Journal of Applied Biomedicine 16 (2018) 40–45

Contents lists available at ScienceDirect

Journal of Applied Biomedicine

journal homepage: www.elsevier.com/locate/jab

Original research article

A novel algorithm for identifying risk factors for rare events: Predicting transient ischemic attack in young patients with low-risk atrial fibrillation

Chieh-Yu Liu[a,b,*], Hui-Chun Chen[c]

[a] National Taipei University of Nursing and Health Sciences, College of Health Technology, Department of Speech Language Pathology and Audiology, Biostatistical Consulting Lab, Taipei, Taiwan
[b] National Taipei University of Nursing and Health Sciences, College of Nursing. Department of Midwifery and Women Health Care, Taipei, Taiwan
[c] National Taipei University of Nursing and Health Sciences, College of Nursing, Department of Nursing, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 14 April 2017
Received in revised form 6 July 2017
Accepted 13 October 2017
Available online 6 November 2017

Keywords:
Multiple correspondence analysis
Hierarchical cluster analysis
Atrial fibrillation
Transient ischemic attack
Stroke
Logistic regression

ABSTRACT

Identification of risk factors for transient ischemic attack (TIA) is crucial for patients with atrial fibrillation (AF). However, identifying risk factors in young patients with low-risk AF is difficult, because the incidence of TIA in such patients is very low, which would result in traditional multiple logistic regression not being able to successfully identify the risk factors in such patients. Therefore, a novel algorithm for identifying risk factors for TIA is necessary. We thus propose a novel algorithm, which combines multiple correspondence analysis and hierarchical cluster analysis and uses the Taiwan National Health Insurance Research Database, a population-based database, to determine risk factors in these patients. The results of this study can help clinicians or patients with AF in preventing TIA or stroke events as early as possible.

Introduction

The prevalence and incidence of atrial fibrillation (AF) have increased worldwide in recent decades (Kim et al., 2017; Wilke et al., 2013; Yiin et al., 2017). Patients with AF have been reported to have a significantly higher risk of ischemic stroke (IS) (Dulli et al., 2003; Novello et al., 1993; Saxena et al., 2001). Currently, stroke is the second highest cause of mortality worldwide (van Rooy and Pretorius, 2015). Because of the improvement in IS treatment in recent years, the overall survival duration of patients with stroke has been prolonged. However, poststroke rehabilitation is expensive and requires more medical resources. Transient ischemic attack (TIA) is defined as a period of focal ischemia, the symptoms of which resemble a thromboembolic stroke, and TIA has been demonstrated as a precursor of IS in patients with AF (Appelros et al., 2017; Cruz-Flores, 2017). Compared with stroke, TIA symptoms typically occur for less than several hours with absence of necrosis, and stroke is often preceded by TIA. Therefore, TIA is a crucial predictor of stroke events. However, the incidence of TIA in young patients with low-risk AF is very low. Thus, identification of risk factors for TIA in young patients with low-risk AF is necessary to more effectively prevent stroke events, because they may have longer survival and require high costs and more medical resources (Abdelhafiz and Wheeldon, 2003).

In the current clinical practice, the risk of IS onset is mainly measured using two risk scoring systems: the congestive heart failure (HF), hypertension (HTN), age (≥75 years), diabetes mellitus (DM), and prior stroke, TIA, or thromboembolism history (CHADS2) score (Rietbrock et al., 2008) and the CHA2DS2-VASc score (Pieri et al., 2011), which is a modification of the CHADS2 score aimed at improving stroke risk prediction in patients with AF by adding three risk factors: age (65–74 years), female sex, and history of vascular disease. Both scoring systems have been useful in predicting stroke events in patients with AF, particularly in older patients (aged 65–74 or ≥75 years). However, a risk scoring system for predicting TIA events in patients with AF is not available, probably because the incidence of TIA in young patients with low-risk AF is very low. These two scoring methods are probably not feasible for predicting the onset of TIA; therefore, a novel algorithm for predicting TIA events in young patients with

* Author for correspondence: National Taipei University of Nursing and Health Sciences, College of Health Technology, Department of Speech Language Pathology and Audiology, Biostatistical Consulting Lab, 365 Min-der Rd., Beitou district, Taipei City, Taiwan.
E-mail addresses: chiehyu@ntunhs.edu.tw, chiehyuliu@gmail.com (C.-Y. Liu).

https://doi.org/10.1016/j.jab.2017.10.005
1214-021X/© 2017 Faculty of Health and Social Sciences, University of South Bohemia in Ceske Budejovice. Published by Elsevier Sp. z o.o. All rights reserved.

low-risk AF is required. Such an algorithm can be developed using a large-scale population-based database, namely the Taiwan National Health Insurance Research Database (NHIRD). In this study, a novel algorithm was developed for identifying risk factors for TIA events in patients with low-risk AF who are younger than 60 years.

Materials and methods

Study database

The claims database of the National Health Insurance (NHI) program in Taiwan, which was launched on March 1, 1995, was used in this study. The NHI program provided comprehensive healthcare services for approximately 99.5% of Taiwanese residents in 2010 (Yeh and Chang, 2015). The NHIRD contains nationwide population-based information including outpatient and inpatient clinical visits, dentistry services, prescription drugs, and traditional Chinese medicine services. The diagnostic and procedural codes of diseases are based on the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and Procedure Coding System (ICD-9-PCS).

Ethics statement

This study was approved by the Institutional Review Board of School of Nursing, National Taipei University of Nursing and Health Sciences (CN-IRB-2011-064). Because personal information that may potentially identify any individual patient was fully encrypted by the National Health Insurance Administration (NHIA), written consent from study patients was not obtained. The NHIA guarantees the confidentiality of the personal and health information of enrolled patients.

Study population

An incidence-based cohort of patients who had received a new diagnosis of AF (ICD-9-CM code 427.31), had at least two outpatient visits with AF being the primary disease, were aged between 20 and 59 years. The exclusion criteria were: (1) had any TIA or stroke event history (including hemorrhagic stroke [HS] and IS) in 2005; (2) AF patients who were aged <20 or >60 years old; (3) AF patients with severe baseline diseases: including cancers, coronary artery disease (CAD, including congenital heart defect (CHD), myocardial infarction, and heart failure), kidney failure (including chronic kidney failure (CKD)) and peripheral artery disease (PAD); (4) AF patients' CHA2DS2-VASc score was ≥2. AF patients who met the above criteria were identified and retrieved from the NHIRD. This present study retrieved only outpatient claims data from the NHIRD. We linked the data of study patients to their 2006–2010 medical claims database to observe if they had onset of TIA. The definition of low-risk AF patients was using AF patients' CHA2DS2-VASc score was <2 (= 0 or 1), which is a widely used definition worldwide in many published studies (Chao et al., 2014; Crivera et al., 2016; Manolis et al., 2016; Piyaskulkaew et al., 2014). The enrollment scheme of study patients is shown in Fig. 1.

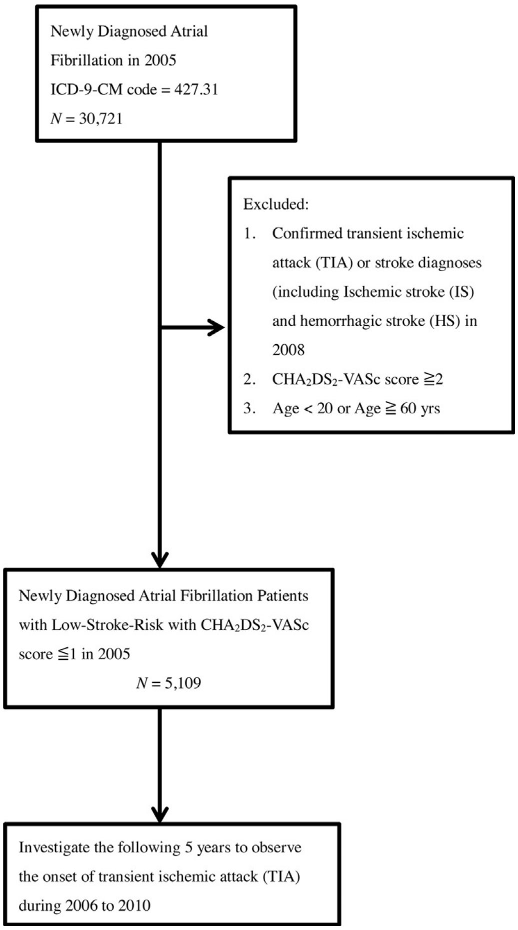
Newly Diagnosed Atrial Fibrillation in 2005
ICD-9-CM code = 427.31
N = 30,721
Excluded:
1. Confirmed transient ischemic attack (TIA) or stroke diagnoses (including Ischemic stroke (IS) and hemorrhagic stroke (HS) in 2008
2. CHA₂DS₂-VASc score ≧2
3. Age < 20 or Age ≧ 60 yrs
Newly Diagnosed Atrial Fibrillation Patients with Low-Stroke-Risk with CHA₂DS₂-VASc score ≦1 in 2005
N = 5,109
Investigate the following 5 years to observe the onset of transient ischemic attack (TIA) during 2006 to 2010

Fig. 1. Enrollment scheme of this study.

Algorithm.

1. Multiple Correspondence Analysis (MCA) (Havranek et al., 2016; Weller and Romney, 1990):

Denoted $A_{I \times K}$ is the raw data matrix with I patients and k nominal variables.

(1) Convert the raw data matrix into a Burt table:

- If one categorical variable has only two levels (or a binary variable), then retain it as the original variable type.
- If one categorical variable has more than two levels, say $J_k > 2$ levels, then convert this variable into an $I \times J_k$ indicator matrix, with each column having all binary outcomes (coding 0 or 1).
- Put all binary variables together as the indicator matrix $\mathbf{X}_{I \times J}$.
- The Burt table = **X'X**

(2) Calculate the column and row coordinates:

- The grand total of $\mathbf{A}_{I \times K}$ is N, and calculate the probability matrix $\mathbf{Z} = \mathbf{N}^{-1}\mathbf{X}$.
- Denote **r** as the vector of the row totals of Z (i.e. $\mathbf{r} = \mathbf{Z1}$, with **1** being a conformable vector of 1's), and denote c as the vector of the column totals, and **Dc** = diag (SPAF III Writing Committee, 1998), **Dr** = diag (SPAF III Writing Committee, 1998).
- The factor scores are calculated from the following singular value decomposition:

$$D_r^{-\frac{1}{2}}(Z - rc^T)D_c^{-\frac{1}{2}} = P\Delta Q^T$$

(Δ is the diagonal matrix of singular values, and $\Delta = \Delta^2$ is the matrix of eigenvalues). The row and column coordinates are thus obtained as

$$\mathrm{F} = D_r^{-\frac{1}{2}} P\Delta \text{and } \mathrm{G} = D_c^{-\frac{1}{2}} Q\Delta$$

(3) Number of dimensions to be determined using inertia:

- The chi-squared (χ^2)-like distances from rows and columns to their respective coordinate centers are obtained as

$$d_r = diag\left\{FF^T\right\} \text{ and } d_c = diag\left\{GG^T\right\}$$

- If we select a subset of F or G , then the inertia for row and column coordinates can be expressed as follows

$$Inertia_r = \frac{diag\left\{F'F'^T\right\}}{N} \text{and } Inertia_c = \frac{diag\left\{G'G'^T\right\}}{N}$$

where F' and G' are the subsets of F and G, respectively.

2. Hierarchical Cluster Analysis (HCA)
The HCA algorithm adopted in this study is as follows:

(1) use the row coordinates obtained as the aforementioned MCA for each level of each categorical variable;
(2) calculate the Euclidean distance between any two of the levels;

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(x_1 - y_1)^2}$$

(3) combine the two nearest levels with shortened Euclidean distances to form starting clusters;
(4) use Ward's method as the agglomerative method to form new clusters;
(5) show the dendrogram of each level of each variable.

At the end of this study, we also conducted a multiple logistic regression analysis of this data. We compared the results obtained using the algorithm proposed in this study and the estimation results obtained from multiple logistic regression. MY Structured Query Language was used for extraction, linkage, and processing of the NHIRD. All statistical analyses were performed using STATISTICA (version 10 for Windows; Statistica company, Tulsa, Oklahoma, USA), and a two-tailed p value of <0.05 was considered statistically significant.

Results

In this study, 5109 patients who had low-risk AF, were aged between 20 and 59 years, had at least two outpatient visits with AF being the primary disease (ICD-9-CM code 427.31), had a CHA2DS2-VASc score of <2 (=0 or 1), and did not have any baseline stroke or TIA event history (including HS and IS) in 2005 were recruited. Of the patients in the study cohort, 44.9% and 55.1% were aged between 20 and 39 years and between 40 and 59 years, respectively; 53.8% were men and 46.2% were women; 4.9% had HTN; 9.7% had DM; 9.8% had hyperlipidemia; and 2.0% had HF (Table 1).

The incidence of subsequent TIA in this study cohort was 2.68% during 2006–2010. The results of MCA were plotted using a three-dimensional scatter plot (Fig. 2).

The coordinates of each level of each variable were further analyzed using HCA (Fig. 3). As illustrated in Fig. 3, three clusters could be identified: (1) The first cluster comprised TIA and heart failure (HF); (2) the second cluster comprised hypertension, hyperlipidemia, and diabetes mellitus (DM); and (3) the third cluster comprised age variables (20–39 and 40–59), sex, and no other risk levels including HF, TIA, hyperlipidemia, DM, or HTN.

The multiple logistic regression analysis results are presented in Tables 2a and 2b. The score test was used to examine whether each independent variable would be significant while being entered into the logistic model, which can also be an important concern when using stepwise logistic regression analysis. As shown in Table 2a, no predictors showed significant results (all $p > 0.05$). When multiple logistic regression analysis was applied to the study dataset, the results (Table 2b) did not show any significant predictor.

Discussion

This study proposes a novel algorithm for predicting a rare but important disease, TIA, in young patients with low-risk AF by using

Table 1
Demographic information of study cohort ($N = 5109$).

		TIA in 2006–2010			χ^2-value	p-value
		Total $N = 5109$ (100%)	Yes ($n = 137$)	No ($n = 4972$)		
Sex					2.586	0.108
	Male	2750 (53.8)	83 (60.6)	2667 (53.6)		
	Female	2359 (46.2)	54 (39.4)	2305 (46.4)		
Age					2.169	0.141
	20–39 yrs	2295 (44.9)	70 (51.1)	2225 (44.8)		
	40–59 yrs	2814 (55.1)	67 (48.9)	2747 (55.2)		
Hypertension					0.873	0.350
	Yes	249 (4.9)	9 (6.6)	240 (4.8)		
	No	4860 (95.1)	128 (93.4)	4732 (95.2)		
Diabetes mellitus (DM)					0.036	0.850
	Yes	498 (9.7)	14 (10.2)	484 (9.7)		
	No	4611 (90.3)	123 (89.8)	4488 (90.3)		
Hyperlipidemia					0.018	0.893
	Yes	502 (9.8)	13 (9.5)	489 (9.8)		
	No	4607 (90.2)	124 (90.5)	4483 (90.2)		
Heart Failure					1.966	0.161
	Yes	102 (2.0)	5 (3.6)	97 (2.0)		
	No	5007 (98.0)	132 (96.4)	4875 (98.0)		

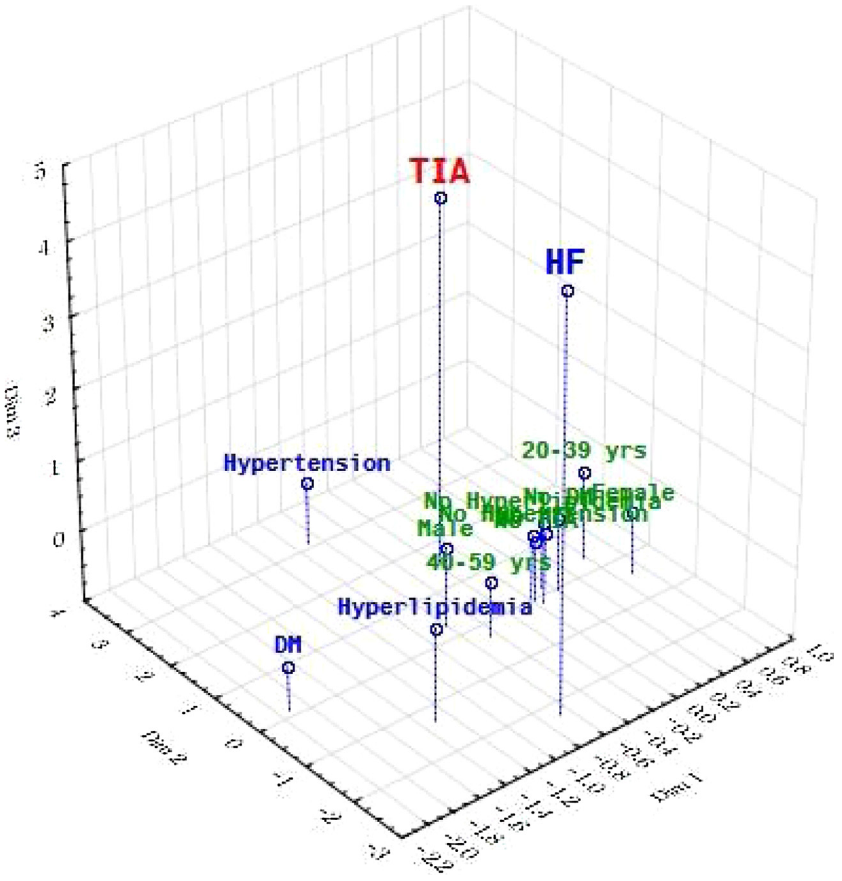

Fig. 2. 3-Dimensional scatter plot of each levels of each variable.

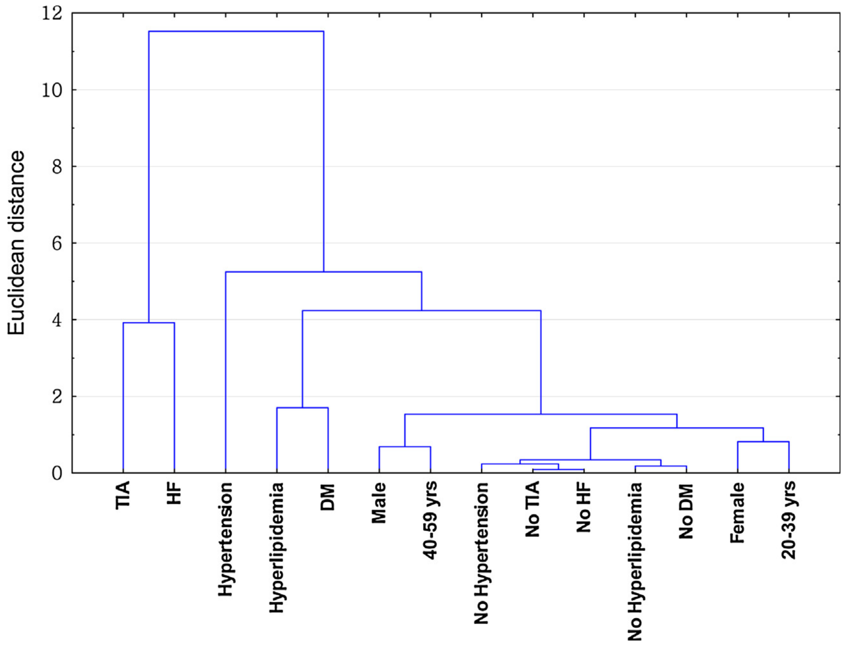

Fig. 3. Hierarchical Cluster Dendrogram of each levels of each variable.

the Taiwan NHIRD, which is a large-scale population-based database. When multiple logistic regression analysis, which is a traditional method, was used to investigate risk factors for TIA in young patients with low-risk AF, risk factors could not be determined successfully (Table 2a, 2b). In this study, heart failure (HF) was closely associated with TIA. A prospective cohort study recruiting patients with coronary heart disease (CHD) aged 45–74 years showed that CHD patients with pre-existing atherosclerotic vascular disease increased the risk of ischemic stroke or TIA (Koren-Morag et al., 2005). However, they recruited middle-aged

Table 2a
Score test of each predictor in this study ($n = 5109$).

	Score test value	df	p-value
Sex: Female vs. male	2.586	1	0.108
Age: 40–59 yrs vs. 20–39 yrs	2.169	1	0.141
Hypertension: Yes vs. no	0.873	1	0.350
DM: Yes vs. No	0.036	1	0.850
Hyperlipidemia: Yes vs. no	0.018	1	0.893
Heart failure: Yes vs. no	1.966	1	0.161
Overall statistics	7.614	6	0.268

Table 2b
Estimation results of multiple logistic regression of this study ($n = 5109$).

	β	S.E.	Odds Ratio (OR)	95% CI for OR		p-value
				lower	upper	
Sex: Female vs. male	−0.278	0.181	0.758	0.531	1.081	0.126
Age: 40–59 yrs vs. 20–39 yrs	−0.273	0.176	0.761	0.539	1.074	0.120
Hypertension: Yes vs. no	0.292	0.354	1.339	0.669	2.679	0.410
DM: Yes vs. no	0.037	0.293	1.037	0.585	1.841	0.900
Hyperlipidemia: Yes vs. no	−0.035	0.298	0.966	0.538	1.733	0.907
Heart failure: Yes vs. no	0.643	0.468	1.901	0.759	4.762	0.170
(Constant)	−3.365	0.152	0.035			<0.001

to old patients and placed less emphasis on younger groups. Besides, published studies using TIA as an outcome variable were still very limited (Rietbrock et al., 2008). The clinical pathways of cardiovascular diseases and TIA development have been proposed (Kate et al., 2012; Koren-Morag et al., 2005); among the possible cardiovascular diseases, including congestive heart failure or other coronary artery diseases (CAD) should be especially noted in young patients with AF. Studies have shown that effective management and monitoring of blood lipid levels may help prevent CHD (Nash, 1982), which is believed to be effective in preventing TIA.

These findings can provide some clinical implications. First, intensive monitoring of TIA is still required in young patients with low-risk AF. Second, although diabetes mellitus (DM), hyperlipidemia and htpertension were not closely clustered with TIA, these diseases must still be noted in young patients with low-risk AF. Third, prevention of heart failure, including monitoring blood pressure, heart rate and keep good lifestyle in these patients is strongly suggested. Although cardiologists may prescribe low-dose aspirin to patients with AF to prevent stroke events, this issue has been controversial in recent years (Hsu et al., 2016; Lip, 2011; Taylor, 2014).

This study has some limitations. First, the NHIRD does not contain information regarding potential confounders, including alcohol drinking, smoking, lifestyle, betel nut chewing, stay up late and un-healthy diet, which are associated with the risk of TIA or stroke events. Second, some young patients with low-risk AF may buy over-the-counter aspirin in drug stores, which is not recorded in the NHIRD. Third, this study retrieved claims database from outpatient medical services, and did not include inpatient medical services. Because the study purpose was to predict transient ischemic attack in young patients with low-risk atrial fibrillation, and we had excluded AF patients who had any TIA or stroke event history (including HS and IS) and severe baseline diseases which may result in possible hospitalizations, the outpatient claims database was believed enough to investigate the research questions of this present study.

Conclusion

The present study proposes a novel algorithm for identifying risk factors in young patients with low-risk AF, which may not be identified using traditional multiple logistic regression. The study results can help physicians and patients with AF in preventing TIA as early as possible.

Conflict of interests

The authors have no conflict of interests to disclose.

Acknowledgements

We thank the National Health Research Institutes (NHRI) for maintaining the National Health Insurance Research Database (NHIRD), and we thank the National Health Insurance Administration for managing the NHIRD.

References

Abdelhafiz, A.H., Wheeldon, N.M., 2003. Use of resources and cost implications of stroke prophylaxis with warfarin for patients with nonvalvular atrial fibrillation. Am. J. Geriatr. Pharmacother. 1 (2), 53–60.

Appelros, P., Hals Berglund, M., Strom, J.O., 2017. Long-term risk of stroke after transient ischemic attack. Cerebrovasc. Dis. 43 (1–2), 25–30.

Chao, T.F., Liu, C.J., Wang, K.L., Lin, Y.J., Chang, S.L., Lo, L.W., et al., 2014. Using the CHA2DS2-VASc score for refining stroke risk stratification in ‘low-risk’ Asian patients with atrial fibrillation. J. Am. Coll. Cardiol. 64 (16), 1658–1665.

Crivera, C., Nelson, W.W., Schein, J.R., Witt, E.A., 2016. Attitudes toward anticoagulant treatment among nonvalvular atrial fibrillation patients at high risk of stroke and low risk of bleed. Patient Prefer Adherence 10, 795–805.

Cruz-Flores, S., 2017. Acute stroke and transient ischemic attack in the outpatient clinic. Med. Clin. North Am. 101 (3), 479–494.

Dulli, D.A., Stanko, H., Levine, R.L., 2003. Atrial fibrillation is associated with severe acute ischemic stroke. Neuroepidemiology 22 (2), 118–123.

Havranek, S., Fiala, M., Bulava, A., Sknouril, L., Dorda, M., Bulkova, V., et al., 2016. Multivariate analysis of correspondence between left atrial volumes assessed by echocardiography and 3-dimensional electroanatomic mapping in patients with atrial fibrillation. PLoS One 11 (3), e0152553.

Hsu, J.C., Maddox, T.M., Kennedy, K., Katz, D.F., Marzec, L.N., Lubitz, S.A., et al., 2016. Aspirin instead of oral anticoagulant prescription in atrial fibrillation patients at risk for stroke. J. Am. Coll. Cardiol. 67 (25), 2913–2923.

Kate, M., Sylaja, P.N., Chandrasekharan, K., Balakrishnan, R., Sarma, S., Pandian, J.D., 2012. Early risk and predictors of cerebrovascular and cardiovascular events in transient ischemic attack and minor ischemic stroke. Neurol. India 60 (2), 165–167.

Kim, T.H., Yang, P.S., Uhm, J.S., Kim, J.Y., Pak, H.N., Lee, M.H., et al., 2017. CHA2DS2-VASc score (Congestive Heart Failure, Hypertension, Age >/=75 [Doubled], Diabetes Mellitus, Prior Stroke or Transient Ischemic Attack [Doubled], Vascular Disease, Age 65–74, Female) for Stroke in Asian Patients With Atrial Fibrillation: A Korean Nationwide Sample Cohort Study. Stroke 48 (6), 1524–1530.

Koren-Morag, N., Goldbourt, U., Tanne, D., 2005. Relation between the metabolic syndrome and ischemic stroke or transient ischemic attack: a prospective cohort study in patients with atherosclerotic cardiovascular disease. Stroke 36 (7), 1366–1371.

Lip, G.Y., 2011. The role of aspirin for stroke prevention in atrial fibrillation. Nat. Rev. Cardiol. 8 (10), 602–606.

Manolis, A.S., Manolis, T.A., Manolis, A.A., Melita, H., 2016. Stroke risk stratification schemes in atrial fibrillation in the era of non-vitamin K anticoagulants: misleading and obsolete, at least for the low-risk patients? Curr. Drug Targets 18 (5) doi:http://dx.doi.org/10.2174/1389450117666160905111822.

Nash, D.T., 1982. Hyperlipidemia therapy: can it prevent coronary atherosclerosis? Postgrad. Med. 72 (2), 207–211.

Novello, P., Ajmar, G., Bianchini, D., Bo, G.P., Cammarata, S., Firpo, M.P., et al., 1993. Ischemic stroke and atrial fibrillation: a clinical study. Ital. J. Neurol. Sci. 14 (7), 571–576.

Pieri, A., Lopes, T.O., Gabbai, A.A., 2011. Stratification with CHA2DS2-VASc score is better than CHADS2 score in reducing ischemic stroke risk in patients with atrial fibrillation. Int. J. Stroke 6 (5), 466.

Piyaskulkaew, C., Singh, T., Szpunar, S., Saravolatz, L., Rosman, H., 2014. CHA(2)DS(2)-VASc versus CHADS(2) for stroke risk assessment in low-risk patients with atrial fibrillation: a pilot study from a single center of the NCDR-PINNACLE registry. J. Thromb. Thrombolysis 37 (4), 400–403.

Rietbrock, S., Heeley, E., Plumb, J., van Staa, T., 2008. Chronic atrial fibrillation: incidence, prevalence, and prediction of stroke using the congestive heart failure, hypertension, age >75, diabetes mellitus, and prior Stroke or transient ischemic attack (CHADS2) risk stratification scheme. Am. Heart J. 156 (1), 57–64.

SPAF III. Writing Committee for the Stroke Prevention in Atrial Fibrillation Investigators, 1998. Patients with nonvalvular atrial fibrillation at low risk of stroke during treatment with aspirin: stroke prevention in Atrial Fibrillation III Study. JAMA 279 (16), 1273–1277.

Saxena, R., Lewis, S., Berge, E., Sandercock, P.A., Koudstaal, P.J., 2001. Risk of early death and recurrent stroke and effect of heparin in 3169 patients with acute ischemic stroke and atrial fibrillation in the International Stroke. Trial Stroke 32 (10), 2333–2337.

Taylor, J., 2014. Aspirin still overprescribed for stroke prevention in atrial fibrillation. Eur. Heart J. 35 (22), 1422.

Weller, S.C., Romney, A.K., 1990. Metric Scaling: Correspondence Analysis. SAGE Pub, Thousand Oaks (CA), USA.

Wilke, T., Groth, A., Mueller, S., Pfannkuche, M., Verheyen, F., Linder, R., et al., 2013. Incidence and prevalence of atrial fibrillation: an analysis based on 8. 3 million patients. Europace 15 (4), 486–493.

Yeh, M.J., Chang, H.H., 2015. National health nisurance in Taiwan. Health Aff. 34 (6), 1067.

Yiin, G.S., Howard, D.P., Paul, N.L., Li, L., Mehta, Z., Rothwell, P.M., Oxford Vascular, S., 2017. Recent time trends in incidence, outcome and premorbid treatment of atrial fibrillation-related stroke and other embolic vascular events: a population-based study. J. Neurol. Neurosurg. Psychiatry 88 (1), 12–18.

van Rooy, M.J., Pretorius, E., 2015. Metabolic syndrome, platelet activation and the development of transient ischemic attack or thromboembolic stroke. Thromb. Res. 135 (3), 434–442.