Predicting Tissue Outcome From Acute Stroke Magnetic Resonance Imaging
Improving Model Performance by Optimal Sampling of Training Data

Kristjana Yr Jonsdottir, PhD, MSc; Leif Østergaard, DMSc, PhD, MD, MSc; Kim Mouridsen, PhD, MSc

**Background and Purpose**—It has been hypothesized that algorithms predicting the final outcome in acute ischemic stroke may provide future tools for identifying salvageable tissue and hence guide individualized therapy. We developed means of quantifying predictive model performance to identify model training strategies that optimize performance and reduce bias in predicted lesion volumes.

**Methods**—We optimized predictive performance based on the area under the receiver operating curve for logistic regression and used simulated data to illustrate the effect of an unbalanced (unequal number of infarcting and surviving voxels) training set on predicted infarct risk. We then tested the performance and optimality of models based on perfusion-weighted, diffusion-weighted, and structural MRI modalities by changing the proportion of mismatch voxels in balanced training material.

**Results**—Predictive performance (area under the receiver operating curve) based on all brain voxels is excessively optimistic and lacks sensitivity in performance in mismatch tissue. The ratio of infarcting and noninfarcting voxels used for training predictive algorithms significantly biases tissue infarct risk estimates. Optimal training strategy is obtained using a balanced training set. We show that 60% of noninfarcted voxels consists of mismatch voxels in an optimal balanced training set for the patient data presented.

**Conclusions**—An equal number of infarcting and noninfarcting voxels should be used when training predictive models. The choice of test and training sets critically affects predictive model performance and should be closely evaluated before comparisons across patient cohorts. *(Stroke. 2009;40:3006-3011.)*

**Key Words:** acute stroke ■ diagnostic imaging ■ magnetic resonance imaging ■ tissue outcome prediction

---

Early identification of tissue at risk of infarction after acute ischemic stroke may aid clinical decision-making and potentially improve long-term patient outcome. Neuroimaging correlates of the ischemic core (irreversibly damaged tissue), the ischemic penumbra (electrically silent yet viable tissue), and (functioning) tissue with benign oligemia is therefore of immense clinical and scientific importance.

The perfusion–diffusion mismatch—tissue showing prolonged bolus transit time characteristics by perfusion-weighted MRI but normal diffusion-weighted imaging (DWI) characteristics—has found widespread use as an operational target for thrombolytic treatment. Although this dichotomization may identify patients who respond favorably to thrombolytic treatment, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.

Although predictive models integrate multimodal voxel-by-voxel image information from cases with known outcome into models that may predict outcome in subsequent patient cases, the underlying training material may critically affect the precision and reliability of such predictions. Intuitively, the model must be trained on a representative range of initial image findings (eg, transit time and diffusion parameter changes) and corresponding outcomes (infarcting, noninfarcting tissue) to gain sufficient precision to guide outcome predictions and warrant group comparisons.

We suggest a clinically informative measure of predictive performance sensitive to correct classification of mismatch tissue. The study aims to use a logistic regression model theory to determine the optimal training set (relative number of infarcting and noninfarcting voxels) and further examine how model performance depends on the proportion of voxels affected by the stroke (mismatch voxels) in the training follow-up imaging outcome (infarct, salvaged) into models predicting tissue outcome have therefore attracted considerable interest and are hypothesized to outperform prediction of outcome based on the mismatch dichotomization.
Finally, we examine how predicted tissue outcome and predicted lesion volume depend on training strategy.

Theory

Training a Predictive Model

Prediction of voxel outcome can be performed by identifying thresholds for a single image modality\(^1\),\(^1\)\(^1\),\(^1\)\(^2\) or by combining several modalities.\(^9\),\(^1\)\(^0\),\(^1\)\(^3\) In the latter case, infarction risk is assumed to depend on a linear combination of the acute modalities \(x_1, \ldots, x_k\) in a particular voxel, eg,

\[
\text{Prob (infarct | } x_1, \ldots, x_k = 1 + \exp[-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)]
\]

in which the coefficients are derived from a training set where voxel outcomes are known. More flexible models such as \(k\)-nearest neighbor\(^1\)\(^4\) and multivariate adaptive regression splines\(^1\)\(^5\) can be applied in a similar fashion.

Results from logistic regression model theory show that optimal models are obtained by using training data with a balanced number of infarcting and surviving voxels.\(^1\)\(^6\) Consequently, a subset of surviving voxels must be chosen to match the patient's final lesion in model training.

Testing a Predictive Model

The ability of a predictive algorithm to correctly predict final outcome is quantified by the size of the regions indicated in Figure 1A. For a given threshold, the number of correctly classified infarcting (true-positives [TP]) and surviving (true-negatives [TN]) voxels must be high, whereas the number of voxels falsely predicted to survive (false-negatives [FN]) or infarct (false-positives [FP]) must be low. This corresponds to simultaneously maximizing sensitivity (TP/TP+FN) and specificity (TN/TN+FP). To obtain a single performance measure independent of thresholds, the area under the receiver operating characteristic curve (AUC) can be calculated. The receiver operating characteristic curve is obtained by plotting (1-specificity) versus sensitivity for all thresholds. It can be shown that AUC is the probability that an infarcting voxel receives a higher score (risk of infarction) than a surviving voxel.\(^1\)\(^7\) This is also known as the Wilcoxon-Mann-Whitney test statistic.

Differentiating infarcting and surviving voxels in the mismatch area (ie, separating benign oligemia from penumbra) is a challenging aspect of predictive modeling—and indeed clinical management—in acute stroke. Consequently, we divide surviving voxels from a patient into 2 distinct subsets: surviving voxels from the mismatch area (R) and surviving voxels outside the mismatch area (S), ie, predominantly contralateral tissue. Performance can now be assessed using infarcting voxels together with surviving voxels from either R (AUC\(_R\)) or S (AUC\(_S\); Figure 1B). It can be shown that the AUC calculated using all voxels (both R, S, and infarcting voxels) is a linear combination of these measures

\[
\text{AUC} = \omega_R \text{AUC}_R + \omega_S \text{AUC}_S,
\]

where AUC\(_R\) is calculated using all infarcted voxels and voxels from R and AUC\(_S\) is calculated using all infarcted voxels and voxels from S (Figure 1B). The weights \(\omega_R\) and \(\omega_S\) are the volume fractions of R and S relative to the volume of all surviving voxels. AUC\(_R\) hence estimates the ability to differentiate infarcting and surviving voxels within the perfusion lesion area, whereas AUC\(_S\) measures the ability to correctly differentiate between the infarct and surviving tissue outside the mismatch area. The volume of R is by far smaller than the volume of S and therefore the AUC value obtained using all voxels will be dominated by performance in regions that trivially never infarct (AUC\(_S\)). Consequently, AUC based on all voxels is insensitive to correct classifica-

---

**Figure 1.** A, All brain voxels can be divided into 4 regions of interest after predicting the final infarct: those that were correctly classified as infarcted (TP), not correctly classified as infarcted (FN), correctly classified as surviving (TN), and those not correctly classified as surviving (FP). B, AUC can be calculated using all voxels (left), infarcting voxels and surviving voxels in mismatch (middle), and infarcting voxels and surviving voxels outside mismatch (right).
tion of mismatch voxels. To simultaneously maximize AUC_S and AUC_R, we consider the mean of AUC_R and AUC_S as a measure of overall predictive performance, AUC_0

\[ AUC_0 = \frac{1}{2} (AUC_R + AUC_S). \]

AUC_0 assigns equal importance to differentiation of infarcting voxels from affected and unaffected noninfarcting voxels.

In the following, voxels used to calculate performance of predictive models are denoted as test data. For illustrating the differentiation of infarcting voxels from affected and unaffected noninfarcting voxels.

To model sensitivity to selection of the noninfarcting voxels in a simulation study: Effect of infarcting voxel proportion in training data

To assess sensitivity to training data selection and verify theoretical results, we simulated MTT values (n=1000) of infarcting and noninfarcting voxels based on MTT values obtained in patient data (noninfarcting: 5±2.5; infarcting: 10±4.5). Logistic regression models were developed with proportion of surviving voxels ranging from 2% to 98%. Independent test samples were simulated (n=1000) using this distribution of MTT and the proportion of surviving voxels was also varied. Mean risk of infarction for infarcting voxels in the simulated test data was calculated for each model.

Effect of training data selection on performance in patient data

To model sensitivity to selection of the noninfarcting voxels in a balanced training sample, we varied the sample fraction of surviving voxels uniformly sampled from the mismatch (R) and unaffected tissue (S) in training material. For each sample fraction, 16 models (each with one patient left out in training) were developed using jackknifing. Using these models, the performance measures AUC, AUC_R, AUC_S, and AUC_0 were calculated for each patient and an optimal training sample was determined as the balance between surviving voxels from R and S, which maximizes AUC_0. The ensuing model is denoted by M_0. For comparison, we developed models using training sets in which all surviving voxels were from S (model M_S) or all surviving voxels were from R (model M_R).

Regional infarct risk dependence on training data selection

Infarct risk in core (apparent diffusion coefficient lesion), recruited tissue (mismatch tissue infarcted on follow-up), salvaged tissue (mismatch tissue normal on follow-up), and normal (ipsilateral) tissue was obtained using models M_R, M_S, and M_0. Predicted

### Table. Patient Demographics, Occlusion Type, Time From Onset to Scan, Admission National Institutes of Health Stroke Scale (NIHSS) Score, Volumes of Acute MTT and ADC Lesions, and Follow-Up Lesion Volume (MLV)

<table>
<thead>
<tr>
<th>Patient No.</th>
<th>Age, Years</th>
<th>Gender</th>
<th>Infarct Side</th>
<th>Occlusion Type</th>
<th>Time From Onset to Scan, Minutes</th>
<th>Admission NIHSS</th>
<th>Acute MTT Volume, mL</th>
<th>Acute ADC Volume, mL</th>
<th>Follow-Up Volume, mL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>M</td>
<td>R</td>
<td>No vessel occlusion</td>
<td>115</td>
<td>10</td>
<td>126.8</td>
<td>19.0</td>
<td>9.6</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>M</td>
<td>L</td>
<td>ICA</td>
<td>140</td>
<td>20</td>
<td>192.4</td>
<td>49.9</td>
<td>85.1</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>F</td>
<td>R</td>
<td>Distal MCA trunk</td>
<td>120</td>
<td>18</td>
<td>200.8</td>
<td>34.8</td>
<td>58.6</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>M</td>
<td>L</td>
<td>ACA</td>
<td>130</td>
<td>16</td>
<td>106.4</td>
<td>44.8</td>
<td>23.7</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>F</td>
<td>R</td>
<td>M2</td>
<td>125</td>
<td>11</td>
<td>136.0</td>
<td>64.8</td>
<td>67.8</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>M</td>
<td>R</td>
<td>ICA</td>
<td>105</td>
<td>12</td>
<td>66.3</td>
<td>11.1</td>
<td>12.7</td>
</tr>
<tr>
<td>7</td>
<td>56</td>
<td>M</td>
<td>R</td>
<td>ICA</td>
<td>127</td>
<td>15</td>
<td>232.2</td>
<td>19.6</td>
<td>16.1</td>
</tr>
<tr>
<td>8</td>
<td>67</td>
<td>F</td>
<td>R</td>
<td>Distal MCA trunk and M2</td>
<td>83</td>
<td>15</td>
<td>190.9</td>
<td>26.9</td>
<td>76.9</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>M</td>
<td>L</td>
<td>Distal branch</td>
<td>63</td>
<td>4</td>
<td>92.0</td>
<td>18.9</td>
<td>19.7</td>
</tr>
<tr>
<td>10</td>
<td>78</td>
<td>M</td>
<td>L</td>
<td>Carotid-T</td>
<td>150</td>
<td>20</td>
<td>73.7</td>
<td>6.6</td>
<td>5.5</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>F</td>
<td>L</td>
<td>M2</td>
<td>50</td>
<td>6</td>
<td>82.1</td>
<td>19.7</td>
<td>16.5</td>
</tr>
<tr>
<td>12</td>
<td>71</td>
<td>F</td>
<td>L</td>
<td>Distal branch</td>
<td>75</td>
<td>5</td>
<td>29.7</td>
<td>5.1</td>
<td>11.7</td>
</tr>
<tr>
<td>13</td>
<td>72</td>
<td>M</td>
<td>L</td>
<td>ICA</td>
<td>110</td>
<td>21</td>
<td>237.2</td>
<td>78.8</td>
<td>155.41</td>
</tr>
<tr>
<td>14</td>
<td>67</td>
<td>M</td>
<td>L</td>
<td>Proximal MCA</td>
<td>110</td>
<td>16</td>
<td>180.6</td>
<td>143.3</td>
<td>114.6</td>
</tr>
<tr>
<td>15</td>
<td>45</td>
<td>M</td>
<td>R</td>
<td>MCA distal branch</td>
<td>138</td>
<td>3</td>
<td>22.3</td>
<td>18.9</td>
<td>15.8</td>
</tr>
<tr>
<td>16</td>
<td>59</td>
<td>M</td>
<td>R</td>
<td>MCA</td>
<td>120</td>
<td>12</td>
<td>160.9</td>
<td>81.6</td>
<td>40.0</td>
</tr>
</tbody>
</table>

ADC indicates apparent diffusion coefficient; M, male; F, female; R, right; L, left; ICA, internal carotid artery; MCA, middle cerebral artery; ACA, anterior cerebral artery.
lesion volume at a threshold of 50% was calculated for the models. The predicted lesion volumes were pairwise compared and compared with the actual measured volumes of the “true” outcome lesion. Moreover, the measured volumes were compared with the volume of the acute MTT lesions. Comparisons were done using a signed Wilcoxon test.

Effect of Test Data Balance on AUC

For further illustrating the dependence of AUC on test balance, we calculated AUC using infarcting voxels and surviving voxels from R (AUCR) for M0. Then, we recalculated AUC after adding surviving voxels from S in increments of 2% until all brain voxels were used for calculation of AUC.

Simulated Data

Median infarction risk increases substantially when the number of infarcting voxels in the training set increases (Figure 2), eg, a voxel with MTT = 10 seconds has a predicted infarct risk of 18% if 4% of the training sample consists of infarcting voxels in contrast to 76% when using a balanced training set. The result demonstrates the importance of sampling equally from infarcting and surviving voxels as indicated by theoretical results.16

Training Data Selection Dependence on Performance in Patient Data

The performance measures AUC, AUCR, and AUCS depend critically on training strategy (Figure 3A). AUC and AUCS do not differ substantially and decrease as the sampling fraction of surviving mismatch voxels increases, ie, the ability to correctly predict unaffected tissue is compromised. This decrease is small if the sampling fraction in R is between 0% and 60%, but as high as 12.4% if it exceeds 60%.

However, AUCR increases continuously as the sample fraction of surviving mismatch voxel increases. Maximum AUC0 is obtained for sampling fraction equal to 60% (Figure 3B), which is seemingly an optimal proportion of noninfarcting mismatch voxels. The corresponding model is denoted by M0. However, sampling fractions from 40% to 60% give similar results.

Regional Infarct Risk Dependence on Training Data Selection

Risk of infarction in core, recruited, salvaged, and normal tissue depends critically on which model (M0, M0, or M0) is applied for prediction (Figure 4). Using the common 50%
threshold for predicting the final infarct results in a large part of salvaged tissue being falsely classified as infarcted for model MC in comparison to model M0. Moreover, interquartile range for risk of infarction in salvaged tissue is larger for MC compared with M0. The predicted lesion volumes of M0 were significantly different from the predicted lesion volumes of MC and MP (P<0.001). The predicted lesion volumes overestimate the final infarct for the 3 models (median differences and interquartile ranges: 78 mL [60 to 116] for MC, 35 mL [24 to 65] for M0, 78 mL [33 to 137] for MP). The acute MTT lesion overestimates the final infarct by 75 mL in median (interquartile range, 66 to 114), which is significantly larger than the bias observed using M0. Figure 5 shows risk maps thresholded at 50% for Patients 11 and 13 using models MC, M0, and MP and the corresponding acute DWI and MTT images. Patient 11 had 50% reperfusion 2 hours after treatment, whereas Patient 13 did not reperfuse. MC clearly overestimates the final infarct of both patients, whereas higher accuracy is observed for M0. Applying MP, areas of false infarct predictions are observed in normal tissue.

Overall Performance
As can be seen in Figure 3C, adding surviving voxels from S to the test set implies an increase in calculated AUC, which we ascribe to different distribution of infarction risks of noninfarcted voxels outside and inside the mismatch. This further illustrates the dependence of calculated performance on test sets. Figure 3D illustrates the interacting effects of testing and training strategies on AUC.

Discussion
Our results confirm earlier findings9,10,13 that logistic regression models may predict final outcome with higher accuracy than the conventional mismatch. Confirming theoretical results, our simulations further demonstrate that a balanced training strategy should be applied to avoid bias in predicted infarct risk. We find that tissue infarct risk estimates depend critically on the choice of the balanced training data (Figures 4 and 5). Furthermore, our results demonstrate how model training may be further refined by determining the ratio of mismatch versus unaffected voxels among noninfarcting tissue to optimize model performance.

As a means of assessing predicted infarct volume, voxels with a predicted infarct risk of >50% has been proposed.9,10,16 Our study confirms this threshold choice yet revealing that predicted lesion volumes based on a 50% threshold vary greatly depending on training strategy. Final lesion volume is consistently overestimated by applying a 50% threshold, but bias may be minimized by the proposed optimal training strategy. Moreover, bias is significantly smaller using the optimal training strategy compared with the conventional mismatch approach. Small susceptibility artifacts caused a part of the observed overestimation. Given the immense potential of predictive algorithms in clinical decision-making and drug efficacy testing demonstrated in pioneering work so far,10,13 our work emphasizes the importance of special attention to optimal selection of training data. In particular, the comparison of observed lesion volumes in patients after a specific treatment with their lesion volume predicted under the condition of no treatment is only possible if the composition of training sets is carefully chosen. This also applies to designs in which treatment efficacy is demonstrated as the decrease in regional infarct risk (Figure 4).

One disadvantage of logistic regression models is that estimating SEs of model parameters requires independence of voxel outcomes. Although this is a common assumption,9,10 the pathophysiology of stroke predicts a regional correlation of infarct risk levels that affects the reliability of derived CIs and subsequent hypothesis testing. To obtain correct error estimates, we suggest that spatial autocorrelation structures21 should be modeled.

AUC is an established measure of predictive ability, although it may be a misleading measure of performance. It avoids subjectivity in threshold selection by summarizing model performance over all possible risk thresholds. In particular, thresholds corresponding to diagnostically unacceptable low sensitivity and specificity contribute to AUC.
Furthermore, the actual scale of infarction risks and the goodness of fit of the model are ignored. Although AUC is theoretically independent of the balance between positive and negative outcomes in a test set, we have shown that AUC changes according to the regional clustering of outcomes. Predicting voxel outcome in the mismatch area is of primary clinical interest and when assessing treatment efficacy. We suggested the measure AUC$_{mp}$, which is sensitive to performance in the mismatch area, because AUC calculated using all brain voxels is insensitive to discrimination of infarcting and noninfarcting mismatch voxels. Other measures of performance used in the literature, eg, the root mean square error (Brier score), accuracy, and Youden’s J-index, would also be affected by the choice of testing material.

The sampling method for selecting training material depends on the specific outlining of the perfusion-weighted imaging lesion and, consequently, optimal sample fraction may depend on the perfusion metric and threshold applied. Using a time-to-peak threshold of 6 seconds as mismatch criterion did not alter the optimal fraction criterion significantly (results not shown). The sampling method can also be applied for other supervised learning methods used for classification of tissue outcome. The optimal choice of training set may vary according to patient cohort, therapy, time from onset, type of stroke, and so on.

There are several limitations to the study. Despite the overwhelming number of image voxels included, the patient cohort was limited, just as we did not include anatomic, stroke type, or reperfusion information, all speculated to improve specificity of outcome prediction. The correct assignment of voxel outcome depends on accurate coregistration of acute and follow-up images and is possibly biased by infarct shrinkage. Means of reducing this potential bias by excluding voxels located at infarct edges as well as means of automatically outlining final lesion volume to reduce bias caused by interrater classification disagreement should hence be pursued.

Acknowledgments

We acknowledge Niels Hjort, Mahmoud Ashkanian, and Christine Sölling for collecting the data used in the manuscript.

Sources of Funding

Supported by the Danish National Research Foundation (L.Ø., K.M.), the EU Commissions Sixth Framework Program (L.Ø., K.Y.J., Contract 027294), and GlaxoSmithKline (K.M., educational grant).

Disclosures

None.

References

Predicting Tissue Outcome From Acute Stroke Magnetic Resonance Imaging: Improving Model Performance by Optimal Sampling of Training Data
Kristjana Yr Jonsdottir, Leif Østergaard and Kim Mouridsen

Stroke. 2009;40:3006-3011; originally published online July 16, 2009;
doi: 10.1161/STROKEAHA.109.552216

Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/40/9/3006

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/