Biomedical Research 2017; 28 (7): 3293-3299

ISSN 0970-938X
www.biomedres.info

Handling imbalanced class problem for the prediction of atrial fibrillation in obese patient.

Cengiz Colak M[1], Erol Karaaslan[2], Cemil Colak[3], Ahmet Kadir Arslan[3*], Nevzat Erdil[1]

[1]Department of Cardiovascular Surgery, Faculty of Medicine, Inonu University, Malatya, Turkey

[2]Department of Anaesthesiology and Reanimation, Malatya State Hospital, Malatya, Turkey

[3]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey

Abstract

Objective: Atrial Fibrillation (AF) is one of the important public health problems with elevated comorbidity, advanced mortality risk, and increasing healthcare costs. In this study, the objective is to explore and resolve the imbalanced class problem for the prediction of AF in obese individuals and to compare the predictive results of balanced and imbalanced datasets by several data mining approaches.
Materials and methods: The retrospective study contained 362 successive obese individuals undergoing Coronary Artery Bypass Grafting (CABG) operation at the cardiovascular surgery clinic. AF developed postoperatively (AF Group) in 42 of the patients, whereas AF did not develop (non-AF Group) in 320 individuals. The Synthetic Minority Over-sampling Technique (SMOTE) was performed to balance the distribution of the target variable (AF/non-AF groups). The LogitBoost and GLMBoost ensemble approaches were constructed with 10-fold cross validation.
Results: After applying SMOTE algorithm, the number of subjects in AF and non-AF was almost balanced (336 in AF and 320 in non-AF groups). The values of accuracy were 0.8812 (0.8433-0.9127) for GLMBoost and 0.9144 (0.8806-0.9411) for LogitBoost on the imbalanced dataset, and 0.8247 (0.7934-0.853) for GLMBoost and 0.9695 (0.9533-0.9813) for LogitBoost on the balanced dataset by SMOTE. The values of the area under the receiver operating curve for GLMBoost and LogitBoost were 0.5088 (0.485-0.5325) and 0.6827 (0.608-0.7573) on imbalanced dataset, and were 0.8259 (0.7971-0.8546) and 0.9696 (0.9564-0.9827) on balanced dataset, respectively.
Conclusions: The predicted results indicated that LogitBoost on the balanced dataset by SMOTE had the highest and most accurate values of performance metrics. Hence, SMOTE and other oversampling approaches may be beneficial to overcome class imbalance issues emerging in biomedical studies.

Keywords: Imbalanced dataset classification, Atrial fibrillation GLMBoost, LogitBoost, Synthetic minority over-sampling technique.

Accepted on January 5, 2017

Introduction

Atrial Fibrillation (AF) is one of the important public health problems with elevated comorbidity, advanced mortality risk, and increasing healthcare costs. AF is one of the most commonly experienced and important cardiac arrhythmias. The explanations behind the increasing prevalence of AF have not been efficiently determined; however, may be associated with improved detection, increasing incidence, and enhanced survival in cardiovascular patients. AF is related significantly with advancing age, with about 1 in 25 people above 60 years and 1 in 10 above 80 years influenced by AF. AF causes increasing morbidity/mortality, advancing risks for death, Congestive Heart Failure (CHF), stroke, and other related diseases. Significant risk factors related to the development of AF contain advanced age, hypertension, smoking, alcohol consumption, obesity, prevalent myocardial infarction, CHF, diabetes mellitus and so forth, described in detail by the studies [1-3].

Obesity is one of the important public health problems influencing both children and adults in the world, and also causes important health risks. Obesity is associated with the development and progression of AF. Hence, obesity can be reduced and prevented by conducting effective public health programs [4].

Knowledge Discovery Process (KDP) is related to reproducing and extracting higher-level comprehensions and knowledge from the database. The methods implemented in KDP are based on knowledge-intense stages and can often take advantage of utilizing supplementary knowledge from different data sources [5]. In biomedical research, the KDP has been applied in the different areas of medicine and has attracted interest [6,7].

Imbalanced classification is one of the important subjects in the knowledge discovery process and data mining. On the imbalanced issues of two classes, a minority class and a majority class were present in the data of interest, which is called imbalanced classification. Class imbalance leads to some troubles for data mining algorithms assuming an almost equal class distribution, and consequently, minority class instances are largely misclassified by the data mining algorithms [8,9].

The primary aim of this study is to explore and resolve the imbalanced class problem for the prediction of AF in obese individuals. The secondary goal of the study is to compare the predictive results of balanced and imbalanced datasets by implementing several data mining approaches.

Materials and Methods

Study design and data

The current retrospective study contained 362 successive obese individuals who underwent Coronary Artery Bypass Grafting (CABG) operation at the cardiovascular surgery clinic of Turgut Ozal Medical Center, Inonu University, Malatya, Turkey between January 2012 and December 2015. AF developed postoperatively (AF Group) in 42 of the patients, whereas AF did not develop (non-AF Group) in 320 individuals. The exclusion criteria for this study were past atrial arrhythmia, requirement for extra procedures, left ventricle and renal dysfunctions, and chronic obstructive pulmonary disease. Table 1 tabulates the details of the target and predictor variables.

Table 1. *The definition of the variables employed in the current study.*

Attribute name	Abbreviation	Attribute type	Explanations	Role
Atrial Fibrillation	AF	Categorical	Present/absent	Target
Age	-	Numerical	Natural number	Input
Gender	-	Categorical	Female/male	Input
Smoking	-	Categorical	Present/absent	Input
Alcohol Consumption	AC	Categorical	Present/absent	Input
Diabetes Mellitus	DM	Categorical	Present/absent	Input
Hypertension	HT	Categorical	Present/absent	Input
Chronic Obstructive Pulmonary Disease	COPD	Categorical	Present/absent	Input
History of Myocardial Infarction	HMI	Categorical	Present/absent	Input
Rhythm	-	Categorical	Present/absent	Input
Emergency Operation	EO	Categorical	Present/absent	Input
Heart Palpitations	HP	Categorical	Present/absent	Input
Early Mortality	EM	Categorical	Present/absent	Input
Blood Surface Area	BSA	Numerical	Positive integer	Input
Blood Urea Nitrogen	BUN	Numerical	Positive integer	Input
Creatinine	CR	Numerical	Positive real number	Input
Haemoglobin	HB	Numerical	Positive real number	Input
Haematocrit	HCT	Numerical	Positive real number	Input
Platelets	PLT	Numerical	Positive integer	Input
Blood Sugar Concentration	BSC	Numerical	Positive integer	Input
Cholesterol	CHL	Numerical	Positive integer	Input
Low Density Lipoprotein	LDL	Numerical	Positive integer	Input
High Density Lipoprotein	HDL	Numerical	Positive integer	Input
Triglyceride	TG	Numerical	Positive integer	Input

Ventilation Time	VT	Numerical	Positive integer	Input
Intensive Care Unit Hospitalization Time	ICUHT	Numerical	Positive integer	Input

Table 2. Summary of the target distribution before and after SMOTE.

The target variable distribution	Before performing SMOTE			After performing SMOTE		
	AF	Non - AF	Total	AF	Non - AF	Total
	42	320	362	336	320	656

Outlier detection

Local Density Cluster-Based Outlier Factor (LDCOF) was used to explore outlier observations. This approach utilizes X-means clustering algorithm which determines heuristically the number of clusters [10,11]. In this study, no outlier observations were detected according to the results of LDCOF analysis.

Synthetic minority over-sampling technique (SMOTE)

The Synthetic Minority Over-sampling Technique (SMOTE) [12] is one of the oversampling methods and constitutes minority class instances. For this reason, it is commonly employed for the class imbalance problems and produces better results than simple oversampling techniques. The SMOTE is a useful and powerful technique used successively in many medical applications. In relation to implementation of this algorithm, artificial data were created according to the attribute space [13,14].

The SMOTE is employed to achieve an artificial class-balanced or almost class-balanced dataset. The percentage of over and under-sampling (%) and the number of nearest neighbours were adjusted respectively to 700:109 and k=5 to constitute the novel synthetic samples. The DMwR package [15] of R was employed to implement SMOTE.

LogitBoost

Boosting technique was essentially suggested to combine many weak learners to increase the prediction performance. LogitBoost, presented initially by [16], is a boosting algorithm employed commonly in many areas [17]. LogitBoost is characterized by presenting a loss function of the log-likelihood to diminish the sensitivity to extreme and outlier observations. Also, it can implement classification and regression through joining a number of weak learners to achieve very powerful and robust learners [18]. Details of this technique can be found [19].

GLMBoost

One of the boosting approaches is GLMBoost based on penalized log-likelihood. This approach can be employed widely for the problems of regression/classification since it is one of the important ensemble learning algorithms. GLMBoost has many implementation advantages. In addition to the ease of calculation, GLMBoost indicates different advantages. It has high calculation capacity, and complex tuning procedures are not necessary [20]. More detailed information on this ensemble learning algorithm can be obtained from [21].

Modelling and performance metrics

In this study, 10-fold cross validation technique was employed for evaluating the models and obtaining unbiased results from the models. LogitBoost and GLMBoost were utilized to compare classification performances of the balanced and unbalanced datasets. All modelling and validation processes were carried out by the caret package of R [22].

In this study, the assessment of the models was performed using accuracy, kappa (κ), Area Under Curve (AUC) of Receiver Operating Characteristic (ROC), recall and precision. These metrics are explained below [23]:

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

$$\kappa=Kappa=(p_0-p_e)/(1-p_e)$$

$$Recall=TP/(FN+TP)$$

$$Precision=TP/(FP+TP)$$

where, *TP* describes the number of true positives, *TN* describes the number of true negatives, *FN* describes the number of false negatives, *FP* describes the number of false positives, p_0 describes the relative observed agreement and p_e describes the hypothetical possibility of chance agreement.

Results

In the current study, there were 320 (88%) subjects in the non-AF group and 42 (12%) subjects in the AF group. When the gender distribution of the study was considered, 170 (47%) subjects were females, and 192 (53%) subjects were males. The mean ages with standard deviations were 59.7 ± 9.4 years and 64 ± 6.4 years for non-AF and AF groups, respectively. Table 2 summarizes the distribution of the target variable before and after SMOTE.

Table 3 indicates the performance metric results of the imbalanced and balanced datasets according to the classification models of LogitBoost and GLMBoost. The performance metrics with 95% Confidence Interval (CI) values are also calculated for accuracy, kappa, AUC of ROC, recall and precision (Table 3). According to Table 3, all of the performance metrics for LogitBoost model were higher than those for GLMBoost model on the imbalanced dataset. Similarly, for the balanced dataset by SMOTE, all of the performance metrics results of LogitBoost model outperformed the results of GLMBoost model. These findings clearly demonstrate that over-sampling with SMOTE ameliorates the deficits raised by the imbalanced structure of the dataset.

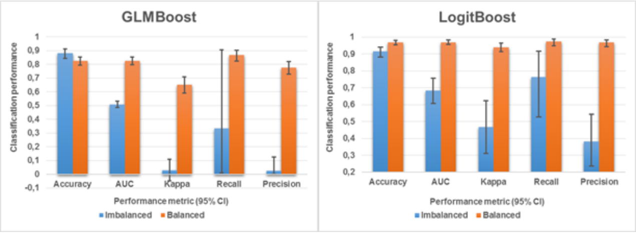

Figure 1. *The performance metric results of the imbalanced and balanced datasets according to the classification models of LogitBoost and GLMBoost.*

Figure 1 demonstrates the performance metric results of the imbalanced and balanced datasets according to the classification models of LogitBoost and GLMBoost. In a similar way, Figure 2 displays the comparative results of performance metrics on the balanced datasets according to LogitBoost and GLMBoost. When Figure 1 is examined, almost all performance metrics for LogitBoost and GLMBoost models are higher in the balanced dataset than in the unbalanced dataset. When the Figure 2 is inspected, LogitBoost model outperforms GLMBoost model in all the performance metrics.

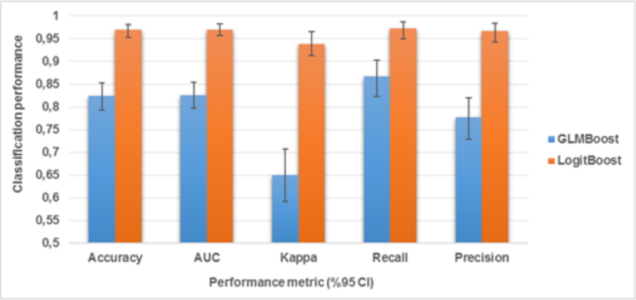

Figure 2. *The comparative results of performance metrics on the balanced datasets according to LogitBoost and GLMBoost.*

ROC curves of LogitBoost and GLMBoost models and imbalanced/balanced datasets are depicted in the Figure 3. According to this figure, each model has more classification power on the balanced dataset as compared to that on the unbalanced dataset.

Table 3. *The performance metric results of the imbalanced and balanced datasets according to the classification models.*

Models	**Performance metric (95% CI) for imbalanced dataset**				
	Accuracy	**AUC**	**Kappa**	**Recall**	**Precision**
GLMBoost	0.8812 (0.8433-0.9127)	0.5088 (0.485-0.5325)	0.0294 (-0.048-0.107)	0.3333 (0.0084-0.9057)	0.0238 (0.0006-0.1257)
LogitBoost	0.9144 (0.8806-0.9411)	0.6827 (0.608-0.7573)	0.4667 (0.311-0.623)	0.7619 (0.5283-0.9178)	0.3810 (0.2357-0.5436)
Models	Performance metric (95% CI) for balanced dataset by SMOTE				
GLMBoost	0.8247 (0.7934-0.853)	0.8259 (0.7971-0.8546)	0.6501 (0.592-0.708)	0.8671 (0.8235-0.9033)	0.7768 (0.7284-0.8202)
LogitBoost	0.9695 (0.9533-0.9813)	0.9696 (0.9564-0.9827)	0.939 (0.913-0.965)	0.9731 (0.9495-0.9876)	0.9673 (0.9422-0.9835)

Table 4. *Predictor importance of the most accurate model.*

Predictor	**Predictor importance**
ICUHT	0.9011
VT	0.6455
Age	0.6448
PLT	0.6337
TG	0.6279
CHL	0.6076
BSA	0.5659
LDL	0.5539
Gender	0.5443
BUN	0.5302
HCT	0.5251
Rhythm	0.5222
EO	0.5196
COPD	0.5181
BSC	0.511
EM	0.5097
HB	0.5052
AC	0.502
HMI	0.5012
HDL	0.4998
HT	0.4851
Smoking	0.4835
DM	0.4767
CR	0.4596

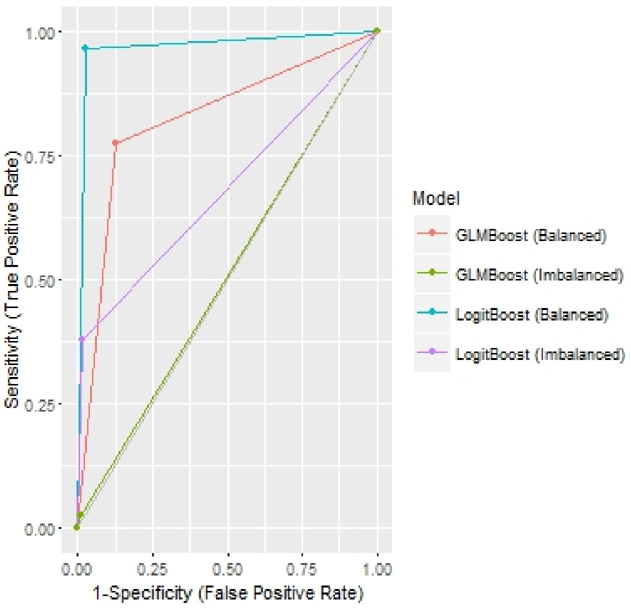

Figure 3. *ROC curves of GLMBoost and LogitBoost models for the imbalanced/balanced datasets.*

Table 4 represents the LogitBoost model based-predictor importance on the balanced dataset by SMOTE given in descending order. Based on the information of Table 4, the highest predictor importance belongs to hospitalization time in the intensive care unit.

Discussion

This research investigates and resolves the imbalanced class problem by a well-known approach of SMOTE for the prediction of AF in obese individuals. Since postoperative AF after cardiac surgery is rarely observed, the imbalanced class problem can arise in biomedical researches. As expected, the dataset used in the analysis was highly imbalanced and the minority class was the AF group (11.6% for AF group versus 88.4% for non-AF group). To cope with this problem, we performed an oversampling method of SMOTE, and afterwards, GLMBoost and LogitBoost models were built to compare the predictive results before and after SMOTE. In the first stage, when we applied the classification models to the imbalanced dataset, the accuracies of each model were observed so high. However, the values of the other performance metrics were considerably lower than expected. For instance, when the AUC value (0.5088) of GLMBoost model was considered, GLMBoost model was not capable of classifying between the AF and non-AF groups due to the imbalanced class problem encountered in this study. Similarly, although the AUC value (0.6827) of LogitBoost was little more than the value of GLMBoost, the AUC was quite low when considering the accuracy of LogitBoost. When only the accuracies of the models are evaluated on the unbalanced data, the obtained results can be very misleading. For this reason, performance evaluation of the constructed models should be performed based on the results of different performance metrics (i.e. AUC, recall, precision and so on.) as suggested and implemented in the present study [24].

Postoperative AF is one of the prevailing complications following cardiac surgery. Postoperative AF leads to higher rates for morbidity and mortality in the patients undergoing cardiac surgery [25]. Important risk factors for postoperative AF can be estimated by predictive models to prevent the development/progression of this complication. In the present research, the most important predictor for postoperative AF was hospitalization time in the intensive care unit (ICUHT), which is associated with this complication [26]. According to the reported works [27], VT is another factor for postoperative AF. Accordingly, we determined Ventilation Time (VT) as the second most important factor related to postoperative AF. In this research, the third most important factor was age, which is an independent factor of the arrhythmia [28]. The other important predictors which are associated with postoperative AF and are determined by the selected model are lengthily given in Table 4. The predictor variables determined in this work for postoperative AF were largely analogous to the risk factors notified by other researches examining the prediction of AF after coronary artery bypass surgery [29-33].

Conclusion

Eventually, the predicted results from this research indicated that LogitBoost on the balanced dataset by SMOTE had the highest and most accurate values of performance metrics. Our results suggest that SMOTE and other oversampling approaches would be so beneficial to overcome class imbalance issues emerging in biomedical studies. As future researches, other sampling techniques incorporating with different ensemble and meta-learning algorithms are planned to handle imbalanced class problems in multi-category classification.

Acknowledgement

We would like to thank to Inonu University Scientific Research Coordination Unit to support by a grant this study. (Project number: 2016/61).

References

1. Bhatt HV, Fischer GW. Atrial fibrillation: pathophysiology and therapeutic options. J Cardiothor Vasc Anesth 2015; 29: 1333-1340.
2. Lloyd-Jones DM, Wang TJ, Leip EP, Larson MG, Levy D. Lifetime risk for development of atrial fibrillation: the Framingham heart study. Circulation 2004; 110: 1042-1046.
3. Schnabel RB, Yin X, Gona P, Larson MG, Beiser AS, McManus DD. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham heart study: a cohort study. Lancet 2015; 386: 154-162.
4. Kwon S, Janz KF, Letuchy EM, Burns TL, Levy SM. Active lifestyle in childhood and adolescence prevents

obesity development in young adulthood. Obesity 2015; 23: 2462-2649.

5. Ristoski P, Paulheim H. Semantic web in data mining and knowledge discovery: A comprehensive survey. Web Seman Sci Serv Agents World Wide Web 2016; 36: 1-22.
6. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed 2015; 119: 181-185.
7. Colak C, Aydogan MS, Arslan AK, Yucel A. Application of medical data mining on the prediction of APACHE II score. Med Sci 2015; 4: 2743-2750.
8. Verbiest N, Ramentol E, Cornelis C, Herrera F. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl S Comput 2014; 22: 511-517.
9. Fernández A, Jesus MJ, Herrera F. Addressing overlapping in classification with imbalanced datasets: a first multi-objective approach for feature and instance selection. Intel Data Eng Autom Learn IDEAL 2015; 36-44.
10. Amer M, Goldstein M. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. Proc 3rd RapidMiner Community Meeting and Conference (RCOMM 2012) 2012.
11. Hofmann M, Klinkenberg R. RapidMiner: Data mining use cases and business analytics applications. CRC Press 2013.
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artific Intel Res 2002: 321-357.
13. Makond B, Wang KJ, Wang KM. Probabilistic modelling of short survivability in patients with brain metastasis from lung cancer. Comput Methods Programs Biomed 2015; 119: 142-162.
14. Majid A, Ali S, Iqbal M, Kausar N. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Comp Meth Progr Biomed 2004; 113: 792-808.
15. Torgo L, Torgo ML. Package DMwR. Comprehensive R archive network (http://cran r-project org/web/packages/DMwR/DMwR pdf) 2013.
16. Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Mach Learn 1999; 37: 297-336.
17. Cai YD, Feng KY, Lu WC, Chou KC. Using LogitBoost classifier to predict protein structural classes. J Theor Biol 2006; 238: 172-176.
18. Feng KY, Cai YD, Chou KC. Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 2005; 334: 213-217.
19. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat 2000; 28: 337-407.
20. Hao M, Wang Y, Bryant SH. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. Analytica Chimica Acta 2014; 806: 117-127.
21. Tutz G, Binder H. Generalized additive modelling with implicit variable selection by likelihood-based boosting. Biometrics 2006; 62: 961-971.
22. Kuhn M. Caret package. J Stat Softw 2008; 28.
23. Karabulut EM, Ibrikci T. Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing. J Med Sys 2014; 38: 1-9.
24. Prati RC, Batista GE, Silva DF. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl Inform Sys 2015; 45: 247-270.
25. Erdil N, Kaynak M, Donmez K, Disli OM, Battaloglu B. Nebivolol in preventing atrial fibrillation following coronary surgery in patients over 60 years of age. Revista Brasileira de Cirurgia Cardiovasc Orgao Oficial da Sociedade Brasileira de Cirurgia Cardiovasc 2014; 29: 581-587.
26. Creswell LL, Schuessler RB, Rosenbloom M, Cox JL. Hazards of postoperative atrial arrhythmias. Ann Thorac Surg 1993; 56: 539-549.
27. Ascione R, Caputo M, Calori G, Lloyd CT, Underwood MJ, Angelini GD. Predictors of atrial fibrillation after conventional and beating heart coronary surgery a prospective, randomized study. Circulation 2000; 102: 1530-1535.
28. Pfisterer ME, Kloter-Weber UC, Huber M, Osswald S, Buser PT, Skarvan K. Prevention of supraventricular tachyarrhythmias after open heart operation by low-dose sotalol: a prospective, double-blind, randomized, placebo-controlled study. The Ann Thorac Surg 1997; 64: 1113-1119.
29. Erdil N, Gedik E, Donmez K, Erdil F, Aldemir M, Battaloglu B. Predictors of postoperative atrial fibrillation after on-pump coronary artery bypass grafting: is duration of mechanical ventilation time a risk factor? Ann Thorac Cardiovasc Surg 2014; 20: 135-142.
30. Hakala T, Hedman A. Predicting the risk of atrial fibrillation after coronary artery bypass surgery. Scand Cardiovasc J 2003; 37: 309-315.
31. Ak K, Akgun S, Tecimer T, Isbir CS, Civelek A. Determination of histopathologic risk factors for postoperative atrial fibrillation in cardiac surgery. Ann Thorac Surg 2005; 79: 1970-1975.
32. Osranek M, Fatema K, Qaddoura F, Al-Saileek A, Barnes ME. Left atrial volume predicts the risk of atrial fibrillation after cardiac surgery: a prospective study. J Am Coll Cardiol 2006; 48: 779-786.
33. Mathew JP, Parks R, Savino JS, Friedman AS, Koch C, Mangano DT. Atrial fibrillation following coronary artery bypass graft surgery: predictors, outcomes, and resource utilization. JAMA 1996; 276: 300-306.

*Correspondence to

Ahmet Kadir Arslan

Faculty of Medicine

Department of Biostatistics and Medical Informatics

Malatya

Turkey