Validating Administrative Data in Stroke Research

David L. Tirschwell, MD; W.T. Longstreth, Jr, MD

Background and Purpose—Research based on administrative data has advantages, including large numbers, consistent data, and low cost. This study was designed to compare different methods of stroke classification using administrative data.

Methods—Administrative hospital discharge data and medical record review of 206 patients were used to evaluate 3 algorithms for classifying stroke patients. These algorithms were based on all (algorithm 1), the first 2 (algorithm 2), or the primary (algorithm 3) administrative discharge diagnosis code(s). The diagnoses after review of medical record data were considered the gold standard. Then, using a large administrative data set, we compared patients with a primary discharge diagnosis of stroke with patients with their stroke discharge diagnosis code in a nonprimary position.

Results—Compared with the gold standard, algorithm 1 had the highest $k$ for classifying ischemic stroke, with a sensitivity of 86%, specificity of 95%, positive predictive value of 90%, and $k=0.82$. Algorithm 3 had the highest $k$ values for intracerebral hemorrhage and subarachnoid hemorrhage. For intracerebral hemorrhage, the sensitivity was 85%, specificity was 96%, positive predictive value was 89%, and $k=0.82$. For subarachnoid hemorrhage, those values were 90%, 97%, 94%, and 0.88, respectively. Nonprimary position ischemic stroke patients had significantly greater comorbidity and 30-day mortality (odds ratio, 3.2) than primary position ischemic stroke patients.

Conclusions—Stroke classification in these administrative data were optimal using all discharge diagnoses for ischemic stroke and primary discharge diagnosis only for intracerebral and subarachnoid hemorrhage. Selecting ischemic stroke patients on the basis of primary discharge diagnosis may bias administrative samples toward more benign, unrepresentative outcomes and should be avoided. (Stroke. 2002;33:2465-2470.)

Key Words: classification ■ data collection ■ records ■ stroke

Administrative data are those data routinely collected, often for billing purposes, on all patients in a given population. Advantages to research based on administrative data include the large number of cases, consistent data across patients, and little expense necessary to obtain and analyze data.1 Disadvantages include the limited range of data elements and inaccuracies of coding.2-3

As expected, and as has been shown in other studies of administrative data and stroke, classifying patients on the basis of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes leads to a tradeoff between sensitivity and specificity3-6; similar data are unavailable for ICD-10. This study was designed to assess the validity of stroke classification using administrative data. Our objectives were (1) to see how different algorithms, using different numbers of diagnoses from hospital discharge administrative data, perform for the classification of stroke types, including ischemic stroke, intracerebral hemorrhage (ICH), and subarachnoid hemorrhage (SAH), and (2) to explore the possibility of bias being introduced into the selected stroke type samples as a result of the different algorithms.

Methods

Data Sources

The Comprehensive Hospital Abstract Reporting System (CHARS) is an administrative database that includes the discharge (or death) diagnoses, age, sex, home zip code, and other variables on all patients hospitalized in Washington State (excluding Veterans Administration hospitals). The discharge diagnoses (up to 9 per hospitalization) are coded in CHARS with ICD-9-CM codes. We obtained a subset of CHARS, which included all hospitalizations for all patients ≥20 years of age in Seattle (Wash) hospitals from 1990 through 1996 with any stroke-related discharge diagnosis (ICD-9-CM 430 to 438). In an attempt to make the study population based, non-Seattle residents were excluded by selecting only those hospitalizations for patients with a home zip code within Seattle city limits. Additionally, some Seattle residents with stroke may have been admitted to hospitals outside of Seattle.

The CHARS data were merged with Washington State death certificates using the first 2 letters of the last name, first 2 letters of the first name, and date of birth. Thus, if a hospital discharge record matched with a death record (from the time of the admission to the end of 1996), we obtained additional information, including death date. Administrative data 30-day case fatality was defined as the CHARS data reporting death at discharge if ≤30 days or, for deaths occurring after hospital discharge, a match found from the merge with death certificates with a date of death ≤30 days from the date of hospital admission.
Classification of Stroke Patients in CHARS

Strokes were identified and classified with specific ICD-9-CM discharge diagnosis codes from the CHARS database. Four previous studies that used either a registry or chart review as the gold standard demonstrated that ICD-9-CM codes 434 and 436 were the most predictive of ischemic strokes. In October 1992, a fifth digit was added to ICD-9-CM codes 433 and 434: 0=without cerebral infarction and 1=with cerebral infarction. Making use of this fifth digit furthers increases specificity and positive predictive value (PPV). Thus, ischemic strokes were identified by ICD-9-CM code 433.x1 ("x" is the fourth digit, can vary to specify a specific arterial distribution), 434 (excluding 434.x0), or 435. The assumptions made in the 4-digit era that code 433 was not an ischemic stroke and that 434 was an ischemic stroke were reasonable (data not shown) but did include some miscoding that would be corrected by use of the fifth digit.

ICD-9-CM code 435 identified transient ischemic attacks (TIAs). ICD-9-CM code 430 identified SAH; code 431 identified ICH. If only ICD-9-CM code 432 (epidural and subdural hematoma), 437 (other and ill-defined cerebrovascular disease), or 438 (late effects of cerebrovascular disease) was present, the case was classified as “not a stroke.” For each stroke type, the case was excluded (classified as not a stroke) if any “traumatic brain injury” ICD-9-CM code (800 to 804, 850 to 854) or “rehabilitation care” as the primary ICD-9-CM code (V57) was present.

We tested 3 algorithms for stroke classification. Algorithm 1 used all available discharge diagnosis codes (up to 9) to identify stroke cases; for each patient, a dichotomous variable was created to indicate the presence or absence of a valid TIA, ischemic stroke, ICH, or SAH code. The overall stroke classification thus included the categories TIA, ischemic stroke, ICH, SAH, and not a stroke. The administrative data classification began with the assignment of all categories TIA, ischemic stroke, ICH, or SAH code. The overall stroke classification thus included the categories TIA, ischemic stroke, ICH, SAH, and not a stroke. The hierarchy in these algorithms was based on the assumption that hemorrhagic strokes (ICH and SAH) are coded more accurately than ischemic strokes; thus, hemorrhagic diagnoses take precedence. Also, if both a stroke and a TIA had occurred, it seemed reasonable to assume that a stroke had occurred. Finally, cases with both ICH and SAH coded were thought, on the basis of the clinical experience of the authors, to be more likely to represent primary SAH. The hierarchy was not adjusted to optimize classification in these data. The hierarchy is not relevant when just the primary discharge diagnosis is used.

Because each patient may have had multiple hospitalizations, we hypothesized that restriction to the first hospitalization during the study interval for any given patient would improve the performance of a stroke classification algorithm and increase the proportion of incident events (defined as no evidence of a history of previous stroke after review of medical records). If a patient had been admitted to the hospital for a stroke before our study interval, the first hospitalization in our data would not represent the incident stroke.

Selection of Cases for Chart Review

A subset of the original 1990 to 1996 data set was identified for medical record review. Only patients with the ICD-9-CM codes most predictive of ischemic stroke, ICH, or SAH (using algorithm 1 as defined above) were selected for review. We deliberately over-sampled hemorrhagic stroke to ensure fairly equal numbers of the 3 main stroke types (ischemic, ICH, SAH). After these stipulations were applied, medical records were randomly selected from 4 convenient hospitals, including 2 community hospitals, 1 health maintenance organization hospital, and 2 university-associated hospitals; abstracted (by a nurse with training in the abstraction of stroke records); and reviewed and classified by a stroke neurologist (D.L.T.) who was blinded to the ICD-9-CM codes.

Statistical Analysis

The performance of the stroke classification algorithms was compared with results from chart review using standard definitions of sensitivity, specificity, and positive predictive value (PPV). For these calculations, the diagnoses of the stroke neurologist after review of the medical record data were considered the gold standard. Percentages are presented with exact binomial 95% confidence intervals (CIs). Alternatively, one might not want to assume that a diagnostic gold standard is present; thus, agreement between administrative stroke classification and reviewing stroke neurologist diagnosis was assessed with unweighted \( \kappa \) statistics. The estimates of \( \kappa \), sensitivity, specificity, and PPV for each of the stroke types were based on a 2×2 table (administrative diagnosis of stroke type yes/no versus chart review diagnosis (gold standard) stroke type yes/no), whereas the overall \( \kappa \) estimates were based on a 5×5 table (administrative diagnosis axis and a chart review diagnosis axis, each with the 5 categories of ischemic stroke, ICH, SAH, TIA, and not a stroke).

Using the first admission for each patient during the study interval and just the administrative data, we compared patients with their stroke diagnosis codes in the primary position with patients whose codes were in a nonprimary position. Nonparametric tests were used to compare age (Wilcoxon rank sum) and sex (\( \chi^2 \)). Using administrative 30-day case fatality as the outcome of interest, we compared patients with primary position and those with nonprimary position stroke discharge diagnosis codes. Unadjusted risk estimates are presented as both relative risks (RRs) and odds ratios (ORs), followed by 95% CIs; the ORs are presented in addition to the preferred RRs to facilitate comparison with multivariate-adjusted ORs derived from logistic regression models. Logistic regression was used to control for age, sex, and multiple comorbid discharge diagnoses.

All analyses were performed with the STATA statistical program (version 6, Stata Corp). The Human Subjects Review Committee of the University of Washington and the Washington State Department of Health approved this study.

Results

The original hospital discharge administrative data set contained 20,703 patient discharges, each having a stroke-related discharge diagnosis (ICD-9-CM 430 to 438). Of these, 13,210 (63.8%) were first admissions for any given patient. We reviewed 206 medical records for this validation study. Characteristics of each data set are shown in Table 1. Restricting from all hospitalizations for each patient (n=206) to the first hospital admission (n=147) raised the proportion of incident strokes from 77% to 88%. The stroke type-specific percentages of incident cases in first admissions were 76% for ischemic stroke, 89% for ICH, and 96% for SAH. This higher proportion of incident cases was the rationale for using just the first admission for each patient in the subsequent analyses.

Merging Hospital Discharge Data and Death Certificates

Of the 34 of 206 patients for whom medical record review data indicated a death during hospitalization, all 34 were coded as dead on discharge in the CHARS data, and 31 of 34 matched with death certificate data. One patient was discharged alive with an incorrect CHARS code for death at discharge. Three patients matched death certificates dated before admission, indicating erroneous matches. Therefore,
with the medical record review as the gold standard for in-hospital death identification, using CHARS data to identify in-hospital deaths had a sensitivity of 34 of 34/100 = 100%, specificity of 171 of 172/100 = 99%, and PPV of 34 of 35/100 = 97%; these parameters reflect the matching for deaths occurring during hospitalization (73% of 30-day case fatalities). Again, with medical record review as the gold standard, using death certificate data to identify in-hospital deaths had a sensitivity of 31 of 34/100 = 91%, specificity of 169 of 172/100 = 98%, and PPV of 31 of 34/100 = 91%; these parameters likely reflect the matching for deaths occurring after discharge (27% of 30-day case fatalities). For deaths that occurred after hospital discharge, there was no way to directly validate the matching with these data.

### Performance of Algorithms for Stroke Classification

Three different algorithms, as described in Methods, were used for stroke type classification of first admissions for both the administrative data and the validation subset. The performance characteristics of each of the 3 algorithms based on 147 first admission medical records reviewed are shown in Table 2. Ischemic stroke classification agreement with the gold standard agreement was maximal with algorithm 1

### TABLE 1. Characteristics of Original Hospital Discharge Administrative Data Set, the Subset With Only First Admissions for Any Given Patient, and the Validation Subset Whose Medical Records Were Reviewed

<table>
<thead>
<tr>
<th></th>
<th>Original Data Set (n=20 703)</th>
<th>Original Data Set First Admissions* Only (n=13 210)</th>
<th>Original Data Set Validation Subset (n=206)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (SD), y</td>
<td>74 (13)</td>
<td>74 (13)</td>
<td>68 (17)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>11 244 (54)</td>
<td>7225 (55)</td>
<td>116 (56)</td>
</tr>
<tr>
<td>Men</td>
<td>9459 (46)</td>
<td>5985 (45)</td>
<td>90 (44)</td>
</tr>
<tr>
<td>Stroke classification,† n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIA</td>
<td>1976 (9.5)</td>
<td>1577 (11.9)</td>
<td>8 (3.9)</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>7360 (35.6)</td>
<td>5992 (45.4)</td>
<td>76 (36.9)</td>
</tr>
<tr>
<td>ICH</td>
<td>1107 (5.4)</td>
<td>949 (7.2)</td>
<td>47 (22.8)</td>
</tr>
<tr>
<td>SAH</td>
<td>390 (1.9)</td>
<td>348 (2.6)</td>
<td>58 (28.1)</td>
</tr>
<tr>
<td>Not a stroke</td>
<td>9870 (47.7)</td>
<td>4344 (32.9)</td>
<td>17 (8.3)</td>
</tr>
</tbody>
</table>

*First admissions during the study interval; a patient might have been admitted to the hospital for stroke before the study interval.

†Based on algorithm 1 as described in Methods in the original data set, first admissions data set, and chart review diagnosis (gold standard) in the validation data set.

### TABLE 2. Sensitivity, Specificity, PPV, and κ for Administrative Classification of Stroke Type Compared With Chart Review Diagnosis (Gold Standard) in 147 First Admission Hospitalizations

<table>
<thead>
<tr>
<th>Algorithm 1: using all discharge diagnoses</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>κ* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ischemic (n=50)†</td>
<td>86 (73–94)</td>
<td>95 (88–98)</td>
<td>90 (77–97)</td>
<td>0.82 (0.65–0.98)</td>
</tr>
<tr>
<td>ICH (n=39)†</td>
<td>82 (66–92)</td>
<td>93 (86–97)</td>
<td>80 (64–91)</td>
<td>0.74 (0.58–0.90)</td>
</tr>
<tr>
<td>SAH (n=51)†</td>
<td>98 (90–100)</td>
<td>92 (84–96)</td>
<td>86 (75–94)</td>
<td>0.87 (0.71–1.0)</td>
</tr>
<tr>
<td>Overall stroke classification κ*</td>
<td>0.79 (0.68–0.90)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm 2: using first 2 discharge diagnoses</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>κ* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ischemic (n=50)†</td>
<td>80 (66–90)</td>
<td>96 (90–99)</td>
<td>91 (78–97)</td>
<td>0.78 (0.62–0.94)</td>
</tr>
<tr>
<td>ICH (n=39)†</td>
<td>85 (69–94)</td>
<td>94 (87–97)</td>
<td>83 (67–93)</td>
<td>0.77 (0.61–0.94)</td>
</tr>
<tr>
<td>SAH (n=51)†</td>
<td>94 (84–99)</td>
<td>94 (87–98)</td>
<td>89 (77–96)</td>
<td>0.87 (0.71–1.0)</td>
</tr>
<tr>
<td>Overall stroke classification κ*</td>
<td>0.77 (0.66–0.87)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm 3: using only primary discharge diagnosis</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>κ* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ischemic (n=50)†</td>
<td>74 (60–85)</td>
<td>95 (88–98)</td>
<td>88 (74–96)</td>
<td>0.72 (0.56–0.88)</td>
</tr>
<tr>
<td>ICH (n=39)†</td>
<td>85 (69–94)</td>
<td>96 (91–99)</td>
<td>89 (75–97)</td>
<td>0.82 (0.66–0.98)</td>
</tr>
<tr>
<td>SAH (n=51)†</td>
<td>90 (79–97)</td>
<td>97 (91–99)</td>
<td>94 (83–99)</td>
<td>0.88 (0.72–1.0)</td>
</tr>
<tr>
<td>Overall stroke classification κ*</td>
<td>0.74 (0.64–0.84)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*κ interpreted as follows: 0.61–0.80 = substantial agreement; 0.81–1.00 = almost perfect agreement.  
†n is the number of chart review diagnoses of each stroke type in first admissions.
ICH and SAH classification agreement was maximal with algorithm 3 (κ=0.82 for ICH and 0.88 for SAH). Overall stroke classification agreement of the administrative data with reviewing stroke neurologist diagnosis was maximal with algorithm 1 (κ=0.79).

Comparing Patients With Primary Position Versus Nonprimary Position Stroke Discharge Diagnoses in the Large Administrative Data Set

There were 993 patients in the first admissions administrative data set with their stroke discharge diagnosis code in a nonprimary position; the proportions by stroke type were as follows: ischemic stroke, 845 of 5992 (14%); ICH, 111 of 949 (12%); and SAH, 37 of 348 (11%) (denominators for each stroke type from Table 1, middle). These patients were compared with patients with a primary discharge diagnosis of stroke; the results are shown in Table 3. Sex proportions did not differ between nonprimary position and primary position stroke patients. Age did not differ between nonprimary position and primary position ischemic stroke patients (median, 77 years) and SAH patients (median, 59 years) but was significantly greater for ICH primary position versus nonprimary position patients (median, 73.3 versus 68.2 years; P=0.004). Nonprimary position ischemic stroke patients were significantly more likely than primary position ischemic stroke patients to have died by 30 days after their hospital admission (30% versus 13% in primary position patients; RR=2.4; 95% CI, 2.1 to 2.7). When age and sex were controlled for, having an ischemic stroke diagnosis code in a nonprimary position versus primary position was associated with an OR of 3.2 (95% CI, 2.7 to 3.8) for 30-day case-fatality; this was reduced to an OR of 2.4 (95% CI, 1.7 to 2.5) after controlling further for comorbid discharge diagnoses (see footnotes in Table 3). Using logistic regression to control for age and sex, having a nonprimary position ICH diagnosis code was associated with an OR of 1.6 (95% CI, 1.0 to 2.4) for 30-day case-fatality; this was reduced to an OR of 1.0 (95% CI, 0.6 to 1.5) after controlling for comorbid discharge diagnoses (see footnotes in Table 3). Compared

| TABLE 3. Comparison of First Admissions for Patients With Primary Position Stroke Discharge Diagnoses With Those With Nonprimary Position Stroke Discharges |
|-----------------|-----------------|-----------------|
|                  | Ischemic Stroke | ICH             | SAH             |
| Nonprimary position, n (%) | 845 (14.1) | 111 (11.7) | 37 (10.6) |
| Primary position, n (%)     | 5147 (85.9)    | 838 (88.3)     | 311 (89.4)    |
| Total, n                   | 5992           | 949            | 348            |

Primary discharge diagnoses present in >3% of those patients with nonprimary position stroke diagnoses, %

- Acute myocardial infarction: 11.0 for ischemic stroke, 15.3 for ICH, and ...;
- Pneumonia*: 8.8 for ischemic stroke, 3.6 for ICH, and ...
- Cardiac arrhythmias†: 5.6 for ischemic stroke, ...
- Heart failure: 5.3 for ischemic stroke, ...
- Coronary artery disease: 4.5 for ischemic stroke, ...
- Hip fracture: 4.0 for ischemic stroke, ...
- Malignant brain tumor‡: ...
- Other intracranial hemorrhage: ...
- Other cerebrovascular disease: ...
- Other aneurysm: ...

30-d Case fatality

- Nonprimary position, %: 30.0 for ischemic stroke, 47.8 for ICH, 35.1 for SAH.
- Primary position, %: 12.6 for ischemic stroke, 39.4 for ICH, 31.2 for SAH.

Values >1 imply increased risk of 30-d case fatality in patients with nonprimary position stroke discharge diagnoses vs patients with primary position diagnoses.

*Including atrial fibrillation.
†Primary or secondary (metastatic).
‡Primary or secondary (metastatic).
with primary position SAH patients, nonprimary position SAH patients were not associated with increased risk of 30-day case-fatality (Table 3).

Discussion

This validation study demonstrates that stroke patients (ischemic, ICH, and SAH) can be identified from administrative data. The algorithm that maximized sensitivity and overall classification agreement corrected for chance ($\kappa$) used all discharge diagnoses (algorithm 1, Table 2). The algorithm that most consistently maximized specificity and PPV was the one based on the primary diagnosis only (algorithm 3, Table 2). These findings were not consistent across stroke subtypes.

With the diagnosis from review of medical record data used for comparison, each stroke type classification was nearly perfect ($\kappa>0.8$) using at least 1 of the 3 prespecified algorithms (Table 2). Ischemic stroke showed best agreement with algorithm 1, for which $\kappa=0.82$. Sensitivity was maximal with algorithm 1 and then steadily decreased with algorithms 2 and 3 (sensitivity, 86%, 80%, and 74%, respectively). At the same time, specificity (95%, 96%, and 95%) and PPV (90%, 91%, and 88%) remained stable. Another argument supporting algorithm 1 as the best classification for ischemic stroke comes from the analyses on the large administrative data set comparing patients with the ischemic stroke discharge diagnosis code in the primary position with those with the diagnosis in a nonprimary position. The nonprimary position patients have a larger comorbidity burden and an elevated risk of 30-day case-fatality (age- and sex-adjusted OR, 3.2). That the increased comorbidity is responsible for much of this increased mortality is suggested by a reduction in OR to 2.0 after adjustment for comorbid discharge diagnoses (Table 3). If one were to use only the patients with an appropriate ICD-9-CM ischemic stroke code in the primary discharge diagnosis position, patients with higher levels of comorbidity and related short-term mortality would be excluded, and estimates related to outcomes in these patients would be overly optimistic. Previous reports using ICD-9-CM discharge data in ischemic stroke have suggested using only the primary discharge diagnosis as a way to optimize the PPV$^3$ or have empirically done so,$^2,^7,^8$ and thus any outcomes reported may have been optimistic.

ICH diagnoses showed best agreement with diagnosis from review of medical record data with algorithm 3, for which $\kappa=0.82$. The PPV steadily increased from 80% to 83% to 89% going from algorithms 1 to 3, whereas there were lesser gains in sensitivity (82%, 85%, and 85%, respectively) and specificity (93%, 94%, and 96%, respectively) (Table 2). Nonprimary position ICH patients were associated with an age- and sex-adjusted OR of 1.6 for 30-day case-fatality; this excess risk disappears (OR=1.0) with adjustment for comorbid discharge diagnoses (Table 3). Much of the increased risk in those removed patients was due to the comorbid, primary discharge diagnoses of malignant brain tumor and acute myocardial infarction. Malignant brain tumor (either primary or secondary) was the primary discharge diagnosis in 16% of ICH patients. Acute myocardial infarction, with associated thrombolytic therapy the likely cause of the ICH, was the primary discharge diagnosis in 15% of ICH patients. In an effort to identify a “pure” ICH population, these diagnoses might be excluded; the estimated RR of 30-day case fatality in nonprimary versus primary position ICH patients then decreases from an OR of 1.6 with them included to a nonsignificant OR of 1.2 with them excluded.

SAH showed nearly perfect agreement with diagnosis from review of medical record data using all 3 algorithms; sensitivity was maximal with algorithm 1 compared with specificity, PPV, and $\kappa$ (just barely) being maximal with algorithm 3 ($\kappa=0.88$) (Table 2). There was no difference in the risk of 30-day case fatality for patients with primary position SAH diagnosis codes compared with those with nonprimary position diagnoses; therefore, excluding those patients would not bias estimates of this outcome. The 30-day case-fatality rate for primary position SAH patients was 31.2%, a figure similar to the 32% reported from a population-based study performed in a similar geographic region in a just-earlier time frame (1987 to 1989).$^9$ This high level of diagnostic agreement ($\kappa$) and the similarity in 30-day case-fatality rates raise the possibility that, at least for SAH, administrative data may be sufficient for descriptive epidemiological research.

The study has a number of potential limitations. The medical records reviewed were chosen from high-probability stroke ICD-9-CM codes; therefore, true stroke patients coded with reportedly less predictive ICD-9-CM codes (ie, 432, 435, 437, or 438) are likely to have been missed, and thus our estimates of sensitivity are optimistic. By excluding most patients who did not have a stroke (those without ICD-9-CM code 430 to 438), we may have made pessimistic estimates of specificity. Avoiding the less predictive ICD-9-CM codes should not affect the estimates of PPV. The number of medical records abstracted (n=206) was small compared with the number of patients in the original administrative data set (n=20703). Moreover, records were reviewed only from a subset of all possible hospitals, making the records reviewed not population based, a convenient rather than a random sample and possibly an unrepresentative sample. If the hospitals sampled did not code discharge diagnoses like the rest of the hospitals in Seattle or throughout the United States, our conclusions might not be generalizable; we have no reason to believe this was the case. The time frame of the validation study was 1990 to 1996, and although these results are directly applicable to our research with the same data set, extrapolation of these results to other time periods and other administrative data sets remains uncertain.

The performance of these algorithms for classifying stroke patients in administrative data is based on the premise that the reviewing stroke neurologist’s diagnosis (heavily influenced by the diagnosis of the treating physician) represents the gold standard—an assumption that may not be completely accurate. If we assume that errors in this gold standard may exist, then the $\kappa$ statistics, used to compare agreement between 2 measures, neither of which is superior, are of value. The medical record sampling scheme used, not reviewing charts for which ICD-9-CM codes (eg, 433 or 435) have been shown to represent few true strokes,$^3,^6$ may have led to optimistic $\kappa$ estimates. The amount of bias introduced would depend on the proportion of disagreement, and if the reviewing stroke
neurologist’s diagnoses were also largely “not a stroke,” it is possible that the bias introduced might be quite small. Also, the κ estimates of overall stroke classification agreement between administrative and medical record sources are artificially heavily weighted by the agreement within the hemorrhagic stroke types. We oversampled these stroke types to have a reasonable number of patients to evaluate. The stroke type-specific κ estimates may therefore be more accurate reflections of the agreement.

In conclusion, patients with stroke can be identified from administrative data. The best algorithm to use to classify the administratively identified potential stroke patients differed by stroke subtype and may differ depending on the research question of interest. For ischemic stroke, the algorithm that considered all possible discharge diagnoses seemed to maximize stroke classification agreement (κ) and may give less biased estimates of 30-day survival. In ICH and SAH, agreement was maximal with the algorithm that used only the primary discharge diagnosis, and these samples seemed unbiased. There are limitations to the conclusions that can be drawn from research results based on administrative data sets. That being said, after careful examination and validation of such sources, the advantages of research based on such sources (large numbers of population-based patients, little cost in obtaining data) justify its continued pursuit. Such research may typically be used to find preliminary answers to complex questions and may help make empirically based decisions about whether such questions are worthy of more labor-intensive and costly additional research. Efforts to improve the validity and increase the breadth of information available in these administrative data sources would allow them to serve as even better sources of hypothesis generation and testing. This inexpensive approach to research may then become increasingly more useful.

Acknowledgments

Dr. Tirschwell’s efforts were supported in part by NINDS grant 1 K23 NS02119-01 and by the Medic One Research Foundation, Seattle, Wash.

References

Validating Administrative Data in Stroke Research
David L. Tirschwell and W.T. Longstreth Jr

Stroke. 2002;33:2465-2470
doi: 10.1161/01.STR.00003240.28636.BD

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/33/10/2465

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/