Supplementary methods: Logistic regression

Logistic regression is a commonly used model for relating a binary outcome to continuous and/or categorical predictors (Vittinghoff *et al.*, 2005), such as in the binary classification problems considered in this study. In the following, we denote the class labels by $\{-1, +1\}$ (corresponding to, e.g., "control" and "prHD$_{\text{near}}$"). The logistic regression model assumes that the probability that a participant with data vector $\mathbf{x}$ (in our case, consisting of the voxel intensities from normalized, segmented grey matter images output by SPM5, each linearly scaled to the interval $[-1, 1]$ across all participants) belongs to class y (either -1 or $+1$) is given by (Lin *et al.*, 2007)

$$P(y = \pm 1|\mathbf{x}, \beta) = \frac{1}{1 + \exp(-y\beta^T\mathbf{x})}$$

where β is a vector of regression weights. This implies that

$$\log\left(\frac{P(y|\mathbf{x}, \beta)}{1 - P(y|\mathbf{x}, \beta)}\right) = y\beta^T\mathbf{x},$$

so on a transformed scale, the model becomes linear in the measured variables. Given a training set of n participants with measured data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and true class labels $y_1, \ldots, y_n$, the regression weights β are estimated by minimizing the negative log-likelihood (Lin *et al.*, 2007), i.e. solving

$$\min_\beta \sum_{i=1}^{n} \log(1 + \exp(-y_i\beta^T\mathbf{x}_i)).$$

In this study, we used an implementation provided by the LIBLINEAR software (Fan *et al.*, 2008) and its Python interface (http://public.procoders.net/liblinear2scipy/src/dist). In this software, the authors added a regularization term to the objective function to obtain good generalization properties and prevent overfitting to the training data (Lin *et al.*, 2007). Hence, the optimization problem is given by

$$\min_\beta \left[\frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n} \log\left(1 + \exp(-y_i\beta^T\mathbf{x}_i)\right)\right].$$

Here, $C > 0$ is a parameter to control the trade-off between getting a good fit to the training data and preventing overfitting. We consider the optimal value of C to be the value giving the best generalization ability of the resulting classifier. The generalization ability is estimated by

five-fold cross-validation. This means that we estimate the regression weights using a training set consisting of 80% of the participants, classify the remaining 20% (the test set) using the estimated model, and repeat the procedure five times, until all participants have been left out from the training set once. We then count how many of the test participants that were classified into the correct class, which gives an estimate of the classification accuracy of the logistic regression model. Furthermore, we report the estimated sensitivity (the proportions of class 1 test participants which were correctly classified into class 1) and specificity (the proportion of class -1 test participants which were correctly classified into class -1) for each classifier. In each binary classification problem, we consider the participants furthest from onset of motor symptoms (with the control group considered to be further from onset than any of the prHD groups) as class -1.

Since the logistic regression model returns a probability for a test participant to belong to each of the two classes, we need to set a cut-off value on these probabilities to obtain a binary classifier (of course, $P(y = 1|\mathbf{x}, \beta) = 1 - P(y = -1|\mathbf{x}, \beta)$, so we need only consider one of the probabilities, say $P(y = 1|\mathbf{x}, \beta)$). Increasing the cut-off value of $P(y = 1|\mathbf{x}, \beta)$ necessary to classify a participant to class 1 will decrease the number of test participants being classified into class 1, hence decreasing the sensitivity and increasing the specificity. To help visualize the overall performance of the classifiers we use ROC (receiver operating characteristic) curves, showing the relationship between the true positive rate (sensitivity) and the false positive rate (1−specificity) for each value of the cut-off probability, ranging from 0 to 1. The reported sensitivities and specificities (Supplementary Table 1) correspond to a cut-off value of 0.5, i.e. the case where a participant is classified to class 1 if it has a higher probability of belonging to that class than to class -1.

References

Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* **9**, 1871-1874.

Lin C-J, Weng RC, Kerthi SS (2007) Trust Region Newton Methods for Large-Scale Logistic Regression. *Proceedings of the 24th International Conference on Machine Learning*

Vittinghoff E, Shiboski SC, Glidden DV, McCulloch CE (2005) *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer, New York.