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 is also significantly associated with stroke incidence (though not included in this dataset), thereby making investigation 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, earlier intervention in potential stroke patients improves outcomes. 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.

This is a public health example; the current study examines whether a moderately effective prediction of stroke outcome (as a binary classification) can be made using a minimal set of predictors, which are largely demographic and noninvasive. This makes for a classification question, 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 could be helpful too.

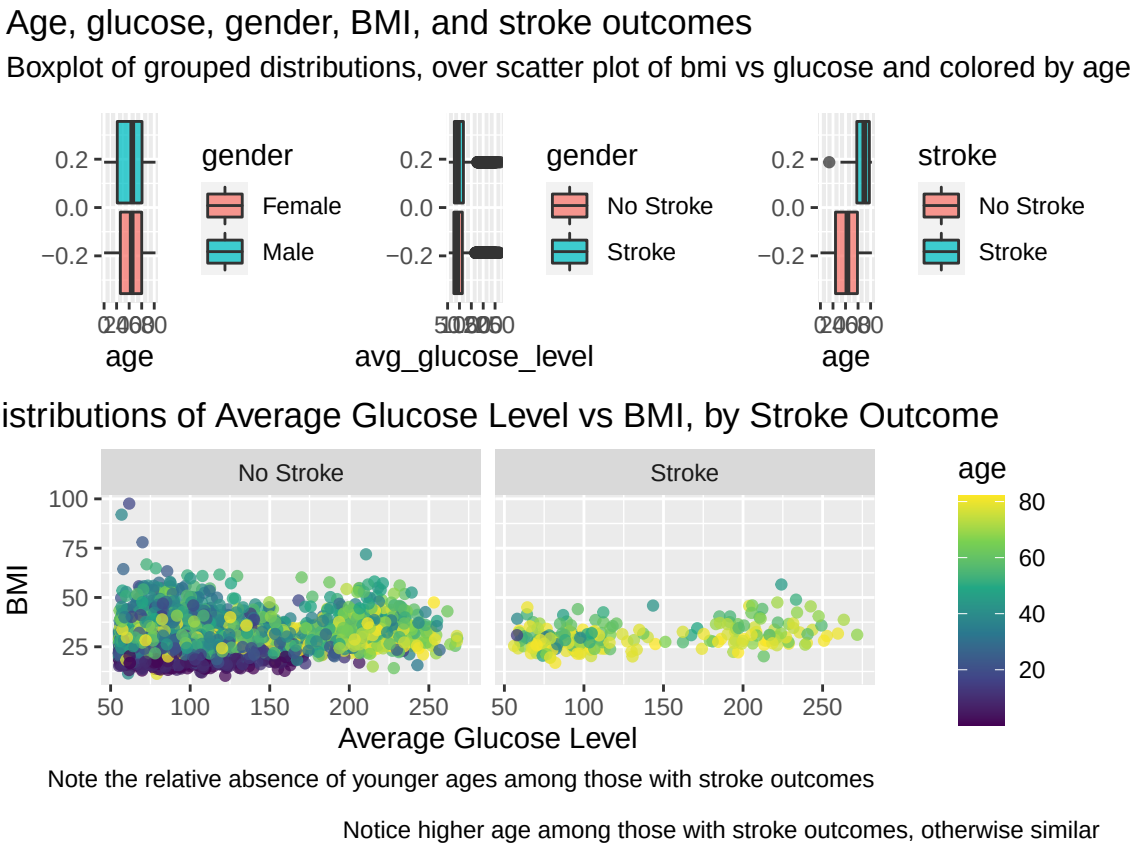
Variables were transmuted into more appropriate data types for analysis (for example, character variables into factor variables). Then, the whole dataset was analyzed for missing data, and tidyverse "data tidying" best practices were implemented to rename and reorder variables. All variables were investigated 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. The smoking has "unidentified" values that could be treated as missing but are not here (without more information from data collection process). 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 some objects (such as a predictor matrix) were created for modeling purposes, as well as any necessary transformations (e.g. centering and scaling).

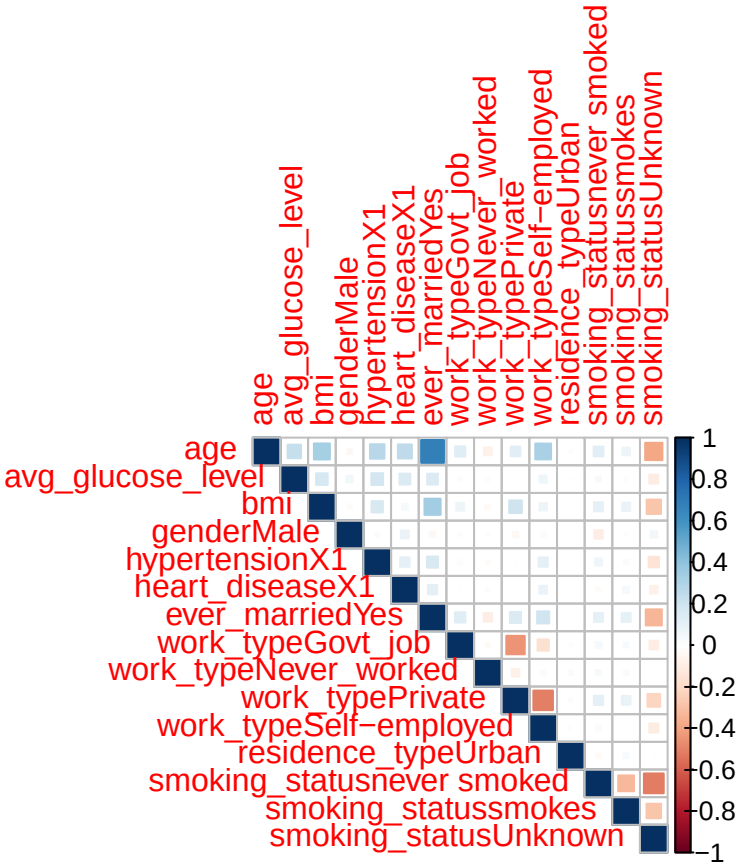
Section 2: Exploratory analysis/visualization

Data summaries, pair-wise visualizations (boxplots between groups), grouped visualizations (scatter plots with continuous variables and groups), and correlation plots were used. There are many 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 with more density around 200 units before reducing again. BMI is 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. No correlations between variables are particularly remarkable, other than a moderate positive association between age and having been ever married, and moderate negative association between private employment and self-employment (both of which are transparent, and denote some collinearity). To get a sense of the data, see below. 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.



Regarding the correlation of predictors, please see the following.

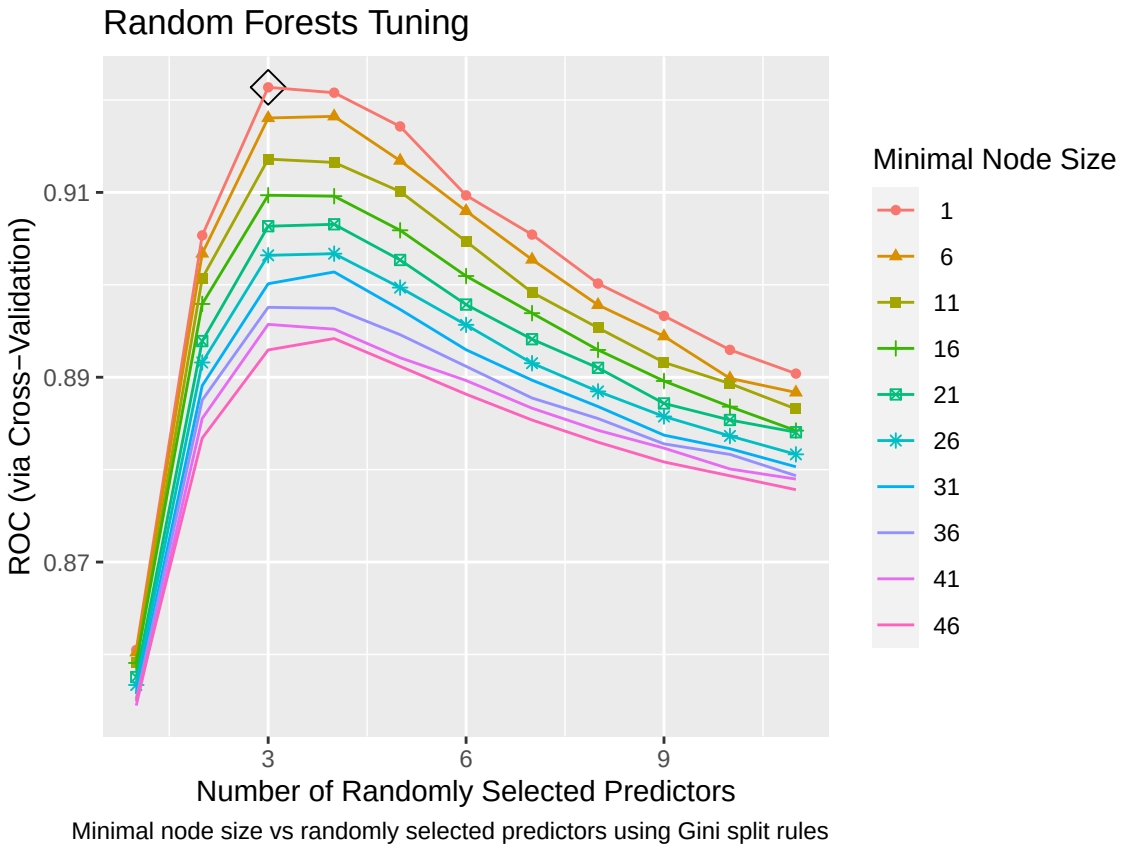


Section 3: Models

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

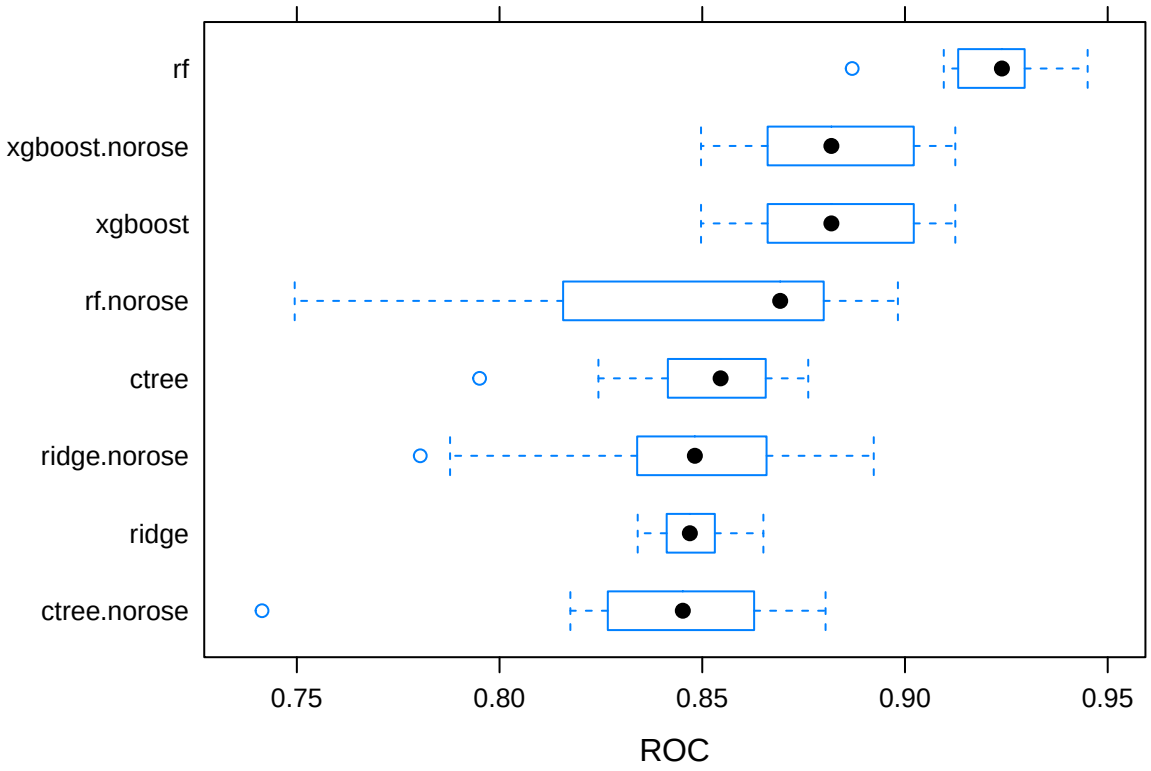
This application focuses on tree-based methods and ensemble methods, given the predictor space. Specifically, techniques used include random forests, conditional inference trees, ridge regression, and xgboost (gradient boosting). Several of these methods are utilized because they have few distributional assumptions and are adaptable to our categorical predictors (though ridge regression makes the same assumptions are MLR: linearity, constant variance, and independence). Random forests 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 the Gini coefficient used in our RF). Gradient boosting are 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 sometimes 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 can be seen here:



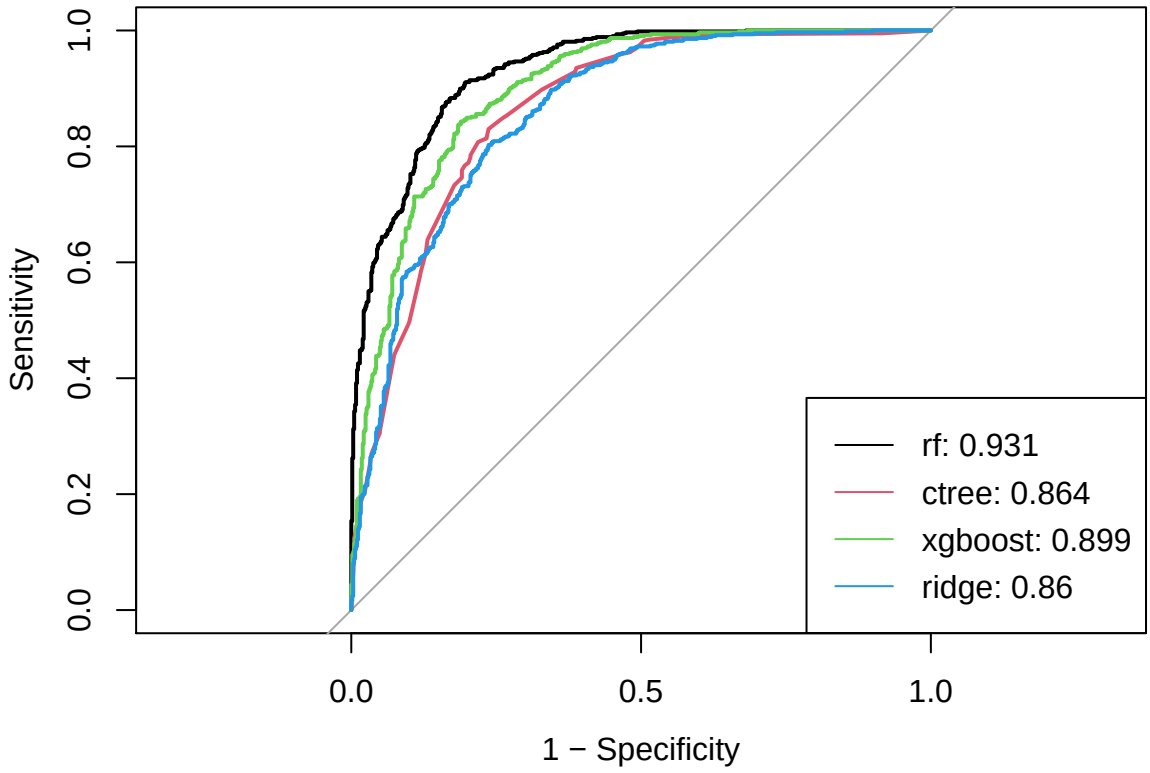
Regarding the training data performance, the random forests algorithm tuned on the oversampled data performs the highest by a relatively wide margin. All methods have ROC between 0.80 and 0.95. Overall, the more flexible and nonparametric methods perform very well. Of 8 models total (4 trained on imbalanced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

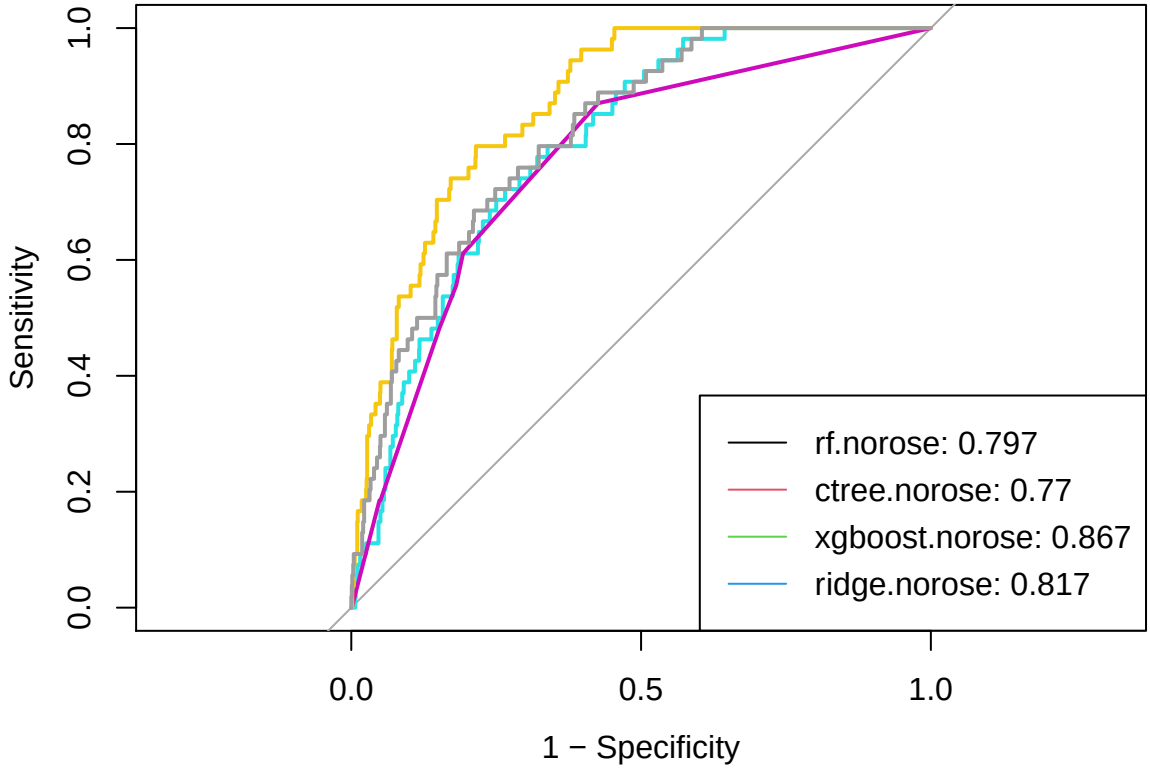
```
bwplot(res, metric = "ROC")
```



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

By order of AUC, the performance is: random forest, xgboost no ROSE, xgboost, ridge no ROSE, conditional inference tree, ridge regression, ctree no ROSE, random forest no ROSE. 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 permutations, 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 these results from each model, one can see there

is variability between model approaches as to the order of variable importance, and the distance 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 ride regression) are 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.

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.