Pak. J. Statist.
2014 Vol. 30(2), 181-196

PREDICTION OF DISEASE STATUS: TRANSITION MODEL APPROACH FOR REPEATED MEASURES

M. Ataharul Islam[1§] and **Rafiqul I. Chowdhury**[2]

[1] Department of Statistics and OR, King Saud University, PO Box 2455, Riyadh 11451, Saudi Arabia. Email: mataharul@yahoo.com

[2] Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada. Email: mchowd23@uwo.ca

[§] Corresponding author

ABSTRACT

This paper develops models for prediction of disease status from longitudinal data. The estimation of area under curve (AUC) is illustrated on the basis of estimates of sensitivity and specificity for repeated binary outcomes of disease status. There are several research papers in this field on cross-sectional data but only a few dealt with the repeated observations. This paper shows the procedures to deal with repeated observations employing Markov models. These procedures employ covariate dependent Markov models for estimating sensitivity and specificity, which in turn, produce the estimates for area under curve. The tests for equality of areas under curve for two models are also suggested. An application is illustrated for depression data from the Health and Retirement Survey, USA. The results indicate that the transition model approach can reveal the matching of disease status very efficiently; an estimate of more than 0.96 was obtained for the AUC for a transition model based prediction of disease from the depression data.

KEY WORDS

Conditional Model; Goodness of fit; Markov Model; Prediction of Disease; Transitions.

1. INTRODUCTION

The prediction of disease status has emerged as an important area of research. The relationship between the underlying risk factors and the disease status at different times produce longitudinal data. In predicting the disease status, the longitudinal data provide the necessary trajectories. During the past years, many attempts have been made to propose models based on the potential risk factors to predict disease status as well as to determine the performance of such models in determining the matches between observed and expected outcomes. Some of the examples of prediction models include: homelessness within three months of discharge among inpatients with schizophrenia (Olfson et al., 1999), nerve function impairment in leprosy patients (Croft et al., 2003), pressure ulcer development (Schoonhoven et al., 2006), risk of depressive episode in adolescents (Van Voorbees et al., 2008), return of spontaneous circulation in intervals without chest compressions during out-of-hospital cardiac arrest (Gundersen et al., 2009).

Similarly, Marwick et al. (2001), Biagini et al. (2005), Boyko et al. (2006) and Zethelius et al. (2008) worked in this area of research. Most researchers in this field employed either the logistic regression or the proportional hazards regression models. Gundersen et al. (2008, 2009) used mixed effects logistic regression models. A multistate transition model with autoregressive logistic regression was proposed by de Vries et al. (1998); earlier Agresti (1997) suggested a model for repeated measurements of multivariate binary response. Some of the more in-depth studies were conducted using the regressive models. The regressive logistic regression models are based on dynamics present in the repeated observations that emerged from the longitudinal data. These models are based on the regressive logistic regression formulation of Bonney (1986, 1987). Sequences of transitions or situations using conditional probabilities are proposed by these models. The transitional models for first order (Muenz and Rubinstein, 1985) or higher orders (Islam and Chowdhury, 2006, 2007; Islam, Chowdhury and Huda, 2009) can also be employed in the prediction of disease status. In other words, it has been customary to employ the longitudinal data to predict the disease status but the research works on performance of predictions are concentrated on the diagnostics for cross-sectional data.

To assess the performance of prediction models, the diagnostic procedures based on the receiver operating characteristic (ROC) curves have been proposed (Delong and Clarke-Pearson, 1988; Pepe, 2000; Rodenberg and Zhou, 2000; Zhang et al., 2002; DeLong, Obuchowski, 2006; Pencina et al., 2008; Qin and Zhou, 2006; Bandos et al., 2009). The tests for area under curve (AUC) have been suggested for both uncorrelated and correlated data. Steyerberg et al. (2010) compared the techniques for assessing the performance of prediction models for binary outcomes. Pepe et al. (2004) showed the limitations of the odds ratio in assessing the performance of a diagnostic, prognostic or screening marker. They demonstrated that merely being associated with outcome does not ensure a good performance, but the measure of assessing performance needs to be based on sensitivity and specificity. However, these tests are provided mostly for cross-sectional data. In longitudinal analysis, the predictions need to be based on repeated measures data. Hence, the test procedures have to be extended for the repeated measures data in order to take account of the transitions in disease status between two or more time points. Islam and Chowdhury (2010) proposed regressive models for a sequence of transitions in longitudinal data. These models are employed to predict the future status of outcome variable of the individuals on the basis of their underlying background characteristics or risk factors. To measure the suitability of the proposed models for predicting the disease status, they have extended the ROC curve approach for repeated measures.

In this paper, we demonstrate a procedure for computation of area under the curve for a fixed time based on cross-sectional data and then it is extended to a Markov chain model to take account of the conditional transition probabilities with covariate dependence. The computation procedure illustrated in this paper displays the estimation of sensitivity and specificity for a first order Markov model which can be extended to a higher order model. The proposed models along with the estimation of AUC and the corresponding test procedures are illustrated employing the Health and Retirement Study (HRS) data on depression among elderly people in the USA.

2. CONSTRUCTION OF ROC CURVE: A LOGISTIC REGRESSION APPROACH

This section illustrates the construction of ROC curve employing sensitivity and specificity on the basis of logistic regression models. Let us consider $Y_i = y_i$ as the value of the actual disease status ($Y_i = 1$ for diseased and $Y_i = 0$ for non-diseased) of the i-th individual $(i = 1,2,\ldots,n)$. The number of non-diseased and diseased individuals are n_0 and n_1 respectively where $n = \sum_{k=0}^{1} n_k$. If we denote $Y_i' = y_i'$ $(i = 1,2,\ldots,n)$ for the predicted value for the same individual ($Y' = 1$ for diseased and $Y' = 0$ for non-diseased), then the following table displays probabilities for the association between actual and predicted values at a fixed time (subsequently for convenience this will be called a zero-order Markov model). It may be noted here that Y can be observed value of the status and the predicted value can be based on diagnostic test. The table below shows the correspondence between the actual and predicted values:

Actual Value	**Predicted Value**	
	$Y' = 1$	$Y' = 0$
$Y = 1$	$P(Y = 1, Y' = 1)$	$P(Y = 1, Y' = 0)$
$Y = 0$	$P(Y = 0, Y' = 1)$	$P(Y = 0, Y' = 0)$

We can define the sensitivity as

$$P(Y_i' = 1 \mid Y_i = 1) = \frac{P(Y_i = 1, Y_i' = 1)}{P(Y_i = 1)}$$

and the specificity as

$$P(Y_i' = 0 \mid Y_i = 0) = \frac{P(Y_i = 0, Y_i' = 0)}{P(Y_i = 0)}.$$

Then, in the logistic regression form these are conditional models which can be reformulated by incorporating a covariate, X, as shown below,

$$P(Y_i' = 1 \mid X_i = x_i, Y_i = 1) = \frac{e^{\beta_{10} + \beta_{11} x_i}}{1 + e^{\beta_{10} + \beta_{11} x_i}}$$

$$P(Y_i' = 0 \mid X_i = x_i, Y_i = 0) = \frac{e^{\beta_{00} + \beta_{01} x_i}}{1 + e^{\beta_{00} + \beta_{01} x_i}}.$$

In the above models, the prediction of disease status is expressed as functions of a risk factor or covariate. These models represent hypothetically that the prediction variable Y' can assume values 0 and 1 for predicting the disease status for given actual disease status 0 and 1 as well as for given values of the risk factor $X = x$. There are two models for actual disease status 1 or 0 indicating two different sets of parameters for these models. In

these models, we replaced the actual disease status by the risk factor or covariate of interest. In other words, the prediction models are functions of risk factors only.

The logit models can be generalized by incorporating a vector of covariates:

$$X = \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_p \end{pmatrix},\ \beta_0 = \begin{pmatrix} \beta_{00} \\ \beta_{01} \\ \vdots \\ \beta_{0p} \end{pmatrix},\ \beta_1 = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \vdots \\ \beta_{1p} \end{pmatrix}.$$

The logistic regression models are

$$P(Y_i' = 1 \mid X_i = x_i, Y_i = 1) = \frac{e^{x_i'\beta_1}}{1 + e^{x_i'\beta_1}}, \tag{2.1}$$

$$P(Y_i' = 0 \mid X_i = x_i, Y_i = 0) = \frac{e^{x_i'\beta_0}}{1 + e^{x_i'\beta_0}}, \tag{2.2}$$

where x_i is the vector of covariate values for i-th individual. We can define for $k = 0,1$ where k is the value for actual disease status:

$$P(Y_i' = y_i' \mid X_{ik} = x_{ik}, Y_i = k) = \frac{e^{x_{ik}'\beta_k \times \delta_i}}{1 + e^{x_{ik}'\beta_k}},\ \text{where } \delta_i = 1,\ \text{if } Y_i' = Y_i = k,\ \text{otherwise, } \delta_i = 0.$$

In the above model, x_{ik} denotes the vector of covariate values for the i-th person with actual disease status k and β_k denotes the corresponding vector of parameters. Hence, the likelihood function is:

$$L = \prod_{k=0}^{1} \prod_{i=1}^{n_k} \left[\frac{e^{x_{ik}'\beta_k \times \delta_i}}{1 + e^{x_{ik}'\beta_k}} \right]^{y_{ik}} \left[1 - \frac{e^{x_{ik}'\beta_k \times \delta_i}}{1 + e^{x_{ik}'\beta_k}} \right]^{1 - y_{ik}} = \prod_{k=0}^{1} L_k.$$

The estimates of the parameters are obtained from the following equations:

$$\frac{\partial \ln L_k}{\partial \beta_{km}} = 0, k = 0,1; m = 0, \ldots p.$$

Hence we may redefine the sensitivity as shown in (2.1) and specificity as (2.2).

3. CONSTRUCTION OF ROC: A MARKOV MODEL APPROACH

The ROC curve can be drawn by plotting the sensitivity in the y-axis against (1-specificity) in the x-axis. The ROC curves have gained importance in recent days in epidemiology for assessing the accuracy of diagnostic tests. In order to assess a prediction model, the ROC curve can assess how good the model is in discriminating

between diseased subjects (True Positives) from non-diseased subjects (True Negatives). In other words, the ROC curve reveals the discriminatory power of the model.

In Section 2, we have described a model for a fixed time point or a zero-order Markov model. In a zero-order Markov model, we compare the actual disease status with the predicted status based on the fitted model at a fixed time. If we consider repeated observations then we need to take account of the first or higher order Markov models for modeling the transition in disease status in subsequent times. In this section, we propose a model such that the outcomes of disease status in subsequent follow-ups are considered and the outcome in the subsequent follow-up compared with that of the prior outcome. In other words, the predictive value indicates whether a change in the status occurs. Any change would indicate a deviation from the previous state due to a transition. In other words, zero-order model provides the goodness of fit of the suitability of the model in diagnosing the disease status and first or higher order models show whether there is any change in the course of disease status over time and then goodness of fit is tested at the endpoint of the order. In other words, the transition probabilities are obtained from status of disease from previous time point to the current time point and the current time point is considered as the end point here for a first order model.

Muenz and Rubinstein (1985), Bonney (1987), Azzalini (1994), Islam and Chowdhury (2006, 2007, 2010), Islam et al. (2009) proposed the regressive logistic models under the Markov assumptions. The joint mass function can be expressed as

$$P(y_{i1}, y_{i2},..., y_{in_i}; x_i) = P(y_{i1}; x_i)P(y_{i2}\,|\,y_{i1}; x_i)P(y_{i3}\,|\,y_{i1}, y_{i2}; x_i)...P(y_{in_i}\,|\,y_{i1},..., y_{in_{i-1}}; x_i)$$

where $x_i = (x_{i1},..., x_{ip})'$, i=1,2,…n, is the vector of covariate values for subject i, n_i is the number of follow-ups for subject i, and y_{ij}, $j = 1,2,..., n_i$, is the value of the outcome variable for the i-th subject at the j-th follow-up. In a first order Markov model for repeated observations, this can be expressed as follows:

$$P(y_{i1}, y_{i2},..., y_{in_i}; x_i) = P(y_{i1}; x_i)P(y_{i2}\,|\,y_{i1}; x_i)P(y_{i3}\,|\,y_{i2}; x_i)...P(y_{in_i}\,|\,y_{in_{i-1}}; x_i).$$

The first order Markov models can be expressed as

$$P(y_{ij}\,|\,y_{i,j-1}; \mathrm{x_i}) = \frac{e^{\theta_m y_{ij}}}{1+e^{\theta_m}} = \pi(\mathrm{x_i}, y_{ij-1}),\ \theta_m = \beta_{m0} + \beta_{m1}x_{i1} + ... + \beta_{mp}x_{ip},\ \mathrm{m}=\mathrm{y}_{ij-1}.$$

The logit is defined as

$$\theta_m = \ln\frac{P(y_{ij}=1\,|\,y_{ij-1}; \mathrm{x_i})}{P(y_{ij}=0\,|\,y_{ij-1}; \mathrm{x_i})},\ \mathrm{m}=\mathrm{y}_{ij-1}$$

where $x_i = (x_{i1},..., x_{ip})'$, order of the Markov chain is 1 and number of covariates is p. The likelihood function can be defined as

$$L = \prod_{i=1}^{n}\prod_{j=1}^{n_i} P(y_{ij}\,|\,y_{ij-1}; x_i) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{e^{\theta_m y_{ij}}}{1+e^{\theta_m}}.$$

Estimates of the parameters can be obtained from the equations of first derivatives of log likelihood function with respect to parameters contained in θ_m as shown below:

$$\frac{\partial \ln L}{\partial \beta_m} = 0.$$

The score vector for β_m where $m = y_{ij-1}$ (0 or 1) are

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} \left[y_{ij} - \pi\left(x_{il}, y_{i1}, ..., y_{ij-1}\right)\right] = 0 \text{ where } l = 0,$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} y_{il}\left[y_{ij} - \pi\left(x_{il}, y_{i1,}, ..., y_{ij-1}\right)\right] = 0, \text{ where } l = 1, 2, ..., p.$$

For a first order Markov model, the sensitivity is

$$\hat{P}(y_{ij} = 1\,|\,y_{i,j-1} = 1; x_i) = \frac{e^{\hat{\theta}_1 y_{i,j-1}}}{1+e^{\hat{\theta}_1}},$$

$$\hat{\theta}_1 = \hat{\beta}_{10} + \hat{\beta}_{11}x_{i1} + ... + \hat{\beta}_{1p}x_{ip} \qquad (3.1)$$

and 1- specificity is:

$$1 - \hat{P}(y_{ij} = 0\,|\,y_{i,j-1} = 0; x_i) = 1 - \frac{e^{\hat{\theta}_0 y_{i,j-1}}}{1+e^{\hat{\theta}_0}},$$

$$\hat{\theta}_0 = \hat{\beta}_{00} + \hat{\beta}_{01}x_{i1} + ... + \hat{\beta}_{0p}x_{ip}. \qquad (3.2)$$

The sensitivity and (1-specificity) in 3.1 and 3.2 are first order Markov model measures comparable to the measures shown in 2.1 and 2.2 respectively for zero order model. In other words, we can construct the area under curve as described in section below for both the zero-order and first order models for examining the goodness of fit using the measures of sensitivity and specificity.

4. AREA UNDER THE CURVE AND TEST PROCEDURES

For both the models proposed in sections 2 and 3, following Obuchowski (2006) we can employ the following function and estimate the probability that a randomly selected positive case will receive a higher score than a randomly selected negative case as:

$$\delta(Y_{it}, Y_{js}) = \begin{cases} 1 & \text{if } Y_{it} > Y_{js} \\ 0.5 & \text{if } Y_{it} = Y_{js} \\ 0 & \text{if } Y_{it} < Y_{js} \end{cases}$$

where Y_{it} = result of the diagnostic test for the ith patient with disease,

Y_{js} = result of the diagnostic test for the jth patient without disease of interest.

We can employ nonparametric method to estimate the ROC area on the basis of all possible pairs of diseased and non-diseased persons as shown below:

$$\hat{\phi} = \frac{1}{n_t n_s} \sum_{i=1}^{n_t} \sum_{j=1}^{n_s} \delta\left(Y_{it}, Y_{js}\right)$$

where Y_{it} = result of the diagnostic test for the ith patient with disease,

Y_{js} = result of the diagnostic test for the jth patient without disease of interest.

n_t = number of persons with disease in the sample,

n_s = number of persons without disease in the sample.

Several studies were conducted during the past 20 years (DeLong et al., 1988; Pepe, 2000; Rodenberg and Zhou, 2000; Zhang et al., 2002; Pencina and Agostino, 2004; Qin and Zhou, 2006; Bandos et al., 2009). We can show following DeLong et al. (1988) and Obuchowski (2006) that the estimated variance of the summary measure $\hat{\phi}$ can be obtained as

$$\hat{V}\left(\hat{\phi}\right) = \frac{1}{n_t - 1} S_t + \frac{1}{n_s - 1} S_s$$

where $S_t = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} \left[V_t(Y_{it}) - \hat{\phi} \right]^2$, $S_s = \frac{1}{n_s - 1} \sum_{i=1}^{n_t} \left[V_s(Y_{js}) - \hat{\phi} \right]^2$, $V_t(Y_{it}) = \frac{1}{n_s} \sum_{j=1}^{n_s} \delta\left(Y_{it}, Y_{js}\right)$

and $V_s(Y_{js}) = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta\left(Y_{it}, Y_{js}\right)$.

We can extend this for different orders of the transitions as indicated below for a transition model (Islam and Chowdhury, 2010):

$$\hat{\phi}_r = \frac{1}{n_{rt} n_{rs}} \sum_{i=1}^{n_{rt}} \sum_{j=1}^{n_{rs}} \delta\left(Y_{rit}, Y_{rjs}\right)$$

where the subscript r is used to indicate the r-th order of the transition. In this case, r=1 shows the areas under ROC curves for Markov models of first order. Using ϕ_r, we can obtain areas under the ROC curve for the conditional models. These areas represent the accuracy of the model predictions as compared to the observed status after each transition as well as after the end of the study.

For testing the equality of areas under the two ROC curves

$$H_0 : \phi_{1r} = \phi_{2r} \quad \text{against } H_1 : \phi_{1r} \neq \phi_{2r}$$

we can use the following test as a generalization of the test for any order for conditional models:

$$z = \frac{\hat{\phi}_{1r} - \hat{\phi}_{2r}}{\sqrt{Var\hat{(}\hat{\phi}_{1r} - \hat{\phi}_{2r})}} = \frac{\hat{\phi}_{1r} - \hat{\phi}_{2r}}{\sqrt{Var\hat{(}\hat{\phi}_{1r}) + Var\hat{(}\hat{\phi}_{2r})}}, \text{ where } r = 0,1,\ldots$$

It is noteworthy that for testing equality of areas under transitional models, we have considered the subjects observed in subsequent follow-ups for obtaining all possible pairs from both the observations.

5. APPLICATIONS

For this study, an application is shown in this section from the Health and Retirement Study (HRS) data (1992-2004). The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG09740) conducted by the University of Michigan (Wave [1-7] Year [1992-2004]). This study was conducted nationwide for individuals over age 50 and their spouses. We used the panel data on depression from the two rounds of the study conducted on individuals over age 50 years in 1992 (Wave I), 1994 (Wave II) data documented by RAND. The depression index is based on the score on the basis of the scale proposed by the Center for Epidemiologic Studies Depression (CESD). As indicated in the documentation of the RAND, the CESD score is computed on the basis of eight indicators attributing depression problem that were based on six negative (all or most of the time: depressed, everything is an effort, sleep is restless, felt alone, felt sad, and could not get going) and two positive indicators (felt happy, enjoyed life). These indicators are yes/no responses of the respondent's feelings much of the time over the week prior to the interview. The CESD score is the sum of six negative indicators minus two positive indicators. Hence, severity of the emotional health can be measured from the CESD score. From the panels of data, we used 9761 respondents for analyzing depression among the elderly in the USA during 1992-1994.

The dependent and explanatory variables were: no depression = 0 (CESD score ≤ 0), depression= 1 (CESD score>0), age (in years), gender (male=1, female=0), marital status (married/partnered = 1, single=0), Body Mass Index (BMI), schooling (number of years), Caucasian (white=1, else 0); Black (black=1, else 0); others= reference category, drinking habit (yes=1, no=0), conditions (number of conditions).

Table 1 displays the descriptive mean and standard deviation of some of the selected variables for Waves I and II. We observed that the number of conditions for subjects with or without depression varied substantially. The average numbers of conditions for Waves I and II are 1.43 and 1.59 as compared to that of 0.88 and 0.96 for with and without depression respectively. The distribution of the background characteristics indicate that the depression increased in Wave II for all categories (Table 2). The transition count displayed in Table 3 shows that 65 percent remains in the depression-free state in both the Waves compared to 71.5 percent in depressed state. As compared to 35 percent

making transition to depressed state from no-depression, 28.5 percent moved from depression to no-depression state.

The logistic regressions are estimated for first two waves on the basis of the formulations given in Section 2 for zero order (Table 4). These estimates are employed to obtain the corresponding sensitivity and specificity for the approach described in Section 2 for Waves I and II. Similarly, the estimates based on the Markov model (as described in Section 3) are displayed in Table 5. The summary measures for area under the curves and the corresponding confidence intervals are presented in Table 6. The areas under curve for Waves I and II are 0.682 and 0.710 respectively using the estimates obtained separately for each wave (Figure 1 and Figure 2). These estimates are quite low in matching the events. On the other hand, the Markov model shows a remarkable increase in the Area Under Curve (Figure 3) estimation (0.968). The transition model is based on the first order Markov chain in this example.

6. CONCLUSION

For assessing the performance of prediction of binary outcome of a disease status, we can employ various procedures. In the literature, there is a large number of research papers for estimating area under curve based on traditional cross-sectional data. On the other hand, the prediction models are mostly based on the longitudinal data but procedures for repeated measures are scanty in the available literature. This paper shows the prediction of disease status employing repeated measures data and then shows the procedure for estimating area under curve for covariate dependent Markov models. We have considered a first order Markov model with covariate dependence for predicting the status of a disease, and then illustrated the procedures for estimating the area under curve. The transition model indicates a remarkable improvement in the area under curve using the sensitivity and specificity estimates. The application to the HRS data on depression illustrates the usefulness of the proposed transition models for repeated measures data.

ACKNOWLEDGEMENT

The authors acknowledge gratefully to the HRS (Health and Retirement Study) which is sponsored by the National Institute of Aging (Grant Number NIA U01AG09740) and conducted by the University of Michigan.

REFERENCES

1. Agresti, A. (1997). A model for repeated measurements of multivariate binary response. *J. Amer. Statist. Assoc.*, 92, 315-321.
2. Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, 8, 767-775.
3. Bandos, A.I., Rockette, H.E., Song, T. and Gur, D. (2009). Area under the free-response ROC curve (FROC) and a related summary index. *Biometrics* 65, 247-256.
4. Biagini, E., Elhendy, A., Schinkel, A.F.L., Rizzello, V., Bax, J.J., Sozzi, F.B., Kertai, M.D., van Domburg, R.T., Krenning, B.J., Branzi, A., Rapezzi, C., Simoons, M.L. and Poldermans, D. (2005). Long-Term Prediction of Mortality in Elderly Persons by Dobutamine Stress Echocardiography. *Journal of Gerontology*, 60A, 1333-1338.

5. Bonney, G.E. (1986). Regressive logistic models for familial disease and other binary trials. *Biometrics*, 42, 611-625.
6. Bonney, G.E. (1987). Logistic regression for dependent binary observations. *Biometrics*, 43, 951-973.
7. Boyko, E.J., Ahroni, J.H., Cohen, V., Nelson, K.M. and Heagerty, P.J. (2006). Prediction of diabetic foot ulcer occurrence using commonly available clinical information: the Seattle Diabetic Foot Study. *Diabetes Care*, 29, 1202-7.
8. Croft, R.P., Nicholls, P.G., Steyerberg, E.W., Richardus, J.H., Withington, S.G. and Smith, W.C. (2003). A clinical prediction rule for nerve function impairment in leprosy patients-revisited after 5 years of follow-up. *Leprosy Review*, 74, 35-41.
9. de Vries, S.O., Fidler, V., Kuipers, W.D. and Hunink, M.G.M. (1998). Fitting multistate transition models with autoregressive logistic regression. *Medical Decision Making*, 18, 52-60.
10. DeLong, E., DeLong, D. and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, 44, 837-845.
11. Gundersen, K., Kvaløy, J.T., Kramer-Johansen, J. and Eftestøl, T. (2008). Identifying approaches to improve the accuracy of shock outcome prediction for out-of-hospital cardiac arrest. *Resuscitation*, 76, 279-84.
12. Gundersen, K., Kvaløy, J.T., Kramer-Johansen, J., Steen, P.A. and Eftestøl, T. (2009). Development of the probability of return of spontaneous circulation in intervals without chest compressions during out-of-hospital cardiac arrest: an observational study. *BMC Medicine*, 6, 1-9.
13. Health and Retirement Study, (Wave [1-7]/Year [1992-2004]) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (Grant Number NIA U01AG09740). Ann Arbor, MI.
14. Islam, M.A. and Chowdhury, R.I. (2006). A higher-order Markov model for analyzing covariate dependence. *Applied Mathematical Modelling*, 30, 477-488.
15. Islam, M.A. and Chowdhury, R.I. (2007). First and higher order transition models with covariate dependence. *In Progress in Applied Mathematical Modeling*, F. Yang (ed), 153-198.
16. Islam, M.A., Chowdhury, R.I. and Huda, S. (2009). *Markov Models with Covariate Dependence for Repeated Measures*. Nova Science, New York.
17. Islam, M.A. and Chowdhury, R.I. (2010). Prediction of Disease Status: A Regressive Model Approach for Repeated Measures. *Statistical Methodology*, 7, 520-540.
18. Marwick, T.H., Case, C., Vasey, C., Allen, S., Short, L. and Thomas, J.D. (2001). Prediction of mortality by exercise echocardiography: a strategy for combination with the duke treadmill score. *Circulation*, 103, 2566-71.
19. Muenz, L.R. and Rubinstein, L.V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 41, 91-101.
20. Obuchowski, N.A. (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine*, 25, 481-493.
21. Olfson, M., Mechanic, D., Hansell, S., Boyer, C.A. and Walkup, J. (1999). Prediction of homelessness within three months of discharge among inpatients with schizophrenia. *Psychiatric Services*, 50, 667-673.

22. Pencina, M.J. and D'Agostino, R.B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23, 2109-2123.
23. Pencina, M.J., D'Agostino Sr., R.B., D'Agostino Jr., R.B. and Vasan, R.S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27, 157-172.
24. Pepe, M.S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56, 352-359.
25. Pepe, M.S., Janes, H., Longton, G., Leisenring, W. and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 159, 882-890.
26. Qin, G. and Zhou, X.H. (2006). Empirical likelihood inference for the area under the ROC curve. *Biometrics*, 62, 613-622.
27. Rodenberg, C. and Zhou, X.H. (2000). ROC curve estimation when covariates affect the verification process. *Biometrics*, 56, 1256-1262.
28. Schoonhoven, L., Grobbee, D.E., Donders, A.R., Algra, A., Grypdonck, M.H., Bousema, M.T., Schrijvers, A.J. and Buskens, E. (2006). Prediction of pressure ulcer development in hospitalized patients: a tool for risk assessment. *Quality & Safety in Health Care*, 15, 65-70.
29. Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. and Kattan, M.W. (2010). Assessing the performance of prediction models. A framework for traditional and novel measures. *Epidemiology*, 21, 128-138.
30. Van Voorhees, B.W., Paunesku, D., Gollan, J., Kuwabara, S., Reinecke, M. and Basu, A. (2008). Predicting future risk of depressive episode in adolescents: the Chicago Adolescent Depression Risk Assessment (CADRA). *Annals of Family Medicine*, 6, 503-11.
31. Zethelius, B., Berglund, L., Sundström, J., Ingelsson, E., Basu, S., Larsson, A., Venge, P. and Arnlöv, J. (2008). Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *New England Journal of Medicine*, 358, 2107-16.
32. Zhang, D.D., Zhou, X.H., Freeman Jr., D.H. and Freeman, J.L. (2002). A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine*, 21, 701-715.

Table 1:
Descriptive Statistics of Background Characteristics by CESD and Waves

Variables	Depression Score (CESD)				
	0		1+		Total
	Mean	SD	Mean	SD	Count
WAVE I					
Age (in years)	55.57	3.22	55.49	3.19	8116
Body Mass Index (BMI)**	26.90	4.84	27.76	5.64	8116
Years of Education**	12.66	2.90	11.32	3.36	8116
Number of Conditions**	0.88	0.96	1.43	1.30	8116
WAVE II					
Age (in years)	57.42	3.20	57.39	3.21	8116
Body Mass Index (BMI)**	26.91	4.67	27.73	5.50	8116
Years of Education**	12.95	2.76	11.32	3.32	8116
Number of Conditions**	0.96	1.01	1.59	1.35	8116

Note: ** Significant at 1% level

Table 2:
Distribution of Background Characteristics by CESD and Waves

Variables	Labels	Depression Score (CESD)				
		0		1+		Total
		Count	Row%	Count	Row%	Count
WAVE I						
Gender**	Female	2745	60.5	1794	39.5	4539
	Male	2324	65.0	1253	35.0	3577
Marital Status**	Single	1009	49.5	1031	50.5	2040
	Married	4060	66.8	2016	33.2	6076
White**	No	812	49.2	839	50.8	1651
	Yes	4257	65.8	2208	34.2	6465
Black**	No	4391	65.1	2351	34.9	6742
	Yes	678	49.3	696	50.7	1374
Whether Drink**	No	1827	57.3	1364	42.7	3191
	Yes	3242	65.8	1683	34.2	4925
WAVE II						
Gender	Female	2161	47.6	2378	52.4	4539
	Male	2003	56.0	1574	44.0	3577
Marital Status**	Single	806	37.7	1333	62.3	2139
	Married	3358	56.2	2619	43.8	5977
White**	No	622	37.7	1029	62.3	1651
	Yes	3542	54.8	2923	45.2	6465
Black**	No	3640	54.0	3102	46.0	6742
	Yes	524	38.1	850	61.9	1374
Whether Drink*	No	1607	43.9	2055	56.1	3662
	Yes	2557	57.4	1897	42.6	4454

Note: ** Significant at 1% level; * Significant at 5% level.

Table 3:
Transition Count and Transition Probability

	Transition Count		Transition Probability		Total
States (Y_{ij})	0	1	0	1	
0	3296	1773	0.650	0.350	5069
1	868	2179	0.285	0.715	3047

Table 4:
Logistic Regression for CESD as Outcome Variable from First Two Waves

Variables	Coeff.	Std. err.	t-value	p-value	95 % CI	
					LL	UL
WAVE 1						
Constant	3.138	.483	42.290	.000		
Age (in years)	-.034	.008	19.692	.000	.952	.981
Gender	.010	.050	.041	.839	.915	1.115
Marital Status	-.555	.057	95.871	.000	.514	.642
Body Mass Index (BMI)	.002	.005	.116	.733	.992	1.011
Years of Education	-.114	.008	195.901	.000	.878	.906
White	-.488	.133	13.413	.000	.473	.797
Black	-.152	.142	1.142	.285	.650	1.135
Whether Drink	-.055	.051	1.150	.284	.855	1.047
Number of Conditions	.376	.023	270.085	.000	1.393	1.523
Model Chi-square (p-value)	885.07 (0.0001)					
WAVE II						
Constant	4.342	.496	76.663	.000		
Age (in years)	-.034	.008	20.268	.000	.952	.981
Gender	-.126	.049	6.458	.011	.801	.972
Marital Status	-.569	.057	99.884	.000	.506	.633
Body Mass Index (BMI)	.000	.005	.003	.959	.991	1.010
Years of Education	-.153	.009	317.420	.000	.844	.873
White	-.567	.141	16.134	.000	.430	.748
Black	-.300	.151	3.921	.048	.551	.997
Whether Drink	-.148	.050	8.725	.003	.782	.951
Number of Conditions	.385	.022	300.194	.000	1.407	1.534
Model Chi-square (p-value)	1192.48 (0.0001)					

Table 5:
Markov Models for CESD as Outcome Variable from First Two Waves

Variables	Coeff.	Std. err.	t-value	p-value	95 % CI	
					LL	UL
State 0 → State 1						
Constant	3.063	0.624	4.906	0.000	1.839	4.286
Age (in years)	-0.025	0.010	-2.636	0.008	-0.044	-0.007
Gender	-0.228	0.063	-3.599	0.000	-0.352	-0.104
Marital Status	-0.348	0.077	-4.512	0.000	-0.499	-0.197
Body Mass Index (BMI)	0.004	0.006	0.619	0.536	-0.009	0.017
Years of Education	-0.139	0.011	12.516	0.000	-0.160	-0.117
White	-0.514	0.186	-2.769	0.006	-0.878	-0.150
Black	-0.248	0.200	-1.239	0.215	-0.639	0.144
Whether Drink	-0.084	0.065	-1.287	0.198	-0.212	0.044
Number of Conditions	0.276	0.032	8.506	0.000	0.212	0.339
State 1 → State 1						
Constant	4.042	0.834	-4.846	0.000	-5.677	-2.407
Age (in years)	-0.023	0.013	1.716	0.086	-0.003	0.049
Gender	-0.060	0.087	0.696	0.487	-0.110	0.230
Marital Status	-0.456	0.097	4.725	0.000	0.267	0.646
Body Mass Index (BMI)	0.003	0.008	-0.339	0.734	-0.019	0.013
Years of Education	-0.132	0.015	9.061	0.000	0.104	0.161
White	-0.375	0.230	1.635	0.102	-0.075	0.826
Black	-0.239	0.245	0.975	0.330	-0.241	0.718
Whether Drink	-0.158	0.089	1.773	0.076	-0.017	0.334
Number of Conditions	0.294	0.039	-7.634	0.000	-0.369	-0.218
Score Chi-square	1549.35(p-value=0.0001, DF=20)					
LRT	1682.79(p-value=0.0001, DF=20)					

Table 6:
Summary Measures of Area under the Curve

Source	**Area**	**Std. Error**	**Asymp. Sig.**	**Asymptotic 95% C.I**	
				Lower Bound	**Upper Bound**
Based on logistic regression for WAVE I	.682	.006	.000	.670	.694
Based on logistic regression for WAVE II	.710	.006	.000	.699	.721
Based on Markov model	.968	.002	.000	.965	.971

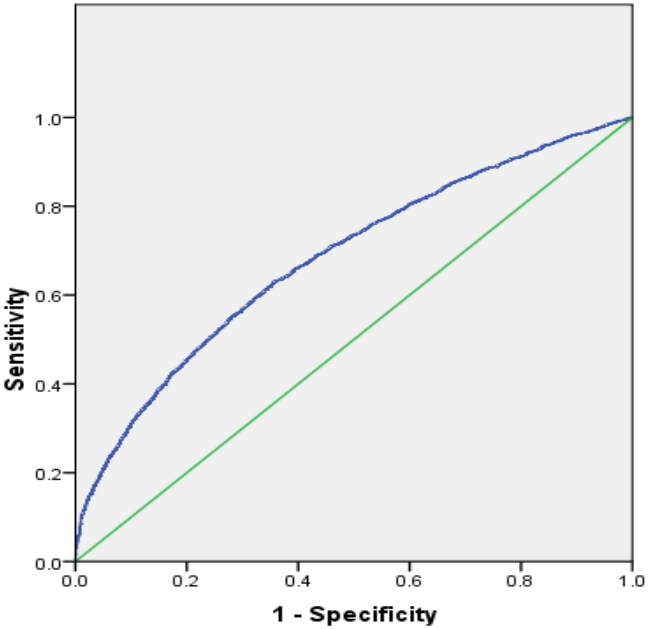

Fig. 1: ROC Curve based on predicted probability of logistic regression for Wave I

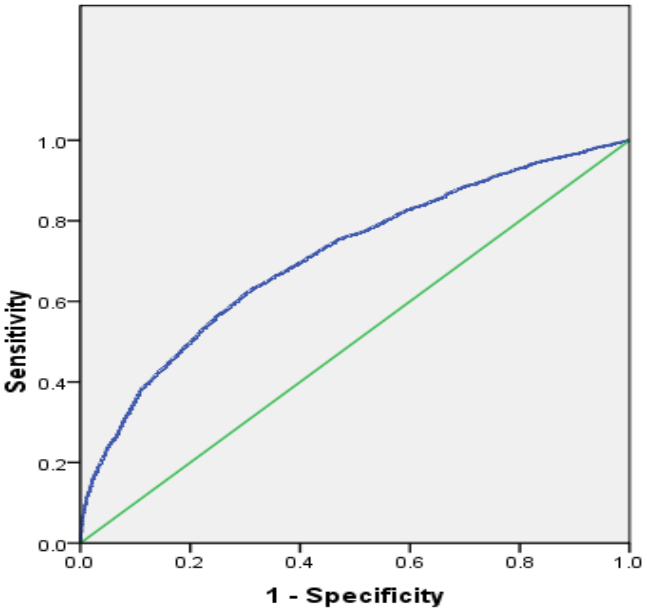

Fig. 2: ROC Curve based on predicted probability of logistic regression for Wave II

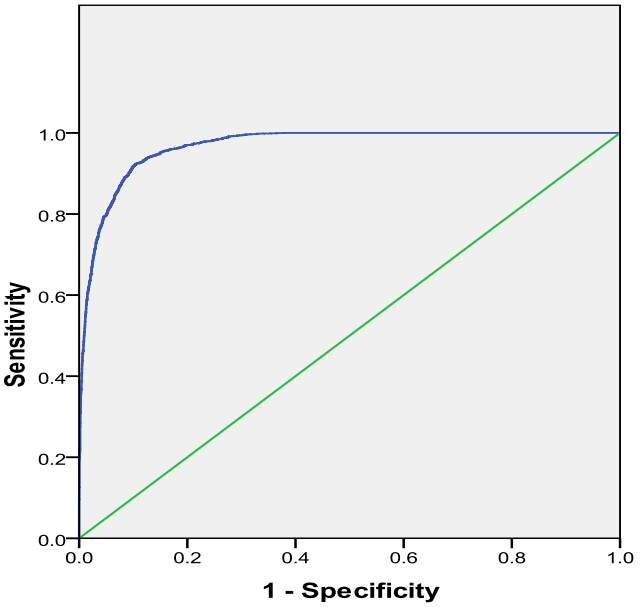

Fig. 3: ROC Curve based on Markov Model