Early functional magnetic resonance imaging activations predict language outcome after stroke

Dorothee Saur,1,2,3 Olaf Ronneberger,4,5 Dorothee Kümmerer,1,2 Irina Mader,2,6 Cornelius Weiller1,2 and Stefan Klöppel2,7

1 Department of Neurology, University Medical Centre Freiburg, Freiburg, Germany
2 Freiburg Brain Imaging, University Medical Centre Freiburg, Freiburg, Germany
3 Department of Neurology, University Medical Center Leipzig, Leipzig, Germany
4 Image Analysis Lab, Department of Computer Science, University of Freiburg, Germany
5 Centre for Biological Signalling Studies (bioss), University of Freiburg, Germany
6 Department of Neuroradiology, University Medical Centre Freiburg, Germany
7 Department of Psychiatry and Psychotherapy. Section for Gerontopsychiatry and Section of Experimental Neuropsychiatry, University Medical Centre Freiburg, Freiburg, Germany

An accurate prediction of system-specific recovery after stroke is essential to provide rehabilitation therapy based on the individual needs. We explored the usefulness of functional magnetic resonance imaging scans from an auditory language comprehension experiment to predict individual language recovery in 21 aphasic stroke patients. Subjects with an at least moderate language impairment received extensive language testing 2 weeks and 6 months after left-hemispheric stroke. A multivariate machine learning technique was used to predict language outcome 6 months after stroke. In addition, we aimed to predict the degree of language improvement over 6 months. 76% of patients were correctly separated into those with good and bad language performance 6 months after stroke when based on functional magnetic resonance imaging data from language relevant areas. Accuracy further improved (86% correct assignments) when age and language score were entered alongside functional magnetic resonance imaging data into the fully automatic classifier. A similar accuracy was reached when predicting the degree of language improvement based on imaging, age and language performance. No prediction better than chance level was achieved when exploring the usefulness of diffusion weighted imaging as well as functional magnetic resonance imaging acquired two days after stroke. This study demonstrates the high potential of current machine learning techniques to predict system-specific clinical outcome even for a disease as heterogeneous as stroke. Best prediction of language recovery is achieved when the brain activation potential after system-specific stimulation is assessed in the second week post stroke. More intensive early rehabilitation could be provided for those with a predicted poor recovery and the extension to other systems, for example, motor and attention seems feasible.

Keywords: aphasia; stroke; outcome prediction; language impairment; functional magnetic resonance imaging; support vector machine

Abbreviations: AAT = Aachen Aphasia Test; DWI = diffusion-weighted imaging; LRS = language recovery score; SVM = support vector machine
Introduction

Prediction of long-term recovery after ischaemic brain damage is an important task for planning rehabilitation and future (professional) life of patients. Early prediction, in particular, is important since it has been suggested that therapeutic interventions are most beneficial when started early after stroke onset (Horn et al., 2005). This idea is fuelled by observations which indicate that the process of brain reorganization starts early, within the first weeks after stroke (Marshall et al., 2000; Saur et al., 2006) and that the highest dynamic of functional recovery is observed in the first months after the event (Demeurisse et al., 1980; Duncan et al., 1992; Laska et al., 2001; Pedersen et al., 2004).

Previous studies evaluating prediction of global (i.e. daily live functioning) rather than system-specific (i.e. motor or language abilities) recovery after stroke typically used demographic (e.g. age, gender), clinical (e.g. vascular risk factors, previous stroke, stroke subtype) and behavioural data (e.g. National Institute of Health Stroke Scale) as baseline parameters. Results show that mainly initial stroke severity is associated with a patient’s global outcome (Henon et al., 1995; Adams et al., 1999; Weimar et al., 2004).

Similarly, studies investigating prediction of system-specific outcome in the language domain indicate that initial severity of language impairment is a highly relevant baseline parameter to distinguish patients with poor from those with favourable language recovery in aphasic stroke (Kertesz and McCabe, 1977; Pedersen et al., 1995, 2004; Laska et al., 2001; Lazar et al., 2008). Although the type of aphasia is predictive for long-term outcome (Demeurisse et al., 1980; Kertesz, 1988), this observation is confounded by the fact that distinct types of aphasia are typically related to a specific degree of impairment (e.g. global aphasia with a severe, conduction aphasia with a less severe disturbance) (Lendrem and Lincoln, 1985).

Adding information from neuroimaging to the prediction [e.g. lesion volume, lesion site (usually dichotomized to cortical versus subcortical)] did not significantly improve accuracy of the models for global outcome (Thijss et al., 2000; Johnston et al., 2002, 2007; Hand et al., 2006; Schiemanck et al., 2006) or for system-specific language outcome in aphasic stroke (Pedersen et al., 2004; Lazar et al., 2008).

Independently from the system of interest (e.g. language or motor functions) outcome prediction after stroke is of limited clinical relevance in subjects who initially present with very mild symptoms. The reasons are that (i) their impairment might not affect every day life activities; and (ii) that patients with a mild impairment in the acute phase have a good chance of complete recovery (Laska et al., 2001; Lazar et al., 2008; Prabhakaran et al., 2008). In contrast, patients with a more severe deficit might show little, moderate or complete recovery (Jorgensen et al., 1999). That is, in this population recovery contains a greater variance which is not explained by initial stroke severity alone (Duncan et al., 1992; Lazar et al., 2008; Prabhakaran et al., 2008; Marshall et al., 2009).

In this study, we sought to identify a diagnostic parameter for early and accurate prediction of language recovery at the individual level in stroke patients with a moderate to severe initial language deficit. To this end, we trained a support vector machine (SVM) with language functional MRI (fMRI) data to differentiate between patients with good and bad language recovery 6 months after stroke. SVMs are multivariate pattern classification techniques (Vapnik, 1998) which are capable of learning the categorization of complex, high-dimensional training data and generalize the learned classification rules to unseen data. So far, this technique has been found useful to predict the future course of neurodegenerative and psychiatric disorders based on structural MRI (Teipel et al., 2007; Davatzikos et al., 2008; Klöppel et al., 2009a; Koutsouleris et al., 2009).

We used language fMRI data as the baseline parameter since this modality has been shown to reflect functionality of the language system even in the early period after stroke (Thulborn et al., 1999; Saur et al., 2006). It can therefore be hypothesized that some of the signal at baseline is also related to language recovery.

In this study we predict absolute language performance 6 months after stroke (referred to as ‘outcome’) as well as the change in performance over the same time span (referred to as ‘improvement’). Prediction of both parameters was based on language fMRI alone and fMRI in combination with behavioural and demographic data. In additional analyses (provided in the online Supplement data) we sought to find the optimal time window and the best imaging modality for baseline evaluation. We therefore explored the usefulness of diffusion-weighted imaging (DWI) and fMRI data acquired in the first days after stroke.

Materials and methods

Study design

We report data from subjects who entered different longitudinal fMRI studies on language recovery. Baseline data were acquired approximately 12 days (mean 12.5, SD 2.25) post-stroke. Outcome was assessed approximately 6 months after stroke (mean 187.5, SD 47.8 days post-stroke).

Baseline parameters included demographic (i.e. age), behavioural, (i.e. language impairment) and fMRI data.

Patient recruitment

We recruited patients from the stroke units of the Departments of Neurology, University Medical Centres Hamburg-Eppendorf and Freiburg. Inclusion criteria included: (i) evidence of aphasia in the Aachen Aphasia Test (AAT); (ii) embolic first time stroke of the left middle cerebral artery-territory; (iii) native language German; and (iv) age ≤ 80 years. Exclusion criteria were: (i) inability to perform the language-task due to severity of aphasia; (ii) reduced general health status; (iii) hearing deficits; (iv) pronounced small vessel disease; (v) recurrent stroke during the study period; (vi) cognitive impairment other than aphasia; and (vii) any contraindication to MRI.

From May 2003 to September 2008, 514 patients with acute aphasic symptoms were screened. 36 consecutive patients met the inclusion criteria and were recruited for the fMRI studies (mean age 56.2, SD 15.1; 11 female). Reasons for exclusion of the other 478 patients were (i) severity of aphasia too mild/too severe n = 75/24 (a ‘too mild’
deficit was given if aphasia was not evident in the AAT and a ‘too severe’ deficit if informed consent of the patient was not possible); (ii) reduced general health status (n = 62); (iii) previous infarcts (n = 61); (iv) large vessel disease with haemodynamic infarctions (n = 12); (v) aetiology (intracerebral haemorrhage, tumour, dementia; n = 107); (vi) age and small vessel disease (n = 54); (vii) hearing deficits (n = 12); (viii) native language other than German (n = 15); (ix) neuropsychological impairments other than aphasia (n = 23); and (x) other, e.g. contraindications for MRI, cooperation, technical problems (n = 31).

Language testing

The challenge of the behavioural evaluation in a longitudinal recovery study is to capture different degrees of language impairments from the acute to the chronic stage. We tried to solve this problem by combining tests which are evaluated for distinct phases and which capture different severities and impairment features. Our standardized aphasia test-battery included the Aachen Aphasia Bedside Test, the AAT subtests repetition, writing, naming and auditory/reading comprehension (Huber et al., 1984), the Token Test, an analysis of spontaneous speech and the Communicative Effectiveness Index. The Aachen Aphasia Bedside Test is a test designed and validated for the assessment of severely impaired patients in the acute phase after stroke (Biniek et al., 1992), while the Communicative Effectiveness Index is a measure of the functional outcome (Lomas et al., 1989). For an analysis of spontaneous speech we recorded a semi-standardized 10-min interview which was analysed according to the AAT criteria of communicative abilities, articulation and prosody, automated speech, semantic, phonemic and syntactic structure.

Scores within each of these five assessments were combined. Task performance in the scanner contributed as a separate language score (see language paradigms for details). Thus, there was a set of six language measures for each patient at each time point. These scores were normalized to a range of 0–1 (score nor) and averaged into a composite score, the overall language recovery score (LRS): LRS = (AABT nor + AAT (without TT and SPS) nor + TT nor + SPS nor + CETI nor + Tasknor)/6 with a resulting range between zero and one (where AABT = Aachen Aphasia Bedside Test; TT = Token Test; SPS = spontaneous speech; and CETI = Communicative Effectiveness Index).

The LRS represents a reasonable univariate index of the overall level of language impairment and allows for separating patients within a broad range of impairments throughout all stages after stroke.

To characterize each patient’s language impairment further, severity (mild (I), moderate (II) and severe (III)), fluency (fluent versus non-fluent) and type of aphasia as indicated by the AAT were evaluated at baseline and outcome. We are aware that the type of aphasia is not stable in the early phase after stroke but this information may provide a basic indication of the patient’s impairment.

Patient cohorts and control subjects

As motivated above, we aimed at including only patients with moderate to severe aphasia. This was defined by a severity score > 1 in the AAT and a LRS ≤ 0.8. 21 of the 36 patients met these criteria at the baseline examination and were finally included in the study (Table 1). The same criteria were used to define bad (severity score > 1, LRS ≤ 0.8) versus good language outcome 6 months after onset. To predict the degree of improvement for each patient we computed the difference in LRS between baseline and outcome. The median of this difference was chosen to separate those with substantial improvement (Δ > 0.28) from those with little change (Δ ≤ 0.28). Predicting the degree of improvement is complementary to the prediction of outcome as a severely impaired patient may show substantial and clinically relevant improvement but may still have moderate or severe aphasia in the AAT rating. Tables 2 and 3 provide an overview of the performance on all language tests separately for both outcome groups. Additional clinical data for each subject are provided in Supplementary Table 1.

Fourteen age-matched control subjects (mean age 48.6, SD 13.9; 3 female) were investigated to define regions of interest from fMRI activation (see below). Full written consent was obtained from all subjects. The study was approved by the local Ethics Committees.

Functional MRI language paradigms

We included data from two similar event-related auditory comprehension experiments.

Paradigm 1 is identical to that used in Saur et al. (2006). In short, we presented a total of 184 stimuli distributed onto six sessions. These stimuli included 46 correct sentences (e.g. ‘The pilot flies the plane’), 46 sentences containing a semantic violation (e.g. ‘The pilot eats the plane’) and 92 time-reversed sentences that were derived from the intelligible sentences. Participants were asked to listen carefully to all stimuli and press a button whenever they detected a mistake. Reversed sentences thus had to be categorized as false.

Paradigm 2 differs from the first concerning number of stimuli and sessions. We presented a total of 90 stimuli in a single session. Stimuli included 15 correct sentences, 15 semantically violated sentences and 30 reversed sentences identical to those used in Paradigm 1. A novel third condition of 30 sentences of pseudo speech was derived from the correct and violated sentences by exchanging phonemes, but entered the analysis only when testing the auditory main effect (see below). Participants were asked to listen carefully to all stimuli and press a button at the end of each stimulus, irrespective of whether they had heard a normal, pseudo or reversed sentence.

In both paradigms, duration of stimuli ranged between 1730 and 2720 ms. Stimuli were presented binaurally in pseudo randomized order with an interstimulus interval varying between 3000 and 6000 ms. The sentences were spoken in German by a female voice. During scanning, subjects kept their eyes open. fMRI paradigms were equally distributed between outcome groups (Table 1).

MRI data acquisition

Structural and fMRI data from all participants were acquired on two 3 Tesla Siemens TIM Trio scanners using a standard head coil.

Functional MRI

In cases when the sequence specifications differed between paradigms, values for language Paradigm 2 are given in parentheses. We acquired 6 × 115 (1 × 256) scans in descending (interleaved) order with 32 (36) axial slices covering the whole brain using a gradient echo echo-planar T2*-sensitive sequence [resolution = 3 × 3 × 3 mm3, repetition time = 1.89 (2.19) s, echo time = 30 ms, flip angle = 70° (90°), matrix = 64 × 64 pixel2].
of view = 160 × 240 × 240 mm³) for spatial processing of the fMRI data. Sequence parameters of DWI for lesion delineation are given in the Supplementary methods.

Pre-processing of functional MRI data

fMRI data were processed using standard procedures implemented in the Statistical Parametric Mapping—5 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/). All slices were corrected for different acquisition times of signals by shifting the signal measured in each slice relative to the acquisition of the middle slice. Resulting volumes were spatially realigned and normalized to the Montreal Neurological Institute reference brain using the normalization parameters estimated during segmentation of the coregistered T1 anatomical scan (Ashburner and Friston, 2005). All normalized images were then smoothed using an isotropic 9 mm Gaussian kernel to account for inter-subject differences.

Statistical analysis of functional MRI data

At first level, the conditions intelligible speech (correct and violated), reversed speech and pseudo speech (in Paradigm 2 only) were modelled as separate regressors. Onset and duration of stimuli were convolved with a canonical haemodynamic response function as implemented in SPM5. Voxel-wise regression coefficients for the three conditions were estimated using weighted least squares.

Table 1 Patient characteristics

<table>
<thead>
<tr>
<th>All</th>
<th>Good outcome (AAT Severity ≤ I)</th>
<th>Bad outcome (AAT Severity &gt; I)</th>
<th>P-value</th>
<th>Good improvement (LRS Δ &gt; 0.28)</th>
<th>Bad improvement (LRS Δ ≤ 0.28)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>21</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>0.05</td>
</tr>
<tr>
<td>Gender (M/F)*</td>
<td>13/8</td>
<td>9/2</td>
<td>4/6</td>
<td>0.05</td>
<td>5/5</td>
<td>8/3</td>
</tr>
<tr>
<td>Age (mean, SD)</td>
<td>58.9 (±14.4)</td>
<td>58.3 (±10.9)</td>
<td>59.5 (±18.8)</td>
<td>0.9</td>
<td>56.8 (+17.7)</td>
<td>60.7 (+10.2)</td>
</tr>
<tr>
<td>Language paradigm (1/2)*</td>
<td>13/8</td>
<td>7/4</td>
<td>6/4</td>
<td>0.9</td>
<td>3/7</td>
<td>5/6</td>
</tr>
</tbody>
</table>

*Chi-squared tests were used to test for an equal distribution between groups (good versus bad outcome/improvement); significant values at P ≤ 0.05 are indicated in bold.

Table 2 Language testing for baseline evaluation

<table>
<thead>
<tr>
<th>Patient</th>
<th>Severity</th>
<th>Fluency</th>
<th>Type</th>
<th>REP [0–150]</th>
<th>WRI [0–90]</th>
<th>NAM [0–120]</th>
<th>COMP [0–120]</th>
<th>TT [0–50]</th>
<th>SPS [0–30]</th>
<th>AABT [0–240]</th>
<th>CETI [0–16]</th>
<th>TASK [0–100]</th>
<th>LRS [0–1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>P01</td>
<td>II</td>
<td>fl</td>
<td>Broca</td>
<td>133</td>
<td>70</td>
<td>85</td>
<td>56</td>
<td>15</td>
<td>16</td>
<td>195</td>
<td>13</td>
<td>82</td>
<td>0.67</td>
</tr>
<tr>
<td>P03</td>
<td>II</td>
<td>n-fl</td>
<td>Broca</td>
<td>144</td>
<td>60</td>
<td>66</td>
<td>57</td>
<td>8</td>
<td>16</td>
<td>147</td>
<td>8</td>
<td>8</td>
<td>0.43</td>
</tr>
<tr>
<td>P04</td>
<td>II</td>
<td>fl</td>
<td>anomic</td>
<td>143</td>
<td>81</td>
<td>106</td>
<td>90</td>
<td>23</td>
<td>20</td>
<td>232</td>
<td>13</td>
<td>100</td>
<td>0.79</td>
</tr>
<tr>
<td>P06</td>
<td>I-II</td>
<td>fl</td>
<td>Wernicke</td>
<td>117</td>
<td>80</td>
<td>100</td>
<td>96</td>
<td>16</td>
<td>23</td>
<td>238</td>
<td>13</td>
<td>88</td>
<td>0.77</td>
</tr>
<tr>
<td>P10</td>
<td>II-III</td>
<td>n-fl</td>
<td>global</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>68</td>
<td>14</td>
<td>13</td>
<td>142</td>
<td>3</td>
<td>83</td>
<td>0.42</td>
</tr>
<tr>
<td>P19</td>
<td>II-III</td>
<td>n-fl</td>
<td>global</td>
<td>136</td>
<td>75</td>
<td>68</td>
<td>87</td>
<td>9</td>
<td>25</td>
<td>223</td>
<td>9</td>
<td>88</td>
<td>0.70</td>
</tr>
<tr>
<td>P22</td>
<td>II-III</td>
<td>n-fl</td>
<td>global</td>
<td>108</td>
<td>53</td>
<td>85</td>
<td>94</td>
<td>35</td>
<td>21</td>
<td>229</td>
<td>10</td>
<td>59</td>
<td>0.72</td>
</tr>
<tr>
<td>P23</td>
<td>II-III</td>
<td>n-fl</td>
<td>Broca</td>
<td>93</td>
<td>30</td>
<td>78</td>
<td>87</td>
<td>31</td>
<td>2</td>
<td>229</td>
<td>8</td>
<td>100</td>
<td>0.55</td>
</tr>
<tr>
<td>P28</td>
<td>II-III</td>
<td>n-fl</td>
<td>global</td>
<td>96</td>
<td>21</td>
<td>52</td>
<td>71</td>
<td>3</td>
<td>2</td>
<td>185</td>
<td>6</td>
<td>100</td>
<td>0.35</td>
</tr>
<tr>
<td>P33</td>
<td>II-III</td>
<td>n-fl</td>
<td>Broca</td>
<td>116</td>
<td>37</td>
<td>78</td>
<td>63</td>
<td>25</td>
<td>20</td>
<td>215</td>
<td>10</td>
<td>100</td>
<td>0.70</td>
</tr>
<tr>
<td>P35</td>
<td>II-III</td>
<td>n-fl</td>
<td>Broca</td>
<td>111</td>
<td>28</td>
<td>85</td>
<td>64</td>
<td>47</td>
<td>15</td>
<td>227</td>
<td>6</td>
<td>76</td>
<td>0.71</td>
</tr>
<tr>
<td>Good outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Subjects showing good improvement (LRS > 0.28) are listed in bold; values in squared brackets indicate range of the respective test score. REP = repetition; WRI = writing; NAM = naming; COMP = auditory and reading comprehension; TT = Token Test; SPS = spontaneous speech; AABT = Aachen Aphasia Bedside Test; CETI = Communicative Effectiveness Index; TASK = task performance in the scanner. Severity and fluency as indicated by the AAT: I = mild; II = moderate; III = severe impairment; fl = fluent; n-fl = non-fluent.
In our earlier work on classification methods (Klöppel et al., 2008, 2009a) we have restricted the separating boundary to a straight line for the 2D case (or a hyperplane in the high dimensional case). The current work represents an extension using a radial basis function kernel with an additional parameter $\gamma$ which allows non-linear relations between the voxel values and the grouping (Vapnik, 1998). A non-linear relationship is likely given that data are very heterogeneous with stroke affecting different parts of the brain in every subject.

Figure 1A also illustrates that two samples are placed on the wrong side of the separating boundary during training. There is a trade-off between a very complex but perfect separation of the two groups and a simpler one that allows a few errors. A very complex separating boundary will be accurate for those samples that were used to define that boundary. However, correct grouping of new data (i.e. a new patient) is more likely to fail when the boundary is too specific. In the context of machine learning, this effect is called over-fitting and a cost parameter regulates a trade-off between an inexact grouping and a simpler one that allows a few errors. A very complex separating boundary will be accurate for those samples that were used to define that boundary. However, correct grouping of new data (i.e. a new patient) is more likely to fail when the boundary is too specific. In the context of machine learning, this effect is called over-fitting and a cost parameter regulates a trade-off between an inexact grouping and a simpler one that allows a few errors.

Support vector classification

A SVM is a supervised, multivariate classification method. SVMs are supervised in the sense that they include a training step to learn about differences between groups to be classified and can apply this knowledge to new data. An intuitive description of the method follows below. A more detailed version can be found in our previous work (Klöppel et al., 2008). For technical information on SVMs, we refer to the textbooks by Vapnik (1998) and Bishop (2006).

Theory

SVM-based grouping is performed in a training and a classification step, which is usually referred to as the testing step as it tests the quality of the training. In the training step, scans from subjects with known outcome are used to train a SVM. During this training process, all image characteristics (i.e. in case of fMRI the voxels coding the parameter estimate in the contrast images) are used to learn a ‘boundary’ that separates diagnostic groups (i.e. patients with good from those with bad language recovery). Figure 1A illustrates the training procedure in an imaginary 2D space: in this example the two groups to be classified are represented by circles and squares. As shown in the figure, a non-linear separating boundary provides best separation of the groups. With real data, there will be as many dimensions as there are input variables (i.e. voxels plus additional information, e.g. age and LRS) and the separating boundary becomes a high dimensional hypersurface.

Data input for classification

We tested classification performance based on fMRI or LRS plus age alone as well as fMRI plus LRS plus age. LRS and age were included as these parameters have been found to influence language recovery (Kertesz and McCabe, 1977; Lendrem and Lincoln, 1985; Laska et al., 2001; Pedersen et al., 2004; Lazar et al., 2008). Before entering the data into the SVM, each modality (i.e. LRS, age, parameter
images) was scaled separately to take values between –1 and +1. This ensures that each input gets the same importance and avoids numerical difficulties.

**Regions of interest**

Selecting regions of interest effectively improves the signal to noise ratio by excluding areas which are unlikely to contribute to the classification. Thus, in addition to analyses across the whole brain, we focused on key language areas identified from the 14 healthy subjects performing Paradigm 1 (see masks Fig. 3). By choosing an independent group of subjects we avoided any circular logic which could have otherwise arisen by defining relevant areas from the same data later used for classification. To define regions of interest, we computed the auditory and language effect in the control group using two separate one sample \( t \)-tests. We applied an uncorrected threshold of \( P < 0.001 \) to the resulting \( t \)-maps. Previous work (e.g. Weiller et al., 1995; Musso et al., 1999; Saur et al., 2006) has shown that the homologue areas of the right hemisphere are involved in the compensatory mechanism following stroke. Therefore, by adding a flipped and non-flipped version of the regions of interest we created two symmetrical masks from the group of healthy controls, which were used in binarized form. The language mask was further subdivided into four language regions in the left and right frontal as well as left and right temporal cortices (Fig. 3 upper right panel).

**Classification procedure and cross-validation**

In practical terms each input variable (i.e. each voxel of the parameter images, LRS, age) is treated as one point in a high dimensional space and its location is determined by the intensity value of each voxel (Fig. 1A). We used an in-house modification of the LIBSVM software (Chang and Lin, 2001), called LIBSVMTL, which is freely available for download (http://lmb.informatik.uni-freiburg.de/lmbsoft/libsvmtl/index.en.html).

A three-way-split cross-validation procedure was used to ensure that data classified in the testing step were independent from data used for training. To this end, a leave-one-out cross-validation procedure with an inner loop of training and testing to optimize the parameters for cost and \( \gamma \) was employed. This cross-validation ensures that the separating boundary can be generalized to correctly separate new data not used in the training process. It represents the clinical situation when predictions have to be performed for scans from a new patient. We report the average accuracy, i.e. the percentage of subjects left out of the training round and assigned to the correct group. The accuracy is reported with 95% confidence intervals (Newcombe, 1998).

**Displaying areas most relevant for the grouping**

In addition to the identification of a hypersurface we determined what voxels contributed most to classification and their distribution in the brain. It is the downside of a non-linear SVM that this visualization is rather complex, because in a certain region an increase in signal intensity can lead to a better recovery rate for a given patient, while it can lead to a worse recovery rate for another patient, and may not influence the recovery rate at all for a third. This is illustrated in Fig. 1A, where an intermediate voxel value (on the \( x \)-axis) indicates good recovery while higher or lower values indicate a poor outcome. Furthermore, due to the multivariate nature of the SVM, there can be cases where the influence of one region to the grouping depends on the signal pattern in another region. For fMRI parameter images this means e.g. if region A is activated to a certain level, then a higher activation in region B will strongly increase the likelihood for good
recovery, but if region A is not activated, increased activation in region B is no longer relevant for the outcome.

To provide insight into the classification process, we have implemented a novel visualization technique. At an intuitive level, the resulting 'weight image' contains information about which voxels are most important for the predicted outcome. They are computed for each new subject in every leave-one-out cycle.

Expressed in more technical language, we visualized relevant regions for each individual patient by a local linear approximation of the non-linear SVM. Thereby, the separating boundary is modelled indirectly by the zero-crossing of a so-called decision function which in turn is binarized at zero to allow grouping. The local linear approximation of this decision function is just the tangent on this function—and the slope of that tangent is the derivative of the decision function at this position. Such partial derivatives have already been used successfully for feature selection (Cao and Francis, 2000) or to incorporate certain invariances into non-linear SVMs (Chapelle and Schölkopf, 2001). The voxel value in the weight images reflects the slope of the tangent. A steeper slope signifies an important voxel as changing its value can change the grouping and thus the predicted outcome. A zero slope indicates that a change in signal intensity does not change the predicted recovery, meaning that this voxel is irrelevant for the grouping of this subject.

Results

Clinical data

Patients all suffered from an embolic left hemispheric infarction caused by a middle cerebral artery-stem or branch occlusion (Supplementary Table 1). Figure 2 displays DWI as mean image averaged across all 21 subjects. Lesion overlay shows that ischaemia affected all parts of the left middle cerebral artery territory with strongest lesion overlap found in subcortical regions.

At admission, 13 patients were treated with intravenous thrombolysis. All patients, except for Patient 11, showed vessel recanalization at time of fMRI. During the 6-month period, all patients received language rehabilitation as an in- and/or out-patient (Supplementary Table 1).

Demographic and behavioural data

Comparing demographics between outcome groups revealed that male subjects tended to show better language outcome (Table 1). An additional classification based only on gender was therefore calculated to exclude any bias resulting from this potential confound (see below).

Patients’ individual language testing, as well as severity and type of aphasia are given in Tables 2 and 3, an overview of the statistics between groups is displayed in Table 4.

Unsurprisingly, patients with good outcome (severity ≤ I in the AAT) yielded significantly better scores in many language tests at baseline resulting in a significantly better baseline LRS. The better outcome of these patients is reflected in significantly higher scores in almost all language tests except for the AAT subtest repetition. No significant differences were found in improvement between patient groups differing in outcome (Table 4).

In contrast, patients with substantial improvement (ΔLRS>−0.28) yielded lower test scores at baseline in most tests, while at outcome, no significant differences were found in most tests. Neither outcome nor improvement is driven by a single subtest but rather by changes in many different subtests (Table 4).

Predictions based on age and LRS

Outcome and improvement prediction based on LRS and age yielded maximal 62% correct classifications (Fig. 4, green dot on y-axis).

Prediction based on functional MRI

Figure 3 displays mean fMRI parameter images for the auditory and language contrast at baseline with typical activation in primary auditory and language areas.

Outcome and improvement prediction based on the auditory main effect from the whole brain or the auditory region of interest was at chance level. Outcome prediction based on the fMRI language effect reached 76% correct classification when using the language region of interest. Predicting the degree of improvement was similarly accurate when based on the right frontal subregion (Table 5).

Combined predictions

Overall best prediction for outcome with 86% correct classifications was achieved by combining LRS and age with the fMRI language effect. A similar accuracy for prediction of improvement was reached when based on LRS and age together with the
activation pattern of left or right hemispheric language regions (Table 5). A graphical overview of all analyses which yielded classification results beyond chance is given in Fig. 4.

**Potential bias caused by gender**

Gender tended to differ between outcome groups with more male patients showing a favourable outcome (Table 1). Therefore, we performed an additional analysis to exclude the possibility that classification was based on differences due to gender rather than language activation per se. In 11 female and 11 male patients, who were matched in baseline LRS, gender should be predicted based on language fMRI activation. The analysis was conducted as described above using the same masks. Accuracy was at chance level (maximum accuracy 59%) making it unlikely that differences in gender between diagnostic groups have had a relevant influence on the classifier.

**Relevant voxels**

Weight images for language fMRI of the winning analyses from an exemplary case are displayed in Fig. 5. The higher the voxel value in the weight image, the more important the voxel value in the underlying image for outcome prediction. This illustrates that a rather complex and heterogeneous pattern of voxels within language areas sub-served prediction of outcome and improvement (see supplementary results for an illustrative example based on whole brain DWI data).

**Discussion**

This study investigates early system-specific prediction of language outcome and improvement in individual stroke patients by applying a multivariate pattern classification technique to language fMRI data. We were able to separate aphasic stroke patients differing in language recovery best by combining language fMRI with baseline language testing and age. Prediction outperformed classification based on language testing and age alone which demonstrates that language fMRI contains substantial outcome-relevant information. The prediction accuracy reported here is encouraging and points to the high discriminative power of the multivariate technique, even in a heterogeneous disorder such as stroke and even with data from different imaging centres and slightly differing fMRI paradigms.

Recovery from stroke depends on a variety of factors including demographics, infarct characteristics, severity of initial impairment, brain status before the disabling stroke (e.g. preceding infarctions, small vessel disease), comorbidity, premorbid representation of the function of interest (e.g. left/right, bilateral representation of language) as well as factors modifying the recovery process (e.g. quality and quantity of rehabilitation therapy, recurrent strokes or newly acquired diseases). Successful outcome prediction can be achieved the more recovery is predetermined by these baseline factors and the
less heterogeneous the effects of post-stroke rehabilitation or complications. In our study, all patients received rehabilitation therapy as in- and/or out-patients and no patient suffered a recurrent stroke or other major complication.

**Functional MRI-based prediction of outcome**

Language fMRI contains system- (or modality-) specific information. Numbers of imaging studies in aphasic stroke patients have shown how language activation correlates with language performance at different stages after stroke (Weiller et al., 1995; Leff et al., 2002; Sharp et al., 2004; Crinion and Price, 2005; Saur et al., 2006). The idea behind fMRI for outcome prediction is that some of the signal related to performance at the time of scanning is related to future recovery too, and thus can be used for outcome prediction.

In the subacute phase approximately two weeks after stroke, a strong language-specific effect predominantly in bilateral frontal and left temporal regions was observed on parameter images. A strong increase of language activation from the acute (Supplementary material) to the subacute phase after stroke, mainly in bilateral frontal brain areas, is well established and was described in detail in our previous work (Saur et al., 2006). Outcome prediction based on the masked language effect at this stage was successful and reached 86% correct classification when combined with age and baseline testing. The pattern of activated and non-activated voxels within bilateral key language areas in this phase after stroke contains most language outcome relevant information (Fig. 5). It is noteworthy that although feeding LRS and age into the analysis improved outcome prediction in most of our analyses, these parameters used in isolation failed to predict the outcome accurately. This indicates that LRS and age
interacted with voxel values without being as informative when used in isolation. This can also be concluded from Table 4. Despite a highly significant difference of the baseline LRS between those with good and bad improvement at the group level, it does provide relatively little information to predict the improvement of the individual.

**Functional MRI-based prediction of improvement**

A similar accuracy was reached when predicting language improvement, which is complementary to the prediction of the absolute outcome. Interestingly, the highest accuracy was reached when based on the key language areas, and this includes areas of the right hemisphere, predominantly the right frontal cortex. As shown in our previous work in a group of 14 aphasic stroke patients, highest activation in the subacute phase was found in the right frontal cortex (Saur et al., 2006). This activation correlated with the early language improvement from the acute to the subacute phase. Our results here suggest that the pattern of activated voxels in the right frontal region is predictive for the degree of improvement from the subacute to the chronic phase.

We can only speculate why improvement rather than outcome prediction benefited from small regions of interest. Unsurprisingly, prediction accuracy for the degree of improvement increased substantially after including age and LRS. It informs the classifier about the level of baseline performance which otherwise is not included when patients are grouped based on improvement. An activation pattern with a low LRS may well indicate a different level of future improvement than the same pattern with a high LRS.

**Time window for prediction of recovery**

As shown in the Supplementary material, in the hyperacute phase after stroke outcome and in particular improvement prediction was less accurate. This was true for fMRI and DWI as well as the combination of both modalities. The only exception was the auditory fMRI effect which showed 83% correct classifications for the prediction of language outcome (although chance level was still included in the confidence interval). In the hyperacute phase, parameter images typically reflected a strong (bilateral) auditory main effect but almost no language specific effect (Supplementary Fig. 1). One important reason for the reduced language activation despite a strong auditory effect in this hyperacute phase is the phenomenon of diaschisis. This is defined as the dysfunction of structurally intact areas remote from the infarction which are functionally dependent from the lesioned area. Primary auditory areas seem to be less vulnerable to this effect and thus the auditory effect is more robust and reliable, detectable also in patients with severe aphasia. Therefore, it is reasonable that in the hyperacute phase, the auditory effect contains more outcome-relevant information compared to the language-specific effect. This is an encouraging result and should motivate further explorations of this phase for outcome prediction. Indeed, in a recent study on motor recovery, Marshall et al. (2009) demonstrated that in patients with hemiparesis activation in a simple motor task obtained within the first two days after stroke was correlated with motor recovery.

| Table 5 Classification performance for outcome prediction and improvement |
|-----------------------------|-----------------------------|-----------------------------|
|                             | Outcome                     | Improvement                  |
|                             | fMRI_LANG                    | fMRI_LANG                    |
|                             | + age/LRS                    | + age/LRS                    |
| Whole brain                 | 67 [65-83] (64/70)           | 52 [32-72] (45/60)           |
| Mask LANG                   | 52 [65-85] (80/91)           | 57 [55-98] (80/73)           |
| Mask R_TEMP                 | 67 [65-83] (70/64)           | 67 [65-83] (70/64)           |
| Mask L_TEMP                 | 50 [60-92] (80/82)           | 50 [60-92] (80/82)           |

The percentage of correct classifications is reported with 95% confidence intervals (Newcombe, 1998) in squared brackets and sensitivity/specificity values (a correctly identified subject with poor outcome or improvement being a true positive) in round brackets for the auditory (fMRI_AUD) and the language contrast (fMRI_LANG). The winning analysis is indicated in bold. Abbreviations for the masks are as indicated in the legend to Fig. 3.
after 3 months. It must be emphasized though that a correlation of baseline fMRI data with a behavioural outcome measure just reflects a relationship between both parameters, while the validity of the results when transferred to new data (as done in our study with the cross-validation scheme) remains unproven. This, however, is necessary when using the information in a clinical scenario to predict outcome on a single-subject level.

Methodological issues

One limitation of our study is that the low number of subjects included resulted in very large confidence intervals which often included chance level performance. In previous work on Alzheimer’s disease we found accuracy to decrease and to become more variable with fewer subjects (Klöppel et al., 2009b). This effect is likely to be even stronger in the current study as stroke lesions are generally more heterogeneous than those resulting from Alzheimer’s disease.

We included data from different imaging centres and from slightly differing fMRI paradigms but could not fully explore these effects due to the low number of subjects. With larger cohorts, the effect on the accuracy could be evaluated by using scans from one centre for training and from the other for testing, as in our previous work (Klöppel et al., 2008).

We expect that even higher accuracy values will be achievable, if outcome or improvement predictions are focussed on a more homogenous patient group. Given a sufficient amount of data, separate classifiers could be trained for patient subgroups which differ in type and severity of language impairment as well as lesion site. In addition, more than two outcome or improvement levels could be defined. This would render prediction more precise and also help to overcome the somewhat arbitrary selection of the cut-off values to create the grouping.

The use of non-linear SVMs makes the visualization and interpretation of the relevant patterns more difficult. The primary aim of the study was to optimize classification accuracy rather than localizing brain regions relevant for outcome prediction. Figure 5 mainly serves to illustrate that a distributed pattern of voxel is responsible for the grouping, speaking to the need to employ multivariate methods. Linear SVMs, which would allow easier visualization rarely performed better than chance level on the data of this study (data not shown).

Application to the clinical setting

This study provides a proof of principle demonstrating that prediction of system-specific recovery after stroke using a SVM based on system-specific imaging is feasible. The cross-validation scheme in this work demonstrates the extension to new data not used for training. It therefore resembles the clinical scenario when the outcome of a new given patient has to be predicted. We envision an application of these methods to the clinical routine. For example, it
seems conceivable that after training with a large sample of fMRI data from different scanners using similar paradigms, a separating boundary can be generated that is sufficiently stable to be applied to newly admitted patients, to evaluate their potential of recovery prospectively. Using a basic 10 min fMRI paradigm in the early phase after stroke should be practicable even in the clinical routine. Important, predictions presented here are based on intensive rehabilitation therapy and by no means justify a reduction of the rehabilitative effort if good recovery is predicted. In contrast, prediction of a bad outcome should motivate early and even more intensive therapy.

Future work may examine the transfer to other domains, e.g. motor function or attention, to evaluate whether similar classification results can be obtained for patients with hemiparesis or neglect. These studies could also explore whether including additional information (e.g. results of language subtests, neuropsychological testing, comorbidity factors) could further improve the prediction. Furthermore, a setting is conceivable in which the response to a specific treatment could be predicted: if the true responsiveness after therapy is known for some patients, SVMs could be trained with pre-treatment data (e.g. fMRI parameter images) to extract information relevant to predict treatment outcome and apply this to new patients. Such an approach could help assigning a given therapy to those who will benefit the most.

Funding

Federal Ministry of Education and Research (BMBF-research collaborations ‘Mechanisms of brain reorganization in the language network’ [01GW0661]); German Research Society (DFG; grant WE1352/14-1); Postdoctoral fellowship of the German Academic Exchange Service (DAAD) to DS; the Excellence Initiative of the German Federal and State Governments (EXC 294) to OR.

Supplementary material

Supplementary material is available at Brain online.

References


Cao L, Francis EHT. Feature selection for support vector machines. IDEAL 2000; 268–73.


