Incidence and Occurrence of Total 
(First-Ever and Recurrent) Stroke

G. Rhys Williams, MS; John G. Jiang, PhD; David B. Matchar, MD; Gregory P. Samsa, PhD

Background and Purpose—It has recently been hypothesized that the figure of approximately half a million strokes substantially underestimates the actual annual stroke burden for the United States. The majority of previously reported studies on the epidemiology of stroke used relatively small and homogeneous population-based stroke registries. This study was designed to estimate the occurrence, incidence, and characteristics of total (first-ever and recurrent) stroke by using a large administrative claims database representative of all 1995 US inpatient discharges.

Methods—We used the Nationwide Inpatient Sample of the Healthcare Cost and Utilization Project, release 4, which contains approximately 20% of all 1995 US inpatient discharges. Because the accuracy of International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) coding is suboptimal, we performed a literature review of ICD-9-CM 430 to 438 validation studies. The pooled results from the literature review were used to make appropriate adjustments in the analysis to correct for some of the inaccuracies of the diagnostic codes.

Results—There were 682 000 occurrences of stroke with hospitalization (95% CI 660 000 to 704 000) and an estimated 68 000 occurrences of stroke without hospitalization. The overall incidence rate for occurrence of total stroke (first-ever and recurrent) was 259 per 100 000 population (age- and sex-adjusted to 1995 US population). Incidence rates increased exponentially with age and were consistently higher for males than for females.

Conclusions—We conservatively estimate that there were 750 000 first-ever or recurrent strokes in the United States during 1995. This new figure emphasizes the importance of preventive measures for a disease that has identifiable and modifiable risk factors and for the development of new and improved treatment strategies and infrastructures that can reduce the consequences of stroke. (Stroke. 1999;30:2523-2528.)

Key Words: diagnosis ■ epidemiology ■ hospitalization ■ incidence ■ stroke, acute

Stroke is the third leading cause of death in the United States, after heart disease and cancer.1 In 1994, Matchar and Duncan2 claimed that each year Americans suffer approximately 550 000 strokes, causing 150 000 deaths and leaving 300 000 survivors disabled. The Heart and Stroke Statistical Update3 of the American Heart Association (1995) states that approximately 500 000 Americans suffer a first-ever or recurrent stroke each year. Both these reports are based on the predominately white cohort study of Framingham, Mass.

Broderick et al4 recently hypothesized that the figure of approximately half a million strokes substantially underestimates the actual annual stroke burden for the United States. They claimed that there were at least 731 000 first-ever or recurrent strokes during 1996.4 This estimate was derived by extrapolating from first-ever strokes among whites in the Rochester, Minnesota Stroke Study. The extrapolation required 2 steps. The first step inflated the baseline to reflect the burden of recurrent stroke. The second step accounts for the higher rate of stroke among African Americans observed in the Greater Cincinnati/Northern Kentucky Stroke Study.4

Population-based stroke incidence studies such as those from Framingham, Mass, and Rochester, Minn, have substantially increased the knowledge about stroke trends, subtypes, risk factors, and incidence rates in men and women. However, these studies were conducted among predominately white populations. Recently, epidemiological studies are focusing on differences in stroke incidence between racial/ethnic groups. Of particular interest are rates for blacks, and few data regarding stroke risk in Hispanics or Asians have been available. Recent data from Northern Manhattan suggest that blacks are not alone in the higher risk category and that Hispanics also appear to be at greater risk than whites.5

The computer revolution has advanced the state of inpatient information databases. These administrative databases are becoming increasingly more important sources of information for epidemiological studies. In addition, the Interna-
ational Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) coding scheme provides an accessible method of identifying certain patients with specific diagnoses within administrative databases. However, as with any administrative database analysis, the interpretation of the results should reflect the varied accuracy of ICD-9-CM diagnosis codes.

Several recent articles have examined stroke incidence rates stratified by subtype, age, sex, and race in great detail. Therefore, the aim of the present study was to estimate the 1995 occurrence of total stroke and the total stroke incidence rate and to examine the characteristics of the stroke population in the United States. This was accomplished by use of a large administrative claims database representing a 20% representative sample of all 1995 US inpatient discharges. The administrative database was supplemented by a literature review of studies reporting the accuracy of ICD-9-CM coding for cerebrovascular disease. This allowed us to correct for some of the inaccuracies in the administrative database.

Subjects and Methods

Nationwide Inpatient Sample Database

The present study used the Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP-3), release 4, which covers calendar year 1995. The fourth release of the NIS contains 6.7 million discharges from a sample of 938 hospitals covering 19 geographically dispersed states. These data represent a 20% stratified sample of all US inpatient discharges. Stratification variables included region, control, location, teaching status, and bed size. The software program SUDAAN was used to convert raw counts generated from the NIS database into weighted counts that represent national estimates.

Inpatient records included clinical and resource use information typically available from discharge abstracts. The NIS database includes most commonly used data elements: patient demographics, admission source, principal and secondary diagnoses and procedures, expected primary source of payment, discharge status, hospital and discharge weights, length of stay, and total charges.

Accuracy of ICD-9-CM 430 to 438 Coding

To analyze a medical condition like stroke, the NIS database, like other large administrative databases, used a standardized classification scheme, ICD-9-CM. However, classification systems such as the ICD-9-CM have a number of potential limitations. For example, their diagnostic codes may not fully encompass the condition under study, or the condition may be distributed across multiple diagnostic codes. Cerebrovascular disease constitutes a broad group of disorders with a common etiology but diverse manifestations.

Specific cerebrovascular conditions may fall under any of several ICD-9-CM disease codes 430 to 438, each of which may describe multiple clinical symptoms. Broderick et al have reported that using primary and all secondary ICD-9-CM codes 430 to 436 will provide nearly complete ascertainment of the occurrence of stroke with hospitalization, referred to as hospital strokes (97%), but will substantially overestimate the number by a factor of 2. Using only primary discharge codes would decrease the false-positive rate but would also decrease the ascertainment of stroke from 97% to 84%. Several studies have recommended using only the primary diagnosis, but there is limited published data documenting the sensitivity and specificity of this strategy. Also, if this approach was used, the analysis would be further complicated by having to estimate the false-negative rate (the percentage of hospital strokes not identified) as well.

Four studies have reported that the use of primary and secondary codes 430 to 438 provide virtually complete ascertainment of all hospital strokes, but the positive predictive value for these specific codes is not optimal. Three of these studies used all the secondary codes, whereas Leibson et al used only up to 5 secondary positions. These 4 studies report positive predictive values for ICD-9-CM codes 430 to 438 that range from 0% to 100%. Fortunately, the positive predictive values for each individual code are remarkably consistent. Therefore, with a reasonable amount of confidence, we can account for this overascertainment by making an appropriate correction in the analysis for each code. We also assumed that using both primary and secondary codes identified all hospital strokes; thus, the false-negative rate would be zero.

Data Analysis

We used only the first code reported to avoid double counting those patients with more than one reported ICD-9-CM code 430 to 438 at discharge. To obtain the estimated number of stroke cases by ICD-9-CM code, we multiplied the number of patients with each ICD-9-CM code by its estimated positive predictive value (PPV) for stroke. To obtain the total number of hospital strokes, we summed across codes. A 95% CI was computed by using Monte Carlo simulation techniques (i.e., the PPV distribution for each code was simulated by using 10 000 iterations of a binomial distribution whose parameters were obtained by pooling data from the 4 literature-based studies).

Technically, an incidence rate should include only the first episode of the disease being studied. But, because this database does not capture strokes that occurred before 1995, total stroke (first-ever and recurrent) incidence rates and occurrences were reported. Stroke (first-ever) incidence rates were estimated by reducing the total stroke rates by the expected number of recurrent strokes. The limited data from population-based cohorts suggest that 25% to 35% of strokes are recurrent. Age- and sex-standardized and stratified incidence rates of total stroke were also estimated. These were estimated by using the 1995 US census population figures.

Nonhospital Stroke

Even though ICD-9-CM primary and secondary discharge codes 430 to 438 provide virtually complete ascertainment of hospital strokes, not all stroke patients are hospitalized. Two population-based stroke incidence studies reported that the proportion of patients with stroke without hospitalization (nonhospital strokes) was 5% and 15%, respectively. Factors such as standards of medical treatment, health care systems, population size, and socioeconomic status may influence the proportion of hospital strokes. Also, nonhospital-stroke patients tend to have milder strokes and lower socioeconomic status. However, some patients who were not hospitalized may have had very severe strokes and may have died before they could be hospitalized.

Therefore, to estimate the total number of strokes (hospital and nonhospital), an appropriate adjustment was made in the analysis. This adjustment conservatively assumed that 10% of strokes are nonhospital strokes. This was a simple mean of the proportions published in 2 stroke studies (5% and 15%). Several international studies of westernized countries have reported proportions ranging from 10% to 30%. The 2 published US proportions (5% and 15%) were also used in a sensitivity analysis.

Results

Table 1 summarizes the PPVs of stroke for ICD-9-CM primary and secondary discharge codes 430 to 438. The results of these 4 studies were combined to produce a more precise estimate of the PPV for each code. These estimates were used to approximate the number of strokes correctly identified by each code.

The NIS database contained 377 544 patient discharges with a primary or secondary (first code reported) diagnosis of cerebrovascular disease. This was extrapolated to 1 977 794 patient discharges in the United States. Table 2 uses the
code-specific pooled PPVs to adjust for false positives and reports that there were 682,000 (95% CI 660,000 to 704,000) hospital strokes in the United States. We believe that at least 68,000 (10%) additional strokes were nonhospital strokes. We thus estimate that there were 750,000 incident and recurrent strokes in the United States in 1995. The sensitivity analysis on the nonhospital stroke rate produced a range from 716,000 to 784,000. Of the estimated 682,000 hospital strokes, 23,400 (3.4%) were subarachnoid hemorrhages, 71,600 (10.5%) were intracerebral hemorrhages, and the remaining 587,000 (86.1%) were ischemic strokes.

The overall incidence rate for total stroke (first-ever and recurrent) was 259 per 100,000 population (age- and sex-adjusted to the 1995 US population). The average annual age- and sex-adjusted incidence rate for first-ever stroke was estimated to be 200 per 100,000. Total stroke incidence rates increased exponentially with age for both men and women. Overall, men had greater age-specific total stroke incidence rates than did women. The incident total stroke rates presented in the Figure are similar to the first-ever stroke incidence rates published by Broderick et al.4 and Brown et al.17 for the 4 youngest groups (aged 25 to 34, 35 to 44, 45 to 54, and 55 to 64 years). But for the older groups (aged 65 to 74, 75 to 84, and 85+ years), the incident total stroke rates were 1.5-, 2-, and 3-fold higher, respectively, than the published first-ever stroke incidence rates. This suggests that although on average 25% to 35% of strokes are recurrent,14,15 for groups aged 75 to 84 and 85+ years, as much as 50% to 70% may be recurrent. This reflects the clinical observation that the older the patient, the more likely they are to have had a recurrent stroke.

Table 3 reports demographic and clinical characteristics for the stroke patients by ICD-9-CM code. The mean age of all patients with stroke was 72.1 years, 45% were male, 80% were white, and 12% of the patients died during the hospitalization. A disproportionately low percentage of patients with subarachnoid hemorrhage were male (37% versus 45%), and as expected, the patients with subarachnoid hemorrhage were much younger than the average stroke patient (56.6 versus 72.1 years). The inpatient mortality rates for subarachnoid and intracerebral hemorrhages were significantly higher than for the overall stroke population, at 26% and 29%, respectively. A surprisingly high percentage of ICD-9-CM code 432 patients were male (61% versus 39%), and most code 433 patients were white (92.5%).

Table 3 also presents data on resource utilization. For the index hospitalization, the stroke patient population had a mean length of stay of 9.8 days (median 5 days) and a mean total charge of $17,711 (median $8,735). Patients with subarachnoid and intracerebral hemorrhages had much longer

### Table 1. PPV of ICD-9-CM Primary or Secondary Codes for Stroke (First Code Reported)

<table>
<thead>
<tr>
<th>ICD-9-CM Code</th>
<th>Leibson et al. (Up to five codes)</th>
<th>Benesch et al. (Ischemic Stroke)</th>
<th>Broderick et al. (Black Patients)</th>
<th>Rosamond et al. (Patients Aged 45–64 y)</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>430</td>
<td>11/11 (100%)</td>
<td>...</td>
<td>9/14 (64%)</td>
<td>19/22 (86%)</td>
<td>39/47 (83%)</td>
</tr>
<tr>
<td>431</td>
<td>20/23 (87%)</td>
<td>...</td>
<td>25/30 (83%)</td>
<td>41/48 (83%)</td>
<td>86/101 (85%)</td>
</tr>
<tr>
<td>432</td>
<td>0/7 (0%)</td>
<td>...</td>
<td>0/6 (0%)</td>
<td>1/11 (9%)</td>
<td>1/24 (4%)</td>
</tr>
<tr>
<td>433</td>
<td>3/20 (15%)</td>
<td>43/295 (15%)</td>
<td>9/26 (35%)</td>
<td>36/266 (14%)</td>
<td>91/607 (15%)</td>
</tr>
<tr>
<td>434</td>
<td>87/102 (85%)</td>
<td>192/226 (85%)</td>
<td>150/187 (80%)</td>
<td>144/186 (77%)</td>
<td>573/701 (82%)</td>
</tr>
<tr>
<td>435</td>
<td>12/78 (15%)</td>
<td>9/99 (9%)</td>
<td>8/69 (12%)</td>
<td>14/105 (12%)</td>
<td>43/351 (12%)</td>
</tr>
<tr>
<td>436</td>
<td>55/64 (86%)</td>
<td>24/29 (83%)</td>
<td>75/117 (64%)</td>
<td>76/108 (70%)</td>
<td>230/318 (72%)</td>
</tr>
<tr>
<td>437</td>
<td>2/9 (22%)</td>
<td>...</td>
<td>8/66 (12%)</td>
<td>1/59 (2%)</td>
<td>11/134 (8%)</td>
</tr>
<tr>
<td>438</td>
<td>0/15 (0%)</td>
<td>...</td>
<td>4/218 (2%)</td>
<td>1/127 (1%)</td>
<td>5/360 (1%)</td>
</tr>
</tbody>
</table>

Data for year and geographical areas of study are as follows: Leibson et al., 8 1994 in Rochester, Minn; Benesch et al., 9 1997 in Calif, Iowa, La, Mass, and Pa; Broderick et al., 4 1998 in Ohio and Ky; and Rosamond et al., 6 1999 in Miss, Minn, Md, and NC.

### Table 2. US Estimates of Number of Hospitalized Strokes in 1995 by ICD-9-CM Code for Primary or Secondary Diagnoses (First Code Reported, No Double Counting)

<table>
<thead>
<tr>
<th>ICD-9-CM Code</th>
<th>Raw Count</th>
<th>Weighted Count</th>
<th>Estimated PPV</th>
<th>US Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>430: Subarachnoid hemorrhage</td>
<td>5580</td>
<td>28 187</td>
<td>39/47 (83%)</td>
<td>23 389</td>
</tr>
<tr>
<td>431: Intracerebral hemorrhage</td>
<td>16 296</td>
<td>84 067</td>
<td>86/101 (85%)</td>
<td>71 582</td>
</tr>
<tr>
<td>432: Other and unspecified intracranial hemorrhage</td>
<td>4833</td>
<td>24 732</td>
<td>1/24 (4%)</td>
<td>1031</td>
</tr>
<tr>
<td>433: Occlusion and stenosis of precerebral arteries</td>
<td>66 001</td>
<td>342 662</td>
<td>91/607 (15%)</td>
<td>51 371</td>
</tr>
<tr>
<td>434: Occlusion of cerebral arteries</td>
<td>83 279</td>
<td>432 458</td>
<td>573/701 (82%)</td>
<td>353 493</td>
</tr>
<tr>
<td>435: Transient cerebral ischemia</td>
<td>45 722</td>
<td>241 255</td>
<td>43/351 (12%)</td>
<td>29 555</td>
</tr>
<tr>
<td>436: Acute but ill-defined cerebrovascular disease</td>
<td>34 065</td>
<td>184 272</td>
<td>230/318 (72%)</td>
<td>133 278</td>
</tr>
<tr>
<td>437: Other and ill-defined cerebrovascular disease</td>
<td>26 273</td>
<td>137 992</td>
<td>11/134 (8%)</td>
<td>11 328</td>
</tr>
<tr>
<td>438: Late effects of cerebrovascular disease</td>
<td>95 495</td>
<td>502 168</td>
<td>5/360 (1%)</td>
<td>6975</td>
</tr>
</tbody>
</table>

Estimated number of hospital strokes in 1995 ... ... ... 682 002
length of stays, averaging 14.0 and 10.5 days, respectively. Their mean total charges were also higher, at $46,711 and $23,097, respectively. The majority of stroke patients had a routine discharge or were discharged to a skilled nursing facility with rates of 39% and 17%, respectively.

Discussion
The epidemiology of total stroke, ie, first-ever and recurrent stroke, has been inadequately studied. We conservatively estimate that there were 750,000 first-ever or recurrent strokes in 1995. The Heart and Stroke Statistical Update of the American Heart Association, 1995, states that ≈500,000 Americans suffer a first-ever or recurrent stroke each year. However, this statement was based on the predominately white cohort study of Framingham, Mass. Broderick et al recently hypothesized that the figure of approximately half a million strokes substantially underestimates the actual annual stroke burden for the United States; they claimed that there were >731,000 first-ever or recurrent strokes during 1996. Having examined a different data source and making appropriate adjustments based on a literature review, we concur with the hypothesis of Broderick et al and believe that the true annual stroke burden in the United States is closer to three quarters of a million strokes.

A number of studies provide important information about stroke incidence rates. Reports consistently show that stroke rates increase exponentially with age and that a greater incidence of stroke is found in men than in women. These sex and age relations were confirmed, but there was a surprisingly low percentage of males identified as having had a stroke (45%). A possible explanation for the lower than expected percentage of strokes in the male population is the effect of...
competiting risks of mortality, leading to a decreased popula-
tion at risk in elderly men. This explanation was supported by
a higher age-standardized total stroke incidence rate for males
than for females (270 per 100,000 versus 248 per 100,000,
respectively).

The poor prognosis for a hemorrhagic stroke is well
known. The present study reinforces the knowledge that
patients with subarachnoid and intracerebral hemorrhages
stay longer in hospital, incur higher charges, and ultimately
have significantly higher inpatient mortality rates than do
patients with ischemic stroke. We also found that patients
with subarachnoid hemorrhages were much younger, but this
was not the case with intracerebral hemorrhages. Previous
population-based stroke incidence studies have observed a
2-fold greater incidence of subarachnoid hemorrhages in
women versus men,\textsuperscript{5,18,19} but the numbers of subarachnoid
hemorrhages were too small to allow any definitive conclu-
sions. The present study supports the earlier findings, but the
data do not allow us to identify any empirical reasons.

It was interesting to observe the much higher percentage of
whites versus blacks (92.5\% versus 4.0\%) with an ICD-9-CM
code 433 compared with other codes. Many patients with
code 433 also receive carotid endarterectomies.\textsuperscript{20,21} Studies
involving the use of carotid endarterectomy consistently
show that blacks receive the procedure at a much lower rate
than do whites.\textsuperscript{20,21} In addition, Veterans Administration data
from 1988 to 1989 identified only 4.2\% of the patients
undergoing carotid endarterectomy as being black.\textsuperscript{22} Thus,
the distribution of race among patients with code 433 pro-
vides indirect support for the low PPV of this code.

The present study has a number of limitations. First, and
potentially the most critical, is the validity of conclusions
drawn from analyses of large administrative databases that
depend on the accuracy of case-defining diagnostic codes.
Therefore, the validity of the present study is highly depen-
dent on the accuracy of the positive predictive values of the
ICD-9-CM codes. Another related limitation was the small
sample size and potential lack of generalizability of the 4
PPV studies. However, even though each individual study
was relatively small, by pooling the results we were able to
attain greater precision. This was appropriate because the
PPVs were consistent by ICD-9-CM code. The population
within each study was fairly homogeneous, but between
studies, the populations were very different. Therefore, we
believe that these results are generalizable to the US
population.

The impact of the uncertainty in the PPV pooled estimates
was examined by constructing a 95\% CI around the number
of hospital strokes. The bounds of this CI were tight (660,000
to 704,000), indicating that the point estimate had reasonable
precision. The issue of bias was addressed earlier, with the
discussion of the consistency of the PPVs across multiple
population-based studies.

A third limitation of the present study is the lack of
documented information on the rate of nonhospital stroke.
Additional data are needed to produce a more reliable
estimate of the proportion of strokes without hospitalization.
By intentionally choosing a low percentage, we were confi-
dent that our estimate of the total annual stroke burden was
not inflated. We used sensitivity analyses to illustrate the
potential impact of a different true percentage.

The methodology used in the present study was different
from the existing literature published on the incidence,
occurrence, and characteristics of stroke. All the previously
published literature used state-of-the-art stroke registries
based in relatively small geographical areas (Framingham,
Mass.; Rochester, Minn.; Rochester, NY; Northern Manhattan,
NY; and Greater Cincinnati/Northern Kentucky). Our ap-
proach might have slightly reduced internal validity, but it
should have far greater external validity. However, we would
be remiss in not adding that the present study would not have
been possible without the publication of very valuable data by
the investigators of the stroke registries.

A major advantage of the present study has been its ability
to eliminate the extensive and expensive review of the
patients’ medical records. Instead, by using the published
literature, accompanied by a large electronic database, the
time and cost of the present study was drastically reduced.
The literature already contains the extensive review of pa-
tients’ medical records along with validation of the patients’
diagnoses, and the large administrative database allows us to
exploit the wealth of information already captured.

In summary, the annual stroke burden is far greater than the
often-quoted figure of half a million first-ever or recurrent
strokes. We conservatively estimate that the true figure is
50\% higher, or closer to three quarters of a million. This new
figure emphasizes the importance of preventive measures for
a disease that has identifiable and modifiable risk factors and
of the development of new and improved treatment strategies
and infrastructures that can reduce the consequences of
stroke.

Acknowledgments

This study was supported by Knoll Pharmaceutical Co. The authors
thank Drs David Levy, David Felson, and George Seage for their
helpful comments on earlier versions of the manuscript. Drs Felson,
Seage, Matchar, and Samsa are members of G.R. Williams’s doctoral
thesis committee.

References

HJ, Stein BM, Mohr JP, Yatsu FM, eds. Stroke: Pathophysiology, Diagnosis,
5:9–12.
3. Heart and Stroke Statistical Update. Dallas, Tex: American Heart Asso-
ciation; 1995.
Minneci L, Shukla R. The Greater Cincinnati/Northern Kentucky Stroke
Study: preliminary first-ever and total incidence rates of strokes among
5. Sacco RL, Boden-Albala B, Gan R, Chen X, Kargman DE, Shea S, Paik
MC, Hauser WA, and the Northern Manhattan Stroke Study Collabo-
rators. Stroke incidence among white, black and Hispanic residents of an
Howard G, Copper LS, Shahar E. Stroke incidence and survival among
middle-aged adults; 9-year follow-up of the Atherosclerosis Risk in
8. Leibson CL, Naessens JM, Brown RD, Whisnant JP. Accuracy of hospital
Incidence and Occurrence of Total (First-Ever and Recurrent) Stroke
G. Rhys Williams, John G. Jiang, David B. Matchar and Gregory P. Samsa

Stroke. 1999;30:2523-2528
doi: 10.1161/01.STR.30.12.2523
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1999 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/30/12/2523