Using Administrative Databases to Calculate Framingham Scores Within a Large Health Care Organization

Olaniyi James Ekundayo, MD, DrPH; Stefanie D. Vassar, MS; Linda S. Williams, MD; Dawn M. Bravata, MD; Eric M. Cheng, MD, MS

Background and Purpose—Framingham calculators are typically implemented in 1-on-1 settings to determine if a patient is at high risk for development of cardiovascular disease in the next 10 years. Because health care administrative datasets are including more clinical information, we explored how well administrative data-derived Framingham scores could identify persons who would have stroke develop in the next year.

Methods—Using a nested case-control design, we compared all 313 persons who had a first-time stroke at 5 Veterans Administration Medical Centers with a random sample of 25,361 persons who did not have a first-time stroke in 2008. We compared Framingham scores and risk using administrative data available at the end of 2007.

Results—Stroke patients had higher risk profile than controls: older age, higher systolic blood pressure and total cholesterol, more likely to have diabetes, cardiovascular disease, left ventricular hypertrophy, and more likely to use treatment for blood pressure ($P<0.05$). The mean Framingham generalized cardiovascular disease score (18.0 versus 14.5) as well as the mean Framingham stroke-specific score (13.2 versus 10.2) was higher for stroke cases than controls (both $P<0.0001$). The c-statistic for the generalized cardiovascular disease score was 0.68 (95% CI, 0.65–0.70) and for the stroke score was 0.64 (95% CI, 0.62–0.67).

Conclusions—Persons who had a stroke develop in the next year had a worse Framingham risk profile, as determined by administrative data. Future studies should examine how to improve the stroke predictive tools and to identify the appropriate populations and uses for applying stroke risk predictive tools. (Stroke. 2011;42:00-00.)

Key Words: administrative database ■ Framingham calculator ■ stroke

Stroke is the third leading cause of death and a major cause of disability in the United States.1,2 Because a history of stroke is the strongest predictor of a future stroke, interventions have rightly focused on secondary stroke prevention.3,4 However, most of the 795,000 strokes that occur in the United States each year are first-time strokes.2 If a health care organization can predict risk of first-time stroke using information available in its administrative database, then population-based strategies can be developed to prevent the occurrence of a first-time stroke.

Although several stroke risk calculators have been developed for predicting first-time stroke, they are typically used in 1-on-1 physician–patient encounters, not in widespread population screening. Health care administrative datasets are starting to include sufficiently detailed clinical information to allow calculation of stroke risk. In this study, we investigated how well stroke calculators based on administrative data could identify persons who would have a first-time stroke develop in the next year.

Materials and Methods

Data Source

The Veterans Administration (VA) Desert Pacific Healthcare Network, otherwise known as Veterans Integrated Service Network 22, encompasses 5 hospitals and 28 community-based clinics serving 1.2 million veterans residing in southern California and southern Nevada. Veterans Integrated Service Network 22 provides research support through abstraction of electronic medical data from individual VA health care facilities through a data warehouse that makes data accessible to investigators for research purposes. The data warehouse has maintained a database that includes separate files for patient demographics, outpatient clinic and inpatient utilization and diagnostic codes, vital signs, pharmacy utilization, laboratory data, as well as data on lifestyle habits, such as smoking and exercise, since 2002. The database also contains admissions to non-VA hospitals if the VA reimburses those hospitals for the care delivered.

The VA Greater Los Angeles Institutional Review Board granted approval for this study.

Study Design

A nested case-control design was used to compare all persons who had a first-time stroke develop during fiscal year (FY) 2008 (October...
The data warehouse contained information for every component of the Framingham Calculators: Stroke Risk Calculators obtained for components of the Framingham Calculators. Persons were excluded if they were assigned a history of stroke up to the end of FY2007. Persons were further excluded if they did not have a primary care visit within the network in FY2007. Excluding persons who had a primary care visit in FY2007 (October 1, 2006–September 30, 2007) because we wanted to evaluate the stroke calculator among persons who could have been potentially enrolled in a hypothetical stroke risk reduction program.

A comparison group was composed of a random sample of all persons who had a primary care visit within the network in FY2007. Persons were excluded if they were assigned a history of stroke up to the end of FY2007. Persons were further excluded if they had a new stroke develop during FY2008 because such persons were already included as cases.

Obtaining Data for Components of the Framingham Calculators: Stroke Risk Calculators

The data warehouse contained information for every component of the Framingham cardiovascular disease (CVD) calculator and the Framingham stroke calculators, but not for other stroke calculators identified. Therefore, for both cases and controls, we obtained all Framingham measurements of age, sex, systolic blood pressure (SBP), blood pressure treatment, total cholesterol, high-density lipoprotein cholesterol, and smoking status during the 12-month period of FY2007. When >1 SBP was available, the average of the last 2 outpatient SBP levels was used. When >1 cholesterol value was available, we used the latest one. Persons were designated to be a smoker if they had smoked anytime in the past year, in accordance with the VA definition of smoking.

We then identified whether persons had atrial fibrillation, cardiovascular disease, left ventricular hypertrophy, or diabetes diagnosed using ICD codes assigned from 2002 to the end of FY2007. In the original Framingham study, left ventricular hypertrophy was diagnosed by electrocardiograms, but because such reports are not included in the data warehouse, ICD codes for left ventricular hypertrophy were used instead. For diabetes ascertainment, prescription of diabetes medications was used in addition to ICD codes, in accordance with the VA method for identifying patients with diabetes.

In addition, all other available components from other identified stroke calculators were abstracted from the data warehouse: diastolic blood pressure, body mass index, low-density lipoprotein, alcohol use, cardiomyopathy, creatinine, estimated glomerular filtration rate, and chronic renal disease.

Analysis

We performed bivariate analyses to compare the stroke cases and the comparison group for all components of all calculators, including ones not fully used in this study. The Framingham calculators can be used either by calculating a score based on assigning points for each component or by calculating an exact 10-year risk based on equations supplied in the original articles. In both methods, missing SBP and cholesterol data were imputed using the normal values as presented in the 2008 generalized CVD calculator study. The Mann–Whitney U and Student t tests were used to compare Framingham scores and risks, and 95% confidence intervals were calculated.

The c-statistic was used to quantify the ability of the calculators to discriminate between persons who did and did not have a first-time stroke develop. The c-statistic represents the area under the receiver-operating characteristic curve and varies between 0.50 (no discrimination) to a maximum of 1.00 (perfect discrimination). In addition, sensitivity, specificity, positive predictive value, and negative predictive value were calculated using different cut-offs of Framingham generalized CVD 10-year risk in our sample after weighting our control group to represent the population from which the random sample was selected. The Framingham calculators were also used among a subset with age older than 55 years and again with a subset with SBP >160 mm Hg, because these subsets concentrate persons at higher risk. Analyses were performed using SAS version 9 (SAS Institute, Cary, NC).

Two sensitivity analyses were performed. First, all persons with hemorrhagic stroke were excluded, and then the primary analyses were repeated for persons who had ischemic stroke develop. Second, ICD codes of medical conditions were restricted to only FY2007 instead of from the inception of the data warehouse, and then the primary analyses were repeated.

Results

In FY2008, of the 909 persons discharged with a stroke, 403 (44.3%) were recurrent strokes. Of the remaining 506 persons with first-time stroke in FY2008, 313 persons were receiving primary care in the network in FY2007. Of these stroke cases, 69 (22%) were admitted to a non-VA hospital. For the comparison group, there were 221 371 persons who had a primary care visit in FY2007. Excluding persons who had a history of stroke or who had a first-time stroke develop in FY2008 reduced the number to 218 876 persons. A random sample of 25 361 persons (11.6% random sample) served as the comparison group (Table 1).

The stroke patients had a higher risk profile than the comparison group on many components of the Framingham calculators: older age, predominantly male, higher SBP, higher total cholesterol, were more likely to have diabetes, CVD, left ventricular hypertrophy, and were more likely to be using treatment for blood pressure (all P<0.05; Table 1). Among non-Framingham components, they were also more likely to have chronic renal disease and cardiomyopathy. Among stroke patients, 8% did not have a recorded SBP and 21% did not have a recorded cholesterol value in FY2007; among controls, 16% did not have a recorded SBP and 27%
did not have a recorded cholesterol value in FY2007 (data not shown).

In Table 2, the mean Framingham generalized CVD score (18.0 versus 14.5) as well as the mean Framingham stroke-specific score (13.2 versus 10.2) was higher for stroke cases than controls (both $P<0.0001$). The c-statistic for the generalized CVD score was 0.68 (95% CI, 0.65–0.70) and for the stroke score was 0.64 (95% CI, 0.62–0.67). The calculated 10-year risk using the generalized CVD calculator was also higher among stroke cases than controls.

Table 1. Characteristics of Patients Who Had a First-Time Stroke in 2008 and Controls Without a First-Time Stroke by 2008

<table>
<thead>
<tr>
<th>Sociodemographics</th>
<th>First-Time Stroke (Ischemic or Hemorrhagic) Developed in FY2008 (N=313)</th>
<th>First-Time Stroke Did Not Develop by FY2008 (N=25,361*)</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD)†</td>
<td>67.2 (11.8)</td>
<td>62.0 (14.9)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Male, %†</td>
<td>97.1</td>
<td>91.6</td>
<td>0.001</td>
</tr>
<tr>
<td>Smoking in the past year, %†</td>
<td>42.2</td>
<td>37.0</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 2. Performance of Framingham Calculators to Identify First-Time Stroke

<table>
<thead>
<tr>
<th>Framingham generalized CVD calculator</th>
<th>Ischemic or Hemorrhagic Stroke in FY2008 (N=313)</th>
<th>No Stroke FY2008 (N=25,361*)</th>
<th>c-Statistic (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score, mean (SD)</td>
<td>18.0 (4.2)†</td>
<td>14.5 (5.9)</td>
<td>0.68 (0.65–0.70)</td>
</tr>
<tr>
<td>Estimated 10-y risk based on score§</td>
<td>&gt;30%</td>
<td>21.6%</td>
<td>...</td>
</tr>
<tr>
<td>Calculated 10-y risk CVD, %</td>
<td>40.1%†</td>
<td>28.0%</td>
<td>0.67 (0.64–0.70)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Framingham stroke-specific score, mean (SD)</th>
<th>Ischemic or Hemorrhagic Stroke in FY2008 (N=313)</th>
<th>No Stroke FY2008 (N=25,361*)</th>
<th>Estimated 10-y stroke risk based on score§</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score, mean (SD)</td>
<td>13.2 (6.1)†</td>
<td>10.2 (6.2)</td>
<td>0.64 (0.62–0.67)</td>
</tr>
<tr>
<td>Estimated 10-y stroke risk based on score§</td>
<td>15.0%</td>
<td>10.0%</td>
<td>...</td>
</tr>
</tbody>
</table>
(40% versus 28%; \( P < 0.0001 \)), and the c-statistic was 0.67 (95% CI, 0.64–0.70).

In a sensitivity analysis, we excluded all 101 persons with hemorrhagic stroke (32%) and re-performed analyses to identify the 212 persons with ischemic stroke (Table 3). The c-statistics were similar to those of the primary analysis. After restricting data on medical conditions to only FY2007, we found the same c-statistic of 0.67 (95% CI, 0.62–0.69) using the calculated 10-year risk using the generalized CVD calculator. Test properties of the Framingham generalized CVD calculator among different thresholds are shown in Table 4. Because the 1-year incidence of first-time stroke was low, the positive predictive value remained low even at the highest Framingham risk thresholds and among older persons and persons with elevated SBP.

**Discussion**

The major finding of this study was that it is feasible to generate Framingham scores for a large number of persons using administrative databases. The mean Framingham score was high for many persons in the control group, but it was still significantly higher for persons who would have a first-time stroke develop in the next year. Therefore, in addition to its current use for individual patients in 1-on-1 encounters with a health care provider, Framingham calculators can also be used for population-wide assessments by health care administrators.

Although the c-statistics reported in our study are comparable to c-statistics of other tools used in stroke care, such as the ABCD2 score and the CHADS2 score,\(^{15,16}\) our results also demonstrate a low positive predictive value because the ratio of cases to controls is low. It is important to emphasize that in this application, the Framingham calculators are not being used as a diagnostic tool that guide use of more definitive tests, but instead are being used as a clinical prediction tool for which no “gold standard” tool for stroke prediction exists. Therefore, a lower positive predictive value and specificity may be an acceptable trade-off to obtain high sensitivity in detecting first-time stroke in the near future, so that resources can be targeted to persons at highest risk.

Differences in the study population, surveillance method, and outcome criteria are established reasons why prognostic indexes tested in a different population perform less well compared to the original derivation study.\(^{17}\) First, our study sample consists of veterans who use the VA, who are predominantly male, and who have lower socioeconomic

---

**Table 3. Performance of Framingham Calculators to Identify First-Time Ischemic Stroke**

<table>
<thead>
<tr>
<th>Ischemic Stroke in FY2008 (N=212)</th>
<th>No Stroke FY2008 (N=25,361*)</th>
<th>c-Statistic (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framingham generalized CVD calculator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>18.3 (4.2)†</td>
<td>14.5 (5.9)</td>
</tr>
<tr>
<td>Estimated 10-y risk based on score§</td>
<td>&gt;30%</td>
<td>21.6%</td>
</tr>
<tr>
<td>Calculated 10-y risk CVD, %</td>
<td>41.4%†</td>
<td>28.0%</td>
</tr>
<tr>
<td>Framingham stroke specific score, mean (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>13.6 (6.1)†</td>
<td>10.2 (6.2)</td>
</tr>
<tr>
<td>Estimated 10-y stroke risk based on score‡§</td>
<td>17.0%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

CI indicates confidence interval; CVD, cardiovascular disease; FY, fiscal year; SD, standard deviation.
*Weighted to represent 218,876 persons.
†\( P < 0.0001 \).
‡Based on information provided in the Framingham study.\(^{8,9}\)
§Insufficient information to calculate the 10-y stroke risk.\(^{4}\)

**Table 4. Performance of Framingham Calculator Among Subsets of Population**

<table>
<thead>
<tr>
<th>Cases/Weighted Controls</th>
<th>c-Statistic</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Framingham generalized cardiovascular disease 10-y risk, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;0</td>
<td>313/218,876</td>
<td>0.675</td>
<td>100.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>&gt;10</td>
<td>297/178,486</td>
<td>0.628</td>
<td>94.9</td>
<td>18.5</td>
<td>0.2</td>
</tr>
<tr>
<td>&gt;20</td>
<td>256/133,506</td>
<td>0.608</td>
<td>81.8</td>
<td>38.4</td>
<td>0.2</td>
</tr>
<tr>
<td>&gt;30</td>
<td>196/90,265</td>
<td>0.617</td>
<td>62.6</td>
<td>58.8</td>
<td>0.2</td>
</tr>
<tr>
<td>&gt;40</td>
<td>148/53,499</td>
<td>0.588</td>
<td>47.3</td>
<td>75.6</td>
<td>0.3</td>
</tr>
<tr>
<td>&gt;50</td>
<td>97/29,352</td>
<td>0.572</td>
<td>31.0</td>
<td>86.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Other cut-offs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age older than 55 y</td>
<td>181/154,519</td>
<td>0.659</td>
<td>57.8</td>
<td>29.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Systolic blood pressure &gt;160 mm Hg</td>
<td>29/69,82</td>
<td>0.540</td>
<td>9.3</td>
<td>96.8</td>
<td>0.4</td>
</tr>
</tbody>
</table>

BP indicates blood pressure; NPV, negative predictive value; PPV, positive predictive value.
status and health status compared to the general population.\textsuperscript{18} Low socioeconomic status was shown to be a predictor in the QRISK cardiovascular disease risk algorithm.\textsuperscript{13} Therefore, it is likely that the Framingham calculator underestimates the risk for stroke in our population because it is more disadvantaged compared to the general population.\textsuperscript{18} Second, the original Framingham cohort studies minimized missing data by making concerted efforts to collect data at all time points. In contrast, persons in our administrative dataset did not undergo such complete surveillance for medical conditions and likely possessed undiagnosed conditions that would have been detected if they were enrolled in a cohort study such as Framingham. By imputing normal values when they were missing, our calculations underestimate the actual cardiovascular risk for all persons. Third, we used administrative codes to identify the outcome variable, development of a first-time stroke. Although this approach maximizes the efficiency and applicability of using such risk prediction tools, it will not be as accurate as a skilled clinician in identifying persons at risk for stroke, as was performed in Framingham.\textsuperscript{5–7} In addition, the data warehouse captures some but not all admissions to non-VA hospitals. One study reported that 76% of dually eligible Medicare and VA patients obtained care for their initial stroke at non-VA hospitals;\textsuperscript{19} as a result, not all veterans admitted to non-VA hospitals may have their data recorded in the data warehouse.

The study has several limitations. We did not have longitudinal data to examine outcomes beyond 1 year. In addition, as with all studies conducted in veteran populations, studies should also be conducted in nonveteran populations before determining whether findings are generalizable to that population. Identified stroke cases and data obtained for the components of the Framingham calculator from the VA administrative database used in this study were not validated by chart abstraction. However, many studies have assessed the validity of information recorded in VA administrative databases using the ICD codes compared to information abstracted from original medical record and reported very high levels of agreement.\textsuperscript{5–7}

Regarding recommendations to future applications of stroke tools using administrative data, the discriminating abilities of the Framingham calculators was similar using the last year of data versus using the previous 5 years of data. The stroke-specific calculator (that includes atrial fibrillation but not cholesterol) performed similarly to the newer CVD calculator, so we believe that either calculator could be used. We will be exploring whether the discrimination of the Framingham calculators can be improved using other clinical data available in VA databases. However, novel markers such as C-reactive protein have not been shown to substantially improve on the already very good discriminating properties of existing Framingham calculators.\textsuperscript{20,21}

### Conclusions

Persons who have a first-time stroke develop do have a significantly higher Framingham risk than controls based on administrative data available in the year before their stroke. Whereas our study shows that the proportion of persons who have a stroke develop in the next year may be too low to be easily identified, the c-statistics of administrative data-derived Framingham scores appear satisfactory enough to attempt predicting a more common outcome, such as stroke in the next 10 years. As clinical datasets become more available, administrative data-derived Framingham scores should be further validated to determine their suitability in studies of population health.

### Sources of Funding

The project reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service Quality Enhancement Research Initiative Project 08-242. Dr. Cheng is supported by a career development award from National Institute of Health/National Institute of Neurological Disorders and Stroke (K23NS058571). Drs. Cheng and Ekundayo are also supported by an award from the American Heart Association Pharmaceutical Roundtable and David and Stevie Spina. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

### Disclosures

None.

### References

Using Administrative Databases to Calculate Framingham Scores Within a Large Health Care Organization

Olaniyi James Ekundayo, Stefanie D. Vassar, Linda S. Williams, Dawn M. Bravata and Eric M. Cheng

*Stroke*. published online May 5, 2011;

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 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/early/2011/05/05/STROKEAHA.110.603340

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/