Sambuceti G, Nobili F. Predicting the transition from normal aging to Alzheimer's disease:a statistical mechanistic evaluation of FDG-PET data. Neuroimage. 2016;141:282–290.
18. Pagani M, Nobili F, Morbelli S, Arnaldi D, Giuliani A, Oberg J, Girtler N, Brugnolo A, Picco A, Bauckneht M, Piva R, Chincarini A, Sambuceti G, Jonsson C, De Carli F. Early identification of MCI converting to AD: a FDG PET study. Eur J Nucl Med Mol Imaging. 2017;44(12):2042–2052.
19. Pearson K. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559–572. https://doi.org/10.1080/14786440109462720.
20. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology. 2010;74(3):201–209.
21. Shi L, Campbell G, Jones EA. The microArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28(8):827–838.
22. Smailagic N, Lafortune L, Kelly S, Hyde C, Brayne C. 18F-FDG PET for prediction of conversion to Alzheimer's disease dementia

in people with mild cognitive impairment: An updated systematic review of test accuracy. J Alzheimers Dis. 2018;64(4):1175–1194.
23. Spetsieris PG, Eidelberg D. Scaled subprofile modeling of resting state imaging data in Parkinson's disease: methodological issues. Neuroimage. 2011;54(4):2899–2914.
24. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull. 1980;87:245–251.
25. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21(11):1389–1393.
26. Trzepacz PT, Yu P, Sun J, Schuh K, Case M, Witte MM, Hochstetler H, Hake A. Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. Neurobiol Aging. 2014;35(1):143–151.
27. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.

Affiliations

Dominik Blum[1] · Inga Liepelt-Scarfone[2,3] · Daniela Berg[4] · Thomas Gasser[2,3] · Christian la Fougère[1] · Matthias Reimold[1] · for the Alzheimer's Disease Neuroimaging Initiative

1 Institute for Nuclear Medicine and Clinical Molecular Imaging, Eberhard Karls University, Tuebingen, Germany

2 German Center of Neurodegenerative Diseases, Eberhard Karls University, Tuebingen, Germany

3 Hertie-Institute for Clinical Brain Research, Eberhard Karls University, Tuebingen, Germany

4 Department of Neurology, Christian-Albrechts-University, Kiel, Germany