Predicting Stroke Outcomes and Comparing Imbalanced versus Oversampled Dataset Training using Random Forests, Conditional Inference Trees, Ridge Regression, and Gradient Boosting

P8106 Data Science 2 Midterm

Emil Hafeez, eh2928

Contents

Section 1: Introduction

This dataset contains 5110 observations, with de-identified individual patients as the unit of analysis. Each patient is characterized by their ID, and a set of numerical and factor variables detailing basic demographic and biomedical information, like their gender, age, and average glucose level. The outcome of interest is a binary indicator of stroke. There is no information regarding prior stroke history. While limited information is provided by the data source, this dataset is thought to consider both ischemic and hemorrhagic types of stroke as stroke outcome. These stroke outcomes are both very severe, and very prevalent; stroke is the leading cause of long-term adult disability and the fifth leading cause of death in the United States[1,2]. Race-ethnicity, though not a variable included in this dataset, is also significantly associated with stroke incidence, thereby making stroke prevention a racial justice concern as well.[3]. As such a major source of disease, stroke is also quite expensive, estimated at about $34 billion per year due to healthcare services and missed work. As with many illnesses, the best approachto reducing the burden of stroke remains prevention[5]. Were stroke outcomes to be predicted from a small set of predictors set as those available, it may be feasible to provide earlier intervention and preventative care, in order to avert these racial and political economic challenges.

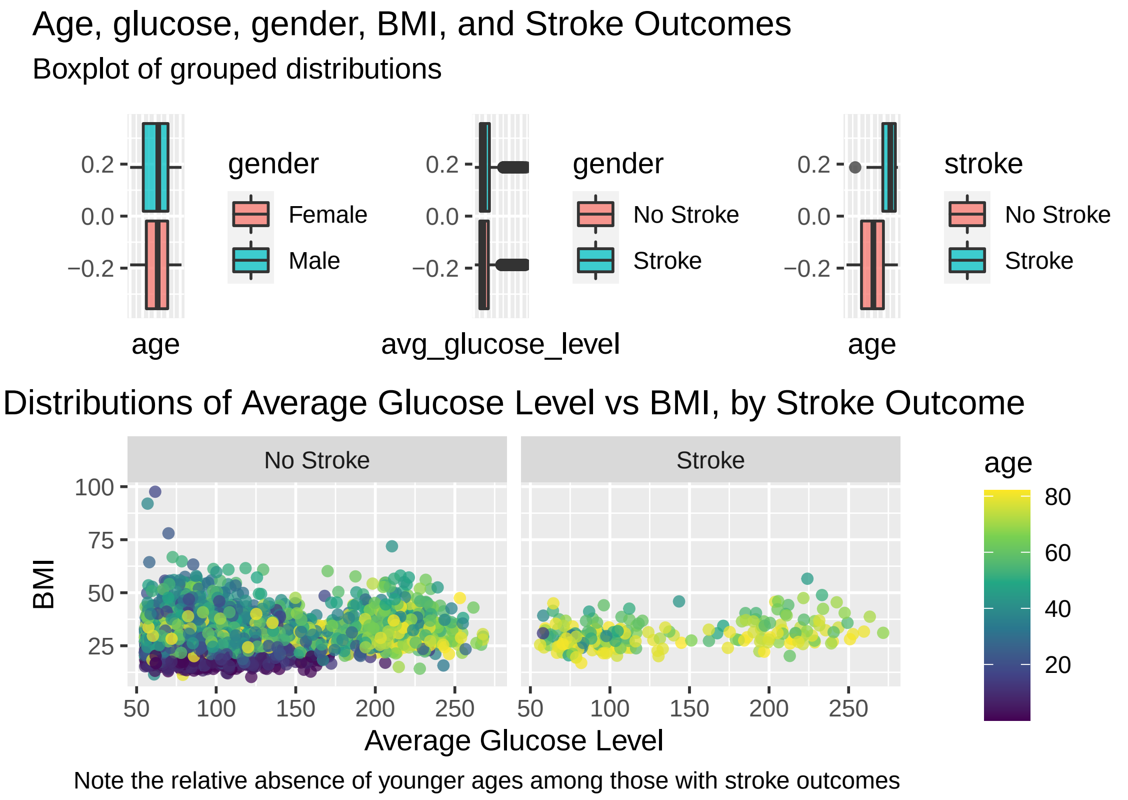
Since the current study examines whether a moderately effective prediction of stroke outcomes (as a binary classification) can be made using a largely demographic and noninvasive minimal set of predictors, this makes for a classification question under investigation, in the service of another question: to what extent can such a dataset be thought of as a decision support system for medical practitioners, and/or serve towards an early-screening tool for stroke prevention and resource allocation (focusing on special precautions for those at-risk)? Ideally, some insight into which of the factors are most relevant for predicting this outcome is made available too.

Variables were transmuted into more appropriate data types for analysis (for example, character variables into factor variables). Then, the whole dataset was analyzed for missing data, and `tidyverse` "data tidying" best practices were implemented to rename and reorder variables. All variables were investigated for abnormal values and uncommon responses, as well as missing data. The BMI variable was missing 201 observations, which were assumed to be due to a missing-at-random and thus omitted. Imputation may also be a good option to consider, to avoid information loss. The smoking categorical variable has values listed as "unidentified" that could be considered as missing data but are not here (without more information from data collection process as to what this means). Regrettably, the gender variable had only one “other” observation, which was dropped to enable model fitting (more comment below). Categorical variables were also subject to some technical adjustments in order to allow usage in prediction algorithms. The `caret` package was used to create testing and training partitions, and perform transformations like centering and scaling as needed. Some objects (such as a predictor matrix) were created for modeling purposes too.

Section 2: Exploratory Analysis and Visualization

Data summaries, boxplots within and between groups, scatter plots with continuous variables and groups, and correlation plots were used. There are several categorical variables, as well as a few nominal and ordinal variables; this presents several challenges for modeling and is a key feature of the data. Other than the missing BMI observations, it is also notable that of the 5110 original observations, only 1 marked “Other” for their gender, which limits the variability modeling can leverage in that regard. Notice also that the dataset is substantially unbalanced (249 strokes versus 4860 no strokes, before missing data removal). Therefore, this analysis also uses the ROSE package to oversample the dataset; model tuning and training is conducted on both the cleaned oversampled dataset and the imbalanced dataset, and the models are compared. “No ROSE” refers to the imbalanced dataset.

It is clear from EDA that average glucose level is not normally distributed, with a heavy right tail. BMI appears bell-shaped though has several high outliers. There is a range of ages in the dataset including high counts from 40 to 60 years old, and a spike near 80 years old. Genders appear approximately equal distributed in age. Interestingly, average glucose level appears approximately the same across the outcome groups (no stroke, and stroke), though the stroke group appears significantly older than the no-stroke group. Regarding correlation, there is a moderate positive association between age and having been ever married, and moderate negative association between private employment and self-employment (and denote some collinearity among predictors, as anticipated). See below before a `patchwork` plot of some predictor distributions; notice the higher ages of those with stroke outcomes in this plot, as well as the distribution of bmi and glucose level (including some extreme values). The correlation plot is also printed.


Age, glucose, gender, BMI, and Stroke Outcomes

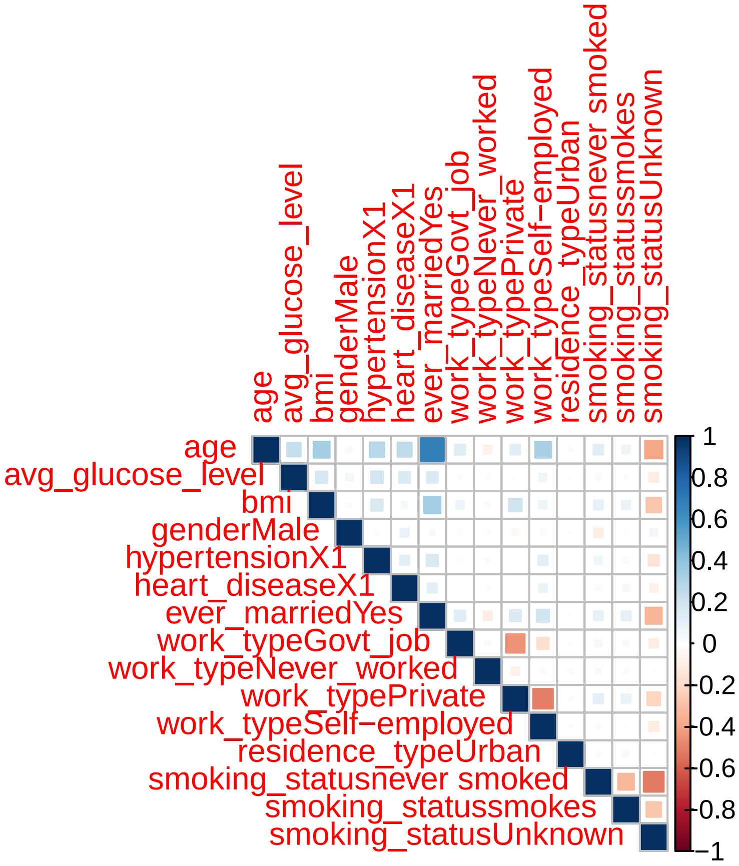
Boxplot of grouped distributions

Distributions of Average Glucose Level vs BMI, by Stroke Outcome

Note the relative absence of younger ages among those with stroke outcomes

Figure 1: Gender, Age, Outcome, and Scatterplot of BMI vs Glucose Colored by Age

Regarding the correlation of predictors, please see the following.

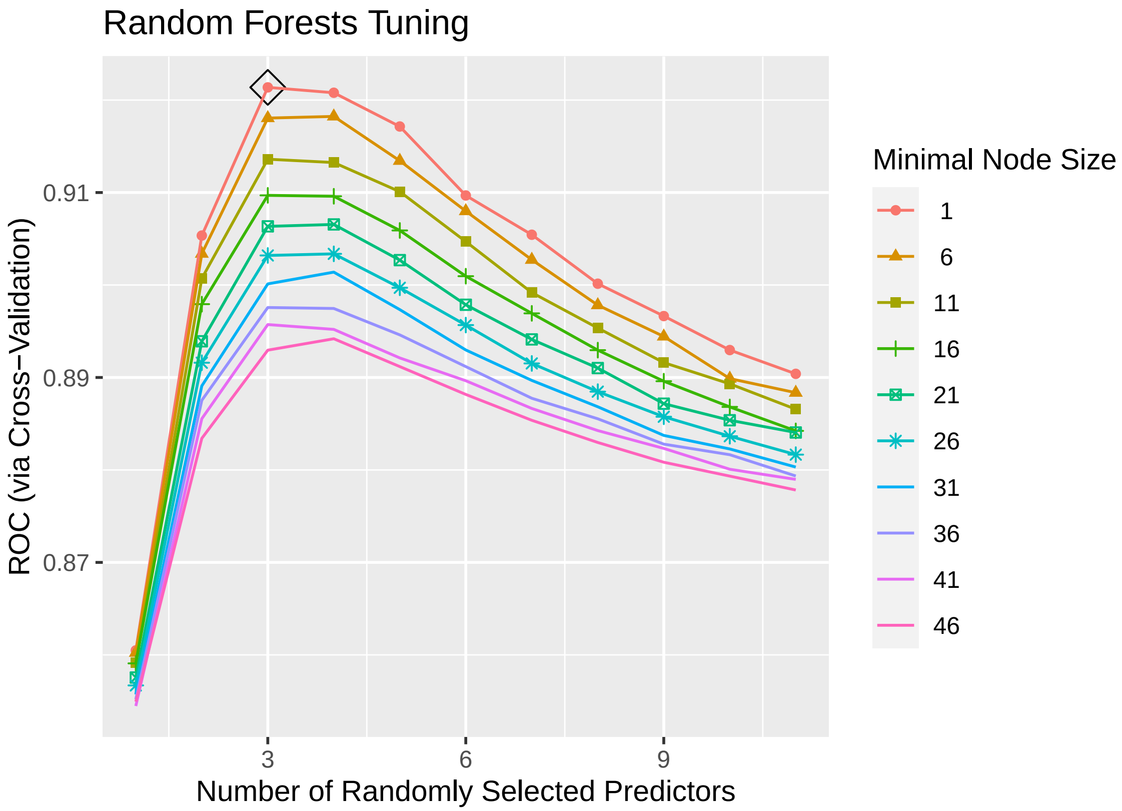


Section 3: Modeling

All variables from the dataset are included in modeling other than the arbitrary numeric identifier. No variable selection procedures prior to modeling were implemented.

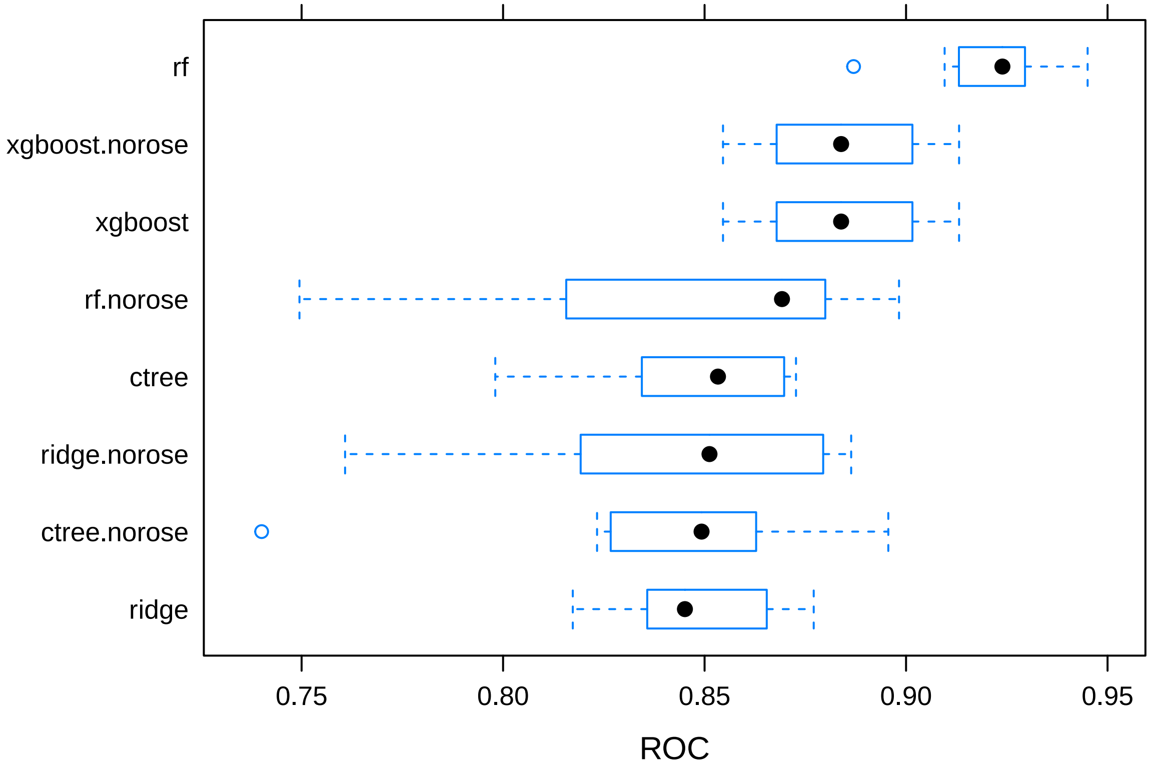
This application focuses on tree-based methods and ensemble methods, given the predictor space. Specifically, techniques used include random forests, conditional inference trees, ridge regression, and gradient boosting (xgboost). Several of these methods are utilized because they have few distributional assumptions and are adaptable to our categorical predictors (though ridge regression makes the same assumptions are MLR: linearity, constant variance, and independence). Random forests are extensions of bagging methods that randomly select subsets of features to use in collections of decision trees and makes no formal distributional assumptions (and are non-parametric). Conditional inference trees are closely related to decision trees, but uses a significance test to select inputs rather than maximizing an information measure (like Gini used in our RF). Gradient boosting is used to address the limits of bagging, using weak classifiers and weighting to find best fits and combine classifiers, assuming that observations are independent, and assumptions about the interaction depth (though this is tuned here).

All models have tuning parameters, and xgboost in particular has many of them. These were tuned using cross-validation, initially using a wider range and coarser search pattern; after locating approximate ranges for the selected parameters, I then iterated the search pattern within a narrower ranger and with more density. A visualization of the random forest tuning results can be seen here:



Randomly selected predictors and minimal node size vs ROC using Gini

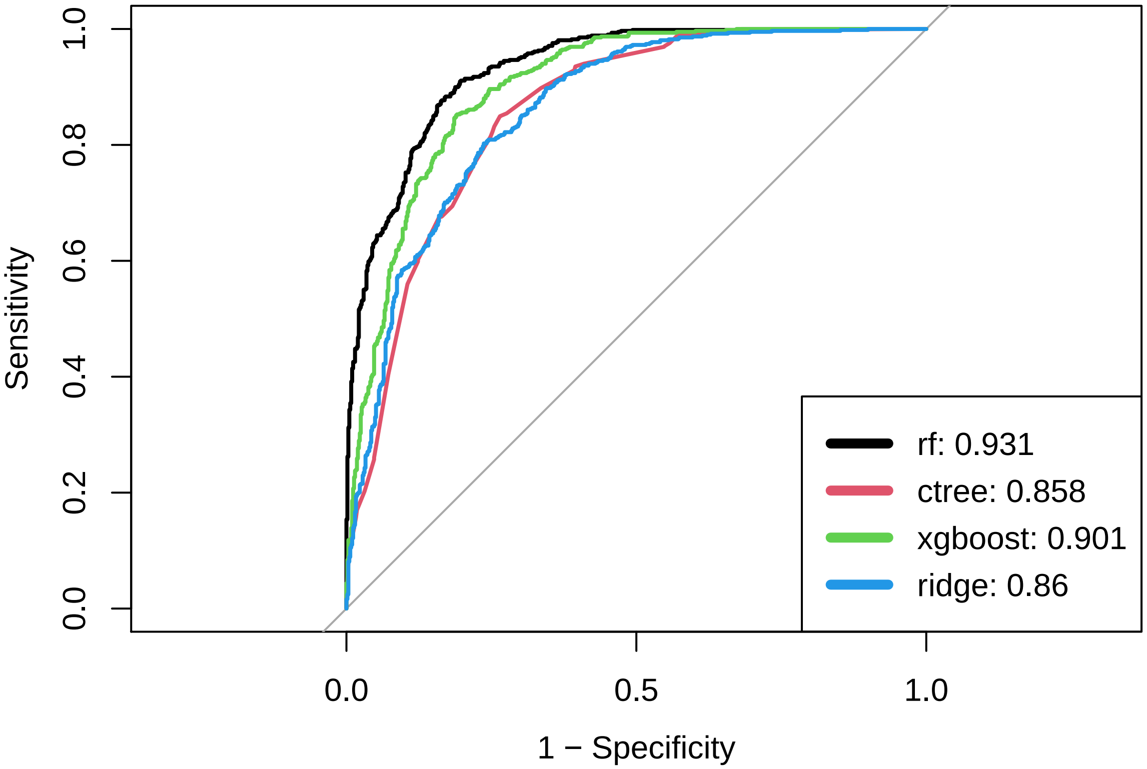
Regarding the training data performance, the random forests algorithm tuned on the oversampled data performs the highest by a relatively wide margin. All methods have ROC between 0.80 and 0.95. Overall, the more flexible and nonparametric methods perform very well. Of 8 models total (4 trained on imbalanced and 4 on ROSEd data): 3 of the top 4 were trained on the oversampled data. Training ROC performance is visualized below.

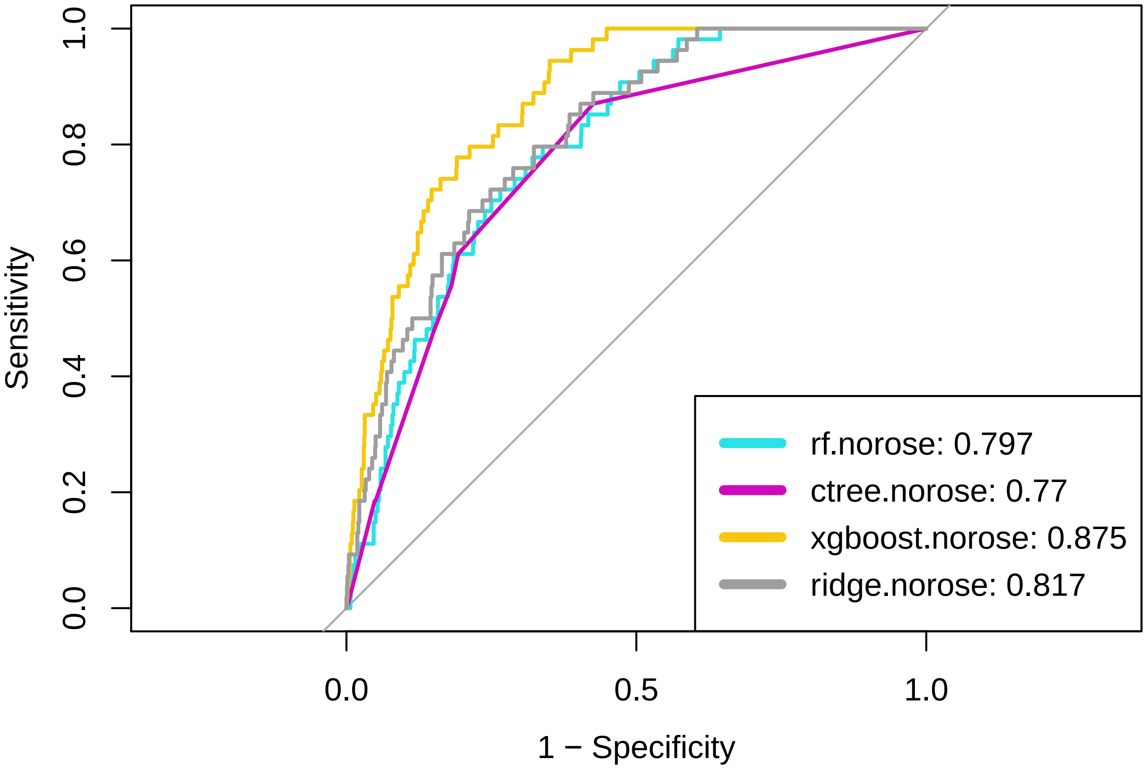


Testing data performance can be seen below, split into 2 plots to accommodate the maximum of plot.roc().

By order of AUC in the ROC curves (where "on no oversampling" is using the original imbalanced data), the test data performance is: random forest, xgboost, xgboost on no oversampling, ridge regression, conditional inference trees, ridge on no oversampling, random forest on no oversampling, conditional inference tree on on no oversampling.

Interestingly, these models are quite close, and there is no clear preference for training on the imbalanced data (contrary to expectation, since training the model on the ROSE data could have been expected to perform when predicting with a new observation).





Variable importance is computed by permutation, implemented during model training. Focusing on the model with the most successful testing performance (RF), we see age as the most important variable by far, followed by average glucose level, hypertension, BMI, and every married. Having extracted the variable importance from each model, one can see there is variability between models as to the order of variable importance, and the margin between each variable. Even between the RF and the RF trained on the imbalanced data, we see that average glucose level is much more important in the former than the latter.

All of these methods (other than perhaps ridge regression) are very limited in their interpretability; while their flexibility is particularly useful for underlying complex truths, they are black-box methods where we have limited control over the models' findings and interpretation. Since, generally speaking, bagging methods reduce variance by applying the same algorithm to a bootstrapped sample (sampling with replacement), and boosting can help to reduce bias by weighting weak learners, it is wise to apply both to better pursue the underlying truth. We know due to the bias-variance tradeoff that these characteristics are in tension, and seek to capture the underlying truth by examining both options separately. This way we can help attain accurate predictions, though we sacrifice the clinical interpretability of variables in the model.

Section 4: Conclusion

Interestingly enough, all of the methods trained on the oversampled dataset have a higher area under the curve using the test data than their counterparts (using the same tuning parameters) trained on the imbalanced dataset. This is perhaps counterintuitive given that training on the oversampled dataset could be expected to perform worse with a new observation (eg, in the test data). Overall, the methods are quite successful at classifying testing observations into the correct classes, as evidenced by the ROC curves. This suggests using such a simple dataset may be helpful in clinical decision support, or even have proactive utility.

Section 5: Bibliography

1. Boehme AK, Esenwa C, & Elkind MSV. Stroke risk factors, genetics, and prevention. Circulation Research. 2017; Retrieved 3/24/2021, from https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.116.308398

2. Roger VL, Go AS, Lloyd-Jones DM, et al.; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2011 update: a report from the American Heart Association. Circulation. 2011; 123:e18–e209. doi: 10.1161/CIR.0b013e3182009701.

3. Howard VJ, Kleindorfer DO, Judd SE, McClure LA, Safford MM, Rhodes JD, Cushman M, Moy CS, Soliman EZ, Kissela BM, Howard G. Disparities in stroke incidence contributing to disparities in stroke mortality. Annals of Neurol. 2011; 69:619–627. doi: 10.1002/ana.22385

4. Giles, MF. Rothwell, PM. 2007. Risk of stroke early after transient ischaemic attack: a systematic review and meta-analysis. Neurology. 6:12, 1063-1072, doi: 10.1016/S1474-4422(07)70274-0.

5. Goldstein LB, Adams R, Becker K, Furberg CD, Gorelick PB, Hademenos G, Hill M, Howard G, Howard VJ, Jacobs B, Levine SR, Mosca L, Sacco RL, Sherman DG, Wolf PH, Zoppo GJ. Primary prevention of ischemic stroke. Stroke. 2001; 32;280-299. doi: http://doi.org/10.1161/01.STR.32.1.280