A Feature Analysis of Risk Factors for Stroke in the Middle-Aged Adults

Focused on Perception of Sudden Speech and Language Problem

Haewon Byeon
Department of Speech Language Pathology & Audiology
Nambu University, Gwangju, Republic of Korea

Hyeung Woo Koh*
Jeju Seogwipo Medical Center
Jeju, Republic of Korea

***Abstract*—In order to maintain health during middle age and achieve successful aging, it is important to elucidate and prevent risk factors of middle-age stroke. This study investigated high risk groups of stroke in middle age population of Korea and provides basic material for establishment of stroke prevention policy by analyzing sudden perception of speech/language problems and clusters of multiple risk factors. This study analyzed 2,751 persons (1,191 males and 1,560 females) aged 40–59 who participated in the 2009 Korea National Health and Nutrition Examination Survey. Outcome was defined as prevalence of stroke. Set as explanatory variables were age, gender, final education, income, marital status, at-risk drinking, smoking, occupation, subjective health status, moderate physical activity, hypertension, and sudden perception of speech and language problems. A prediction model was developed by the use of a C4.5 algorithm of data-mining approach. Sudden perception of speech and language problems, hypertension, and marital status were significantly associated with stroke in Korean middle aged people. The most preferentially involved predictor was sudden perception of speech and language problems. In order to prevent middle-age stroke, it is required to systematically manage and develop tailored programs for high-risk groups based on this prediction model.**

Keywords*—C4.5; stroke; decision tree; risk factor; speech problem

I. Introduction

Stroke is a generic term for both cerebral infarction caused by the blockage of blood vessel in the brain and cerebral hemorrhage caused by the rupture of blood vessel (in the brain). As of 2013, death rate from cerebrovascular diseases was 50.2 persons per 100,000, which is the second highest right after cancer [1]. This order of death rate has not changed over the last 10 years and especially, stroke is serous in that it takes the second place in the cause of death regardless of gender.

Incidence of stroke is high in old age. According to 2013 Annual Report on the Cause of Death Statistics, death rate of cerebrovascular disease was 10.1 persons per 100,000 for people in their 40s compared to 277.4 for 70s, which is approximately 27 times higher [1]. In terms of life cycle, however, death rate of stroke skyrockets from 40s and over the recent 20 years, increase rate of stroke is the highest in 40s and 50s [2]. In addition, it has been reported that health risk behaviors causing stroke is most frequent in middle age [3]. Therefore, in order to maintain health during middle age and achieve successful aging, it is important to elucidate and prevent risk factors of middle-age stroke.

In particular, in the case of stroke, even though operation is performed successfully, not only is the disease highly likely to accompany disabilities such as speech impediment during rehabilitation process but the patients also are likely to experience loss of labor. Middle age is the period when one accomplishes his/her goal of life. Acute diseases such as stroke not just are the direct cause of loss of job but cause enormous economic loss as well [4]. As of 2011, socio-economic loss from stroke (e.g. medical cost, transportation, nursing care, loss of production, etc.) in Korea surpassed U$ 3.5 billion and among them, social cost for middle-aged people from age 40 to 50 (45% of total cost) was reported to be the greatest [5].

Although it is important to comprehend and systematically manage high-risk groups of middle-age stroke, risk factors of middle-age stroke are less known than old-age stroke and there is also lack of studies on its risk groups. So far, chronic diseases such as diabetes, hyperlipidemia and high blood pressure and life style factors such as smoking, drinking, eating habits and exercise and social and economic status are known to be risk factors of middle-age stroke [6][7][8][3].

However, since preceding studies which investigated risk factors of stroke did not adjust socio-economic factors such as occupation and level of income, it is difficult to find out social factors of middle-age stroke [9][10]. Moreover, as health risk behaviors tend to cluster together rather than individually exist (separate from other factors) [11], investigation on individual risk factor has a limitation in identifying high-risk groups of cardiocerebrovascular diseases with various characteristics.

Especially, recent studies reported that perception of sudden speech/language problems are major warning signs of stroke and in a survey on Korean adults, 80% of stroke patients perceived speech/language problems as a warning sign of stroke and 98% of stroke patients visited medical institutions due to speech/language problems as a warning sign, which is translated that perception of speech/language problem is a major factor of warning sign for stroke [12]. If high-risk groups are comprehended and managed by considering risk factors and warning signs of stroke, significant portion of strokes can be prevented and the time required to respond to emergency situation can also be reduced.

Recently, as a method of exploring multiple risk factors of diseases, data-mining analysis such as decision tree is being

used [13]. Use of data-mining can facilitate comprehension of attributes of diseases as well as multiple risk factors.

Since tendency of occurrence and risk factors of stroke differ depending on ethnicity and culture, in order to prevent stroke in Korea, it is necessary to develop a stroke prediction model reflecting demographic characteristics of middle age population of Korea and, based on it, manage them systematically.

This study investigated high risk groups of stroke in middle age population of Korea and provides basic material for establishment of stroke prevention policy by analyzing sudden perception of speech/language problems and clusters of multiple risk factors. Organization of this study is as follows; chapter 2 explains data resources and definition of variables and chapter 3 explains procedure for development of prediction model; chapter 4 suggested results of developed prediction model and chapter 5 presents results and suggests direction for future studies.

II. Methods

A. Sources of data

Study subjects were adults aged 40–59 who participated in the 2009 Korea National Health and Nutrition Examination Survey (KNHANES), a nationwide representative survey of the non-institutionalized population in the Republic of Korea, and who then participated in an health survey [14].

The KNHANES is a nationwide cross-sectional survey conducted annually by The Korea Centers for Disease Control and Prevention. It employs a rolling sampling design that uses a complex, stratified multistage probability cluster survey of representative non-institutionalized civilians. The KNHANES sampling process is described in detail elsewhere [14]. Briefly, the creators of the survey redesign the KNHANES from once every years to once every year in order to provide timely health statistics for monitoring changes in health risk factors and diseases and developing associated public health policies and health programs. The 2009 KNHANES, conducted in January to December, was composed of three component surveys: a health interview, health examination, and nutrition survey. Trained medical staff and interviewers performed the health interview and health examination at a mobile examination center and at participants' households. The 2009 KNHANES was conducted on 12,722 persons out of 4,000 households with a participation rate of 82.8% (n=10,533).

This study targeted 2,885 persons who completed both the health survey and examination. Of these, 134 persons whose nonrespondents were excluded from the research, and data from 2,751 persons (1,191 males and 1,560 females) were analyzed.

B. Measurements

Outcome was defined as prevalence of stroke. Explanatory variables were included as age (40~49, 50~59), sex, final education (high school and lower, over college), Occupation (economically inactive, manual workers, non-manual workers), income (quartiles), marital status (living with spouse, living without spouse, unmarried person), at-risk drinking (yes, no), smoking (non-smoker, past smoker, current smoker), subjective health status (good, fair, poor), moderate physical activity (yes, no), Diabetes (yes, no), hypertension (yes, no), sudden perception of speech and language problems (yes, no).

High-risk drinking was classified into normal (less than 12 points) and high-risk drinking (over 12 points) by using alcohol use disorder identification test (AUDIT) [15]. Regular moderate physical activity was defined as practicing moderately breathless exercise for more than 30 minutes per session over 5 days a week. Occupations classified based on the Korean Standard Classification of Occupations (KSCO-06)[16] were reclassified into economically inactive (unemployed person, homemaker), non-manual (managers & professionals, clerical support workers, service & sales workers), and manual (skilled agricultural & forestry & fishery workers, craft & plant and machine operators and assemblers, and unskilled laborers) occupations.

III. Statistical Analysis

A. Exploration on factors related to the stroke

For general characteristics, mean and percentage were presented and difference between groups based on stroke was analyzed by Chi-square test.

B. C4.5 algorithm

C4.5 is a decision tree algorithm developed by Quinlan [17], purpose of which is to create a tree which can exactly classify outcomes even with small number of tests. This algorithm constructs the simplest decision tree by using the concept of entropy based on information theory [18] (Figure 1).

In general, entropy means numbers representing disorder. As data sources are mixtures of proper cases and improper ones, they are very high in the degree of disorder. However, degree of disorder becomes 0 since terminal nodes are decided with one grade after decision tree is learned. Thus, it calculates information gain of each factor while it classifies data, keeping entropy close to 0.

Then, if the attribute with highest discerning power is selected as standard of classification, it makes as many branches as the number of kinds of given attribute values. Cases are divided according to the value of each branch and same processes are repeated in each branch. If there is no more decrease in information, the division stops [19].

Method of dividing tree by C4.5 algorithm is as follows; First, information gain of root node is acquired at input variables where target variables are composed of p and n.

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad (1)$$

Second, the gain is acquired which decreases degree of disorder in the case it is divided by attribute A or variable A at root node.

$$gain(A) = I(p,n) - E(A) \quad (2)$$

Third, among various attributes, node is divided by the attribute with greatest gain. If the divided node is composed only of either p or n, the node stops multiplying.

In case incidence rate is low as the outcome of this study, (which is) prevalence rte, there may be problems due to unbalanced data distribution [20]. In order to complement this unbalanced distribution, this study adjusted data balance by asymmetrically setting weight of misclassification costs considering prevalence rate of middle-age stoke in Korea [21]. Validity of the developed model was assessed with 10-fold cross-validation method.

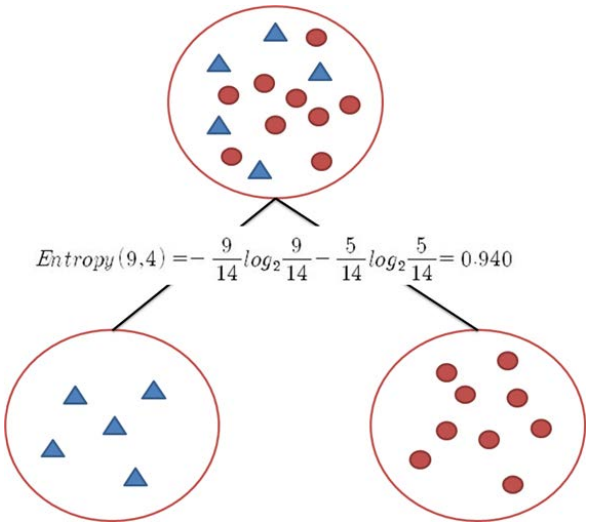

Fig. 1. Calculation of entropy

IV. Results

A. Characteristics characteristics of subjects and potential factors related to stroke

General characteristics of subjects and factors related to stroke are presented in Table 1. Among the total of 2,751 subjects, number of those who have stroke was 33 (1.2%).

As the result of chi-square test, prevalence of stroke has statistically significant difference in age, gender, final education, income, marital status, diabetes, hypertension, and sudden perception of speech and language problems ($p<0.05$).

The prevalence of stroke was higher in aged 50~59 (1.8%), man (1.8%), high school and lower (1.5%), Groups the lowest income (3.4%), unmarried person (3.7%), those with bad subjective health (2.4%), diabetes (4.2%), hypertension (4.5%), and sudden perception of speech and language problems (83.4%).

TABLE I. General Characteristics of the Subjects Based on Stroke (Univariate Analysis), N (%)

Characteristics	Stroke		p
	No (n=2,718)	*Yes (n=33)*	
Age			0.014
40~49	1,488 (99.3)	11 (0.7)	
50~59	11,230 (98.2)	22 (1.8)	
Sex			0.018
Male	1,170 (98.2)	21 (1.8)	
Female	1,548 (99.2)	12 (0.8)	
Education			0.036
High school and lower	2,032 (98.5)	30 (1.5)	
Over college	676 (9.6)	3 (0.4)	
Occupation			0.070
Economically inactive	740 (98.0)	15 (2.0)	
Non-manual workers	1,084 (99.1)	10 (0.9)	
Manual workers	877 (99.1)	8 (0.9)	
Income (quartiles)			<0.001
Q1	315 (96.6)	11 (3.4)	
Q2	583 (98.3)	10 (1.7)	
Q3	819 (99.0)	8 (1.0)	
Q4	977 (99.6)	4 (0.4)	
Marital status			0.002
Living with spouse	2,366 (99.1)	22 (0.9)	
Living without spouse	300 (97.1)	9 (2.9)	
Unmarried person	52 (96.3)	2 (3.7)	
At-risk drinking			0.769
No	1,812 (99.0)	19 (1.0)	
Yes	586 (98.8)	7 (1.2)	
Smoking			0.284
Non-smoker	1,626 (99.0)	33 (1.2)	
Past smoker	475 (98.1)	9 (1.9)	
Current smoker	617 (98.7)	8 (1.3)	
Moderate physical activity			0.250
Yes	449 (99.3)	3 (0.7)	
No	2,261 (98.7)	30 (1.3)	
Subjective health status			0.009
Good	1,175 (99.0)	12 (1.0)	
Fair	966 (99.3)	7 (0.7)	
Poor	568 (97.6)	14 (2.4)	
Diabetes			<0.001
Yes	160 (95.8)	7 (4.2)	
No	2,558 (99.0)	26 (1.0)	
Hypertension			<0.001
Yes	425 (95.5)	20 (4.5)	
No	2,293 (99.4)	13 (0.6)	
Sudden perception of speech and language problems			<0.001
Yes	2 (16.6)	10 (83.4)	
No	2,718 (99.2)	21 (0.8)	

B. Prediction model for stroke using C4.5 algorithm

Prediction model for stroke using C4.5 algorithm is presented in Figure 2. As the result of constructing statistical classification model using C4.5 algorithm after including variables set as factors related to stroke through chi-squared test, factors having significant effect were sudden perception of speech and language problems, hypertension, and marital

status. The most preferentially involved predictor was sudden perception of speech and language problems.

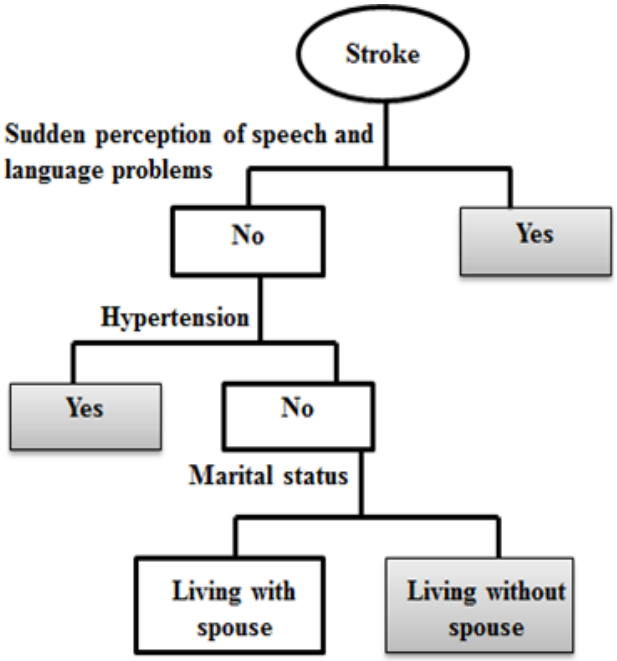

Fig. 2. Prediction model for stroke among Korean middle aged people

Table 2 is a profit chart of prediction model for stroke by C4.5 algorithm suggested in the higher order of path for subjects' improved gain. When this study drew out profit indicator for each node to seek out prediction paths for stroke, 3 nodes were confirmed as significant paths which effectively predict the stroke.

The first path with the biggest profit indicator for the prediction of the stroke was “middle-aged persons from the age of 30 to 58 who currently perceive sudden language problems” and its profit indicator was 8336.4%.

The second path was “middle-aged persons from the age of 30 to 58 who currently do not have sudden language problems or high blood pressure and do not live with spouse due to divorce or bereavement” and its profit indicator was 735.6%.

The third path was “middle-aged persons from the age of 30 to 58 who currently do not have sudden speech/language problems but have high blood pressure” and its profit indicator was 210.3%.

When the analysis on the prediction model by CART algorithm was completed, this study conducted 10-fold cross-validation test to assess developed prediction model. As the result of the 10-fold cross-validation test to compare stability of drawn-out model, drawn-out risk index was 0.360 and misclassification rate was 36% for cross classification model, showing the same risk index 0.352 and misclassification rate 35% of prediction model.

TABLE II. GAINS CHART OF PREDICTOR VARIABLE BY C4.5 ALGORITHM

Node no	Node n (%)[1]	Gain n (%)[2]	Response %[3]	Gain Index %[4]	Group
2	12 (0.4)	12 (36.4)	83.4	1336.4	Middle-aged persons from the age of 30 to 58 who currently perceive sudden language problems
6	34 (1.2)	3 (9.1)	8.8	735.6	Middle-aged persons from the age of 30 to 58 who currently do not have sudden language problems or high blood pressure and do not live with spouse due to divorce or bereavement
3	436 (15.8)	11 (33.3)	2.5	210.3	Middle-aged persons from the age of 30 to 58 who currently do not have sudden speech/language problems but have high blood pressure

[1] Node n(%); node number, % to 2,751
[2] Gain n(%); gain number, % to 33
[3] Response (%): The fraction of the stroke
[4] Gain index (%):= 1336.4 in total 4 node

V. CONCLUSION

Early detection and management of high-risk groups of stroke enables healthy and happy aging. This study developed prediction model for middle-age stroke by using C4.5 algorithm. As the result of constructing stroke prediction model considering multiple risk factors, perception of sudden speech/language problems, high blood pressure and marital status were significant prediction factors for middle-age stroke and among them, perception of sudden speech/language problem was the most prioritized prediction factor. Numerous preceding studies have reported that perception of sudden speech/language problems is a major risk factor of stroke and those who perceived speech problem had higher rate of stroke [22][23]. However, these studies were limited to exploring individual risk factors while this study confirmed as the result of exploring multiple risk factors that combination of individual risk factors causes a synergy effect.

Another finding of this study was that marital status is major prediction factor for stroke. This study found out that middle-aged people from the age of 30 to 58 who do not live with spouse due to divorce, bereavement or separation are high-risk group for stroke. It is supposed that middle-aged

people who do not live with spouse have high risk of stroke since the middle-aged living alone not only have frequent health risk behaviors such as smoking but also are more vulnerable in health management. According to studies which researched on the relationship between marital status and health, married men who lived away from family had higher risk of accidents, alcohol and substance addiction, depression, death and cardiocerebrovascular diseases than men with stable marriage life and had 2.3 times more suicide rate, 4.7 times more death rate from alcohol and alcohol addiction and 1.7 times more death rate from cardiocerebrovascular diseases [24].

Especially, it has been reported that unstable marriage states such as divorce, separation and bereavement have negative effect on cardiocerebrovascular system by causing depression, which in turn increases death risks [25]. Hence, in order to prevent middle-age stroke, it is necessary to develop health management programs for the middle-aged without spouse. Furthermore, it is also necessary to prescribe guidelines for the prevention of middle-age stroke so that they will immediately visit medical institutions when they perceive sudden speech/language problems even if they do not have stroke-related diseases such as high blood pressure and diabetes.

Results of this study are expected to be an important ground to be considered in the strategy to prevent and manage stroke. In order to prevent middle-age stroke, it is required to systematically manage and develop tailored programs for high-risk groups based on this prediction model.

ACKNOWLEDGMENT

The author wish to thank the Korea Centers for Disease Control and Prevention that provided the raw data for analysis.

REFERENCES

[1] Statistics Korea, Cause of death statistics 2013. Daejeon, Statistics Korea, 2013.

[2] Ministry of Health and Welfare, 2001 National Health and Nutrition Survey. Seoul, Ministry of Health and Welfare, 2002.

[3] S. Kaffashian, A. Dugravot, E. J. Brunner, S. Sabia, J. Ankri, M. Kivimäki, and A. Singh-Manoux, Midlife stroke risk and cognitive decline: A 10-year follow-up of the Whitehall II cohort study. Alzheimer's & Dementia, vol. 9, no. 5, pp. 572–579, 2013.

[4] H. J. Lee, and M. Yi, Adjustment of middle-aged people with hemiplegia after a stroke. Journal of Korean Academy of Nursing, vol. 36, pp. 792–802, 2006.

[5] National Rehabilitation Center, Report on the socio-economic costs of disorders. Seoul, National Rehabilitation Center, 2015

[6] R. Behrouz, and C. J. Powers, (2015). Epidemiology of classical risk factors in stroke patients in the Middle East. European Journal of Neurology, E-pub, DOI: 10.1111/ene.12742, 2015.

[7] N. Allen, J. D. Berry, H. Ning, L. Van Horn, A. Dyer, and D. M. Lloyd-Jones, Impact of blood pressure and blood pressure change during middle age on the remaining lifetime risk for cardiovascular disease: the cardiovascular lifetime risk pooling project. Circulation, vol, 125, no. 1, pp. 37–44, 2012.

[8] J. Addo, L. Ayerbe, K. M. Mohan, S. Crichton, A. Sheldenkar, R. Chen, C. D. Wolfe, and C. McKevitt, Socioeconomic status and stroke an updated review. Stroke, vol. 43, no. 4, pp. 1186–1191, 2012.

[9] J. B. Olesen, G. Y. Lip, M. L. Hansen, P. R. Hansen, J. S. Tolstrup, J. Lindhardsen, C. Selmer, O. Ahlehoff, A. M. Olsen, G. H. Gislason, and C. Torp-Pedersen, Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. British Medical Journal, vol. 342, doi: http://dx.doi.org/10.1136/bmj.d124, 2011.

[10] L. Friberg, L. Benson, M. Rosenqvist, and G. Y. Lip, Assessment of female sex as a risk factor in atrial fibrillation in Sweden: nationwide retrospective cohort study. British Medical Journal, vol. 344, doi: 10.1136/bmj.e3522, 2012.

[11] H. Byeon, and Y. Lee, Laryngeal pathologies in older Korean adults and their association with smoking and alcohol consumption. Laryngoscope vol. 123, no. 2, pp. 429–433, 2013.

[12] Y. H. Lee, Y. T. Kim, G. J. Oh, N. H. Kim, K. H. Cho, H. Y. Park, H. S. Lee, Y. S. Ha, J. Cheong, J. K. Park, K. S. Lee, and H. S. Kim, Effects of community-based education and advocacy intervention on public awareness about the warning signs of stroke and the golden window of time. Korean Journal of Health Promotion, Vol.32, No.1, pp.1–10, 2015.

[13] H. Wimmer, and L. Powell, A comparison of the effects of K-anonymity on machine learning algorithms. International Journal of Advanced Computer Science and Application. Vol. 5, No. 11, pp. 155–160, 2014.

[14] Korea Centers for Disease Control and Prevention. The Korea national health and nutrition examination survey 2008, Seoul, Korea Centers for Disease Control and Prevention, 2009.

[15] D. F. Reinert, and J. P. Allen, The alcohol use disorders identification test (AUDIT): a review of recent research. Alcoholism: Clinical and Experimental Research, vol. 26, no. 2, pp. 272–279, 2012.

[16] Korea National Statistical Office. The Korean standard classification of occupations, Daejeon, Korea National Statistical Office, 2007.

[17] J. R. Quinlan, C4. 5: programs for machine learning. Burlington, Elsevier, 2014.

[18] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, vol. 4, no. 2, pp. 13–19, 2014.

[19] H. Byeon, Development of prediction model for endocrine disorders in the Korean elderly using CART algorithm. International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 125–129, 2015.

[20] P. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, Boston, Addison Wesley, 2006.

[21] H. Byeon, The risk factors of laryngeal pathology in Korean adults using a decision tree model. Journal of Voice, vol. 29, no. 1, pp. 59–64, 2015.

[22] D. W. Dietrich, N. J. Okon, D. V. Rodriguez, and A. M. Burnett, J. A. Russell, M. J. Allen, C. C. Fogle, S. D. Helgerson, D. Gohdes, T. S. Harwell. Rural community knowledge of stroke warning signs and risk factors. Preventing Chronic Disease, vol. 2, no. 2, pp. 1–8, 2005.

[23] T. G. Robinson, A. Reid, V. J. Haunton, A. Wilson, and A. R. Naylor, The face arm speech test: does it encourage rapid recognition of important stroke warning symptoms?. Emergency Medicine Journal, vol. 30, no. 6, pp. 467–471, 2013.

[24] W. G. Ringback, B. Burstom, and M. Rosen, Premature mortality among lone fathers and childless men. Social Science & Medicine, vol. 59, no. 2, pp. 1449–1459, 2004.

[25] P. L. Morris, R. G. Robinson, and J. Samuels, Depression, introversion and mortality following stroke. Australian and New Zealand Journal of Psychiatry, Vol. 27, No. 3, pp. 443–449, 1993.