Supplementary materials

Classification analyses using SVM

SVMs are an established, supervised machine learning technique and widely used across various disciplines. They have been used successfully to determine the diagnostic and prognostic potential of structural MRI in a range of neurological and psychiatric disorders(Davatzikos et al. 2008; Fan et al. 2008; Kawasaki et al. 2007; Teipel et al. 2007a; Teipel et al. 2007b; Marzelli et al. 2011; Koutsouleris et al. 2011; Kloppel et al. 2009), including MS(Weygandt et al. 2011; Bendfeldt et al. 2012).

What makes SVMs particularly attractive is their generally good performance in comparison to other classifiers as well as their relatively easy extension to non-linear problems using kernels. In addition, careful tuning of regularisation and kernel parameters make them less prone to overfitting.(Cawley and Talbot 2007)

Linear SVMs are well suited for classification problems where the number of input features is very large, potentially much larger than the number of samples in the data set. However, as the number of features decreases non-linear SVM approaches often perform better due to an increased flexibility in fitting the data. This comes at the cost of having to estimate an additional parameter. Thus, our approach of employing both linear and non-linear SVMs for two separate analyses of the same data is driven by the markedly different sizes of the respective feature sets. For a comprehensive overview of support vector machines, see e.g.(Bishop 2006; Vapnik 1999; Campbell 2002; Ivanciuc 2007).

Details on the classification analysis using geometric features and non-linear SVMs

Lesion-based geometric measures

Individual lesions were identified on binary lesion masks via lesion clustering using FSL's cluster command, i.e. a cluster is defined as contiguous, lesion-marked voxels based on an order-26 neighbourhood. We then used Minkowski functionals(Legland et al. 2011) to extract and describe the geometry of each lesion. In 3D-Eucledian space, Minkowski functionals are directly related to the following geometric quantities: volume, surface area, mean breadth and Euler-Poincare (EP) characteristic. The measure of mean breadth is related to the integral of mean curvature(Hawkes 1988) and represents the mean extension of the lesion taken over all possible rotations in three dimensions(C. H. Arns et al. 2001; Lang et al. 2001), and gives a pose-independent measure of breadth. Finally, for a single cluster of voxels the EP number is the most basic Minkowski functional and provides a connectivity parameter of the object(C. Arns et al. 2010). It is equivalent to 1 minus the number of holes plus the number of handles in the geometric object. Furthermore, the whole-brain sum of EP numbers of single lesions is closely related to the total lesion count. A general discussion of high dimensional Minkowski functionals can be found, for example, in(Legland et al. 2007).

Preprocessing and feature reduction

Affine **rigid-body** co-registration of the images and lesion masks was used to preserve volumetric information. The linear part of the transformation matrix from a DARTEL registration of the images to standard space was then used to affine register the lesion masks to the MNI template brain. Based on these binary images, geometric measures for each individual lesion were extracted (using Minkowski functionals) and whole-brain as well as

ROI-based summary features were computed. From the SPM/VBM8 pipeline given in Figure 1, grey matter volume (whole-brain and ROI proportions) were included in the feature set. A schematic summary of the preprocessing pipeline is given in Figure 2. After extracting the geometric measures for all lesions from the MRI data, we transformed volume and area measures by taking their cubic and square roots, respectively(Sormani and Filippi 2007). This leads to all geometric features being comparable in scale. Even after power-transformation of volume and surface area, the three geometric measures are highly correlated as they are all influenced by the gross size of the lesion. To reduce redundancy and to gain greater sensitivity to (size-independent) lesion shape, square-root surface area and mean breadth were divided by their respective value assuming the lesion was a sphere. These adjusted measures thus reflect the topological information contained in 1D (mean breadth), 2D (surface area) and 3D (volume) measures and indicate how far the shape of a lesion deviates from a perfect sphere.

Based on lesion specific values, whole-brain summaries were computed to be used as input features. The full geometric feature set thus included the sum total, mean, median, maximum, minimum and standard deviation for each measure, i.e. lesion volume, surface area and mean breadth, as well as the EP characteristic. Additionally, grey matter volume measurements were included as ratio of grey matter volume to whole brain volume.

In order to encode spatial information about the location of individual lesions, we also split the whole brain measures according to 13 regions of interest (ROI) based on white matter (WM) track segmentations derived from the Johns-Hopkins brain atlas (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases). The 13 ROIs include anterior thalamic radiation, corticospinal tract, cingulum (cingulate gyrus), cingulum (hippocampus), posterior forceps, anterior forceps, inferior fronto-occipital fasciculus, inferior longitudinal fasciculus, superior

longitudinal fasciculus, uncinate fasciculus, superior longitudinal fasciculus (temporal), other cerebrum, other cerebellum.

We used principal component analysis (PCA) to transform the high dimensional feature sets and reduce the number of potentially redundant features. The number of principal components was optimised during the training stage of the classifier.

All features were standardized to have zero mean and unit variance. This not only makes the features comparable but ensures that the PCA components correspond to directions of maximum variance.

Evaluation and Interpretation

For model evaluation we carried out nested, stratified k-fold (k=10) cross-validation (CV). The stratification ensured that the number of subjects from both classes was balanced. Within each CV fold, a three-dimensional parameter optimization via grid search was performed on the SVM box-constraint (soft margin parameter), the RBF kernel parameter and the number of principal components. To ensure independence of the test fold, the training data consisting of k-1 folds was once again split into a training (k-2 folds) and a test set (1 fold) to perform the parameter optimization. The overall classification accuracy was estimated at the end by using the so far unseen data from the initially held out fold.

Finally, predictive performance of the models was evaluated by computing a confusion matrix. We report average classification accuracies, as they are more informative than total accuracies in cases where the two classes are of different size. Average accuracy (also called balanced accuracy(Combrisson and Jerbi 2015)) is given by the mean of correctly classified samples in each class, whereas total accuracy is given as the ratio of overall correctly predicted outcomes to total number of test samples.

For both the linear and non-linear SVM analyses we determined statistical significance classification accuracy by permutation testing(Nichols and Holmes 2002; Golland and Fischl 2003). This involved repeating the classification procedure with training group labels randomly allocated multiple – in our case 1000 – times in order to generate an empirical estimate of the null distribution of accuracies. The p-value is computed as the proportion of permutations in this null distribution achieving an accuracy as large or larger than the observed accuracy. Statistical inferences were made at a significance level of $p < 0.05$.

Details on the classification analysis using cortical grey matter segmentations and linear SVMs

Leave-one-out cross-validation

As above, the SVM was trained to predict conversion to CDMS.

Unbalanced group sizes can lead to a bias of the hyperplane weighting towards the larger group. Here we took a different approach and carefully balanced the two groups to avoid any possible bias in sensitivity and specificity. 500 random (bootstrapped) samples were selected from the larger group with size equal to the smaller group.

In the case of the placebo group this means that 25 converters were randomly selected from the whole set of 44 converters in order to match the group size of the non-converters. This procedure was repeated 500 times to give a better idea of how the results will generalise to the whole cohort. The resulting cohorts of 50 (25 converters and 25 non-converters) and 98 (49 non-converters and 49 converters) for placebo and interferon beta-1b, respectively, were then used to train and test an SVM using the common leave-one-out cross-validation(Young et al. 2013).

Within the whole dataset, we further focused on a subgroup of placebo-treated patients optimally matched according to clinical/demographic data, scanner and image quality (n=2*25 placebo). Weight vector maps were created to show the brain regions that best discriminated between groups.

References

Arns, C., Knackstedt, M., & Mecke, K. (2010). 3D structural analysis: sensitivity of Minkowski functionals. *Journal of microscopy, 240*(3), 181-196.

Arns, C. H., Knackstedt, M. A., Pinczewski, W. V., & Mecke, K. R. (2001). Euler-Poincaré characteristics of classes of disordered media. *Physical Review E, 63*(3), 031112.

Barkhof, F., Polman, C. H., Radue, E.-W., Kappos, L., Freedman, M. S., Edan, G., et al. (2007). Magnetic Resonance Imaging Effects of Interferon Beta-1b in the BENEFIT Study: Integrated 2-Year Results. *Arch Neurol, 64*(9), 1292-1298, doi:10.1001/archneur.64.9.1292.

Bendfeldt, K., Kloppel, S., Nichols, T. E., Smieskova, R., Kuster, P., Traud, S., et al. (2012). Multivariate pattern classification of gray matter pathology in multiple sclerosis. *NeuroImage, 60*(1), 400-408, doi:10.1016/j.neuroimage.2011.12.070.

Bishop, C. M. (2006). Pattern Recognition. *Machine Learning*.

Campbell, C. (2002). Kernel methods: a survey of current techniques. *Neurocomputing, 48*(1), 63-84.

Cawley, G. C., & Talbot, N. L. C. (2007). Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research, 8*, 841-861.

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods, 250*, 126-136, doi:10.1016/j.jneumeth.2015.01.010.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., & Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging, 29*(4), 514-523.

Fan, Y., Batmanghelich, N., Clark, C. M., & Davatzikos, C. (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage, 39*(4), 1731-1743.

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. *Inf Process Med Imaging, 18*, 330-341.

Hawkes, P. W. (1988). Image-Analysis and Mathematical Morphology, Vol 2, Theoretical Advances - Serra,J. *Nature, 336*(6199), 531-532.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in computational chemistry, 23*, 291.

Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S. Y., Nakamura, K., et al. (2007). Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *NeuroImage, 34*(1), 235-242.

Kloppel, S., Chu, C., Tan, G. C., Draganski, B., Johnson, H., Paulsen, J. S., et al. (2009). Automatic detection of preclinical neurodegeneration: presymptomatic Huntington disease. [Multicenter Study

Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *Neurology, 72*(5), 426-431, doi:10.1212/01.wnl.0000341768.28646.b6.

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Müller, H.-J., & Riecher-Rössler, A. (2011). Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. *Schizophrenia Bulletin*, doi:10.1093/schbul/sbr145.

Lang, C., Ohser, J., & Hilfer, R. (2001). On the analysis of spatial binary images. *Journal of microscopy, 203*(3), 303-313.

Legland, D., Kiêu, K., & Devaux, M.-F. (2007). Computation of Minkowski measures on 2D and 3D binary images. *Image Analysis and Stereology, 26*, 83-93.

Legland, D., Kiêu, K., & Devaux, M.-F. (2011). Computation of Minkowski measures on 2D and 3D binary images. *Image Analysis & Stereology, 26*(2), 83-92.

Marzelli, M. J., Hoeft, F., Hong, D. S., & Reiss, A. L. (2011). Neuroanatomical spatial patterns in Turner syndrome. *NeuroImage, 55*(2), 439-447.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp, 15*(1), 1-25.

Sormani, M. P., & Filippi, M. (2007). Statistical issues related to the use of MRI data in multiple sclerosis. [Review]. *Journal of neuroimaging : official journal of the American Society of Neuroimaging, 17 Suppl 1*, 56S-59S, doi:10.1111/j.1552-6569.2007.00139.x.

Teipel, S. J., Born, C., Ewers, M., Bokde, A. L., Reiser, M. F., Moller, H. J., et al. (2007a). Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *NeuroImage, 38*(1), 13-24.

Teipel, S. J., Stahl, R., Dietrich, O., Schoenberg, S. O., Perneczky, R., Bokde, A. L., et al. (2007b). Multivariate network analysis of fiber tract integrity in Alzheimer's disease. *NeuroImage, 34*(3), 985-995.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Trans Neural Netw, 10*(5), 988-999, doi:10.1109/72.788640.

Weygandt, M., Hackmack, K., Pfüller, C., Bellmann–Strobl, J., Paul, F., Zipp, F., et al. (2011). MRI Pattern Recognition in Multiple Sclerosis Normal-Appearing Brain Areas. *PloS one, 6*(6), e21138, doi:10.1371/journal.pone.0021138.

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., Ourselin, S., et al. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's

disease in patients with mild cognitive impairment. *NeuroImage: Clinical, 2*, 735-745.