**Topical Review**

Section Editors: Amanda G. Thrift, BSc (Hons), PhD, PGDipBiostat and Barbara G. Vickrey, MD, MPH

---

**Stroke Risk Calculators in the Era of Electronic Health Records Linked to Administrative Databases**

Adam Richards, MD, MPH; Eric M. Cheng, MD, MS

_electronic health records (EHRs) linked to administrative databases can unlock the potential of clinical data by overcoming limitations of current cohort-derived risk calculators. EHRs can facilitate implementation of risk-based approaches, whereas current manual or web-based risk calculators pose prohibitive practical barriers to widespread use. We discuss how EHRs can be used: (1) To derive increasingly precise risk equations calibrated to the populations in which they will be applied; (2) To facilitate communication of individual risk to patients; and (3) To provide population-based risk information for researchers, administrators, and policymakers._

---

**Application and Derivation of Risk Calculators: A Primer**

Several risk calculators directly predict the outcome of stroke. Stroke is also included as part of a composite outcome in other calculators, an approach that may grow in popularity as the concept of stroke as a cardiovascular disease (CVD) risk equivalent becomes more widely accepted.

Guidelines state that the level of stroke and CVD risk should inform decision-making about initiating treatments such as aspirin or lipid-lowering agents. These strategies are based on the observation that the respective relative risk reduction of aspirin and statin therapies is similar for most subpopulations, and therefore the absolute benefit of treatment is proportional to the absolute risk of stroke or coronary heart disease. Because clinicians do not accurately estimate cardiovascular risk, adhering to these guidelines requires the use of explicit risk calculators.

At least 110 stroke and CVD risk scoring methods exist. Early systems typically relied on a points-based system that required clinicians to manually calculate the sum of points associated with various risk factors. More recent calculators allow clinicians to input parameters directly into a website that uses a multivariate equation to predict risk. A recent study comparing 2 formats of the Framingham calculator suggests that the equation is more accurate than the points-based approach.

---

**Derivation**

Most risk calculators are derived from data in prospective cohort studies. The quintessential cohort study initiated in Framingham, MA in 1948 revolutionized the preventive approaches to CVD and pioneered the logistical and statistical methods of risk prediction. Most calculators apply to primary prevention populations without a history of cerebrovascular disease, although some calculators have also been developed to predict stroke in the setting of atrial fibrillation, and after a recent transient ischemic attack or stroke. Randomized controlled trials are also used to derive CVD calculators such as the Reynolds Risk Score.

EHRs linked to administrative databases are increasingly used to derive and validate risk prediction models. For example, the Anticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study used 13,559 patients included in a clinical database of Kaiser Permanente of Northern California to predict the risk of warfarin-associated hemorrhage.

---

**Large Sample Sizes Facilitate Study of Uncommon Risk Factors**

The primary limitation of relying on prospective cohort studies to derive risk calculators is the relatively small number of outcomes. The cohorts used to derive the Framingham and Atherosclerosis Risk in Communities (ARIC) stroke scores were based on 472 and 434 stroke events, respectively, numbers insufficient to evaluate multiple risk factors and their interactions. For example, age modifies the relative risk of stroke associated with numerous risk factors, including blood pressure, smoking, and atrial fibrillation but few cohort-based CVD risk scores, and none of the stroke-specific risk scores, include parameters for these age-related interactions. Atrial fibrillation is excluded from the ARIC stroke score (Table) because men with atrial fibrillation at baseline experienced only 2 stroke events; and no women with atrial fibrillation had a stroke. In another example, the prevalence of rheumatoid arthritis is 1% in the general population and studies much larger than the Framingham cohort have established that rheumatoid arthritis is associated with a 30% increased risk of stroke. However, we would expect fewer than 60 Framingham participants to have rheumatoid arthritis based on the original study size of 5,734, and the expected

---

Received May 30, 2012; final revision received September 5, 2012; accepted September 10, 2012.

From the David Geffen School of Medicine, University of California, Los Angeles, CA (A.R., E.M.C.); and VA Greater Los Angeles Healthcare System, Los Angeles, CA (E.M.C.).

Correspondence to Eric M. Cheng, MD, MS, VA Greater Los Angeles, 11301 Wilshire Blvd, Department of Neurology, ML 127, Los Angeles, CA 90073. E-mail eric.cheng@va.gov

(Stroke. 2013;44:564-569.)

© 2012 American Heart Association, Inc.

_Stroke is available at http://stroke.ahajournals.org_

DOI: 10.1161/STROKEAHA.111.649798
number of strokes in this subgroup is <15. Even if the risk calculator perfectly predicted risk in this group it would have negligible impact on overall metrics of discrimination, calibration, or reclassification.

When a country uses a national EHR platform linked to an administrative database then the entire population effectively participates in a registry. For example, all 73,538 patients with atrial fibrillation not treated with vitamin K antagonists in Denmark in the period 1997–2006 contributed to an analysis to conclusively show that CHA$_2$DS$_2$-VASc is more valid for stroke prediction in patients categorized as being at low and intermediate risk by CHADS$_2$. The QRSK2 score to predict a composite outcome of CVD and stroke was derived using 1.5 million patients in the QRESEARCH database of 551 clinical practices in the United Kingdom. Compared with the cohort studies and randomized controlled trials used to derive Framingham and Reynolds Risk Scores (Table) the number of patients and events contributing to the derivation and validation of QRSK2 is vast: nearly 3.9 million patients.

### Table. Stroke and Cardiovascular Disease Risk Calculators

<table>
<thead>
<tr>
<th>Stroke-Specific Scores</th>
<th>Cardiovascular Disease Scores</th>
<th>Administrative Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observational Cohort Studies</strong></td>
<td><strong>Randomized Controlled Trials</strong></td>
<td><strong>Database</strong></td>
</tr>
<tr>
<td>Framingham Stroke$^1$</td>
<td>Reynolds Risk Score$^{10}$</td>
<td>QRESEARCH Database</td>
</tr>
<tr>
<td>ARIC Stroke$^2$</td>
<td>Physicians' Health Study</td>
<td></td>
</tr>
<tr>
<td>CHS Stroke$^3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Derivation cohort</strong></td>
<td><strong>Sample size</strong></td>
<td><strong>No. of strokes</strong></td>
</tr>
<tr>
<td>Framingham Heart Study</td>
<td>5734</td>
<td>472</td>
</tr>
<tr>
<td>ARIC</td>
<td>14,685</td>
<td>434</td>
</tr>
<tr>
<td>CHS</td>
<td>5711</td>
<td>399</td>
</tr>
<tr>
<td><strong>Age at baseline</strong></td>
<td><strong>No. of events</strong></td>
<td><strong>Primary outcome</strong></td>
</tr>
<tr>
<td>55–84</td>
<td>1174</td>
<td>Any stroke</td>
</tr>
<tr>
<td>45–65</td>
<td>766</td>
<td>Ischemic stroke</td>
</tr>
<tr>
<td>65+</td>
<td>1294</td>
<td>Any stroke</td>
</tr>
<tr>
<td><strong>Observational Cohort Study</strong></td>
<td><strong>Administrative Databases</strong></td>
<td><strong>Sample size</strong></td>
</tr>
<tr>
<td>30–74</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td>140,115</td>
</tr>
<tr>
<td>45+</td>
<td>CHD, coronary revascularization, ischemic stroke</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>1,535,583</td>
<td></td>
</tr>
<tr>
<td>35–74</td>
<td>CHD, TIA/stroke</td>
<td></td>
</tr>
</tbody>
</table>

ARIC indicates Atherosclerosis Risk in Communities; BP, blood pressure; CHD, coronary heart disease; CHF, congestive heart failure; CVD, cardiovascular disease; Hs-CRP, high-sensitivity C-reactive protein; PAD, peripheral arterial disease; TIA, transient ischemic attack; and SES, socioeconomic status.

<table>
<thead>
<tr>
<th>Traditional risk factors</th>
<th><strong>Observational Cohort Study</strong></th>
<th><strong>Administrative Databases</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Framingham Heart Study</td>
<td>QRESEARCH Database</td>
</tr>
<tr>
<td>Sex</td>
<td>ARIC</td>
<td></td>
</tr>
<tr>
<td>Systolic BP</td>
<td>CHS</td>
<td></td>
</tr>
<tr>
<td>BP medications</td>
<td>5734</td>
<td>1174</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>14,685</td>
<td>766</td>
</tr>
<tr>
<td>Cigarette smoking (Y/N)</td>
<td>5711</td>
<td>1294</td>
</tr>
<tr>
<td>Total number of cigarettes</td>
<td>472</td>
<td>1294</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>434</td>
<td>1294</td>
</tr>
<tr>
<td>Left ventricular hypertrophy</td>
<td>399</td>
<td>1294</td>
</tr>
<tr>
<td>Prior CVD</td>
<td>177</td>
<td>1294</td>
</tr>
<tr>
<td>Family history of CHD</td>
<td>CHS Stroke$^3$</td>
<td>1,535,583</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>Reynolds Risk Score$^{10}$</td>
<td>140,115</td>
</tr>
<tr>
<td>Laboratory biomarkers and clinical measurements and diagnoses</td>
<td>Physicians' Health Study</td>
<td></td>
</tr>
<tr>
<td>Hs-CRP</td>
<td>QRESEARCH Database</td>
<td></td>
</tr>
<tr>
<td>Creatinine</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Chronic renal disease</td>
<td>CHD, coronary revascularization, ischemic stroke</td>
<td></td>
</tr>
<tr>
<td>Impaired fasting glucose</td>
<td>CHD, TIA/stroke</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>15-foot walk time</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Demographic factors</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
<tr>
<td>Neighborhood SES</td>
<td>CHD, angina, TIA/stroke, PAD, CHF</td>
<td></td>
</tr>
</tbody>
</table>
or 14% of UK population aged 34 to 74 years, experienced 211,580 CVD events. The large size facilitates inclusion of parameters for uncommon risk factors, minority ethnic populations, low socioeconomic status (SES) populations, and interaction terms to account for important effect modification by age for 8 risk factors.

**Administrative Datasets Address Threats to External Validity**

Systematic over- or underestimation of risk occurs whenever the Framingham Risk Score (FRS) is applied in populations with different risk factor prevalence or different 10-year risk of stroke/CVD from the original Framingham cohort. The FRS discriminates reasonably well in other countries and ethnic groups but recalibration is required. Recalibration, in turn, requires knowledge of absolute 10-year event rates, as well as the mean values for continuous model parameters (such as age, blood pressure, and cholesterol) measured in the population in which the score will be applied. Practicing clinicians rarely have access to event rates and risk factor values specific to the populations they serve, so analysis of local EHR or administrative databases would be needed for recalibration.

The large secular decline in CVD risk in the United States has lowered the absolute CVD risk from the era when cohort studies were initiated. To predict the shifting target of absolute risk in different populations and over time, risk equations ideally would be derived using contemporary data from the actual populations in which they will be applied. For example, QRISK2 in the United Kingdom is updated at least annually. Just as there is an ongoing need to recalibrate equations, there is also an ongoing need to validate them using local EHR or administrative databases.

**Addressing Missing Data in Administrative Databases**

The primary strength of cohort studies and randomized controlled trials is their ability to minimize misclassification of exposures and outcomes by achieving high ascertainment of baseline risk factors and by adjudicating cardiovascular events and deaths. Conversely, the perceived Achilles heel of administrative databases is incomplete or inaccurate data. For example, in a study that assessed the discrimination of FRS using Veterans Administration (VA) databases, 16% of patients were missing data on blood pressure and 27% did not have a recorded cholesterol level in an administrative database from 5 VA medical centers. In addition, the dataset underestimates outcome events because patients can be admitted with a stroke to non-VA hospitals. Despite this extent of missing data, we found discrimination of the FRS based on VA administrative data was comparable with other risk prediction tools.

In the future, the Meaningful Use program run by the Centers for Medicare and Medicaid Services should minimize missing data in EHR-linked administrative databases. Some of the Meaningful Use objectives require entry of information such as race, ethnicity, vital signs, and laboratory in discrete data fields instead of existing as free text in a progress note. Until then, there are several approaches in limiting validity threats because of missing data. One approach leverages existing data from the medical literature. For example, the proprietary Archimedes model combines information from a clinical database of a large group-based health maintenance organization with data from cohort studies as well as randomized controlled trials. Another approach is to impute missing data if the database is sufficiently large. The QRESEARCH database is missing cholesterol values for over two third of participants. However, the number of participants with complete data far exceeded the number included in traditional cohort studies, and this permits the use of robust multiple imputation methods. Multiple imputation formally assumes that data are missing at random, and this assumption may not hold in administrative datasets that rely on clinically available information. Methods exist to address the possible bias introduced when data are not missing at random. For example, variables can be included in the imputation model that do not appear in the risk equation, imputations can be weighted to reflect their plausibility under specified mechanisms of not missing at random, and weighted imputations can be combined with inverse-probability methods. Additional research may be necessary to determine the ideal methods for handling missing data. Risk equations should be externally validated in independent datasets and their performance should be compared with existing tools. QRISK2 has been validated in a large independent dataset and appears to improve calibration and reclassification compared with the FRS in the United Kingdom.

In addition to the problem of missing data, some risk score components are unlikely to be stored as discrete data in administrative datasets. For example, the Cardiovascular Health Study stroke score includes 15-foot walk time and the lifetime number of cigarettes smoked, and Reynolds Risk Score includes family history of premature coronary heart disease (see Table). However, it is possible to improve ascertainment of important risk factors. For example, a dedicated assessment of family history using a patient questionnaire identified 5 times as many patients with a family history of premature coronary heart disease as a review of the clinic EHR in 1 study, and this assessment resulted in reclassification of 5% of the clinic population into a higher risk category eligible for treatment.

Most administrative databases also do not include emerging risk factors that are not routinely measured in clinical practice such as coronary artery calcium. However, such components are no more likely to be available for practicing clinicians who do not use EHRs. Researchers interested in prospective evaluation of new risk factors may find it simpler and cheaper to perform novel assessments on patients already integrated into an existing EHR than to conduct a new dedicated cohort study.

**Facilitating Patient Communication About Risk**

The Institute of Medicine Comparative Effective Research Prioritization project included among its list of 100 priority topics to compare the effectiveness of adding information about new [risk factors]… with standard care in motivating behavior change and improving clinical outcomes. EHRs can instantly display predicted risk derived from the original equation models without manual data entry, saving precious time in clinical encounters. EHRs can incorporate patient
preferences to facilitate informed choice. For example, patients can view risk scenarios by selecting combinations of care options and choose from multiple state-of-the-art risk communication formats to complement simple absolute and relative risks; such as number-needed-to-treat, ratios, and percentages that may improve comprehension.25

Because EHR-derived risk scores can be quickly calculated at the point of care, there is flexibility in choosing what to present. Multiple risk scores from different calculators can be shown. Even when calculators produce divergent risk estimates,26 clinicians can still take into account this lack of consensus in their decision-making. EHRs can also accommodate alternative time horizons such as lifetime risk29 which may be particularly relevant to younger patients.

EHRs do not definitively establish whether to treat patients whose predicted risk is near treatment thresholds. Guidelines currently recommend lipid-lowering therapy for a patient with a 10-year predicted risk of CVD of 10.1%, but would not recommend treatment for a similar patient with a predicted risk of 9.9%, even though the prediction intervals around these estimates would sufficiently overlap to render them indistinguishable (a prediction interval for an individual observation is related to, but wider than, a confidence interval for a population parameter). Currently available risk calculators do not provide information related to the uncertainty of their predictions. Large linked administrative databases can improve precision of risk prediction and reduce the number of people whose prediction intervals approximate clinical decision thresholds, but additional research is necessary to explore whether and how to integrate information on uncertainty of risk prediction into clinical guidelines and facilitate patient-centered decision-making.

EHRs may also facilitate use of risk calculators by the public. Just as clinicians do not accurately estimate cardiovascular risk,6,7 patients’ perceived stroke risk does not correlate well with either their Framingham-calculated or observed stroke risk.28 In particular, patients at higher risk for stroke (and most other diseases) are more likely to underestimate their risk, a phenomenon described as unrealistic optimism.30 Patients state that they prefer seeing their individualized stroke risk, but evidence is lacking to demonstrate that this additional information improves risk factor control or medication adherence.31,32

Facilitating Documentation of Disparities and Implementation of Interventions to Improve Population Health

Systematic undertreatment of entire subpopulations holds ethical implications for clinical practice. For example, the current FRS underestimates CVD risk among low SES groups and this systematic bias may lead clinicians to undertreat low SES groups (and to overtreat high SES groups), thus inadvertently exacerbating disparities. In the United States, 15% of persons of low SES status are classified into a higher risk category when education and income are added to the FRS,34 and people whose predicted 10-year CVD risk is 6% or 13% according to FRS will experience a true risk of 10% and 20% if they live in a low-income neighborhood.35 Efforts to incorporate patient-reported measures such as education into EHRs may soon improve the quality of SES data available in administrative databases.36 QRI SK2 accounts for excess CVD risk among the poor by including a measure of neighborhood deprivation linked to each patient using their home address. EHRs facilitate the automated linkage of patients to similar indices of neighborhood deprivation developed in the United States.

By contributing clinical parameters to administrative databases, EHRs facilitate precise risk estimation for millions of patients at a time. Identifying the highest risk individuals permits targeting of intervention to those most likely to benefit. The Medicare Shared Savings Plan incentivizes Accountable Care Organization to reduce healthcare utilization such as hospitalizations by identifying and treating a list of high-risk patients.37

If a risk tool is used to prioritize enrollment of patients in a care intervention to improve population health, an alternative to using an identical risk threshold for every individual patient is to base patient selection criteria on the resource constraints of a care intervention. For example, if resources exist to enroll 1000 patients in an intervention then an organization may elect to replace criteria based on an absolute risk threshold (eg, 10% risk of CVD in 10 years) with alternative ranking criteria that selects individuals with the highest 1000 risk scores. EHRs by themselves do not address the normative and political questions required to define high risk thresholds or other treatment criteria but they provide a flexible platform to derive, evaluate, and implement these myriad approaches.

Facilitating the Use of Risk Calculators

To improve clinical decisions, risk calculators must first be used. Data are lacking, but utilization of risk prediction tools is likely to be low in most settings. The Joint Commission’s new Tobacco Cessation Performance Measure-Set38 provides an example of how to incentivize and track the implementation of risk calculators. The Tobacco Measure-Set requires hospitals to identify and document the smoking status of all patients, provide eligible patients with cessation counseling and medication, prescribe cessation medication at discharge, and document tobacco-use status 1 month later. Similarly, clinicians could be incentivized to use risk calculators through quality measures. Such a measure could state that clinicians should provide patients with their own CVD and stroke risk and provide counseling on how to reduce the risk.
An EHR could facilitate performance of such a measure by automatically calculating level of risk.

Current outcome measures in control of atherosclerotic risk factors are based on categorical thresholds of continuous risk factors. There is an ongoing vigorous debate about the relative merits of simpler so-called treat-to-target strategies that target a specific low-density lipoprotein or blood pressure level versus more accurate risk-based39 or individualized guideline40 strategies informed by risk calculators. Future measures should strongly consider augmenting (or replacing) specific treatment goals with rewards proportional to achieved reductions in overall risk.

Conclusions

The accurate prediction of risk for atherosclerotic events such as stroke presents clinicians and patients with an opportunity to practice patient-centered, personalized medicine, and provides administrators and policymakers with a powerful tool to efficiently target care interventions to individuals and populations most likely to benefit. EHRs may facilitate a transition to a future in which patient- or clinician-selected risk scores are automatically calculated for clinicians to inform treatment decisions, and used by hospitals, accountable care organizations and insurers for risk adjustment and the prioritization of high-cost interventions. Health systems such as the VA or others located in the Stroke Belt may choose to develop their own stroke-specific risk scores based on their own unique populations. Eventually EHR-linked databases that currently reside in clinical information silos could be linked so that every person contributes in real-time to the derivation and recalibration of risk calculators. Once privacy concerns are addressed, an EHR-linked database that includes the US population of 311 million would contribute ≈850 000 years of person-time and over 1600 first stroke events every single day.

Acknowledgments

We thank France Nguyen, PhD, and Miriam Ayad, MPH, for assistance in formatting the paper.

Sources of Funding

Dr Cheng is supported by a Career Development Award from National Institutes of Health/National Institute of Neurological Disorders and Stroke (K23NS058571). Drs Cheng and Richards are supported by the UCLA Outcomes Research Center, funded through 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 reflect the position or policy of the Department of Veterans Affairs or the United States government.

Disclosures

None.

References


Key Words: cardiovascular disease □ database management systems □ electronic health records □ epidemiology □ risk assessment □ stroke
Stroke Risk Calculators in the Era of Electronic Health Records Linked to Administrative Databases
Adam Richards and Eric M. Cheng

Stroke. 2013;44:564-569; originally published online November 29, 2012;
doi: 10.1161/STROKEAHA.111.649798

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/44/2/564

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/