Validation of an Acute Ischemic Stroke Model
Does Diffusion-Weighted Imaging Lesion Volume Offer a Clinically Significant Improvement in Prediction of Outcome?

Karen C. Johnston, MD, MSc; Douglas P. Wagner, PhD; Xin-Qun Wang, MS; George C. Newman, MD; Vincent Thijs, MD, PhD; Souvik Sen, MD, MS; Steven Warach, MD, PhD; for the GAIN, Citicoline, and ASAP Investigators

Background and Purpose—Prediction models for ischemic stroke outcome have the potential to contribute prognostic information in the clinical and/or research setting. The importance of diffusion-weighted magnetic resonance imaging (DWI) in the prediction of clinical outcome, however, is unclear. The purpose of this study was to combine acute clinical data and DWI lesion volume for ischemic stroke patients to determine whether DWI improves the prediction of clinical outcome.

Methods—Patients (N=382) with baseline DWI data from the Glycine Antagonist In Neuroprotection and citicoline (010 and 018) trials were used to develop the prediction models by multivariable logistic regression. Data from prospectively collected patients (N=266) from the Acute Stroke Accurate Prediction Study were used to externally validate the model equations. The models predicted either full recovery or nursing home–level disability/death, as defined by the National Institutes of Health Stroke Scale, Barthel Index, or modified Rankin Scale.

Results—The full-recovery models with DWI lesion volume had areas under the receiver operating characteristic curves (AUCs) of 0.799 to 0.821, and those without DWI lesion volume had AUCs of 0.758 to 0.798. The nursing home–level disability/death models with DWI had AUCs of 0.832 to 0.882, and those without DWI had AUCs of 0.827 to 0.867. All models had mean absolute errors ≤0.4 for calibration.

Conclusions—All 12 models had excellent discrimination and calibration, with 8 of 12 meeting prespecified performance criteria (AUC ≥0.8, mean absolute error ≤0.4). Although DWI lesion volume significantly increased model explanatory power, the magnitude of increase was not large enough to be clinically important.

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

Clinicians and researchers must frequently anticipate the potential outcomes of the patients they encounter. Prediction models for ischemic stroke outcomes have the potential to contribute prognostic information to the clinical and/or research setting.1–3 Models that use CT infarct information have been limited by the fact that CT does not measure acute stroke infarct volume. For models that include acute clinical information and 1-week infarct volume, clinical and imaging information together predict outcome better statistically than does clinical information alone,4 although the clinical relevance is less clear. The use of 1- to 2-week-old clinical information plus CT or magnetic resonance imaging variables has not demonstrated added value to the prediction.5,6 Because these data suggest that imaging information from the acute stroke setting may be most useful in the prediction of outcome, acute diffusion-weighted imaging (DWI) data have been considered as predictors. The magnitude of the contribution of acute DWI in predicting outcome, however, is unclear because the data are mixed. Several models have suggested the value of early DWI imaging in the prediction of recovery,2,3 whereas others have not supported a strong relation.7,8

The purpose of this study was to combine acute clinical information and DWI data in 2 large acute stroke datasets to determine whether DWI improves the prediction of 3-month outcome: either full/nearly full recovery or nursing home–level disability/death. We hypothesized that the models with DWI would predict outcome better than the models without DWI. Additionally, we hypothesized that the models with DWI would be valid as determined by an area under the receiver operating characteristic curve.

Received November 30, 2006; final revision received January 10, 2007; accepted January 11, 2007.

From the Departments of Neurology (K.C.J.) and Public Health Sciences (K.C.J., D.P.W., X.-Q.W.), University of Virginia, Charlottesville; the Department of Neurology (G.C.N.), Albert Einstein Medical Center, New York, NY; University Hospitals Leuven (V.T.), Leuven, Belgium; the Department of Neurology (S.S.), University of North Carolina at Chapel Hill; and the National Institute of Neurological Disorders and Stroke (S.W.), Bethesda, Md.

This work was presented in part at the 30th International Stroke Conference of the American Heart Association, Kissimmee, Fla, February 16, 2006. Correspondence to Karen C. Johnston, MD, MSc, Department of Neurology, University of Virginia Health System, #800394, Charlottesville, VA 22904-0394. E-mail kj4v@virginia.edu

© 2007 American Heart Association, Inc.

Stroke is available at http://www.strokeaha.org DOI: 10.1161/STROKEAHA.106.479154
TABLE 1. Model Variables

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>DWI Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Variables</td>
<td></td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
</tr>
<tr>
<td>Initial NIHSS score</td>
<td>Time to DWI scan, h*</td>
</tr>
<tr>
<td>History of diabetes</td>
<td>Time by DWI interaction, h×cm³</td>
</tr>
<tr>
<td>Prestroke disability</td>
<td></td>
</tr>
<tr>
<td>t-PA use</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent Outcome†</td>
<td></td>
</tr>
<tr>
<td>NIHSS score = 0, 1</td>
<td>NIHSS score ≥15, death</td>
</tr>
<tr>
<td>BI score = 95, 100</td>
<td>BI score ≤60, death</td>
</tr>
<tr>
<td>mRS score = 0, 1</td>
<td>mRS score = 5, 6</td>
</tr>
</tbody>
</table>

*Time from symptom onset to DWI scan. †Excellent outcome suggests full or nearly full recovery. ‡Devastating outcome suggests severe disability or death.

Model Variables

The independent variables for the models are shown in Table 1. All variables were prespecified before the development or validation of datasets were examined, based on previous analyses that had demonstrated strong relations with full recovery and/or devastating outcome. Age, NIHSS score, DWI volume, and time were used as continuous variables. The interaction term was the cross product of time to DWI and DWI volume. All other variables were dichotomized. All variables listed were retained in the internal and external validation schemes.

The dependent variables were prespecified and included excellent outcome (reflecting full or nearly full recovery) or devastating outcome (reflecting nursing home–level disability or death) as previously defined and measured by the NIHSS, BI, or mRS14,17 (Table 1). Each patient was assessed for excellent outcome and devastating outcome on each individual scale. In some cases, a patient was categorized as having an excellent outcome on 1 scale (eg, BI) but not on another (eg, NIHSS) and similarly for devastating outcome, because the 3 scales measure different outcomes. Baseline reliability was assured by using double data entry and a review of source documents in all cases of a discrepancy.

Missing Data and Exclusions

Seven subjects from the GAIN trials were missing time to DWI. In an attempt to maximize the data and because we assumed that these data were missing at random, time to DWI was imputed. Because the median time from DWI to study therapy initiation for the GAIN subjects was 1.05 hours, the time of therapy initiation minus 1 hour was imputed for these 7 subjects.

Patients in the ASAP trial who were missing DWI volume or all 3-month outcome data were excluded from the analysis. A total of 35 of the 301 subjects were excluded from all analyses. Thirteen were missing DWI volume (9 missing scans, 2 patients too large to be scanned, 1 could not tolerate scan, and 1 scan was unreadable), 8 subjects did not have strokes (1 hyperglycemia, 2 seizures, 2 transient ischemic attacks that resolved before enrollment, 1 migraine, 2 peripheral nerve conditions), 8 subjects withdrew before outcome assessment, 5 subjects had confounding neurologic or psychiatric disease (1 previous stroke with same deficit, 2 psychiatric conditions, 1 brain metastases, 1 Parkinson disease and multiple previous strokes), and 1 was lost to follow-up (incarcerated and prohibited from in-person or telephone contact). All missing DWI data were assumed to be missing at random. In addition, 5 subjects did not have 3-month NIHSS scores, and 22 subjects had NIHSS scores estimated by phone. All 27 of these subjects were excluded from the NIHSS score model analyses.

Statistical Methods

The models were estimated on the combined GAIN/citicoline dataset by multivariable logistic regression including all prespecified independent variables to predict each of the 6 clinical outcomes to create 6 equations. The models were then reestimated without the DWI variables to create 6 more equations. Internal validation was completed in the development dataset for all 12 models according to bootstrap techniques18,19 with resampling occurring 300 times. The model equations were then frozen and used with fixed coefficients to forecast outcomes in the ASAP dataset. Model performance was assessed by using measures of discrimination and calibration. Discrimination was determined from the AUC, which measures the ability of the model to discriminate between outcomes across the entire range of predicted probabilities. AUCs vary from 0.50 for a model that correctly predicts outcome no better than chance to 1.0 for a model that perfectly discriminates between the 2 outcomes. The prespecified AUC for acceptable discrimination was ≥0.8, which has been suggested as adequate for individual prediction.19 Calibration was assessed from calibration curves that were created by calculating mean predicted and observed outcome rates by quintiles of the data. The prespecified acceptable absolute mean error for the calibration curves was ≤0.4. A likelihood ratio χ² statistic19 was used to compare the overall performance of the model with DWI versus the model without DWI for each pair in the estimation dataset. Analyses were done with SAS version 9.1 (SAS Institute, Cary, NC) or S-Plus Version 6.1 (Insightful Corp, Seattle, Wash).
studies (bootstrapping) indicated that 270 patients would provide 80% power to detect a 0.05 difference in the AUC at a type I error rate of 0.05.

Sensitivity Analyses and Assessment for Potential Bias

To assess for a possible confounding effect by different experimental treatments in different clinical trials (either gavestinel or citicoline), the logistic-regression models were reestimated by adjusting for an additional indicator of different treatments. To assess for potential bias of those who were excluded from the ASAP trial who did not return for follow-up NIHSS score but were assessed for the other outcomes (BI and mRS), we excluded against those included. To assess for potential bias due to treatment effect, the Wald $\chi^2$ test indicated that there was no overall treatment effect in the prediction of outcome. Compared with the reference group (control group), there was no evidence to indicate that the other treatment groups differed significantly, because all 95% CIs for the odds ratios included unity. The model equation information is available in supplemental Table II, available online at http://stroke.ahajournals.org.

In the assessment of potential bias due to treatment effect, the Wald $\chi^2$ test indicated that there was no overall treatment effect in the prediction of outcome. Compared with the reference group (control group), there was no evidence to indicate that the other treatment groups differed significantly, because all 95% CIs for the odds ratios included unity. The model equation information is available in supplemental Table II, available online at http://stroke.ahajournals.org.

The external validation results are shown in Table 3. Five of the 6 models with DWI data, age, NIHSS score, and DWI volume consistently contributed the most to the models ($P<0.05$). Diabetes and tissue-type plasminogen activator (t-PA) treatment contributed significantly to some of the models. All variable odds ratios and 95% CIs are available in the supplemental Table I, available online at http://stroke.ahajournals.org.

The external validation results are shown in Table 3. Five of the 6 models with DWI had an AUC $\geq 0.8$. The sixth model missed this cutoff by the narrowest of margins (AUC=0.799). The models without DWI performed less well than did those with DWI data ($P=0.01$ by the likelihood-ratio $\chi^2$ test), although 3 of the models met the prespecified criteria and a fourth missed by a very narrow margin (AUC=0.798).

The calibration curves for models with DWI are shown in the Figure. The mean absolute error for each was $\leq 0.4$, with the curves for the models with DWI more closely resembling perfect calibration. The models without DWI also had mean absolute errors $\leq 0.4$ but were less well calibrated (data not shown). The model equation information is available in supplemental Table II, available online at http://stroke.ahajournals.org.

In the assessment of potential bias due to treatment effect, the Wald $\chi^2$ test indicated that there was no overall treatment effect in the prediction of outcome. Compared with the reference group (control group), there was no evidence to indicate that the other treatment groups differed significantly, because all 95% CIs for the odds ratios included unity. The forecasts from equations that adjusted for possible effects of treatment were virtually identical to those without the adjustment variable.

Of the 35 validation dataset patients excluded from the analysis, 13 subjects did not have stroke. The 22 subjects with stroke who were excluded from analysis were compared with the analysis sample (266) to assess for bias, and no clinically or statistically relevant differences were found in demographic characteristics. Additionally, those subjects with a missing NIHSS score at 3 months were compared with those

### Table 2. Predictor and Outcome Variable Characteristics

<table>
<thead>
<tr>
<th>Predictor Variables</th>
<th>GAIN/Citicoline (Development) N=382</th>
<th>ASAP (Validation) N=266</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age (IQ range), y</td>
<td>69 (58, 77)</td>
<td>70 (58, 78)</td>
</tr>
<tr>
<td>Female sex</td>
<td>176 (46)</td>
<td>125 (47)</td>
</tr>
<tr>
<td>White race</td>
<td>299 (78)</td>
<td>203 (76)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>92 (24)</td>
<td>84 (32)</td>
</tr>
<tr>
<td>Previous disability</td>
<td>8 (2)</td>
<td>48 (18)</td>
</tr>
<tr>
<td>Median NIHSS score (IQ range)</td>
<td>14 (10, 18)</td>
<td>5 (3, 10)</td>
</tr>
<tr>
<td>Minimum/maximum</td>
<td>5, 34</td>
<td>0, 30</td>
</tr>
<tr>
<td>Median DWI volume (IQ range), cm$^3$</td>
<td>32 (10, 76)</td>
<td>2 (0.2, 15)</td>
</tr>
<tr>
<td>Minimum/maximum</td>
<td>0, 410</td>
<td>0, 182</td>
</tr>
<tr>
<td>Time to DWI (IQ range), h</td>
<td>7 (5, 15)</td>
<td>15 (10, 21)</td>
</tr>
<tr>
<td>Minimum/maximum</td>
<td>1, 83</td>
<td>0*, 47</td>
</tr>
<tr>
<td>t-PA</td>
<td>29 (8)</td>
<td>41 (15)</td>
</tr>
</tbody>
</table>

### Table 3. Validation Results: AUCs

<table>
<thead>
<tr>
<th>Model</th>
<th>Internal Validation</th>
<th>External Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIHSS</td>
<td>BI</td>
<td>mRS</td>
</tr>
<tr>
<td><strong>Excellent Outcome</strong></td>
<td><strong>Devastating Outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Models with DWI</td>
<td>0.802</td>
<td>0.803</td>
</tr>
<tr>
<td>Models without DWI</td>
<td>0.761</td>
<td>0.794</td>
</tr>
</tbody>
</table>

Results

A total of 382 subjects were included in the development (GAIN/citicoline) dataset and 266 in the validation (ASAP) dataset. Baseline characteristics for the 2 populations are shown in Table 2. The 2 datasets were comparable for baseline demographic characteristics, including age, sex, race, and comorbid disease. The 2 datasets differed substantially, however, in stroke severity as assessed by baseline NIHSS score, DWI volume, and 3-month outcome. The difference in prestroke disability reflects the restriction of level of disability at enrollment by the GAIN and citicoline
with the outcome, and no clinically or statistically relevant differences were found (data not shown).

**Discussion**

Prediction models including 5 commonly captured clinical variables plus DWI information are valid in predicting clinically relevant extreme 3-month outcomes in acute ischemic stroke patients, as demonstrated by our internal and external validation. Excellent outcome, which reflects full or nearly full recovery, is highly clinically relevant to patients and clinicians as well as in clinical research. Devastating outcome, which reflects nursing home–level disability or death, is also highly clinically relevant to all involved and generally reflects the outcome most feared by patients and loved ones. Five of the 6 models that included DWI performed at the prespecified level of success (AUC ≥0.8) for external validation, and the sixth had an external validation AUC of 0.799. The models without DWI performed less well. Only the 3 devastating-prediction models without DWI met our prespecified requirement for discrimination, but 2 of the 3 excellent-outcome models had an AUC >0.79, suggesting very acceptable performance. Our data suggest that DWI does contribute to the prediction but that the magnitude of the contribution is actually quite small and may not be clinically relevant.

Conclusions regarding the prediction value of acute DWI information in ischemic stroke models have been mixed. Baird et al validated a 3-point model including NIHSS score, DWI lesion volume, and time from stroke onset to imaging and demonstrated the added value of the imaging information. In a population of nonlacunar stroke patients, Thijs et al demonstrated that DWI volume was an independent predictor of outcome in a multivariable analysis, whereas Wardlaw et al did not find that DWI was an independent predictor of outcome in a mixed acute ischemic stroke population. Most recently, Hand and colleagues also suggested that acute DWI information did not add to the prediction beyond that of clinical variables in a broad acute ischemic stroke population.
The small sample sizes used in the Wardlaw model (total \( N = 108 \)) and the Hand model (total \( N = 82 \)) and specifically the small number of least-frequent outcomes in the models likely limited the power to identify a relation. Our larger sample may have allowed the statistically significant result but again, without clear evidence of a clinically relevant improvement in prediction. Another consideration, as suggested by Bang and colleagues, is that the lesion pattern and not the volume may be most predictive of outcome. Additionally, the differences in populations considered in the various studies may indicate a differential effect of DWI information in predicting outcome. The lack of a substantial impact of acute infarct volume is somewhat surprising. The absence of localization data with infarct volume or the variability of time to DWI may have confounded the relation identified. Additionally, NIHSS score and DWI volume may be too closely correlated. NIHSS score and CT infarct volume at 1 week are known to be correlated (Spearman correlation 0.64).\(^5\) as are baseline DWI and 24-hour NIHSS (Pearson’s correlation 0.67).\(^2\) and baseline NIHSS score and baseline DWI were correlated in both our development dataset (Spearman correlation 0.60) and our validation dataset (Spearman correlation 0.62), suggesting that the overlap between the 2 measures is likely contributing to the small magnitude of effect of DWI.

Stroke prediction models may be very clinically useful. The value of a validated prediction model may be evident to any clinician who has used the rule proposed by Levy et al.\(^2\) to predict outcome in a patient with hypoxic-ischemic coma and has become more apparent with the recent validation of individual prediction in transient ischemic attack patients.\(^23\) These models allow the physician to capture individual information about a patient and project the probability of outcome to the patient, the family, and often to involved healthcare team members. In addition, such prediction models can be useful in clinical trial design for stratification or severity-adjusted analysis and may ultimately allow a more individualized approach to care.\(^2.25\)

The strengths of our study include the large number of patients involved in both the development dataset and the validation dataset. The larger sample size allowed us to identify relations with greater power and precision. Also, the rigorous methods, including prespecified independent variables, avoidance of overmodeling, formally defined internal and external validation criteria, and the use of clinically relevant outcomes that are easily translatable to patients, families, and healthcare workers, all support our findings. In addition, our study included a multicenter, multinational development dataset, and the models were validated in a prospectively collected external dataset that included a broad stroke patient population, thereby increasing the generalizability of these models.

Limitations of this work include the fact that our numbers of devastating outcomes in the validation dataset were smaller than expected and may reflect selection bias. Although these models were validated, the calibration charts for these models demonstrate the lack of data available for a >60% likelihood (predicted or observed) of a devastating outcome. Some variability in the performance of these models in other datasets is possible. For this reason, it would be reasonable to validate the devastating models in a second validation dataset that included a much sicker acute stroke population to assess the discrimination and calibration in those highly likely to have devastating outcomes. Differences in baseline characteristics in the 2 populations likely reflect enrollment differences. Despite these patient selection differences, the models performed quite well in the validation dataset (a broader population), suggesting a potential increase in generalizability of the results. Our models also may not perform as well in the posterior circulation subset of stroke patients. Our analysis did not include localization data, and the development dataset had only a small percentage of posterior infarct patients, based on the eligibility criteria that were intended to capture large cortical infarcts. Perfusion data, growth of volume data, and other imaging data were not captured in these datasets. Each of these factors may have limited the models’ ability to predict outcome. The possibility that gavestinel and/or citicoline had an effect on outcome cannot be completely dismissed. Both primary studies were negative for treatment effect and our analysis suggested no substantial treatment effect, but we recognize that a small treatment effect on outcome could have been missed by both analyses. In addition, we examined only the interaction between time to DWI and DWI infarct volume. Numerous other interactions could have occurred. The excellent discrimination of the models, however, suggests that any interaction effect that was missed is likely to be small. Because of our rigorous criteria for model development (limiting independent variables), other potentially contributing independent variables were not included. On the basis of our previous analyses\(^4\) and the results of this current work, we suggest that any easily captured single acute stroke variable would not be likely to have a strong relation with outcome. Finally, a selection bias in the ASAP validation dataset is possible, because only 266 subjects of the 301 had complete data for analysis. The analysis for potential bias lends support to our assumption that the missing data were likely to have been missing at random and did not affect outcome. We recognize, however, that if these missing data were not missing at random, then bias could exist.

In summary, we have demonstrated and validated a strong predictive relation between 5 commonly collected clinical acute ischemic stroke patient variables with or without the addition of DWI data and 3-month outcome. Although addition of the imaging information was statistically significant, the addition of infarct volume information, time to DWI, and the interaction between the 2, on average, may not add a clinically significant improvement to the prediction. Challenges in capturing these imaging variables (eg, the interaction term), especially in the clinical setting, may limit the usefulness of these models without predeveloped software. For this reason and because the clinical models performed nearly as well, we anticipate that the models without DWI may be more useful in an acute stroke setting, in which there is no standardized infrastructure to consider the complicated relation between imaging information and 3-month outcome. These models have potential to inform both clinical and research decisions.
Acknowledgments
The authors gratefully acknowledge the contribution of the GAIN, citicoline, and ASAP trial investigators and patients, without whose efforts this work would not have been possible.

Disclosures
The GAIN trials were funded by Glaxo-Wellcome Inc, and Drs Johnston, Newman, Thijs, Sen, and Warach all received research support and/or consulting income from Glaxo-Wellcome Inc during the planning and conduct of the clinical trials. The citicoline trials were funded by Interneuron Pharmaceutical Inc, and Dr Warach received research support and consulting income from Interneuron Pharmaceutical Inc during the planning and conduct of the clinical trials. These clinical trial sponsors had no role in ASAP study design, data collection, analysis, interpretation, or writing of the report. The ASAP study was funded by the National Institutes of Health-National Institute of Neurological Disorders and Stroke (K23NS02168), and Drs Johnston and Wagner received support from this grant.

References
Validation of an Acute Ischemic Stroke Model: Does Diffusion-Weighted Imaging Lesion Volume Offer a Clinically Significant Improvement in Prediction of Outcome?
for the GAIN, Citicoline, and ASAP Investigators

Stroke. 2007;38:1820-1825; originally published online April 19, 2007;
doi: 10.1161/STROKEAHA.106.479154

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/38/6/1820