Predicting Outcome in Ischemic Stroke
External Validation of Predictive Risk Models

Karen C. Johnston, MD, MSc; Alfred F. Connors, Jr, MD; Douglas P. Wagner, PhD; E. Clarke Haley, Jr, MD

Background—Six multivariable models predicting 3-month outcome of acute ischemic stroke have been developed and internally validated previously. The purpose of this study was to externally validate the previous models in an independent data set.

Summary of Report—We predicted outcomes for 299 patients with ischemic stroke who received placebo in the National Institute of Neurological Disorders and Stroke rt-PA trial. The model equations used 6 acute clinical variables and head CT infarct volume at 1 week as independent variables and 3-month National Institutes of Health Stroke Scale, Barthel Index, and Glasgow Outcome Scale as dependent variables. Previously developed model equations were used to forecast excellent and devastating outcome for subjects in the placebo tissue plasminogen activator data set. Area under the receiver operator characteristic curve was used to measure discrimination, and calibration charts were used to measure calibration. The validation data set patients were more severely ill (National Institutes of Health Stroke Scale and infarct volume) than the model development subjects. Area under the receiver operator characteristic curves demonstrated remarkably little degradation in the validation data set and ranged from 0.75 to 0.89. Calibration curves showed fair to good calibration.

Conclusions—Our models have demonstrated excellent discrimination and acceptable calibration in an external data set. Development and validation of improved models using variables that are all available acutely are necessary. (Stroke. 2003;34:200-202.)

Key Words: cerebral ischemia ■ models, statistical ■ prognosis ■ stroke outcome

A multivariable model that could predict outcome after stroke would be useful in clinical trials to assess the balance of treatment groups and to predict expected outcomes of patients who are lost to follow-up. We developed and internally validated a series of predictive risk models for 229 ischemic stroke patients from the Randomized Trial of Tirilazad Mesylate in Patients with Acute Stroke (RANT-TAS)1. However, the models have not been validated in an external data set. The purpose of this study was to assess the validity of those predictive models in an independent data set.

Methods

Study Population
Two hundred ninety-nine patients from the placebo group of the National Institute of Neurological Disorders and Stroke (NINDS) rt-PA trial were used for the validation analysis.2 The NINDS rt-PA trial population has been described in detail previously.2 Briefly, this was an ischemic stroke population treated with intravenous tissue plasminogen activator (tPA) or placebo within 3 hours from symptom onset. Only the placebo group was used for this analysis because intravenous tPA is known to improve clinical outcome, and the predictive model being validated was designed to predict outcome without an intervention.

Three hundred twelve patients were treated with placebo in the NINDS rt-PA trial. Thirteen patients were excluded from our analysis for missing variables: 6 were missing 7- to 10-day CT scan infarct volume, 5 were missing stroke history information, and 2 were missing diabetes history information. The remaining 299 were used for this analysis.

Independent Variables
Baseline clinical information including age, National Institutes of Health Stroke Scale (NIHSS) score,3 history of previous stroke, history of diabetes mellitus, and history of prestroke disability were collected acutely. Infarct volume and stroke subtype were collected at 7 to 10 days after stroke onset.

Outcome Variables
The NIHSS, Barthel Index (BI),4 and Glasgow Outcome Scale (GOS)5 were used as outcome measures 3 months after stroke symptom onset. Each was dichotomized into excellent outcome (NIHSS ≤1, BI ≥95, GOS =1) or devastating outcome (NIHSS ≥20 or death, BI <60 or death, GOS ≥2) as previously defined.1
TABLE 1. Clinical and Imaging Characteristics of the 2 Populations

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Original (RANTTAS Data), n=229</th>
<th>Validation (NIHDS Data), n=199</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (median, IQ range)</td>
<td>70 (61–77)</td>
<td>67 (59–74)</td>
</tr>
<tr>
<td>Sex (M, %)</td>
<td>56</td>
<td>59</td>
</tr>
<tr>
<td>Race, %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>83</td>
<td>63</td>
</tr>
<tr>
<td>Black</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>NIHSS (median, IQ range)</td>
<td>10 (5–16)</td>
<td>15 (10–20)</td>
</tr>
<tr>
<td>CT volume, cm³ (median, IQ range)</td>
<td>9 (6–65)</td>
<td>28 (6–104)</td>
</tr>
</tbody>
</table>

Excellent outcomes, %

| NIHSS     | 43 | 20 |
| BI        | 56 | 39 |
| GOS       | 48 | 31 |

Devastating outcomes, %

| NIHSS     | 15 | 24 |
| BI        | 29 | 42 |
| GOS       | 32 | 46 |

IQ indicates interquartile.

Statistical Analysis

The previously defined models were forecast to the study population using the previously defined weights. Model discrimination was assessed using area under the receiver operating characteristic (ROC) curve, which was computed by a nonparametric method. An area under the ROC curve of 0.5 indicates no ability to discriminate and an area of 1.0 indicates perfect discrimination. We prespecified an acceptable area under the ROC as ≥0.8. Calibration was assessed using calibration curves. The perfect 45° line demonstrates ideal calibration. The closer the model calibration is to the line of identity, the better the calibration. Hosmer-Lemeshow tests of the calibration accuracy indicate that 5 of the 6 models have excellent discrimination (ROC area ≥0.8). The model for devastating outcome defined by the NIHSS (ROC area=0.75) likely discriminates less well due to the small number of devastating outcomes in the model development data set. Five of the 6 models appeared to be reasonably well calibrated, as shown in the calibration curves. The Hosmer-Lemeshow test demonstrates that these are not perfectly calibrated but the calibration appears to be adequate.

The validation population clearly had more severe strokes as demonstrated by both the NIHSS and infarct volume. Although the model discrimination was affected little by this, the poorer calibrations in the 6 models compared with the original models may be related to this difference. The fact that the models had such good discrimination and adequate calibration in 5 of the 6 models, even in such a different population, supports the generalizability of these models.

A predictive tool that would allow an accurate prediction of individual outcome at 3 months could be very useful in clinical research. The great heterogeneity among stroke patients may contribute to difficulty identifying treatment effects in clinical trials. Heterogeneity in a randomized clinical trial is mathematically expected to result in an underestimate of the treatment effect. A predictive model that could adjust for heterogeneity would allow a less biased estimate of the treatment effect and demonstrate a larger treatment effect in the same sized sample.

The small number of subjects and least frequent outcomes in both data sets limits how well the models can predict outcome. These models are also limited by the fact that only 5 of the 7 variables used in the prediction were collected acutely. Infarct volume and stroke subtype were collected at 1 week. Although a prediction of 3-month outcome is still valuable to the clinician at 1 week, this limits the use of these models in acute stroke clinical research. The dichotomized outcomes, though identifying the extreme outcomes (excellent outcome suggesting full or nearly full recovery; devastating outcome suggesting nursing home level disability or death), are not designed to predict the clinically relevant recovery levels in between. These extreme outcomes using the NIHSS, BI, and GOS are most useful in the clinical research realm in that they allow more reliable comparisons of standardized outcomes. These models, therefore, function as the proof of concept very little decline in area under the ROC curve in all 6 data sets and demonstrate excellent discrimination in 5 of the 6 models.

Model calibration is demonstrated in the Figure. Five of the 6 models have calibration curves that are very similar to the line of identity. The excellent outcome as measured by the NIHSS model calibrated less well and predicted a greater probability of excellent outcome than was observed in the validation data set for most of the range of predicted probabilities. For example, in the calibration curve “Excellent NIH Outcome,” at a model prediction of 70% probability of excellent outcome (fourth point on the curve), only about 40% of patients were actually observed to have excellent outcome by the NIHSS. Hosmer-Lemeshow tests of the calibration accuracy indicate that 5 of the 6 models were detectably (P<0.05) less than perfect, with all of the models tending to predict better outcomes than were observed.

Discussion

These 6 models have now been validated in an external data set. Five of the models have excellent discrimination (ROC area ≥0.8). The model for devastating outcome defined by the NIHSS (ROC area=0.75) likely discriminates less well due to the small number of devastating outcomes in the model development data set. Five of the 6 models appeared to be reasonably well calibrated, as shown in the calibration curves. The Hosmer-Lemeshow test demonstrates that these are not perfectly calibrated but the calibration appears to be adequate.

The validation population clearly had more severe strokes as demonstrated by both the NIHSS and infarct volume. Although the model discrimination was affected little by this, the poorer calibrations in the 6 models compared with the original models may be related to this difference. The fact that the models had such good discrimination and adequate calibration in 5 of the 6 models, even in such a different population, supports the generalizability of these models.

A predictive tool that would allow an accurate prediction of individual outcome at 3 months could be very useful in clinical research. The great heterogeneity among stroke patients may contribute to difficulty identifying treatment effects in clinical trials. Heterogeneity in a randomized clinical trial is mathematically expected to result in an underestimate of the treatment effect. A predictive model that could adjust for heterogeneity would allow a less biased estimate of the treatment effect and demonstrate a larger treatment effect in the same sized sample.

The small number of subjects and least frequent outcomes in both data sets limits how well the models can predict outcome. These models are also limited by the fact that only 5 of the 7 variables used in the prediction were collected acutely. Infarct volume and stroke subtype were collected at 1 week. Although a prediction of 3-month outcome is still valuable to the clinician at 1 week, this limits the use of these models in acute stroke clinical research. The dichotomized outcomes, though identifying the extreme outcomes (excellent outcome suggesting full or nearly full recovery; devastating outcome suggesting nursing home level disability or death), are not designed to predict the clinically relevant recovery levels in between. These extreme outcomes using the NIHSS, BI, and GOS are most useful in the clinical research realm in that they allow more reliable comparisons of standardized outcomes. These models, therefore, function as the proof of concept very little decline in area under the ROC curve in all 6 data sets and demonstrate excellent discrimination in 5 of the 6 models.

Model calibration is demonstrated in the Figure. Five of the 6 models have calibration curves that are very similar to the line of identity. The excellent outcome as measured by the NIHSS model calibrated less well and predicted a greater probability of excellent outcome than was observed in the validation data set for most of the range of predicted probabilities. For example, in the calibration curve “Excellent NIH Outcome,” at a model prediction of 70% probability of excellent outcome (fourth point on the curve), only about 40% of patients were actually observed to have excellent outcome by the NIHSS. Hosmer-Lemeshow tests of the calibration accuracy indicate that 5 of the 6 models were detectably (P<0.05) less than perfect, with all of the models tending to predict better outcomes than were observed.

Discussion

These 6 models have now been validated in an external data set. Five of the models have excellent discrimination (ROC area ≥0.8). The model for devastating outcome defined by the NIHSS (ROC area=0.75) likely discriminates less well due to the small number of devastating outcomes in the model development data set. Five of the 6 models appeared to be reasonably well calibrated, as shown in the calibration curves. The Hosmer-Lemeshow test demonstrates that these are not perfectly calibrated but the calibration appears to be adequate.

The validation population clearly had more severe strokes as demonstrated by both the NIHSS and infarct volume. Although the model discrimination was affected little by this, the poorer calibrations in the 6 models compared with the original models may be related to this difference. The fact that the models had such good discrimination and adequate calibration in 5 of the 6 models, even in such a different population, supports the generalizability of these models.

A predictive tool that would allow an accurate prediction of individual outcome at 3 months could be very useful in clinical research. The great heterogeneity among stroke patients may contribute to difficulty identifying treatment effects in clinical trials. Heterogeneity in a randomized clinical trial is mathematically expected to result in an underestimate of the treatment effect. A predictive model that could adjust for heterogeneity would allow a less biased estimate of the treatment effect and demonstrate a larger treatment effect in the same sized sample.

The small number of subjects and least frequent outcomes in both data sets limits how well the models can predict outcome. These models are also limited by the fact that only 5 of the 7 variables used in the prediction were collected acutely. Infarct volume and stroke subtype were collected at 1 week. Although a prediction of 3-month outcome is still valuable to the clinician at 1 week, this limits the use of these models in acute stroke clinical research. The dichotomized outcomes, though identifying the extreme outcomes (excellent outcome suggesting full or nearly full recovery; devastating outcome suggesting nursing home level disability or death), are not designed to predict the clinically relevant recovery levels in between. These extreme outcomes using the NIHSS, BI, and GOS are most useful in the clinical research realm in that they allow more reliable comparisons of standardized outcomes. These models, therefore, function as the proof of concept very little decline in area under the ROC curve in all 6 data sets and demonstrate excellent discrimination in 5 of the 6 models.

Model calibration is demonstrated in the Figure. Five of the 6 models have calibration curves that are very similar to the line of identity. The excellent outcome as measured by the NIHSS model calibrated less well and predicted a greater probability of excellent outcome than was observed in the validation data set for most of the range of predicted probabilities. For example, in the calibration curve “Excellent NIH Outcome,” at a model prediction of 70% probability of excellent outcome (fourth point on the curve), only about 40% of patients were actually observed to have excellent outcome by the NIHSS. Hosmer-Lemeshow tests of the calibration accuracy indicate that 5 of the 6 models were detectably (P<0.05) less than perfect, with all of the models tending to predict better outcomes than were observed.
that predictive models can be developed and then internally and externally validated for the prediction of 3-month outcome. Future models must now be developed using variables that are all available in the acute setting.

Acknowledgments

Dr Johnston is supported by the National Institutes of Health, National Institute of Neurologic Disorders and Stroke (grant No. K23NS02168-01).

The NINDS rt-PA Stroke Trial was funded by the National Institutes of Health, National Institute of Neurologic Disorders and Stroke through contracts to the participating sites. The RANTTAS study was supported, in part, by the National Institutes of Health, National Institute of Neurological Disorders and Stroke (grant No. R01-NS31554), and Pharmacia and Upjohn Company (Kalamazoo, Mich).

The authors gratefully acknowledge the contribution of the NINDS rt-PA Stroke Trial investigators and the RANTTAS investigators, without whose efforts this work would not have been possible.

References


Predicting Outcome in Ischemic Stroke: External Validation of Predictive Risk Models
Karen C. Johnston, Alfred F. Connors, Jr, Douglas P. Wagner and E. Clarke Haley, Jr

*Stroke*. 2003;34:200-202; originally published online December 5, 2002;
doi: 10.1161/01.STR.0000047102.61863.E3

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 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/34/1/200

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/