Machine Learning for Predicting Delayed Onset Trauma Following Ischemic Stroke

Anthony Ma[1], Gus Liu[1]
Department of Computer Science, Stanford University, Stanford, CA 94305

Stroke is currently the third leading cause of death in the United States. Interestingly, however, only a small percentage of stroke patients (~15%) die immediately from the initial trauma. Some of the leading causes that eventually lead to death may be initial ischemic infarction, recurrent ischemic stroke, recurrent hemorrhagic stroke, pneumonia, coronary artery disease, pulmonary embolism, and other vascular or nonvascular causes. Most studies that apply machine learning to stroke focus on predicting the risk of having a stroke or the likelihood of survival given attributes of a patient, but not so much on likely outcomes of patients that do survive the initial stroke attack. Therefore, the goal of our project is to apply principles of machine learning over large existing data sets to effectively predict the most probable life threatening risks that may follow the first incident. In this paper, we show our models generating predictions with over four-fold accuracy compared to determining outcomes by chance. Further refinement in these algorithms could provide immense utility in clinical settings and stroke therapy.

Introduction

In this paper, we apply both unsupervised and supervised machine learning methodologies to patient profile data. First we will demonstrate (i) that features from differential diagnoses and medical interviews can be used in building classifiers that discriminate between likely outcomes of fatality. (ii) Crucial features in determining outcome can be identified through Principle Components Analysis (PCA). (iii) Unsupervised learning principles such as K-Means Clustering can be applied to group individuals into canonical patient "profiles". Appending data on cause of death, we can then gain insight on the most likely cause of death for a new patient fitting one of these profiles.

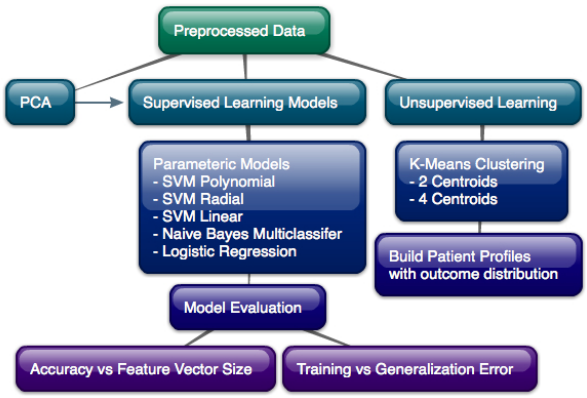


Figure 1: Outline of tools used in study

By demonstrating how the cause of fatality following ischemic stroke is highly connected to patient profile extracted from initial examination, our results in this paper serve not only as an effective classification tool, but also lay the foundations for creating more robust prediction tools for clinical applications.

Methods

Data

Patient profile data was obtained from a 6-year trial retrieved from the International Stroke Trial Database. We started with over 19,000 data points, but performed a refinement process to remove patients with incomplete patient profile, and those that remain alive, since our goal is predicting outcome of death if one were to die. In the end we generated preprocessed data set of ~4000 patients. For each patient, there are 14 features that we are focusing on: sex, age, atrial fibrillation, visible infarct under CT, aspirin, systolic blood pressure, facial, arm, leg deficit, dysphasia, hemianopia, visuospatial disorder, brainstem signs, and other deficits. Possible outcome of death DEAD1, DEAD2, … DEAD8 correspond to initial stroke, recurring ischemic, recurring hemorrhagic, pneumonia, heart disease, pulmonary embolism, other vascular, and non-vascular causes. To avoid comparing binary and continuous features, we set Age>65 and BP>150 to 1 and -1 otherwise.

Feature	Metric	Description
SEX	1, 0	Gender of patient (Male = 1, Female = 0)
AGE	1, -1	Age in years (>=65 = 1, <65 = -1)
RATRIAL	1, 0	1 = Presence of atrial fibrillation
RVISINF	1, 0	1 = Infarct visible on CT imaging
RASP3	1, 0	1 = Aspirin taken within 3 days of randomization
RSBP	1, -1	1 = Systolic blood pressure >=150
RDEF1	1,0,-1	Face Deficit (Yes, No, Can't Access)
RDEF2	1,0,-1	Arm/hand Deficit (Yes, No, Can't Access)
RDEF3	1,0,-1	Leg/foot Deficit (Yes, No, Can't Access)
RDEF4	1,0,-1	Dysphasia (Yes, No, Can't Access)
RDEF5	1,0,-1	Hemianopia (Yes, No, Can't Access)
RDEF6	1,0,-1	Visuospatial Disorder (Yes, No, Can't Access)
RDEF7	1,0,-1	Brainstem/Cerebellar signs (Yes, No, Can't Access)
RDEF8	1,0,-1	Other Deficits (Yes, No, Can't Access)

Table 1 – Feature description and quantification

Outcome	Metric	Description
DEAD1	1,0	1 = Initial Stroke
DEAD2	1,0	1 = Recurrent ischemic stroke
DEAD3	1,0	1 = Recurrent hemorrhagic stroke
DEAD4	1,0	1 = Pneumonia
DEAD5	1,0	1 = Coronary heart disease
DEAD6	1,0	1 = Pulmonary embolism
DEAD7	1,0	1 = Other vascular or unknown
DEAD8	1,0	1 = Non vascular causes

Table 2 – Outcome descriptions

Feature Selection

Principal Components Analysis (PCA) maps data of original feature dimension *n* to smaller dimension *k*. These new principal components or PCs are linear combinations of original features that carry maximal variance when data is projected onto it. Original data set is represented by only 14 features, which happen to be predictive of stroke risk according to literature. Therefore most algorithms were done on full feature dimension. Feature selection techniques such as PCA, however, can give intuition on the most important factors in determining patient outcome.

K – Means

K-Means clustering algorithm was implemented in MatLab with 2 and 4 centroids. K-means is an unsupervised learning algorithm, which clusters patient profiles into k centroids by minimizing weighted norms between data point and centroid position. We represented each patient with $\boldsymbol{p_j}$ representing a 14-dimensional vector containing profile information, and $\boldsymbol{\mu_j}$ denotes the mean of points in cluster $\boldsymbol{G_i}$.

$$\text{argmin}_G \sum_{i=1}^{k} \sum_{\boldsymbol{p}_j \in \boldsymbol{G}} ||\boldsymbol{p_j} - \mu_j||^2$$

Our algorithm runs until convergence after 2000 iterations.

Supervised Learning Algorithms

After performing PCA analysis and seeing how the original 14 features were indeed highly representative and predictive of eventual patient outcome, all supervised learning algorithms were performed on full feature dimension (n=14). Each learning algorithm was performed on varying sample size after applying randomization algorithm to choose subset of size m = {*50, 100, ... 1000*} from full data set. In addition to generalization error, training error was computed as well for KNN and multiclass logistic regression.

Naïve Bayes

Multiclass Naïve Bayes was implemented in MatLab based on frequency of observed features values and corresponding outcomes. Laplace smoothing of smoothing parameter = 1.0 was applied. Assumption of independence and Gaussian distribution was made despite some features having high correlation (i.e. facial, arm, and leg deficit usually come together).

Support Vector Machine (SVM)

Support Vector Machine was implemented in R using e1071 package. A variety of linear, Gaussian, sigmoid, and polynomial kernels were applied to predict patient fatality outcome. Training was performed on 950 patients, which generated 880 support vectors, $\gamma = 0.0714$ and a C=1 constant of the regularization term in the Lagrange formulation.

Multinomial Logistic Regression (SoftMax)

The Multinomial Logistic Regression, is a supervised learning algorithm where output can take on arbitrary k outcome classes. It requires significantly more training time than Naïve Bayes since iterative algorithms are necessary in parameter estimation. Most of the computation was done with **mnrfit** function on MatLab. In order to build the multiclass model, we estimate $\theta_1, \theta_2, \ldots \theta_k \in R^{n+1}$ parameters, where θ_i vector stores coefficients of i^{th} outcome for each n feature and intercept term. Probability of a patient being classified into certain outcome equals:

$$p(y^{(i)} = j | x^{(i)}; \theta) = \frac{\exp(\theta_j^T x^{(i)})}{\sum_{1 \le l \le k} \exp(\theta_l^T x^{(i)})}$$

Applying maximum a posteriori decision rule, we classify a new patient into outcome of highest probability.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{\exp(\theta_j^T x^{(i)})}{\sum_{1 \le l \le k} \exp(\theta_l^T x^{(i)})}\right]$$

We use the cost function as defined by and determine corresponding theta parameters.

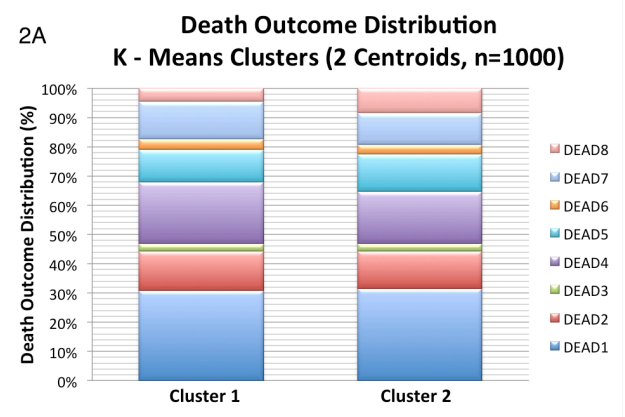
K Nearest Neighbors (KNN)

KNN classifies each new test patient based on the most popular labeling of k-nearest neighbors, as determined by the weighted norm of Euclidean distances. For our model, we have K = 3 since it is large enough to reduce noise on classification but avoids making boundaries between classes indistinguishable.

Results

	Profile 1 (n=844)	Profile 2 (n=156)	*p* - value
DEAD1	30.81 %	31.41 %	0.4919
DEAD2	13.39 %	12.82 %	0.5137
DEAD3	2.61 %	2.56 %	0.9543
DEAD4	21.09 %	17.95 %	***3.90E-04***
DEAD5	11.14 %	12.82 %	***0.0552***
DEAD6	3.67 %	3.21 %	0.59816
DEAD7	12.68 %	10.9 %	***0.0423***
DEAD8	4.62 %	8.33 %	***3.09E-05***

Table 3 – 2 Cluster K-Means profile outcome distribution



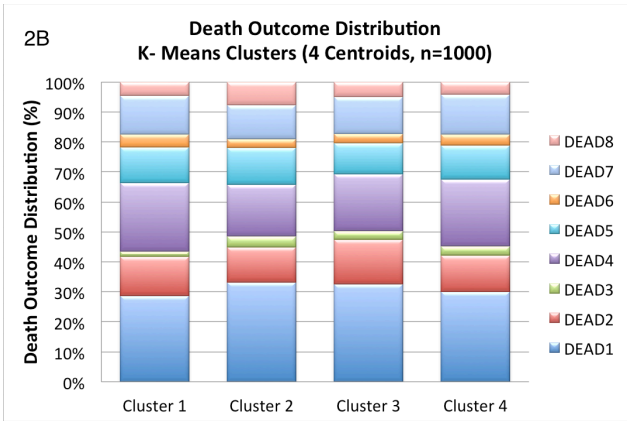


Figure 2: a.) Patients are clustered on feature profiles with 2 centroids. Outcome distributions shown below b.) Patients are clustered on feature profiles with 4 centroids. Outcome distributions shown below. Different profiles (i.e. male vs female, presence/absence of facial, arm, and leg deficit) have distinct outcome distribution.

2 Clusters:
Profile 1: M >65 N N N BP>150 Y Y Y Y N N N N
Profile 2: F >65 N N N BP>150 N Y N Y N N N N

4 Clusters:
Profile 1: F >65 N Y N BP>150 Y Y Y Y Y Y N N
Profile 2: F >65 N N N BP>150 Y Y Y N N N N N
Profile 3: M >65 N N N BP>150 Y Y Y N N N N N
Profile 4: M >65 Y N N BP>150 Y Y Y Y Y Y N N

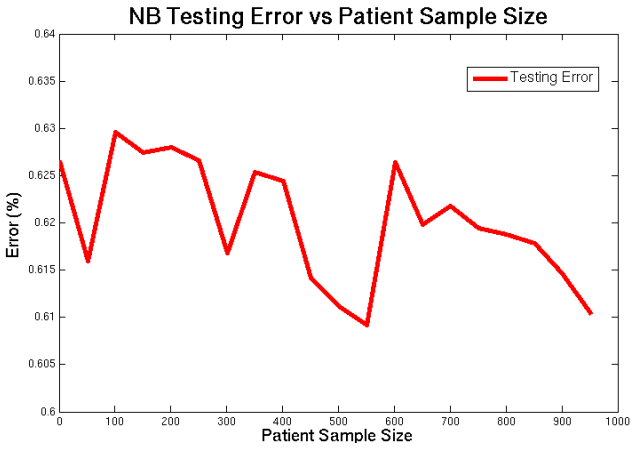


Figure 3: Testing error for Naïve Bayes multi-classifier in relation to patient sample size.

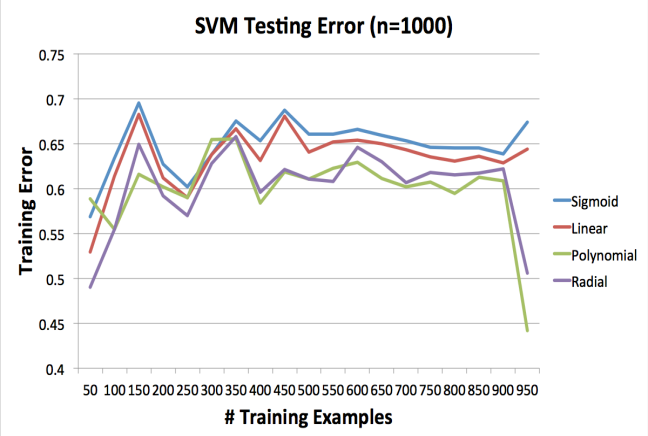


Figure 4: Testing accuracy for (n=1000) on SVM with sigmoid, linear, polynomial, and radial kernels. It is evident that polynomial and radial kernel leads to drastic decrease in generalization error as number of training examples increase past the threshold of n=1000.

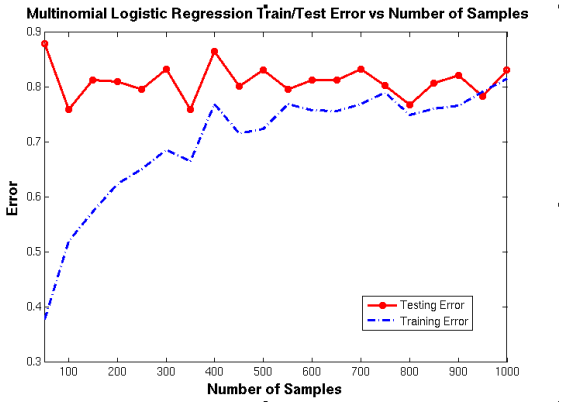


Figure 5: Training/Testing error vs number of samples. Relatively high error percentages (78%) due to failure of convergence in determining theta parameters.

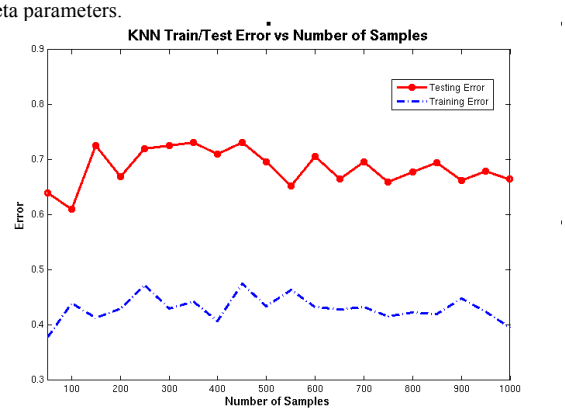


Figure 6: Training/Testing error vs number of samples. Average 67% generalization error with significantly lower testing error (41%)

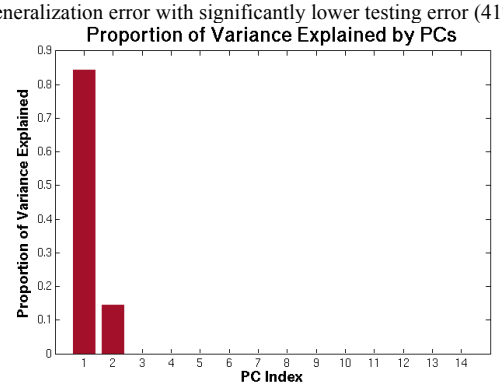


Figure 8: Percent of variance over training data accounted for by each principal component.

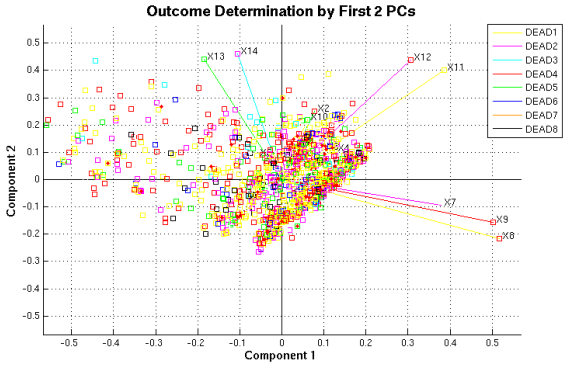


Figure 7: Principal component scores in lower dimension by first two PCs. Lack of distinct clusters observed implies poor linear separation of data.

Model	Testing Error	N Iterations
Naïve Bayes	0.60	1000
SoftMax	0.80	1000
KNN	0.66	1000
SVM (Sigmoid)	0.68	1000
SVM (Linear)	0.63	1000
SVM (Polynomial)	0.44	1000
SVM (Radial)	0.51	1000

Table 4 – Generalization error for different supervised learning algorithms

Discussion

Feature Selection

Based on literature studies on predictive factors of ischemic stroke, it is well known that features such as age, sex, blood pressure, infarct size, and craniofacial deficits are very important in determining patient outcome. Therefore, our initial hypothesis was that running our learning algorithms on full feature dimension produces lowest generalization error, which is true. Although our data is not characterized by large feature dimension, principal component analysis could still reveal some important relationships between different features as well as the predictive value each individual feature has on outcome of fatality. As shown in Figure - 8, over 99% of all variance is accounted for just by the first two principle components. Features: age, sex, and blood pressure were quite indicative of death outcome. Finally, there seems to be high correlation between arm/hand deficit and leg/foot deficit as well as hemianopia and visuospatial disorder.

Unsupervised Learning

In this project, we used K-Means clustering algorithm as both an exploratory tool to determine underlying structure within data points as well as a way to generate canonical "patient profiles" for important clinical applications. Resultant clustering representation confirms the presumed idea that gender is one of the most predictive measures for eventual outcome of death. This is evident in both the case with 2 centroids and 4 centroids. From Figure - 2A, we gain insight on the characteristics of the common stroke victims. They are profile 1: Male, age 65 or older, high-blood pressure, with facial, arm, leg deficit, and signs of dysphasia. And profile 2: Female, age 65 or older, high blood pressure, with arm/hand deficit and dysphasia. There were 844 patients corresponding to profile 1, and 156 corresponding to profile 2. After appending supervised data for these patients, calculating distribution of death outcome, and conducting a 2-sample *t*-Test to compare the differences in mean patients falling into each outcome category, we determined *p*-values for differences in outcome distribution. From Table - 3, it is evident how profile 1 patients have a statistically significant higher chance of eventually dying of pneumonia/immune system failure or other vascular causes: p = 3.90E-04 and p = .0423 respectively. Similarly profile 2 patients face much higher risks of coronary heart disease and non-vascular causes of death than former candidates: p = .0552 and p = 3.09E-05 respectively. Figure – 2B illustrates K-Means algorithm applied in generating four profiles of distinct outcome distribution.

Supervised Learning

All results from learning algorithms were performed on full feature set, which led to minimization of training and generalization error. We first implemented a Multiclass Naïve Bayes classifier as our baseline supervised, parametric model. As seen in Figure - 3, the model performed fairly well, achieving approximately 40% testing accuracy (60% error), considering there were 8 outcomes to choose from. Comparing this to percentage accuracy of random decision 12.5%, we achieved a greater than 3-fold prediction accuracy increase. As the number of classification outcomes increase, generalization error generally rises as well. Furthermore, in the context of predicting likely outcomes following initial ischemic attack, even minor increases in prediction accuracy carries high clinical utility. There are future steps to take in reducing error reducing error percentages. It was also discovered that most death outcomes corresponded to initial stroke, pneumonia, and non-vascular causes (DEAD1, DEAD4, and DEAD8 respectively) and our Naïve Bayes model almost exclusively predicted those three outcomes, thus resulting in a fairly high generalization error.

The next parametric supervised learning algorithm we explored was the multiclass logistic regression (SoftMax). As seen in Figure – 5, the algorithm's generalization error was quite high at approximately 78% for all sample sizes. Training error starts relatively low at 37% and asymptotically increases to match generalization error as sample size increases towards n=1000. This larger error value was largely due to failure upon converging on true theta parameters during training for large sample sizes with **mnrfit** MatLab software. Overall, multinomial logistic regression can only serve as a reference point, and has less capability in outcome prediction.

Taking a different approach with the non-parametric KNN classifier (with K=3), we achieved a 34% testing accuracy (67% generalization error), which is slightly worse than Naïve Bayes. Figure - 6 depicts the significantly smaller training error.

Finally, to gain more insight into the data, we used support vector machines with multiple kernel options (Figure - 4). Our SVM model using polynomial kernels provided the best accuracy of 56% when using at least 1000 training examples. This leads to the optimal 4.5 fold increase in prediction accuracy. Radial kernels performed quite well as well with 49% accuracy. Finally, linear and sigmoid kernels performed with only 37% and 36% accuracy respectively. Given that linear kernels had a relatively poor performance, we confirm the fact that our data is not linearly separable as indicated previously by PCA results. For the polynomial and radial SVMs, the generalization error decreases with increasing training examples, specifically past the threshold of n=1000. Our SVM algorithm was limited because of the extremely high

feature vector dimensions, resulting in over-fitting and inaccurate generalization to other data. Though this issue could have been mediated by extensive parameter adjustment or feature reduction, we chose not to apply these techniques because of the complexity of our data, the significance of all of our utilized features, and already having reduced our error significantly.

Conclusion

In this article, we demonstrated how Naïve Bayes and support vector machines (polynomial) leads to maximal outcome prediction accuracy of 40 %, and 56% respectively in classifying 8 different death outcomes following initial ischemic trauma, using 14 crucial features. Comparing to the average predictive accuracy following randomization (12.5%), our best algorithm achieves up to a 4.5 fold prediction accuracy increase, which carries immense clinical utility in improving a patient's chance of survival and quality of life. With the combination of unsupervised learning algorithms such as K-Means and supervised outcome data, we also built canonical "profiles" of the most common patients doctors are likely to encounter following initial stroke attack. Each one represents a corresponding distribution of death outcomes. New patients can then be fitted into the most representative profile and plan of action will be taken to minimize chances of the most likely ensuing risks.

Future Directions

Future work involves discovering predictive value of individual features and their relationship/correlation strength with each other. We also plan on extending our classification model to patients that do not die immediately after initial ischemic infarction. Furthermore, our models can take into account the likelihood of outcomes at different timespans after initial attack (14 days, 6 months, 1 year, etc…) such that physicians can gain intuition on optimal treatment plans based on particular stage of patient recovery. Finally, many more related studies may be done using similar learning tools but starting with different sorts of initial trauma (i.e. hemorrhagic stroke, thrombotic stroke, transient ischemic attack, …) as well as discovering the role of pre and post-conditioning factors on survival rates.

References

[1] Ishikawa, H., N. Tajiri, J. Vasconcellos, Y. Kaneko, O. Mimura, M. Dezawa, and C. V. Borlongan. "Ischemic Stroke Brain Sends Indirect Cell Death Signals to the Heart." *Stroke* 44.11 (2013): 3175-182. Web.

[2] "Machine Learning Blog & Software Development News." *Datumbox*. N.p., n.d. Web. 24 Nov. 2014.

Noback, Charles R. *The Human Nervous System: Structure and Function*. Totowa, NJ: Humana, 2005. Print.

[3] Sandercock, Peter Ag, Maciej Niewada, and Anna Członkowska. "The International Stroke Trial Database." *Trials* 12.1 (2011): 101. Web.

[4] Shan, Li-Yang, Ji-Zhao Li, Ling-Yun Zu, Chen-Guang Niu, Albert Ferro, Ying-Dong Zhang, Le-Min Zheng, and Yong Ji. "Platelet-Derived Microparticles Are Implicated in Remote Ischemia Conditioning in a Rat Model of Cerebral Infarction." *CNS Neuroscience & Therapeutics* 19.12 (2013): 917-25. Web.

[5] Stetler, R. Anne, Rehana K. Leak, Yu Gan, Peiying Li, Feng Zhang, Xiaoming Hu, Zheng Jing, Jun Chen, Michael J. Zigmond, and Yanqin Gao. "Preconditioning Provides Neuroprotection in Models of CNS Disease: Paradigms and Clinical Significance." *Progress in Neurobiology* (2014): n. pag. Web.

[6] "Types of Stroke." Johns Hopkins, n.d. Web. 24 Nov. 2014.

[7] "Understanding Stroke Risk." *Understanding Stroke Risk*. N.p., n.d. Web. 24 Nov. 2014.