http://www.cjmb.org

Open Access Original Article

Crescent Journal of Medical and Biological Sciences
Vol. 5, No. 2, April 2018, 149–154
eISSN 2148-9696

Detection of Functional Limitation in Diabetic Patients Based on the Optimal Combination of Care Indicators Using Ramp AUC and Comparing its Performance With the Existing Methods

Parvin Sarbakhsh[1,2*], Leili Faraji Gavgani[1], Mohamad Asghari Jafarabadi[3,1], Seyed Morteza Shamshirgaran[1]

Abstract

Objectives: The area under the ROC curve (AUC) is a common criterion to assess the overall classification performance of the markers. In practice, due to the limited classification ability of a single marker, we are interested in combining markers linearly or nonlinearly to improve classification performance. Ramp AUC (RAUC) is a new statistical AUC-based method which can find such optimal combinations of markers. In this study, RAUC was used to find the optimal combinations of care indicators related to functional limitation as a complication of diabetes and accurately discriminate this outcome based on its underlying markers.

Materials and Methods: This cross-sectional study was conducted on 378 diabetic patients referred to diabetic centers in Ardebil and Tabriz during 2014 and 2015. To have an accurate classification of diabetic patients according to their functional limitation status, RAUC method with RBF kernel was employed to look for an optimal combination of care indicators. Classification performance of the model was evaluated by AUC and compared with logistic regression, support vector machine (SVM) and generalized additive model (GAM) via training and test validation method.

Results: Out of 378 diabetics, 67.46% had functional limitation. RAUC had an AUC of 1 for the test dataset and outperformed logistic (AUC = 0.079), GAM (AUC = 0.082), SVM with linear kernel (AUC = 0.67) and was slightly better than SVM with RBF kernel (AUC = 0.98).

Conclusions: There was a strong nonlinearity in data and RAUC with RBF kernel which is a nonlinear combination of markers could detect this pattern

Keywords: Ramp AUC model, SVM, GAM, Diabetes, Functional limitation, Classification, Kernel function, RBF kernel

Introduction

Diabetes is the most common metabolic disorder and one of the most important causes of death in the world (1). In type 2 diabetes, due to its chronicity, morbidity and disability and the need for lifelong patient care, the quality of life is severely affected (2) which would have different consequences for a diabetic patient (3).

Functional limitation is one of the most important health-related concerns among people with diabetes. Many studies have shown the effect of diabetes on the daily functioning (4-7). Functional limitation is defined as existing any impairment in physical function, performing various daily activities such as walking, taking shower, shopping and so on (5,8,9).

Identifying factors which affect the functional limitation caused by type 2 diabetes can be effective for diagnosis and second-level preventive planning in these patients. In addition, by reducing the amount of functional limitation, the quality of life and the quality of care can be improved.

On the other hand, we are often interested in biomedical studies to predict or classify an outcome based on its underlying markers. In practice, a single marker has limited power to classify or predict an outcome, and there may exist some special linear or nonlinear combinations of markers in relationships between them and outcome and considering such combinations in classification process can improve the performance of the classifiers.

In the ordinary regression models, we do not have enough knowledge about the nature and form of the relationship between the dependent and independent variables and usually, variables are entered into the model just with their linear form, while in the case of more complex associations and nonlinear relationships in data, ignoring such complex patterns in the model leads to a decrease in efficiency and increase in misclassification rate of the model.

Received 25 March 2017, **Accepted** 10 September 2017, **Available online** 1 October 2017

[1]Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran. [2]Road Traffic Injury Research Center, Tabriz University of Medical Sciences, Tabriz, Iran. [3]Medical Education Research Center, Tabriz University of Medical Sciences, Tabriz, Iran.
*Corresponding Author: Parvin Sarbakhsh, Tel: +9841-33355781 (392), Email: p.sarbakhsh@gmail.com

In such cases, we would like to find optimal combinations of marker to have a better classification performance. AUC is a commonly used criterion to assess classification ability of the classifiers (10, 11). Therefore, in our attempt to find optimal combinations, the combination that led to the greatest AUC, was desirable.

Up to now, several statistical methods have been introduced to maximize AUC based on combinations of markers (12-16), but these methods have 2 major limitations:

First, most of the existing method can only find the best linear combination of the markers, which will not be effective if there is a nonlinear and complex association between them and outcome. Second, they use gradient-based algorithms to search for optimal combinations. This algorithm often leads to suboptimal local solutions. To solve these 2 limitations, a new kernel-based AUC optimization method called ramp AUC (RAUC) was proposed in 2016 (17).

Due to the importance of the accurate classification of diabetic patients according to their functional limitation status, in this study, RAUC method was used to identify the optimal combination of the care indicators and its performance was compared with some existing method including logistic regression, GAM and SVM.

Materials and Methods

The data for this study are obtained from 2 cross-sectional studies that have been done in the northwest of Iran. Individuals who had a definite diagnosis of diabetes, those who referred to referral diabetes clinics of Ardabil and Tabriz during 2014 to 2015 and those who were eligible to enter the study were selected by convenience sampling method.

Further details on these studies are provided in several articles (18,19). In the present study, by integrating the data of these 2 studies and removing cases with missing data for the required variables, 378 subjects with complete data remained for the modeling process.

According to the rule of thumb, the sample size needed for the classification of functional limitation was sufficient (20).

Some demographic and clinical characteristics that were available and have been shown were related to poor control of diabetes (21,22) and consequently related to functional limitation (6), which is mentioned in this study as “care indicators”, including age, gender, duration of diabetes, body mass index (BMI), hemoglobin A1C (HbA1C), systolic blood pressure (SBP), diastolic blood pressure (DBP), cholesterol (Chol), triglycerides (TG), fasting blood sugar (FBS) and high-density lipoprotein (HDL) that were collected through a checklist and patient care records. The physical function subscale of 36-item short form health survey questionnaires (SF36) was used to measure functional capacity in patients. This questionnaire consisted of 10 questions, and participants were asked to indicate whether they had any problem and limitation in doing moderate and severe physical activities such as lifting, shopping, climbing, walking, bending, kneeling, bathing and wearing clothes. Based on 3 options and assigning 2 points to the option “I have no problem at all”, one point to “I have a little problem” and zero for the “I have a lot of problems”, the total performance score for each person was calculated as percentages. Based on the score, the status of functional disorder was divided into 2 groups, so that total score ≥90 showed normal functional capacity and score <90 indicated functional limitation in the diabetic patient (9,23).

Statistical Analyses

The RAUC method, logistic regression, GAM and SVM procedures were applied to find the optimal combination of the care indicators (sex, age, duration of diabetes, BMI, FBS, hemoglobin A1C, SBP, DBP, Chol, TG and HDL) in detecting the functional limitation. AUC, as the overall classification performance of the models, was assessed for all models via training and test validation method. To do this, from total 378 sample data, 70% were randomly selected as the training set and the remaining 30% were considered as the test set. Models were fitted to the training set and then the fitted models were evaluated on the test set. The AUC criteria of the methods were compared with each other using Delong test. Statistical analysis was performed by R 3.3.2 software (“mgcv”, “e1071”, “aucm” and “AUC” packages) and *P* values <0.05 were considered to indicate statistical significance.

Brief Explanations About Criterions and Methods Employed in This Study

AUC

The plot of the true positive rate (TPR) against the false positive rate (FPR) for different cut-points of a marker is called ROC (receiver operating characteristic curve). Each point on the ROC indicates a sensitivity-specificity pair related to a specific decision threshold(s). The area under the curve (AUC) is given by:

$$AUC = \int_{-\infty}^{+\infty} TPR(s)(-FPR(s))ds$$

AUC is a measure of the discrimination ability of a marker to discriminate between 2 groups of subjects (11).

Logistic Regression

Logistic regression is a statistical method to find the best model to describe the relationships between independent variables and binary outcome. Logistic regression is a member of generalized linear model (GLM) family in which response variable has a probability density function from the exponential family and the relationships between independent variables and outcome are considered as linearity and dependent variable assumed to be affected by independent variable only through their linear combination (24).

Generalized Additive Models

Generalized Additive Model (GAM) extends the parametric form of the predictors in GLM to nonparametric forms using a link function to establish an association between the mean of the response variable and a smoothed function of the independent variables. The GAM approach replaces the simple products of parameter values, time and the predictor values with a spline smoother of each predictor. The degrees of freedom is specified for the spline smoothers by GAM (25).

In this study, GAM with logit link function and generalized cross validation (GCV) function based on the expected prediction error was used to choose the smoothing parameters of the model.

Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression approaches (26,27). Much of the benefit of SVMs comes from the fact that they are not restricted to being linear classifiers. In classification approach, in contrast to the parametric models, which depend on a pre-determined and specific form of relationship between independent variables and outcome, SVM distinguishes between groups by generating hyperplanes that can separate groups after transformation of the input variables into a high-dimensional space via kernel function. The linear kernel employed by SVM which means the decision boundary is a straight line defined as: $K(x_i,x_j)=(x_i,x_j)$ and RBF kernel as a nonlinear transformation of the input variables is:

$$K(x_i, x_j) = \exp(-\gamma \, || \, x_j - x_i \, ||^2)$$

where $|| \, x_j - x_i \, ||^2$ is the squared Euclidean distance between . γ is a parameter to control the complexity of the model and C is the penalty for misclassifying a data point.

In this study, SVM with classification approach and both linear and nonlinear kernel (RBF kernel) was used to classify functional limitation. Gamma and C Parameters were set as 1.

The RAUC Method

RAUC is a kernel-based method that approximates the empirical AUC loss function with a ramp shape function which finds the best combination by a difference of convex functions algorithm (DCA). If the best combination is not among the linear combinations of the markers, the feature space can be expanded by basis expansion and finds the best linear combination in the enlarged feature space by mapping the input vector of markers to a feature space without having to specify the mapping explicitly (kernel trick) (17).

Results

Of the 378 diabetic patients, 255 (67.46%) had functional limitation. This outcome was more prevalent in females compared to males as 72.3% of the females (249) had functional limitation while this proportion was 58.1% in males ($P = 0.005$). The basic clinical characteristics of diabetic patients according to functional limitation status are presented in Table 1. According to the results of ordinary logistic regression model that consider only degree of 1 form of the explanatory variables in linear model, sex ($P < 0.001$), age ($P < 0.001$), BMI ($P = 0.046$), SBP ($P = 0.043$), HDL-cholesterol ($P = 0.043$) and TG ($P = 0.019$) were significantly associated with functional limitation. Although other variables were not significantly associated with the outcome, they were considered in the overall linear combination with coefficients yield from

Table 1. Demographic and Care Indices According to Different Subgroups of Functional Limitation

Variables	Without Functional Limitation (n=123)	With Functional Limitation (n=255)	*P* Value
Sex[a]			0.005
Female	69 (27.7)	180 (72.3)	
Male	54 (41.9)	75 (58.1)	
Age*(y)	51.02±8.67	58.65±7.72	<0.001
BMI[b] (kg/m2)	27.76±3.86	29.74±4.87	<0.001
Hemoglobin A1C[b] (%)	7.85±1.85	8.03±1.94	0.390
FBS§ (mg/dL)	150 (67)	150 (78)	0.958
SBP[b] (mm Hg)	127.19±17.01	125.52±17.38	0.380
DBP[b] (mm Hg)	75.16±10.59	75.13±9.53	0.981
Chol[b] (mg/dL)	174.55±40.87	174.38±43.86	0.971
HDL[c] (mg/dL)	45 (14)	48 (16)	0.100
TG[c] (mg/dL)	150 (93)	164 (106)	0.185
Diabetes duration[b] (y)	7.80±5.60	9.45±6.28	0.014

[a] Sex was described as number (%) and compared between 2 groups via chi-square test.

[b] Normally distributed variables have been described as mean ± SD and were compared using independent *t* test between 2 groups.

[c] Non-normally distributed variables have been described as median (interquartile range) and compared using Mann-Whitney U test between 2 groups.

logistic regression. According to the results of training –test validation method, AUC for the test dataset, as the overall discrimination ability of this linear combination was 0.79.

Furthermore, the results of GAM model that can find nonlinear forms of association between variables show that age ($P < 0.001$) and TG ($P = 0.02$) with degree of 1, BMI with the cubic form ($P = 0.02$), and SBP with a quadratic form ($P = 0.01$) were significantly associated with the risk of functional limitation. The degree of freedom for each variable shows the shape of association of that variable with the outcome. AUC for the nonlinear combination obtained from GAM model in the test set was 0.82.

To detect functional limitation using SVM, accuracy, sensitivity and specificity for this model with RBF kernel (that is a nonlinear combination of underlying variables) were 0.99, 1 and .97 respectively. AUC in SVM with RBF kernel was 0.98. This value for SVM with linear kernel was obtained 0.67 (Table 2).

Regarding RAUC method, the results showed that AUC in the RAUC model with RBF kernel was gained 1, which was significantly higher than the logistic model (AUC = 0.79), GAM (AUC = 0.82), SVM with linear kernel (AUC = 0.67) ($P = 0.001$) but was not significantly different from SVM with RBF kernel (AUC = 0.98).

Discussion

The aim of this study was to search for the optimal combination of the independent variables to have a more accurate classification of diabetic patients according to their functional limitation. To do this, the new RAUC method with penalized loss function was used and its classification performance was compared with other existing methods such as logistic, SVM and GAM.

The results showed that the RAUC method can find the best combination of underlying variables and reveal the pattern of the relationship between variables using RBF kernel. According to the results, the linear classifiers including logistic regression and SVM with linear kernel had lower classification ability rather than nonlinear classifiers including GAM, SVM with RBF and RAUC with RBF kernel and these nonlinear methods outperformed the linear classifiers. Based on this data, the best nonlinear classifier was RAUC with RBF kernel which had an AUC of 1 for the test dataset and can perfectly classify subjects according to their functional limitation status without any misclassification. GAM had the lowest AUC among the nonlinear classifiers, but still had better performance than the linear classifiers.

Although both of the SVM and RAUC methods with RBF kernel could find a consistent nonlinear combination of markers, and the AUC criteria of them were close to each other, the DCA algorithm used by RAUC is less likely to be stuck in suboptimal local solutions compared to the algorithm employed by SVM. Furthermore, RAUC is much robuster than SVM in the presence of outliers (17).

In our data, the RAUC method with the RBF kernel function, which is a nonlinear combination of markers, outperforms logistic, GAM and SVM with linear kernel methods and is slightly better than SVM with RBF kernel. It shows that there is a strong non-linear relationship in the data and RAUC with nonlinear kernels could detect this nonlinear pattern and improve classification accuracy compared to the existing alternative methods. As the result showed, employing more efficient statistical methods could improve classification performance from 0.79 (for logistic as the routine and simple method) to 1 (for RAUC as the kernel-based method). This shows that we have to search in the data to discover real nature of associations. Keeping relationships very simple (for example, only linear relationship) leads to weak classification or prediction performance and high rate of misclassification while detection pattern of the relationship can lead to perfect performance (AUC = 1) as we could see in the present study.

The acceptable ability of underlying variables in detecting the functional limitation is expected because most of these variables are care indicators and previous studies about the risk factors of functional limitation in the diabetic patients have been shown that poor control of diabetes (based on Hemoglobin A1C) is related with disorder in functional performance, and people with worse controls and higher Hemoglobin A1C, are more likely to suffer from daily dysfunction and disability than those with better control of the disease (6).

The results of a study by Salehi et al in Iran on determining risk factors for retinopathy using GAM in diabetic patients show significant non-linear relationships between diabetes duration, hemoglobin A1C and systolic blood pressure with diabetic retinopathy. Therefore, the duration of diabetes and hemoglobin A1C (degree 5) and systolic blood pressure (degree 2) were associated with retinopathy outcome (28).

In addition, several studies have shown the efficiency of the SVM method in the correct classification of data. A systematic review of the applications of machine learning algorithms and data mining techniques in the various field of diabetes research consisting of prediction and diagnosis, complications, genetic background and environment, and health care and management showed that SVM was

Table 2. The AUC criteria as the Overall Classification Performance of the Considered Methods

Model	Area under the curve (AUC)				
	Logistic	GAM	SVM (Linear Kernel)	SVM (RBF Kernel)	RAUC (RBF Kernel)
AUC	0.79	0.82	0.67	0.98	1

the most successful and widely used algorithm (29). In addition, in one study that has examined the capability of machine learning methods in diabetes research, SVM and other data mining techniques were used to diagnose and predict diabetes; the results showed better performance of the SVM algorithm rather than other methods (30).

There is no study which used RAUC method to assess functional limitation or other complications of diabetes, however, it was employed for finding the combinations of biomarkers from blood samples to assess vaccine-induced protection (17).

In another study, the extended form of the RAUC algorithm has been used to select the best variables in the presence of a large number of candidate markers (31).

Conclusions

Due to the importance of the accuracy of the result in medical studies, identification and consideration of complex and nonlinear relationships in the data can be so helpful to have accurate prediction or classification of an outcome.

Limitations

Although, the sample size for analysis is appeared to be sufficient, and the results indicate an acceptable precision of the estimations, missing data and the lack of complete information of all patients are the limitations of this study.

Conflicts of Interests

Authors declare that they have no conflict of interests.

Ethical Issues

The present study was approved by the Ethics Committee of Tabriz University of Medical Sciences, (Ethics number: TBZMED.REC.1395.794). Here are Ethics approvals for the first 2 studies as the source of data for the present study: (Ethics numbers TBZMED.REC.1392.207 and TBZMED.REC.1394.55).

Financial Support

This research was financially supported by the Vice Chancellor for Research of Tabriz University of Medical Sciences.

References

1. Roglic G. WHO Global report on diabetes: A summary. Int J Noncommun Dis. 2016;1(1):3-8.
2. Schram MT, Baan CA, Pouwer F. Depression and quality of life in patients with diabetes: a systematic review from the European depression in diabetes (EDID) research consortium. Curr Diabetes Rev. 2009;5(2):112-119.
3. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2004;27 Suppl 1:S5-s10. doi:10.2337/diacare.27.2007.S5
4. Volpato S, Ferrucci L, Blaum C, et al. Progression of lower-extremity disability in older women with diabetes: the Women's Health and Aging Study. Diabetes Care. 2003;26(1):70-75.
5. Gregg EW, Beckles GL, Williamson DF, et al. Diabetes and physical disability among older U.S. adults. Diabetes Care. 2000;23(9):1272-1277.
6. De Rekeneire N, Resnick HE, Schwartz AV, et al. Diabetes is associated with subclinical functional limitation in nondisabled older individuals: the Health, Aging, and Body Composition study. Diabetes Care. 2003;26(12):3257-3263.
7. Manas LR, Sinclair AJ. Diabetes and functional limitation. In: Sinclair AJ, Dunning T, Rodríguez Mañas L, Munshi M. Diabetes in Old Age. John Wiley Sons, Ltd; 2017:213-224. doi:10.1002/9781118954621.ch16
8. Maty SC, Fried LP, Volpato S, Williamson J, Brancati FL, Blaum CS. Patterns of disability related to diabetes mellitus in older women. J Gerontol A Biol Sci Med Sci. 2004;59(2):148-153.
9. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care. 2007;45(5 Suppl 1):S32-38. doi:10.1097/01.mlr.0000246649.43232.82
10. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. John Wiley Sons, Inc; 2008:i-vii. Wiley Series in Probability and Statistics. doi:10.1002/9780470317082
11. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861-874. doi:10.1016/j.patrec.2005.10.010
12. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997;55(1):119-139. doi:10.1006/jcss.1997.1504
13. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. Biostatistics. 2000;1(2):123-140. doi:10.1093/biostatistics/1.2.123
14. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics. 2006;62(4):961-971. doi:10.1111/j.1541-0420.2006.00578.x
15. Komori O. A boosting method for maximization of the area under the ROC curve. Ann Inst Stat Math. 2011;63(5):961-979. doi:10.1007/s10463-009-0264-y
16. Xu-hui W, Ping S, Li C, Ye W. A ROC Curve Method for Performance Evaluation of Support Vector Machine with Optimization Strategy. International Forum on Computer Science-Technology and Applications; 2009.
17. Fong Y, Yin S, Huang Y. Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. Stat Med. 2016;35(21):3792-3809. doi:10.1002/sim.6956
18. Shamshirgaran SM, Mamaghanian A, Aliasgarzadeh A, Aiminisani N, Iranparvar-Alamdari M, Ataie J. Age differences in diabetes-related complications and glycemic control. BMC Endocr Disord. 2017;17(1):25. doi:10.1186/s12902-017-0175-5
19. Ataei J, Shamshirgaran S, Iranparvar Alamdari M, Safaeian A. Evaluation of Diabetes Quality of Care Based on a Care Scoring System among People Referring to Diabetes Clinic in Ardabil, 2014. J Ardabil Univ Med Sci. 2015;15(2):207-219.
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-1379.

21. Kamuhabwa AR, Charles E. Predictors of poor glycemic control in type 2 diabetic patients attending public hospitals in Dar es Salaam. Drug Healthc Patient Saf. 2014;6:155-165. doi:10.2147/dhps.s68786
22. Sazlina SG, Mastura I, Cheong AT, et al. Predictors of poor glycaemic control in older patients with type 2 diabetes mellitus. Singapore Med J. 2015;56(5):284-290. doi:10.11622/smedj.2015055
23. Gubhaju L, Banks E, MacNiven R, et al. Physical Functional Limitations among Aboriginal and Non-Aboriginal Older Adults: Associations with Socio-Demographic Factors and Health. PLoS One. 2015;10(9):e0139364. doi:10.1371/journal.pone.0139364
24. Dobson AJ, Barnett A. An Introduction to Generalized Linear Models. 3rd ed. Boca Raton, FL: Chapman Hall/CRC Press; 2008.
25. Wood SN. Generalized Additive Models: An Introduction With R. CRC Press; 2006.
26. Vladimir NV. The Nature of Statistical Learning Theory. New York: Springer-Verlag, Inc; 1995.
27. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273-297. doi:10.1023/a:1022627411411
28. Salehi M, Vazirinasab H, Khoshgam M, Rafati N. Application of the generalized additive model in determination of the retinopathy risk factors relation types for Tehran diabetic patients. Razi J Med Sci. 2012;19(97):1-9.
29. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104-116. doi:10.1016/j.csbj.2016.12.005
30. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10:16. doi:10.1186/1472-6947-10-16
31. Huang Y. Identifying optimal biomarker combinations for treatment selection through randomized controlled trials. Clin Trials. 2015;12(4):348-356. doi:10.1177/1740774515580126