Can Self-Reported Strokes Be Used to Study Stroke Incidence and Risk Factors? Evidence From the Health and Retirement Study

M. Maria Glymour, ScD; Mauricio Avendano, PhD

Background and Purpose—Most stroke incidence studies use geographically localized (community) samples with few national data sources available. Such samples preclude research on contextual risk factors, but national samples frequently collect only self-reported stroke. We examine whether incidence estimates from clinically verified studies are consistent with estimates from a nationally representative US sample assessing self-reported stroke.

Methods—Health and Retirement Study (HRS) participants (n=17 056) age 50+ years were followed for self- or proxy-reported first stroke (1293 events) from 1998 to 2006 (average, 6.8 years). We compared incidence rates by race, sex, and age strata with those previously documented in leading geographically localized studies with medically verified stroke. We also examined whether cardiovascular risk factor effect estimates in HRS are comparable to those reported in studies with clinically verified strokes.

Results—The weighted first-stroke incidence rate was 10.0 events/1000 person-years. Total age-stratified incidence rates in whites were mostly comparable with those reported elsewhere and were not systematically higher or lower. However, among blacks in HRS, incidence rates generally appeared higher than those previously reported. HRS estimates were most comparable with those reported in the Cardiovascular Health Study. Incidence rates approximately doubled per decade of age and were higher in men and blacks. After demographic adjustment, all risk factors predicted stroke incidence in whites. Smoking, hypertension, diabetes, and heart disease predicted incident stroke in blacks.

Conclusions—Associations between known risk factors and stroke incidence were verified in HRS, suggesting that misreporting is nonsystematic. HRS may provide valuable data for stroke surveillance and examination of classical and contextual risk factors. (Stroke. 2009;40:873-879.)

Keywords: epidemiology ■ incidence ■ prevention ■ public health ■ risk factors ■ stroke

The lack of a national reporting system for incidence of stroke poses major challenges for surveillance and epidemiological investigations. With few exceptions,1 the best data on stroke incidence are collected in a handful of geographically localized studies with designs that allow thorough medical verification of strokes.2–7 Such data are ideal for investigations into certain individual risk factors, stroke subtypes, and short-term trajectory of stroke survivors, but the detailed medical information comes at the expense of information about variability across US populations. National cohort studies with self-reported stroke such as the Health and Retirement Study (HRS), may provide valuable supplements to local studies. HRS provides an ongoing data source, includes detailed sociodemographic assessments, covers diverse national subpopulations, and can provide insight into long-term trends and contextual risk factors. Because of uncertainty in the validity of self-reported stroke, it is important to verify that known patterns in stroke incidence rates and risk factors can be replicated in HRS. Demonstrating that previously documented stroke risk factors operate similarly in predicting medically verified and self-reported strokes will help validate self-reported data. If risk factors identified in studies with medically verified strokes cannot be replicated in HRS, this strongly suggests that using self-reported strokes introduces important biases in etiological research. If results are similar, this bolsters confidence in the usefulness of self-reported stroke for surveillance and epidemiological investigations.

In this article, we present results on the incidence of first stroke, estimated by self- or proxy report of doctor-diagnosed stroke, in HRS, a nationally representative cohort of Americans aged 50+ years. HRS is of special interest because the design and many questions are being replicated in studies across the world, including Europe,8,9 Mexico,10 China,11 and
South Korea with the goal of fostering cross-national comparisons. We compare estimates of incidence rates in HRS with previously published findings from geographically localized samples and examine whether previously established stroke risk factors also predict stroke in the HRS sample.

Methods

Study Population

HRS is a longitudinal survey of a national sample of US adults aged 50+ years and their spouses. Details of the study are provided elsewhere. Enrollment was staggered by birth cohort with enrollments in 1992, 1993, and 1998, but we begin our follow-up in 1998. This is the earliest year when the sample is representative of all birth cohorts 1900 to 1947. Biennial interviews (or proxy interviews for decedent participants) were conducted through 2006. We included HRS participants born 1900 to 1947 reported stroke-free in 1998. For the purposes of comparison with other US studies, we included only respondents who self-identified as black or white. Original survey response rates varied across enrollment cohorts from 70% to 81%, and retention rates through 2006 ranged from 82% to 86%. The HRS is approved by the University of Michigan Health Sciences Human Subjects Committee.

From 20,002 age-eligible respondents interviewed in 1998, 703 (3.5%) were excluded due to reporting race other than black or white, and 1483 (7.4%) were excluded due to reporting a stroke that occurred before the baseline 1998 interview. From these 17,816 potentially eligible participants, 380 (2.1%) were excluded due to missing information on risk factors and 380 (2.1%) were excluded due to missing follow-up information, leaving 17,056 individuals contributing time to the primary analyses. For age-stratified analyses, some people contribute time to more than one age stratum.

As a sensitivity analysis, we also estimated incidence rates (IRs) excluding baseline and follow-up proxy reports unless the respondent was dead. This analysis excluded 1,401 respondents with proxy baseline interviews.

Stroke Outcomes

Incident events were defined as first nonfatal or fatal strokes based on self- or proxy-report of a doctor’s diagnosis (“Has a doctor ever told you that you had a stroke?”). Reports of transient ischemic attacks were not systematically assessed and so were not coded as strokes. No information on stroke subtypes was obtained. For participants who had died or those unavailable for a direct interview, interviews were conducted with proxy informants, predominantly spouses. At each assessment, respondents reported stroke month and year. Stroke events for which the exact month in the 2-year interview interval was unknown (n=242) were assigned the median stroke month for events reported by other participants in the same interview wave.

Exposures and Demographic Characteristics

We stratify IRs by age group (55 to 64, 65 to 74, 75 to 84 years), sex, and race (black versus white); strata were defined to facilitate comparison with prior studies. Additional controls include years of completed education, household income and wealth (each divided by the square root of the number of household members and natural logged to reduce skew in the distribution), birth in a southern state, and birth outside the United States. We also consider the associations between stroke risk and first available report of: current smoking status (current versus all other); overweight (body mass index [BMI] 25 to <30 kg/m²), obese (BMI 30 to <35 kg/m²), or very obese (BMI 35+ kg/m²); vigorous physical activity (“On average over the last 12 months, have you participated in vigorous physical activity or exercise 3 times a week or more?”); vigorous physical activity, we mean things like sports, heavy housework, or a job that involves physical labor?%; dichotomized as yes/no; and self-reported baseline diagnoses of hypertension, diabetes, heart disease, or alcohol use (dichotomized as any versus none). We could not use the conventional criteria for extreme obesity (BMI 40+ kg/m²) because of inadequate sample size. For each risk factor, we present hazard ratios (HRs) for risk of first stroke separately and then simultaneously adjusted for all other risk factors.

Methods of Analysis

Incidence rates are calculated within each stratum based on the number of incident events divided by person-years at risk. We present overall IRs and rates within age, sex, and race strata.

Elevation in hazard of stroke associated with each risk factor was estimated with Cox proportional hazard models. Survival was defined as time from baseline interview to self or proxy reported month of first stroke, proxy-reported death, or last interview. Basic models were adjusted only for age, age-squared, sex, and education in addition to the risk factor in question. We show results using this minimal list of covariates to facilitate comparison with other published studies, many of which do not have comprehensive sociodemographic information. We used education as an indicator of socioeconomic status in these models because it is frequently used in prior studies even when income or wealth is unavailable and it is most likely to be temporally before the risk factors considered. Fully adjusted models were additionally adjusted for birth in a southern state, birth outside the United States, income, wealth, and all of the risk factors simultaneously. Combined model coefficients for behavioral risk factors should not be interpreted as “causal,” because they are simultaneously adjusted for hypertension, diabetes, and heart disease, which are presumably key mediators linking behavioral risk factors to stroke.

Analyses were conducted using SAS 9.1. We present 95% CIs in lieu of probability values. HRS used a multistage, clustered sample design, which may lead to artificially narrow CIs. We applied the HRS sampling weights to make the population representative of the US population aged 50+ years. When we compared conventional CIs with CIs obtained by a bias-corrected bootstrap with resampling at the cluster level (1000 resamples) for selected analyses, accounting for clustering of the sample had negligible consequences for CI width (detailed results available from the authors). We therefore report conventional CIs.

Comparison Studies

We compare results from HRS with those from studies reporting comparable stroke rates for blacks or whites: Rochester population studies, Framingham Heart Study (FHS), Greater Cincinnati/Northern Kentucky (GCNKY), Cardiovascular Health Study (CHS), and Northern Manhattan Stroke Study (NOMASS). Data are from either the original reports of these studies or the National Heart Lung and Blood Institute summary report on stroke incidence, which included estimates based on previously unpublished data. Other studies that provide high-quality surveillance information such as the Atherosclerosis Risk in Communities Surveillance Study could not be compared because comparably stratified rates were unavailable. SEs for published incidence rates were calculated as:

\[ SE_{IR} = \sqrt{\frac{\text{number of events}}{\text{person years of follow-up}}} \]

Data on the raw number of events stratified by age and race were not available for NOMASS or Rochester, so no SEs are presented. When possible, 95% CIs for the parameter estimates are presented, calculated as the IR ± 1.96*SEIR. We also present hypothesis tests for whether IRs estimated from HRS differ significantly differ from those derived from comparison studies by assessing whether the CIs overlap when the CIs are defined using the IR ± 1.41*SEIR. This scaling parameter was chosen as 1.96/1.41 = 1.4. This scaling parameter was chosen as 1.96/1.41 = 1.4. The actual value of k for our comparisons ranged from 1.39 to 1.41, but for simplicity, we have used a single value. This provides an approximate \( \alpha = 0.05 \) size test of the null hypothesis that the IRs are the same.
Results

Incidence

The unweighted sample characteristics are described in Table 1. Over an average of 6.8 years of follow-up on 17,056 sample members, 1293 incident first strokes were reported. After weighting to the 1998 US population aged 50 years, the estimated rate was 10.0 per 1000 person-years. Figure 1 compares HRS IRs within age/sex strata for white men and women to estimates from FHS and CHS. Age- and sex-stratified results were not available for Rochester or NOMASS, so we compare rates for both sexes combined. HRS estimates tend to be comparable with those from medically verified studies with no clear pattern of over- or underestimation. Estimates in HRS seemed most similar to those from CHS, and estimates from the 2 studies did not differ significantly for any stratum. Women aged 55 to 64 in FHS had significantly lower stroke rates than comparable women in HRS, although older women had nonsignificantly higher stroke rates. The Rochester estimates were significantly lower than HRS rates (combined for both sexes) for the 55 to 64 and 75 to 84 age strata. We were not able to formally test whether HRS estimates differ from those in NOMASS.

Figure 2 compares HRS IRs within age and sex strata for black men and women with results from CHS. Sex- and age-stratified rates were not available for GCNKY or NOMASS, so we present rates for both sexes combined. Incidence rates for blacks in HRS are generally higher than those reported for other studies. The CIs for the HRS estimates include the estimates from CHS and GCNKY, but not from NOMASS. There were no statistically significant differences between HRS estimates and those from any CHS stratum, the only other study for which SEs could be estimated.

In sensitivity analyses, we calculated IRs excluding individuals interviewed by proxy at baseline and events that were reported by proxy for a living study member (Figures 1 and 2). Excluding such proxy-reported events resulted in slightly decreased IRs for both blacks and whites.

Risk Factors

The largest single risk factor for stroke is age.20 As shown in Figure 1, stroke risk in FHS and CHS approximately doubles with every decade of additional age. This is consistent with results in HRS, in which estimated IRs increase from 4.6 to 10.0 to 18.0 for white men ages 55 to 64, 65 to 74, and 75 to 84, respectively. Similar patterns prevail among white women (4.0, 6.8, and 16.4), black men (9.8, 13.9, 23.4), and black women (9.4, 16.9, 20.7), respectively.

In addition to age, prior research demonstrates that obesity, physical activity, smoking status, alcohol use, hypertension, diabetes, and heart disease diagnoses are major risk factors for stroke among whites20, and hypertension, smoking, and diabetes are consistently linked with elevated stroke in blacks.21 Consistent with these results, Table 2 shows that white respondents who were very obese, currently smoked, or had diagnosed hypertension, diabetes, or heart disease were at increased risk of incident stroke. Vigorous physical activity...
and alcohol use were associated with reduced risk of incident stroke among whites. Each of these risk factors maintained an independent relationship with stroke even after simultaneous adjustment for the other risk factors. Among blacks, only current smoking and diagnosed hypertension, diabetes, or heart disease significantly predicted stroke incidence. After simultaneously adjusting for all risk factors, hypertension was no longer statistically significant.

Discussion

HRS may not be fully comparable to studies with medical verification because of differences in geographic regions, demographic composition, time period covered, and precise definition of a qualifying event. Nonetheless, the overall patterns are promising and suggest that HRS may be a valuable resource for research on stroke risk factors. HRS...
The black population in the United States is extremely heterogeneous and stroke rates differ substantially across US regions.22–26 For example, the black population in Northern Manhattan includes a large number of Caribbean-born individuals. Caribbean-born blacks may have lower stroke rates than US-born blacks.27–29 In fact, the previously published rates more than estimates for whites. One possible explanation is that the black population in the United States is extremely heterogeneous and stroke rates differ substantially across US regions.22–26 For example, the black population in Northern Manhattan includes a large number of Caribbean-born individuals. Caribbean-born blacks may have lower stroke rates than US-born blacks.27–29 In fact, the previously reported studies we consider here included only one site in the “Stroke Belt” (Forsyth County, NC, in CHS) and no sites in what is traditionally considered the “deep South.” Black stroke mortality and the black mortality excess compared with whites are both greater in southern states.30 This highlights the potential value of having a national data source that facilitates examination of a variety of distinct population subgroups. In this respect, it is encouraging to note that the HRS rates are generally most similar to those in CHS. CHS is the only study we considered that draws samples from multiple sites in different US regions.

Incorrect reporting is another possible explanation for the discrepancy between the incidence rates estimated in HRS and those based on medically verified strokes. Prior evidence suggests that self-reported strokes correspond well but imperfectly to medically verified strokes. Okura reported a sensitivity of 78.4% and a specificity of 98.6% for prevalent self-reported stroke or transient ischemic attack in a sample from Olmsted County, Minn, with a corresponding 67.4% positive predictive value. This is fairly consistent with findings in other studies, for example the positive predictive value of self-reported stroke was also found to be 67% in a Dutch sample31 and in the First National Health and Nutrition Examination Survey Epidemiological Followup Study.32 A study in Tromso, Norway, indicated a positive predictive value of 79% and the authors suggested a sensitivity of approximately 80% and specificity over 99%.33

The net bias in prevalence estimates introduced by this level of misreporting depends on the true prevalence in the sample. Given a sensitivity and specificity of 78.4% and 98.6%, respectively (from Olmsted County), the net bias in estimated prevalence is 0 if the true prevalence is 6.1%. Self-reported prevalence will be too high if the true prevalence is under 6.1% and too low if the true prevalence is higher than 6.1%. For example, if the true stroke prevalence is 5%, then at these sensitivity and specificity values, 5.3% of the sample is expected to report a stroke; if the true prevalence is 10%, only 9.1% will report a stroke. In the HRS sample, 6.8% of baseline eligibles ever reported a stroke. Thus, it is not surprising that the IRs in HRS correspond well to those from medically verified strokes and, among whites at least, do not appear to systematically over- or underestimate stroke rates. However, even modest reductions in specificity lead to substantial overestimates of stroke, so we cannot rule out this possibility without more detailed individual-level substudies.

Even more encouraging is the replication of individual risk factor findings previously demonstrated in studies with medically verified strokes. For example, analyses in CHS estimated the HR for any alcohol use (calculated as a weighted average of the HR in all alcohol use categories) to be 0.9034 compared with 0.83 in HRS. CHS estimated HRs for diabetes ranged from 2.1 to 2.535 compared with 1.7 and 2.1 for blacks and whites in HRS, respectively. We find that whites who were very obese (with BMI ≥35 kg/m²) were at elevated risk of stroke. Interestingly, obesity did not appear to predict stroke onset in blacks, consistent with some prior evidence.36 Simultaneous adjustment for all other risk factors rendered hypertension only marginally significant among blacks, but this is likely because heart disease is an important mediator between hypertension and stroke. The prevalence rates of several major risk factors such as smoking and obesity have changed in recent decades, and this may contribute to divergence between the IRs in HRS and those in previously

### Table 2. HRs for Stroke Risk Factors in Black and White HRS Participants

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Separate Models* HR (95% CI)</th>
<th>Combined Models† HR (95% CI)</th>
<th>Separate Models* HR (95% CI)</th>
<th>Combined Models† HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overweight</td>
<td>0.79 (0.58–1.08)</td>
<td>0.86 (0.63–1.19)</td>
<td>1.04 (0.90–1.20)</td>
<td>0.99 (0.85–1.15)</td>
</tr>
<tr>
<td>Obese</td>
<td>0.90 (0.63–1.27)</td>
<td>0.91 (0.63–1.31)</td>
<td>0.98 (0.80–1.20)</td>
<td>0.83 (0.67–1.02)</td>
</tr>
<tr>
<td>Very obese</td>
<td>0.72 (0.45–1.14)</td>
<td>0.71 (0.43–1.16)</td>
<td>1.75 (1.35–2.27)</td>
<td>1.31 (1.00–1.72)</td>
</tr>
<tr>
<td>Vigorous activity</td>
<td>0.80 (0.60–1.05)</td>
<td>0.96 (0.67–1.23)</td>
<td>0.75 (0.66–0.87)</td>
<td>0.85 (0.74–0.98)</td>
</tr>
<tr>
<td>Drinks alcohol</td>
<td>0.88 (0.66–1.18)</td>
<td>0.91 (0.67–1.23)</td>
<td>0.74 (0.64–0.84)</td>
<td>0.83 (0.73–0.96)</td>
</tr>
<tr>
<td>Current smoker</td>
<td>2.01 (1.50–2.68)</td>
<td>1.87 (1.38–2.55)</td>
<td>1.57 (1.32–1.87)</td>
<td>1.69 (1.41–2.02)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.40 (1.07–1.83)</td>
<td>1.25 (0.94–1.65)</td>
<td>1.65 (1.45–1.87)</td>
<td>1.45 (1.27–1.66)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.79 (1.36–2.36)</td>
<td>1.70 (1.27–2.26)</td>
<td>2.43 (2.09–2.84)</td>
<td>2.05 (1.74–2.40)</td>
</tr>
<tr>
<td>Heart disease</td>
<td>1.83 (1.37–2.44)</td>
<td>1.57 (1.16–2.12)</td>
<td>1.74 (1.52–2.00)</td>
<td>1.51 (1.31–1.74)</td>
</tr>
</tbody>
</table>

All models adjusted for age, age squared, sex, and years of education.
*Each cardiovascular risk factor entered into the model individually in “separate” models.
†All risk factors entered into the model simultaneously. Models additionally adjusted for income, wealth, birth outside the United States, and southern birth.
reported studies. At this point, we cannot fully evaluate the consequences of the changing population patterns of risk factors on stroke incidence, but as HRS waves accumulate, we should be able to do so. For more common outcomes, HRS is already being used for cross-cohort comparisons.37

Non-systematic misreporting of stroke generally attenuates estimated HRs for risk factors. The range of bias introduced by self-reports can be estimated in sensitivity analyses based on prior evidence about the sensitivity and specificity of self-reports of stroke. Based on the sensitivity and specificity estimated in the Olmsted County study, with stroke prevalence of 7.0%, exposure prevalence of 50%, and no relationship between exposure and incorrect reporting, a true relative risk of 2 would be attenuated by approximately 15% to 20% (calculations available from the authors). Because of the large sample size in HRS, there may be sufficient power to detect a statistically significant effect even under pessimistic assumptions about the sensitivity and specificity; however, we should anticipate some attenuation in estimated HRs. The results we present here suggest that for known risk factors, the attenuation is not severe. Although we interpret the similarity of results between HRS and comparison studies as evidence that using self-reported strokes does not bias results, one concern is that people who falsely report stroke may in fact have been diagnosed with a related cardiovascular condition that has similar risk factors.

The most important questions for extending this research focus on the extent to which misreporting differs across relevant exposure groups. Plausible factors potentially associated with misreporting might be geographic, racial/ethnic, and educational characteristics, because these influence patterns of accessing medical care and perhaps quality of clinician–patient communication. These are all measured in HRS, and stratification or statistical adjustment for these is likely to partially mitigate bias introduced by differential misreporting. Future analyses might integrate self-reported stroke and stroke symptoms checklists to improve surveillance, although this would likely capture numerous undiagnosed strokes.38

The HRS has important limitations and cannot be used to address certain research questions. No information on stroke subtypes or lesion characteristics is available. This may be especially limiting when risk factors are thought to have different effects on stroke subtypes. Because results will provide a weighted average across all stroke subtypes and ischemic strokes predominate in the population, the parameter estimates will generally be weighted toward the effects for ischemic strokes. Individual-level stroke triggers, short-term outcomes, and short-term recovery trajectories cannot be examined because interviews are only scheduled biennially.

On the other hand, HRS is unusually well suited to address certain research questions that currently constitute important gaps in our understanding of stroke epidemiology. Many of these questions are difficult or impossible to address with other data sources. Risk factors that vary at large geographic scales, eg, policies, geographic differences in medical resources, or environmental or social toxins, cannot be adequately studied in the geographically localized studies that form the backbone of current stroke research. Such factors can be studied in HRS. Major initiatives currently underway such as the Reasons for Geographic and Racial Disparities in Stroke study reflect the growing interest and commitment to understanding contextual factors that influence stroke risk.3,38,39 Similarly, because of the long follow-up, many HRS participants are interviewed several times before stroke onset. HRS data are thus well suited to study pre-stroke factors that influence stroke incidence and long-term recovery. HRS is among the most valuable sources of information on socioeconomic, racial, and geographic disparities in stroke onset and consequences, because it includes exceptionally comprehensive, well-validated socioeconomic assessments.40 Because the HRS enrollment strategy included both spouses in married couples, it is also appropriate to study how spousal characteristics influence risk and recovery.41 Finally, HRS may be valuable for surveillance of secular trends in incidence and prevalence assessing cohort differences in IRs and testing whether the effects of major risk factors such as hypertension42,43 or obesity44 have evolved in recent years. Addressing such questions, however, depends on the time invariance of reporting patterns.

In conclusion, HRS offers a promising source of information for etiologic and surveillance research on stroke incidence. Data sources that allow us to examine contextual and social risk factors will be crucial as we focus efforts on reducing incidence of stroke after the remarkable successes in reducing stroke mortality over recent decades.5,45–48 Although imperfect, HRS and similar large population surveys with self-reported strokes offer the promise of substantially broadening the research agenda for population determinants of stroke risk.

Sources of Funding
The authors gratefully acknowledge financial support from the US National Institute of Aging (AG023399) and the Robert Wood Johnson Foundation Health and Society Scholars Program. M.G. is a Robert Wood Johnson Health and Society Scholar at Columbia University. M.A. is supported by a grant from the Netherlands Organisation for Scientific Research (NWO, grant number 451-07-) and a Fellowship from the Erasmus University.

Disclosures
None.

References
Glymour and Avendano


Can Self-Reported Strokes Be Used to Study Stroke Incidence and Risk Factors?: Evidence From the Health and Retirement Study

M. Maria Glymour and Mauricio Avendano

Stroke. 2009;40:873-879; originally published online January 15, 2009; doi: 10.1161/STROKEAHA.108.529479

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

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/40/3/873

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/