Semiparametric Estimation of AUC from Generalized Linear Mixed Model

Okeh UM[1]

[1] Ebonyi State University,Abakaliki,Nigeria.

Received: 13 February 2015 Accepted: 1 March 2015 Published: 15 March 2015

Abstract

Methods of evaluating the performance of diagnostic tests are of increasing importance in medical science. When a test is based on an observed variable that lies on a continuous scale, an assessment of the overall value of the test can be made through the use of a Receiver Operating Characteristic (ROC) curve. The ROC curve describes the discrimination ability of a diagnosis test for the diseased subjects from the non-diseased subjects. The area under the ROC curve (AUC) represents the probability that a randomly chosen diseased subject will have higher probability of having disease than a randomly chosen non-diseased subject. Semi-parametric being a ROC curve estimation method is widely used in making inferences from diagnostic test results that are at least measurements on ordinal scale. In this paper, we proposed a method of semi-parametric estimation in which predicted probabilities of discordant pairs of observation are obtained from generalized linear mixed model (GLMM) and used in modeling ROC and AUC. The AUC obtained which is time dependent is equivalent to the Mann-Whitney statistic (Hanley and McNeil, 1982) often applied for comparing distributions of values from the two samples.

Index terms— AUC, ROC, GLMM, GDM, semi-parametric, mann-whitney.

1 Introduction

n health studies, the diagnosis of a patient are very often based on some classification errors calibrated based on the sensitivity and specificity. An individual presenting for a screening test for a disease, is discriminated based on a cut-off value c whether he/she is healthy or diseased when test results are measurements on at least the ordinal scale. Many procedures exist for estimating the accuracy of test measurements such as the parametric, nonparametric and semi-parametric methods and their associated summary measures. In this paper, we will propose a semi-parametric regression type method of obtaining predicted probabilities from the Generalized Linear Mixed Model (GLMM) and using them to model the receiver operating characteristic (ROC) curve and area under the ROC curve(AUC) for continuous binary test results that are time dependent.( ) { } (.) ( ), ( ) , (,)**(1)**ROC FPR c TPR c c = ? ?? ?

The accuracy of ROC is summarized by the AUC given as ( )1 0 ( ) . (**2**) AUC P X Y ROC t dt = > = ?

This is the probability that a randomly chosen diseased subject will have higher probability of having disease than a randomly chosen non-diseased subject.

Since different estimation methods can provide a span of estimated AUC values on the same data set, their properties are always examined in order to provide a recommendation as to the preferred approach. Dorfman and Alf (1969) proposed a parametric iterative method for obtaining the maximum likelihood estimates of the parameters of a bi-normal ROC curve to model ordinal data. They assumed that test results for the diseased (X) and non-diseased (Y) populations are normally distributed respectively as I Suppose Y and X denotes the test result of subjects with and without disease respectively. Let c be cut-off value. Then P(X > c) = G(c) and P

(Y > c)= F(c) where F(c) is sensitivity and 1-G(c) represents specificity. Therefore ROC is a plot of F(c) versus G(c) for all possible thresholds, c. In terms of TPR and FPR at c, (
, , .X X Y Y X N and Y N µ ? µ ? ? ?**(3)**

While parametric binormal ROC curve is given as( ) 1 ( ) ( ) ,0 1, ROC t a b t t ? =? + ? ? ? , .**(5)**X Y Y X X where a b µ µ ? ? ? ? = =

Here a and b are parameter estimates which gives the statistical inference while denotes the standard normal cumulative distribution function. By algebraic simplification, the AUC is given as:( ) ( ) 2 2 2 (6) 1 X Y X Y a AUC b µ µ ? ? ? ? ? ? ? ? ? ? = ? = ? ? ? ? ? ? ? ? + + ? ? ? ?

Reiser and Faraggi(2002) and Goddard and Hinberg (1990) proposed the transformation (say logarithmically) of test results and making it normal due to the violation of the normality assumption. They proposed the transformed normal (TN) approach which is a parametric estimation method based on the normal theory. It involves applying a Box-Cox power transformation **??**Box and Cox,1964) to the data and subsequently using the N estimator to the transformed data.

In general, the problems identified with maximum likelihood method of estimating parameters in parametric method is the inability of the parameter estimates to quickly attain convergence because it is an of iterative method. There exists also the restrictive assumptions of normality or transformation to normality of the parametric method about the distribution of test results making the estimates inconsistent thereby giving a misleading picture of the regression relationship when the assumption is violated **??**Pepe,2003).

According to Hanley and McNeil (1982), the empirical non-parametric method uses the MW statistic in estimating ROC curves. As usual, they are used when the normality assumption for test results is violated. Here AUC is calculated using the MW version of the twosample rank-sum statistic of Wilcoxon as ( ) **??**)0 1 1 1 1 0 1 ?, (**7**)n n i j i j AUC Y Y n n + ? = = = ? ? ? ( ) 1 1 , (**8**) 2 0 i j i j i j i j if Y Y where Y Y if Y Y if Y Y + ? + ? + ? + ? ? > ? ? ? = = ? ? ? < ?0 1 1 1 1 0 1 1 ?(10) 2 n n i j j i i j AUC P Y Y P Y Y n n + ? ? + = = = > + = ? ?

In general, nonparametric estimation method does not yield smooth curve, especially in small samples (Zou et al, 1998). They models avoid restrictive assumptions of the functional form of the regression function. There is also lack of a one to one correspondence between TPR and FPR values makes inference awkward (Zou et al, 1998). Dodd and Pepe (2003) proposed a semiparametric AUC regression model for data with a nonnormally distributed response variable which can adjust for continuous and discrete covariates. Assume that one needs to adjust the AUC for a covariate X, the covariatespecific AUC can be expressed as ( ), (**11**)D D ij i j i j AUC P Y Y X X = >

Where is the ith response in diseased (or treatment) group with covariate value and is the jth response in non-diseased (or control) group with covariate value Often one is interested in estimating the AUC at a specified covariate level, i.e.

2 (

). (**12**)D D i j i j P Y Y X X X > = =

Dodd and Pepe applied this model to the GLM framework which allows one to model the AUC with covariates, in which case their model can be written as, **??**), (**13**) T ij ij g AUC X ? =

where g is a monotone link function such as the probit or logit link, Xij is a vector function of , and is a vector fixed and unknown parameters to be estimated. Note that ( )( ) . (**14**)D D i j ij ij E I Y Y X AUC > =

Thus, for estimating the parameters in the model, Dodd and Pepe proposed the use of the logistic regression model where the response variable is a Bernoulli variable Dodd and Pepe demonstrated that the estimates of parameters are found as solution to the usual score equations given by ( ) ( ), (**15**) D D N N ij ij ij i j ij I AUC AUC V I ? ? ? ? ? ? ? Where ( ) . D D ij i j I I Y Y = >

Therefore, one obtains this estimate using standard statistical software.

According to Colak et al (2012) as well as Wolfgang et al(2004),the most preferred method of estimation is the semi-parametric method because it combines the flexibility of the nonparametric method with the advantages accruable to the parametric procedure in achieving better results. Semi-parametric (SP) approach is an intermediate strategy between parametric and non-parametric methods for estimating the ROC curve in the sense that it assumes a parametric bi-normal form for the ROC curve, but does not assume that the diagnostic test results follow any particular distribution. This informed the choice of the method in this work.

II. that on the average a randomly selected subject from the population test or respond positive to the condition under study while the variance is given as 2 I ? , where I is an n x n identity matrix. The estimation of ? can be carried out using the least square method by obtaining ? as the best estimate of ? through the minimization of the sum of squared errors. The result is

3 Linear Regression Model

( ) 1 ?(17) X X X Y ? ? ? ? = Where ( ) 1 2 ?, ( ) N X X ? ? ? ? ? ? and 1 ( ) X X ? ?

is the inverse of the nonsingular variance-covariance matrix.

4 III.

Generalized Linear Model (GLM) GLM is an extension of the linear regression model and for modeling binary data, GLM is made up of a linear predictor given as **??**) ( )1 1 ( ) ( ) (20) Va rY V g X V g ? ? ? ? = =

Meanwhile, GLMM is a model extension of GLM in which the linear predictor contains both fixed effects and random effects (McCullagh and Nelder, 1989). In matrix notation, it is given as(21) Y X Zu ? ? ? ? = + = + + ( ) ( ) 0, ; 0, ; (,) 0; (,) 0.

whereu N G N R E u Cov u ? ? ? = = ? ?

As defined previously for Y, ? is a p x 1 column vector of fixed effects, u is a q x 1 vector of random effects, ? is a n x 1 vector of random error terms, X is the n x p design matrix for the fixed effects relating to ?, Z is the n x q design matrix for the random effects relating to u. The structure of the covariance matrices of G and R specifies the structure of correlation among the random effects and error term respectively. The variance of Y for GLMM is given as:( ) (22) V Y ZGZ R ? = +

Where Z is a diagonal matrix and A is a diagonal matrix that contains the variance functions of the model.

5 IV.

6 The Proposed Method

To obtain the predicted probability from GLMM, we incorporate the time of measurement of binary data for subjects having n observations. Since the binary logistic model is a linear relationship between the natural logarithm and the linear component. Then(23) 1 it it it it i it In X Z u ? ? ? ? ? ? ? = = + ? ? ? ? ? ? it

where ? is the predicted probability of the positivity of ith randomly selected subject at time t for 1, 2,..., ; 1, 2,..., i n t T = = . Here T is total time period and it ? is the linear predictor for ith subject at time t. 1 1 1 1 1 1 (25) X RY X R X X R Z u Z R Y ZR X Z R Z G ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? = ? ? ? ? ? ? ? ? ? + ? ? ? ? ? ?

These estimates are respectively obtained and the solution is given as ( ) ( )1 1 1 1 ?, (26) X V X X V Y u GZ V Y X where V ZGZ R ? ? ? ? ? ? ? ? ? ? = = ? ? = + V.

7 Constructing Roc Curve

The estimated predicted probability will then serve as a bio-marker for constructing the ROC curve for discriminating a diseased subject from a non-diseased subject longitudinally. The procedure is first to obtain estimates of sensitivity and specificity from a four-fold table so as to have insufficient pairs of sensitivity and 1specificity that are incapable of producing the actual ROC curve analysis. To obtain sufficient pairs capable of generating the actual smooth ROC curve, a series of pairs of sensitivity and 1-specificity up to the sample size under consideration (sn(1),1-sp(1)),...,(sn(n),1-sp(n)) is calculated from varying cuts of positivity escalated by increments of 0.005 in predicted probability. The ROC curve is created by plotting for n number of subjects at t time, n pairs of sensitivity and 1specificity data points starting with the strictest positive criterion of 1 to the loosest positive criterion of 0.005.

The AUC is given in a closed form for the purpose of this study as:( ) 1 , , 0 , (**27**) X Z X Z AUC ROC t dt = ?

This is the ROC value with false-positive rate t that is associated with the fixed effect predictor X and random effects predictor Z where the integration limits run from 0 to 1. Due to the difficult nature of obtaining the result as seen by other authors **??**Dorfman et al,1969), we will alternatively construct AUC based on predicted probabilities from binary measure models, by adapting the MW method to compare the size of the predicted probabilities of each discordant pair. This is achieved by dichotomizing the predicted probability so that two probabilities given as ( ) Estimating auc from Estimated Predicted Probability represent predicted probability of the diseased and nondiseased responses for the ith subject respectively at time t for the binary measure design. The MW method is the choice because under the GLMM framework, there is no simple closed-form solution of the ROC curve and the MW method yields ROC estimates with a good precision. Here the AUC is given as11 1 1 (28) n T it i t D D AUC u n n = = = ? ? Where D D

n and n are the numbers of observed values for the diseased and non-diseased subjects respectively while t and T are time of test measurement and total time period of measurement respectively.

Also it u is a function comparing the test result of ith subject with and without disease at time t. The total number of (discordant pairs) sample observations, n as:

(29)D D n n n = +

The difference between the AUC given above and that suggested by other authors such as Hanley and McNeil (1982) is that here AUC is calculated from predicted probabilities that are time dependent instead of test scores. For each discordant pair, ordering of the corresponding predicted probabilities are compared in relation to the observed outcome values, and the AUC is calculated based on these ordering results so as to compare the size of the predicted probabilities of each discordant pair. In binary measure design, where there exist complete discrimination of health status, each subject has two possible mutually exclusive outcomes either Yes (diseased coded1) or No (non-diseased usually coded 0) whose values may vary from time to time. This is represented as 1, 0,**(30) 1, 2,..., ; 1, 2,...**? = ? ? ? = =

The values of 0 and 1 as outcomes of this function shows that the subjects health status are well discriminated ??Bernd et al, 2003; ??olak et al, 2012). Evaluation of this function through the ordering procedure gives the unbiased estimate suitable for use in calculating the AUC.

8 VII.

9 Illustrative Example

The data for this study were obtained from the medical record units of five randomly selected hospitals in Ebonyi State, Nigeria. The data represents binary test results of 1114 pregnant women susceptible for gestational diabetic mellitus (GDM).These are measurements taken at various time periods (trimesters).

10 Data Analysis and Results

The data analysis was assisted using SAS version 8 software and the results of semi-parametric roc analysis with their graphs are shown in table 2 below. 2 ? value at one (1) DF and the 95% C.I indicates highly statistically significant relationship(strong degree of association) between screening test results and state of nature or condition (GDM) for all the trimesters. For all the trimesters, ROC curve analysis showed that (see Fig.

11 Discussion

In the present study the cutoff values of GCT in 1st, 2nd, 3rd, and all trimesters were 184, 177, 179, and 179 mg/dl respectively. These values were higher than the previous reports obtained outside Nigeria that recommended the use of 50g GCT level at 130-140 mg/dl for screening of GDM in pregnant women at risk for GDM between 24-28 weeks of gestation (Friedman et al, 2006; ??erger et al, 2002;Miyakoshi et al, 2003; ??itoratos et al,1997). Also Vitoratos et al (1997) and Tanir et al (2005) recommended 126 mg/dl and 185 mg/dl respectively in their study. These are due to differences in race and nutrition of the populations involved. This study also showed that semi-parametric GLMM method provided reliable, unbiased, and consistent estimates for the parameters and AUC. Similar results were obtained by Colak et al (2012).

X.

12 Summary and Conclusions

ROC analysis revealed varying cut-off values of 184,177, 179 and 179 mg/gl for the I st , 2 nd ,3 rd and all trimesters and a common cut-off value of 177 mg/dl is chosen for screening 50 grams GCT irrespective of the trimester and is rather suitable for high BMI or obese pregnancy. These variable cutoff values of 50g GCT for screening of GDM is because of increasing weight as pregnancy progresses. Race and nutrition of the population causes differences in cut-off values of 50g GCT for screening women at risk for GDM. High values of NPV such as 92.73-94.82%, indicates the existence of low false negative. Semi-parametric procedure of obtaining predicted probabilities from GLMM because the predicted probabilities of this method have a high statistical efficiency since for all the trimesters, there exist statistical significance. These estimators showed high statistical efficiency. A common cut-off value of 177 mg/dl is recommended for screening 50 grams GCT irrespective of the trimester. Based on the findings in this study, pregnant women from thirty years of age, have greater number of risk of getting GDM at their 2 nd and 3 rd trimester than those in their 1st trimester of gestation age. It is advised that such category of women should start living healthy life style. Semi-parametric method is preferred to other methods for estimating ROC and constructing AUC because it is more superior in terms of simplicity and accuracy of results .It is therefore recommended.

Figure 1:

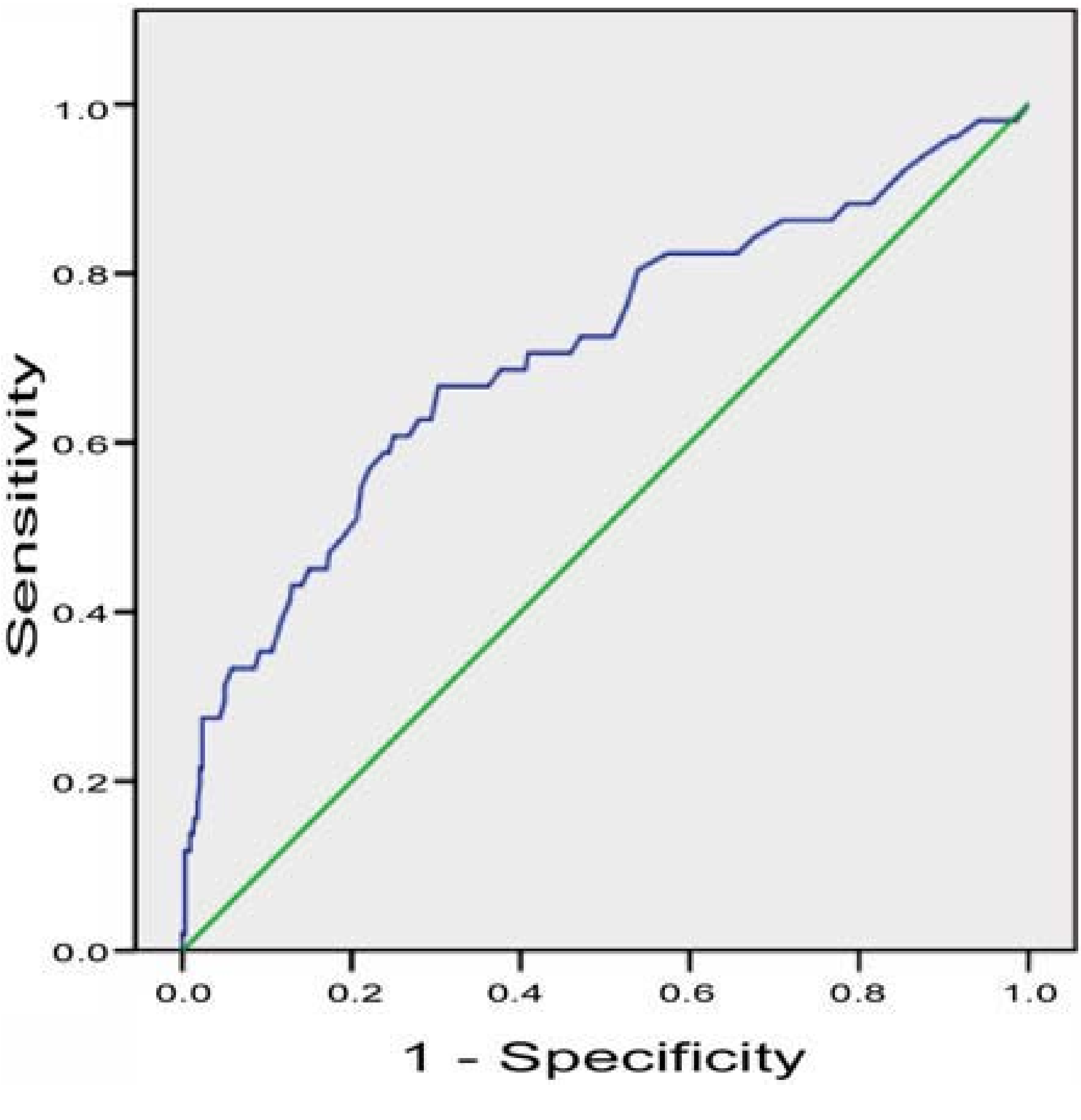


Figure 2: =

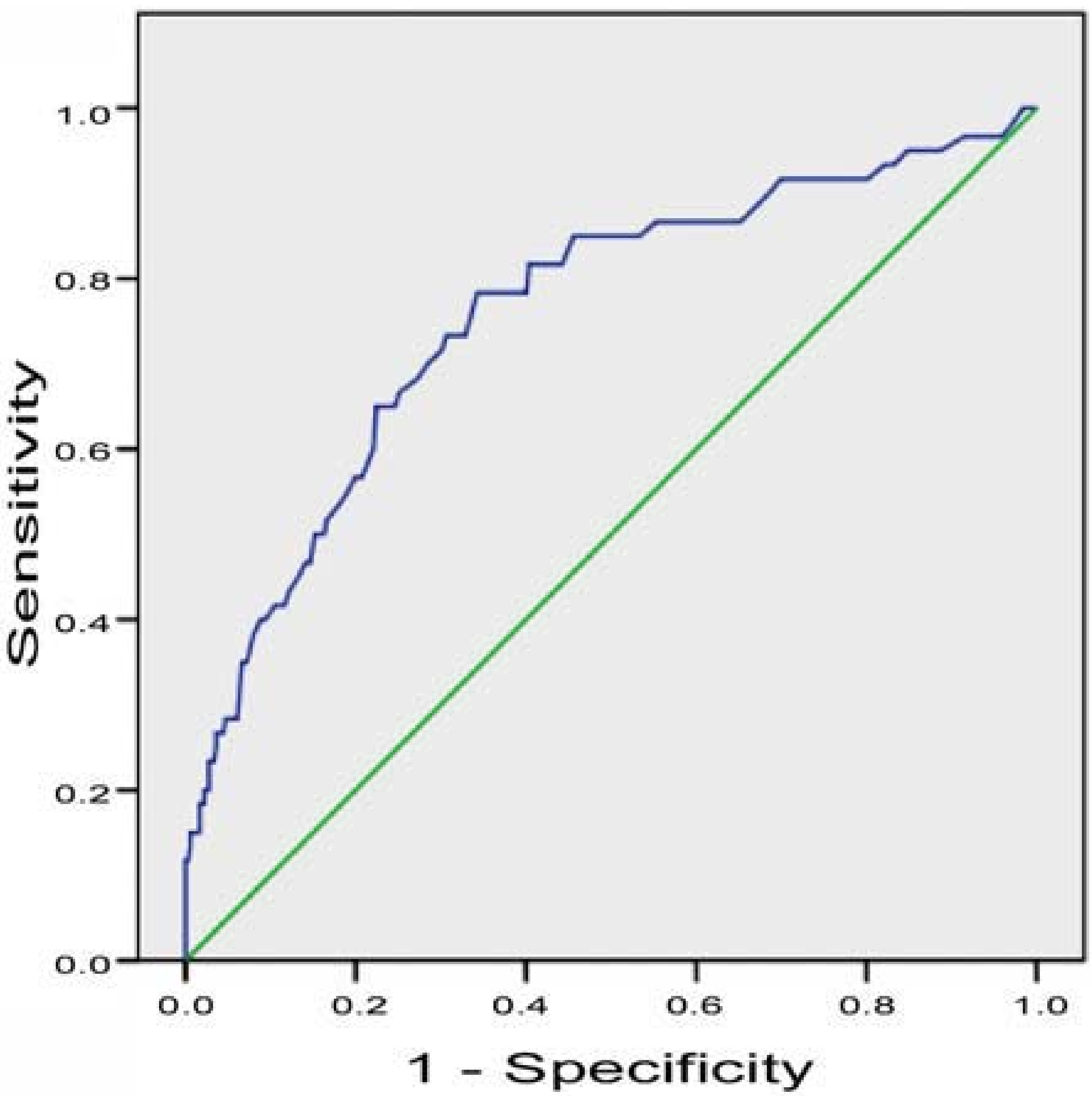


Figure 3:

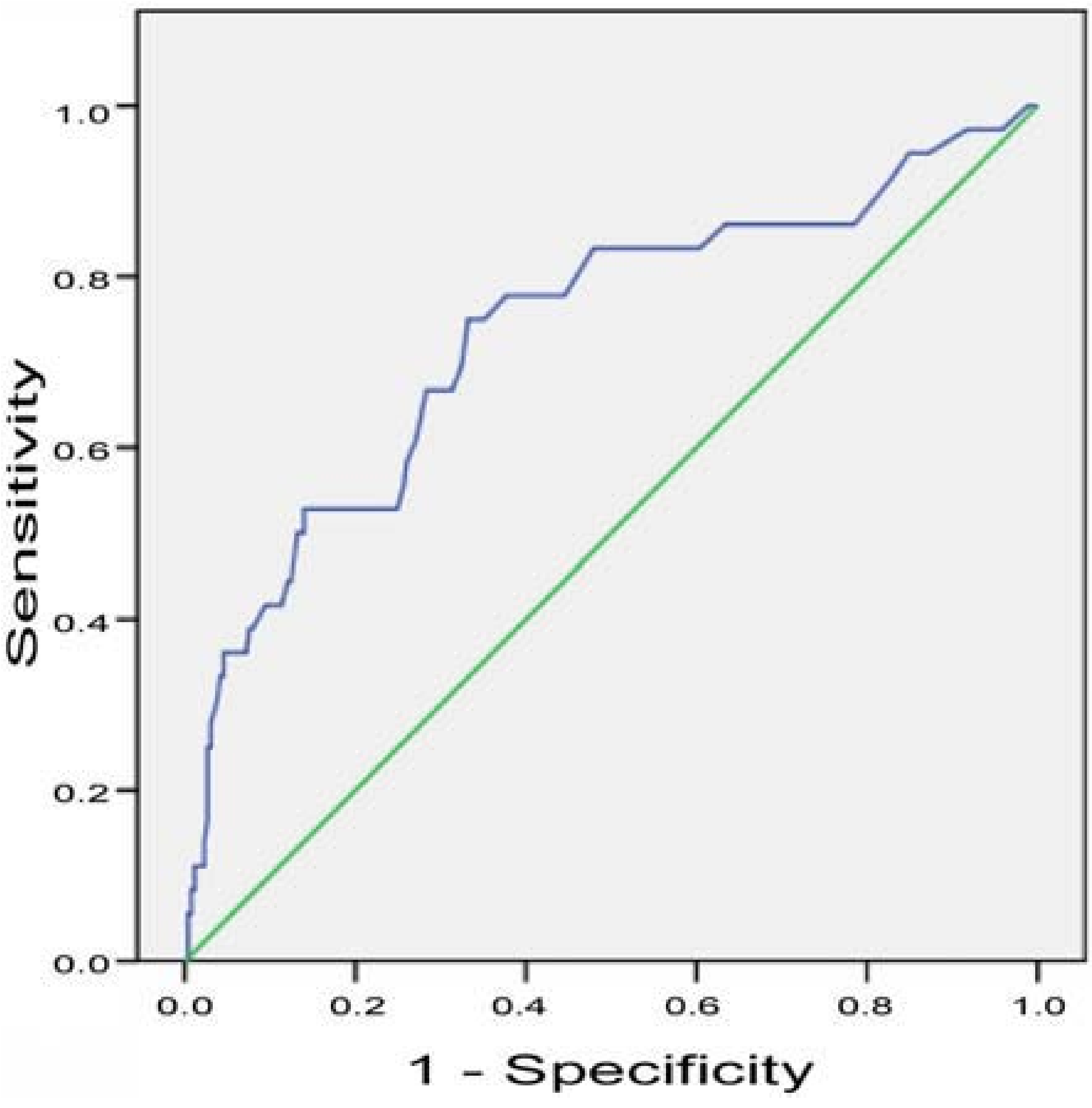
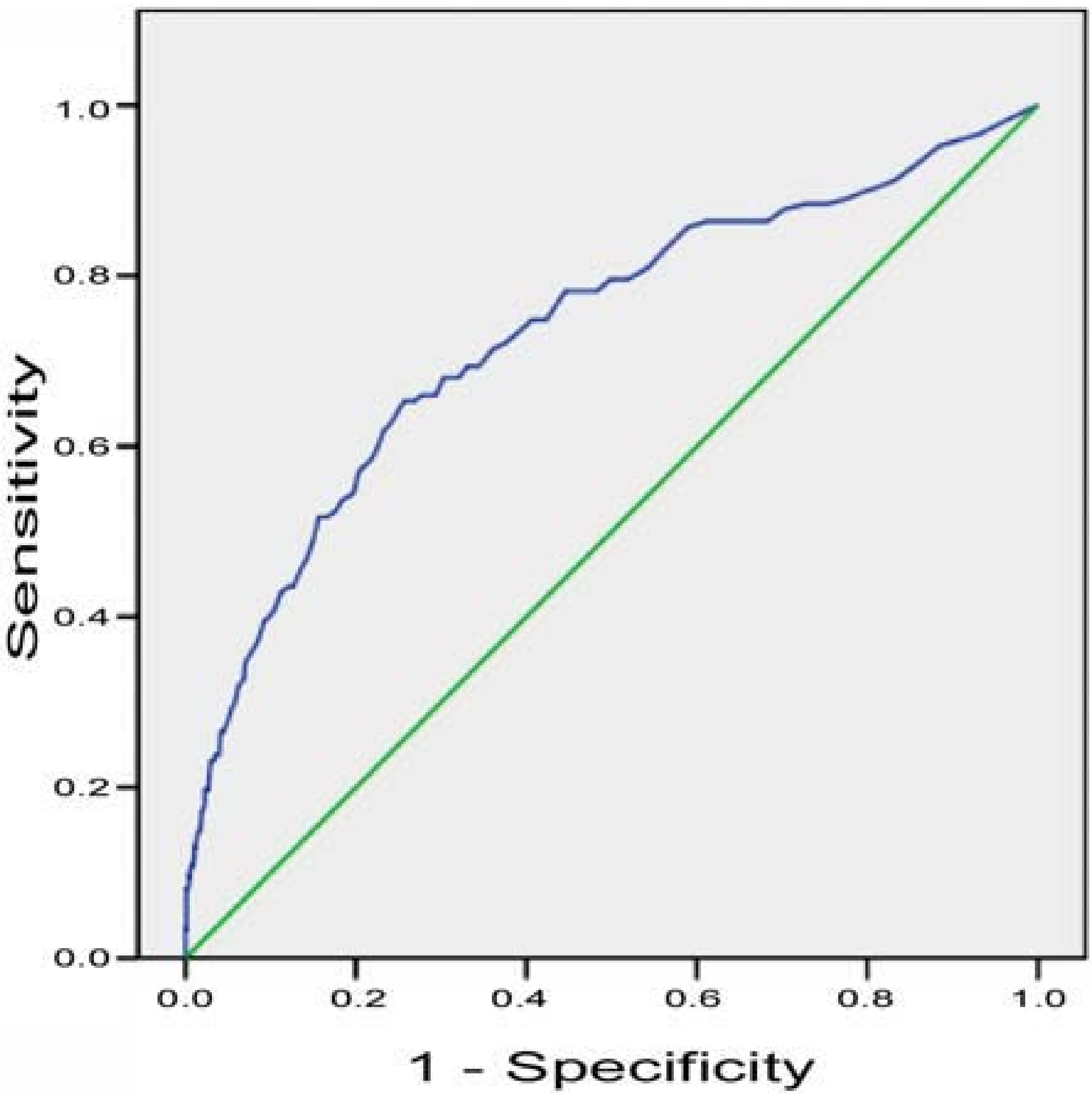


Figure 4: Semiparametric



1123

Figure 5: 1 -Figure 1 :Figure 2 :Figure 3 :

	?	it if x is the test score in the ith subject screened at
it u	time t that tested positive	
		otherwise
for i		n t T

Figure 6:

1

Year 2015

Figure 7: Table 1 :

2

Trimesters	1 st	2 nd	3 rd	All
Cutoff value of GCT with max AUC	184	177	179	179
Sensitivity with 95% CI	50.00 (44.35-55.65)	60.78 (55.94-65.62)	78.33 (74.4-82.26)	65.31 (62.51-68.1)
Specificity with 95% CI	86.79 (82.97-90.62)	75.00 (70.71-79.29)	65.75 (61.22-70.27)	74.35 (71.79-76.92)
PPV with 95% CI	33.96 (28.61-39.31)	26.72 (22.34-31.11)	27.49 (23.23-31.74)	27.91 (25.27-30.54)
NPV with 95% CI	92.74 (89.81-95.67)	92.73 (90.15-95.3)	94.82 (92.71-96.94)	93.38 (91.92-94.84)
Max. AUC with 95% C.I.	0.684(0.59-0.77)	0.6789(0.61-0.75)	0.7204(0.65-0.77)	0.6983(0.66-0.74)
D n	265	340	362	967
D n	36	51	60	147
?	1.578	1.446	1.430	1.409
û	1.170	1.007	0.966	0.932
Predicted Probability(it ?)	0.6857	0.7101	0.8234	0.9210

Figure 8: Table 2 :

[Engelmann et al.] , Bernd Engelmann
RISK JANUARY 2003 WWW.RISK.NET , Evelyn Hayden
RISK JANUARY 2003 WWW.RISK.NET , Dirk Tasche
RISK JANUARY 2003 WWW.RISK.NET .

[Tanir et al. ()] 'A tenyear gestational diabetes mellitus cohort at a university clinic of the mid-Anatolian region of Turkey'. H M Tanir , T Sener , H Gurer , M Kaya . *Clinical & Experimental Obstetrics & Gynecology* 2005. 32 (4) p. .

[Box and Cox ()] 'An analysis of transformations'. Gep Box , D R Cox . *Journal of the Royal Statistical Society, Series B* 1964. 26 p. .

[Miyakoshi et al. ()] 'Cutoff value of 1 h, 50 g glucose challenge test for screening of gestational diabetes mellitus in a Japanese population'. K Miyakoshi , M Tanaka , K Ueno , K Uehara , H Ishimoto , Y Yoshimura . *Diabetes Res Clin Pract* 2003. 60 p. .

[Faraggi and Reiser ()] 'Estimation of the Area Under the ROC Curve'. D Faraggi , B Reiser . *Statistics in Medicine* 2002. 21 p. .

[Mccullagh and Nelder ()] *Generalized linear models*, P Mccullagh , J A Nelder . 1989. New York: Chapman Hall.

[Friedman et al. ()] 'Glucose challenge test threshold values in screening for gestational diabetes among black women'. S Friedman , F Khoury-Collado , M Dalloul , D M Sherer , O Abulafia . *Am J Obstet Gynecol* 2006. 194 p. .

[Dorfman and Alf ()] 'Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-ratingmethod data'. D D Dorfman , E Alf . *J Math Psych* 1969. 6 p. .

[Hardle et al. ()] 'Non-parametric and Semiparametric models'. Wolfgang Hardle , Marlene Muller , Stefan Sperlich . *Axel Werwa Z* 2004. Springer.

[Mann and Whitney ()] 'On a test of whether one of two random variables is stochastically larger than the other'. H Mann , Whitney . *Annals of Mathematical Statistics* 1947. 18 p. .

[Zou et al. ()] 'Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones'. K H Zou , C M Tempany , J R Fielding , S G Silverman . *Academic Radiology* 1998. 5 p. .

[Goddard and Hinberg ()] 'Receiver operator characteristic (ROC) curves and non-normal data: An empirical study'. M J Goddard , I Hinberg . *Statistics in Medicine* 1990. 9 p. .

[Berger et al. ()] 'Screening for gestational diabetes mellitus'. H Berger , J Crane , D Farine , A Armson , R S De La , L Keenan-Lindsay . *J Obstet Gynaecol Can* 2005. 24 p. .

[Semiparametric Estimation of AUC from Generalized Linear Mixed Model] *Semiparametric Estimation of AUC from Generalized Linear Mixed Model*,

[Dodd and Pepe ()] 'Semiparametric regression for the area under the receiver operating characteristic curve'. L E Dodd , M S Pepe . *Journal of the American Statistical Association* 2003. 98 p. .

[Hanley and Mcneil ()] 'The meaning and use of the area under a receiver operating characteristic ROC curve'. J A Hanley , B J Mcneil . *Radiology* 1982. 143 p. .

[Pepe ()] *The Statistical Evaluation of Medical Tests for Classification and Prediction*, M S Pepe . 2003. New York, NY, USA: Oxford University Press.

[Vitoratos et al. ()] 'Which is the threshold glycose value for further investigation in pregnancy?'. N Vitoratos , E Salamalekis , P Bettas , D Kalabokis , A Chrisikopoulos . *Clin Exp Obstet Gynecol* 1997. 24 p. .