Journal of Statistical Research *ISSN 0256 - 422 X*
2019, Vol. 53, No. 1, pp. 63-77

REVIEW AND EVALUATION OF THE CONCORDANCE MEASURES FOR ASSESSING DISCRIMINATION IN THE LOGISTIC REGRESSION MODELS

Biplab Biswas*
Department of Statistics
Bangabandhu Sheikh Mujibur Rahman Science and Technology University,Gopalganj-8100
Email: bbiswas@isrt.ac.bd

Maidul Husain
Department of Statistics
Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100
Email: mhusain@isrt.ac.bd

M. Shafiqur Rahman
Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000
Email: shafiq@isrt.ac.bd

SUMMARY

Concordance statistic (C-statistic), which is equivalent to the area under a receiver operating characteristic curve (AUC), is frequently used to quantify the discriminatory power (the ability of the model to distinguish low and high risk patient) of a risk prediction model developed in the logistic regression framework. Several methods for estimating concordance statistics including both non-parametric and parametric have been proposed in the literature. Despite the several proposals of the C-statistic, it is still unclear to the practical users which approaches should be applied in practice. This paper reviewed and evaluated some commonly used C-statistics by illustrating them using two datasets with different prognostic abilities and an extensive simulation study and compared their results to make some practical recommendations. Several simulation scenarios were considered by varying the sample size, prevalence of the binary outcome, and distribution of prognostic index (or log-odds) derived from the model, to mimic the scenarios in practice. The results revealed that both non-parametric and Kernel-smoothing based methods showed comparable results in most simulation scenarios but performed better than the parametric approach particularly for small sample situation and skewed distribution of the prognostic index. Based on the findings of the study, some practical recommendations are discussed.

Keywords and phrases: Risk prediction model, binary data, C-statistic, area under ROC curve.

* Corresponding author

1 Introduction

Logistic regression models are frequently used in various clinical settings such as cardiology and oncology to predict the probability of the occurrence of a binary outcome given a specific vector of predictors (Royston and Altman, 2010; Austin and Steyerberg, 2012). For example, in cardiology the models may be used to predict the risk of developing a cardiovascular disease or death due to the disease using his/her clinical and demographic characteristics. Predictions based on these models have an important role in classifying the patients with low-and high-risk and hence in guiding their future courses of treatment(Steyerberg et al., 2010). Given their important role in clinical research, it is very essential to evaluate the predictive performance of the models before using these for clinical predictions (Wyatt and Altman, 1995; Steyerberg, 2008). To characterize the predictive performance, two key aspects of a model are usually evaluated (Altman and Royston, 2000; Moons et al., 2009). These include (i) 'calibration'-the agreement between the observed and predicted outcome of interest (ii) 'discrimination'-the ability of the model to distinguish between high and low risk patients. Good discrimination of a model does not necessarily imply good calibration or vice-verse. As suggested by Harrell et al. (1996) and Pencina and D'Agostino (2004) good discrimination should be a primary focus. This is because re-calibration is always possible, which is not true for discrimination. This paper focuses on discrimination in the risk models developed using logistic regression framework.

Several methods have been proposed in the literature to quantify the discrimination ability of the logistic regression models (Steyerberg, 2008; Metz, 1978). These includes 'discrimination slope', 'integrated discrimination index' (Pencina et al., 2017), Cohen and Hedges distance (known as 'effect size') and 'overlap' between two distributions associated with diseased and healthy population (Royston and Altman, 2010). Of them concordance statistic (C-statistic), which is equivalent to the area under a receiver operating characteristic curve (AUC): the graph of sensitivity (true-positive rate) versus one minus specificity (true-negative rate) (Harrell et al., 1996), is widely used because of its straightforward clinical interpretation (Antolini et al., 2004; Austin and Steyerberg, 2012). It quantifies the probability that, for a randomly selected pair of subjects (event vs non-event), the subject who developed the event has higher predicted probability (or risk score) than those who didn't develop the event. It ranges between 0.5 and 1: a value of 0.5 suggests no discriminatory ability of the model between low and high risk patients and a value of 1 suggests perfect discrimination. There are two main approaches for estimating concordance statistics (Faraggi and Reiser, 2002): non-parametric and parametric. The former is based on Mann-Whitney U statistic (Mann and Whitney, 1947) while the latter is based on comparison of distribution of prognostic index or log-odds derived from the model for subjects with event and those without event. An alternative to the Mann-Whitney statistic approach is suggested by Lloyd (Lloyd, 1998), which is based on Kernel smoothing, to obtain an estimate of C-statistic equivalent to the area under a smooth ROC curve.

Despite the proposal of several approaches of the C-statistic, it is still unclear to the practical users which approaches should be applied in practice, particularly when developing a

prediction model for binary outcome using logistic regression framework. This is because the estimation and interpretation of C-statistic (or AUC) for a multivariable logistic regression model are not straightforward like a continuous diagnostic marker (or biomarker) that frequently used to determine the state of a disease in medicine. Generally a multivariate logistic models yield a continuous risk score or log-odds, which is a linear combination of several predictors (mixture of both continuous and categorical) weighted by the estimated regression coefficients, and a transformation of which gives the estimated event probability for an individual patient. This is generally known in literature as 'risk model' or 'prognostic model' (Moons et al., 2009). The estimated risk score is used as the proxy of continuous biomarker when estimating C-statistics. This paper reviewed and evaluated some popular approaches of the C-statistic for assessing discriminatory power of the logistic regression model by illustrating them using two real datasets of different prognostic abilities and an extensive simulation study to make some practical recommendations.

The paper is organized as follows. Section 2 describes use of logistic regression model in risk prediction and methods for estimating C-statistics. An illustration of the methods using two practical datasets is described in Section 3, and Section 4 describes the simulation study. Section 4 ends the paper with discussion of the major findings and providing recommendations for practical use.

2 Methodology

2.1 The Model

Let $Y_i, (i = 1, 2, \ldots, n)$, be a binary outcome (0/1) for the ith subject which follows Bernoulli distribution with the probability $\pi_i = \Pr(Y_i = 1)$. The logistic regression model can be used to model the relationship between the outcome and predictors and to predict the probability of the positive outcome and is defined as

$$\text{logit}[\Pr(Y_i = 1|\boldsymbol{x}_i)] = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i,$$

where $\boldsymbol{\beta}^T$ is a vector of regression coefficients of length $(p+1)$, and $\boldsymbol{x}_i$ is the ith row vector of the predictor matrix $\boldsymbol{x}$ which has order $n \times (p+1)$. The term $\eta_i = \boldsymbol{\beta}^T\boldsymbol{x}_i$ is called as risk score or 'prognostic index (PI)'. The parameters of the model, $\boldsymbol{\beta}$, can be estimated using maximum likelihood estimation technique. Once the estimates, $\hat{\boldsymbol{\beta}}$ are available, the prediction can be made using the following equation:

$$\pi(\hat{\boldsymbol{\beta}}|\boldsymbol{x}_i) = [1 + \exp(-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)]^{-1}.$$

2.2 C-statistics

The C-statistic quantifies the probability that, for a randomly selected pair of subjects (event vs non-event), the predicted event probability is higher for the subject who experienced the event of interest than those who did not experience the event. Now for a pair of subjects

(i, j), where i and j correspond to subject who experienced the event and those who did not, respectively, with event probabilities $\{\pi(\boldsymbol{\beta}|\boldsymbol{x}_i), \pi(\boldsymbol{\beta}|\boldsymbol{x}_j)\}$, the C-statistic can be defined as

$$C = \Pr[\pi(\boldsymbol{\beta}|\boldsymbol{x}_i)|Y_i = 1 > \pi(\boldsymbol{\beta}|\boldsymbol{x}_i)|Y_i = 0].$$

Since there exists a one-to-one transformation between π and $\boldsymbol{\beta}^T\boldsymbol{x}$, the above probability expression can be written as

$$C = \Pr[(\boldsymbol{\beta}^T\boldsymbol{x}_i|Y_i = 1) > (\boldsymbol{\beta}^T\boldsymbol{x}_j|Y_j = 0)].$$

The C-statistic for the logistic regression models can be estimated using both parametric and nonparametric approaches (Molodianovitch et al., 2006). The widely used nonparametric approach to estimate the C-statistic is based on the Mann and Whitney U statistic (Mann and Whitney, 1947) and does not require any distributional assumptions regarding the prognostic index. Let $\eta_i^{(1)} = \boldsymbol{\beta}^T\boldsymbol{x}_i|Y_i = 1$ and $\eta_j^{(0)} = \boldsymbol{\beta}^T\boldsymbol{x}_j|Y_j = 0$ be the prognostic index or log-odds derived by the model for subject i who had experienced the event and for subject j who did not, respectively. Further, let n_1 and n_0 be the number of events and non-events, respectively. Considering all pairs, the concordance statistic can be estimated by analogy to the U statistic formulation (Hanley and McNeil, 1982) as

$$C_U = \frac{1}{n_1 n_0}\sum_{i=1}^{n}\sum_{j=1}^{n} I\big(\eta_i^{(1)} > \eta_j^{(0)}\big) + \frac{1}{2}I\big(\eta_i^{(1)} = \eta_j^{(0)}\big),$$

where $n_1 = \sum_{i=1}^{n} I(Y_i = 1)$ and $n_0 = \sum_{j=1}^{n} I(Y_j = 0)$ and $I(\cdot)$ is the indicator function.

An asymptotic confidence interval for true C_U can be derived assuming that $\big(\hat{C_U} - C_U\big)/\sqrt{\widehat{\text{var}}[\hat{C_U}]}$ is asymptotically $N(0, 1)$. The $100(1-\alpha)\%$ confidence interval for C_U can be obtained as

$$\Big(\hat{C}_U - Z_{\alpha/2}\sqrt{\widehat{\text{var}}[\hat{C}_U]}, \hat{C}_U + Z_{\alpha/2}\sqrt{\widehat{\text{var}}[\hat{C}_U]}\Big),$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of standard normal distribution. Several approaches have been proposed to estimate the variance of the area under ROC curve (Hanley and McNeil, 1982; DeLong et al., 1988). However, it can be showed that all these approaches are approximately equivalent when sample size is large. In this paper, the method of DeLong et al. (1988) is adapted to derive the variance expression of C_U for logistic regression model as:

$$\widehat{\text{var}}[\hat{C}_U] = \frac{1}{n_1}S_{10}^U + \frac{1}{n_0}S_{01}^U,$$

where

$$S_{10}^U = (n_1 - 1)^{-1}\sum_{i=1}^{n_1}\Big(V_{10}^U - \hat{C}_U\Big)^2 \qquad S_{01}^U = (n_0 - 1)^{-1}\sum_{j=1}^{n_0}\Big(V_{01}^U - \hat{C}_U\Big)^2$$

$$V_{10}^U = n_0^{-1}\sum_{j=1}^{n_0} I\Big(\eta_i^{(1)}, \eta_j^{(0)}\Big) \text{ for all } \eta_i^{(1)} \qquad V_{01}^U = n_1^{-1}\sum_{i=1}^{n_1} I\Big(\eta_j^{(0)}, \eta_i^{(1)}\Big) \text{ for all } \eta_j^{(0)}.$$

To obtain C-statistic from the area under a smooth ROC curve, an alternative to the above estimator suggested by (Lloyd, 1998) is based on standard normal Kernel smoothing. The resulting Kernel estimate of the C-statistic can be written as

$$C_K = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \Phi\left(\frac{\eta_i^{(1)} - \eta_j^{(0)}}{\sqrt{h_1^2 + h_0^2}} \right),$$

where h_0 and h_1 are the bandwidth that control the degree of smoothing in estimating CDF of $\eta_i^{(0)}$ and $\eta^{(1)}$, respectively. There are several choices of bandwidth selection (Zhou and Harezlak, 2002; Hall and Hyndman, 2003; Pulit, 2016). The general choice is $h_1 = 0.9 \min(s_1, IQR_1/1.34) n_1^{-1/5}$, where s_1 and IQR_1 are the standard deviation and inter quartile range of risk score $\eta_i^{(1)}$ and $\Phi(.)$ is the standard normal cumulative distribution function. Similarly one can define h_0 for risk score $\eta_i^{(0)}$. The asymptotic confidence interval for the true C_K can be obtained using the same approach discussed for C_U.

In addition to the non-parametric approaches, the C-statistics can be estimated parametrically as follows. Based on the central limit theorem, the prognostic score η_i is likely to follow normal distribution as the dimension of the parameter vector β increases (Choodari-Oskooei et al., 2012). The parametric approach is based on the assumption of bi-normal distribution of the prognostic score. Mathematically, assume that

$$\eta_i^{(1)} = (\boldsymbol{\beta}^T x_i | Y_i = 1) \sim N(\mu_1, \sigma_1^2) \text{ and } \eta_j^{(0)} = (\boldsymbol{\beta}^T x_j | Y_j = 1) \sim N(\mu_0, \sigma_0^2).$$

Therefore, $\eta_i^{(1)} - \eta_j^{(0)} \sim N(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2)$. The parametric C-statistic

$$C_P = Pr[\eta_i^{(1)} > \eta_j^{(0)}]$$

can be obtained after standardizing the term $\eta_i^{(1)} - \eta_j^{(0)}$ as

$$C_P = Pr\left[Z < \frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}} \right] = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}} \right),$$

where $Z \sim (0, 1)$ and $\Phi(.)$ is the standard normal CDF. The estimate of the C_P can be obtained by replacing μ_1, μ_0 and σ_1^2, σ_0^2 by their sample estimates (MLEs) $\bar{x}_1$, $\bar{x}_0$ and s_1^2, s_2^2, respectively. An approximate $100(1-\alpha)\%$ confidence interval interval for the true C_P is given by

$$\left\{ \Phi\left(\hat{\delta} - Z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\delta})} \right), \Phi\left(\hat{\delta} - Z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\delta})} \right) \right\},$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ standard normal percentile and $\delta = (\mu_1 - \mu_2)(\sigma_1^2 + \sigma_0^2)^{-1/2}$ and $\text{Var}(\hat{\delta})$ given below can be estimated using Delta method provided that the variances of $\hat{\mu_1}$, $\hat{\mu}_0$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are available from the inverse of the Fisher information matrix obtained from the maximum likelihood procedure for bi-normal distribution:

$$\widehat{\text{Var}}(\hat{\delta}) = \left[\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0} \right] \times (\hat{\sigma}_1^2 + \hat{\sigma}_0^2)^{-1} + \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{4(\hat{\sigma}_1^2 + \hat{\sigma}_0^2)^3} \times \left[\frac{2\hat{\sigma}_1^4}{n_1 - 1} + \frac{2\hat{\sigma}_0^4}{n_0 - 1} \right].$$

For more details on using Delta method for estimating variance see elsewhere (Reiser, 2000; Faraggi, 2000).

3 Illustration using Practical Data

First we describe illustration of the C-statistics using two datasets with different prognostic abilities to see if there is any difference in the estimate of the C-statistics for each dataset and between the datasets. One case study is based on data for low birth weight of new born and the other is based on data for a patients at ICU in hospital. Details of the data and analysis are discussed below.

Low Birth Weight Data

Low birth weight (LBW<2500 gm) is an adverse pregnancy outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's lifestyle during pregnancy including diet, smoking habits, and receiving prenatal care can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. The predictors of interest were age of mother in years (AGE), weight in pounds at the last menstrual period (LWT), RACE (white, black, other), smoking status during pregnancy (SMOKE: yes, no), history of premature labor (PTL: none, at least one), history of hypertension (HT: yes, no), presence of uterine irritability (UI: yes, no). For more details about the data, see elsewhere (Hosmer Jr et al., 2013). The main focus here is to illustrate the C-statistics under study for assessing discriminatory power of the risk model developed in the logistic regression framework to predict the risk of having child with LBW.

Based on the literature model on low birth weight data and exploratory analysis (results not shown), the model we developed for predicting risk of having child with LBW has the following prognostic index:

$$\begin{aligned}\hat{\eta}(\boldsymbol{x})_{lbw} = \hat{\beta}\boldsymbol{x} = 0.633 - 0.038 * \text{AGE} - 0.015 * \text{LWT} + 1.212 * \text{RACE (black)} \\ +0.805 * \text{RACE (other)} + 0.846 * \text{SMOKE (yes)} + 1.222 * \text{PTL (one or more)} \\ +1.837 * \text{HT (yes)} + 0.711 * \text{UI (yes)}.\end{aligned}$$

Prediction (the risk of having child with LBW) can be made as $\hat{\pi}(\boldsymbol{x}) = [1+\exp(-\hat{\eta}(\boldsymbol{x})_{lbw})]^{-1}$ for a subject with values for each of the predictors, $\boldsymbol{x}$. The distribution of the prognostic scores derived from the model for women having child with LBW and those having child without LBW indicates that there is quite difference between two distributions suggesting a certain amount of discrimination between the groups (Figure 1a). The corresponding C-statistics suggest that the model has strong ability to discriminate the subjects with event from those without event (Table 1).

The ICU Data

This dataset (Hosmer Jr et al., 2013) contains the information of 200 patients following admission to an adult intensive care unit (ICU), of which 40 patients were died in ICU. Vital status (alive/died) of patients after the admission in the ICU depends mostly in some clinical and demographic factors such as age of the patients, service at ICU admission, history of chronic renal failure, sex, systolic blood pressure at ICU admission etc. The predictors of interest were AGE, SEX (male, female), service at ICU admission (SER: medical, surgical), cancer part of present problem (CAN: no, yes), history of chronic renal failure (CRN: no, yes), infection probable at ICU admission (INF: no, yes)), CPR prior to ICU admission (CPR: no, yes)), systolic blood pressure at ICU admission (SYS), heart rate at ICU admission (HRA), previous admission to an ICU within 6 months (PRE: no, yes), type of admission (TYP:elective,emergency), fracture (FRA: no, yes) creatinine from initial blood gases (CRE:cre$\leq$2.0, cre>2.0) and level of consciousness at ICU admission (LOC:no, deep stupor, coma). The main objective here is to develop a risk model to predict the risk of mortality in ICU and assess its discriminatory ability using C-statistics. Based on literature model and exploratory analysis (results not shown), the model developed for predicting the risk of ICU mortality consists of the following prognostic index:

$$\begin{aligned}\hat{\eta}(\boldsymbol{x})_{icu} = \hat{\beta}\boldsymbol{x} = -3.608 + 0.034 * \text{AGE} 0.759 * \text{CRN (yes)} + 1.185 * \text{CPR (yes)} \\ -0.014 * \text{SYS} + 2.048 * \text{TYP (emergency)} + 0.268 * \text{FRA (yes)}\end{aligned}$$

The distribution of the prognostic index derived from the model for the patient experienced event and those who didn't suggest that there is a quite strong discrimination between two groups of the patients (Figure 1b). The estimates of the C-statistics suggest that the model has strong ability to discriminate the patients who experienced the event from the counter group (Table 1).

Table 1: Estimated C-statistics for the model for both LBW and ICU data

	LBW			ICU		
C-stat.	Est.	SE	90% CI	Est.	SE	90% CI
C_U	0.7462	0.0375	[0.6844, 0.8079]	0.7901	0.0429	[0.7195, 0.8607]
C_K	0.7464	0.0375	[0.6845, 0.8081]	0.7902	0.0429	[0.7196, 0.8608]
C_P	0.7505	0.1155	[0.6865, 0.8068]	0.7858	0.1365	[0.7148, 0.8454]

Comparing the results for two datasets, the model for ICU patients is reported to show stronger discriminatory power than that for LBW data. For the both models, the estimates of C_U and C_K and their standard error are observed to be quite similar, however, the estimates of C_P and its standard error is quite different from those for C_U and C_K, given the distribution of the prognostic index between the subjects with event and those without

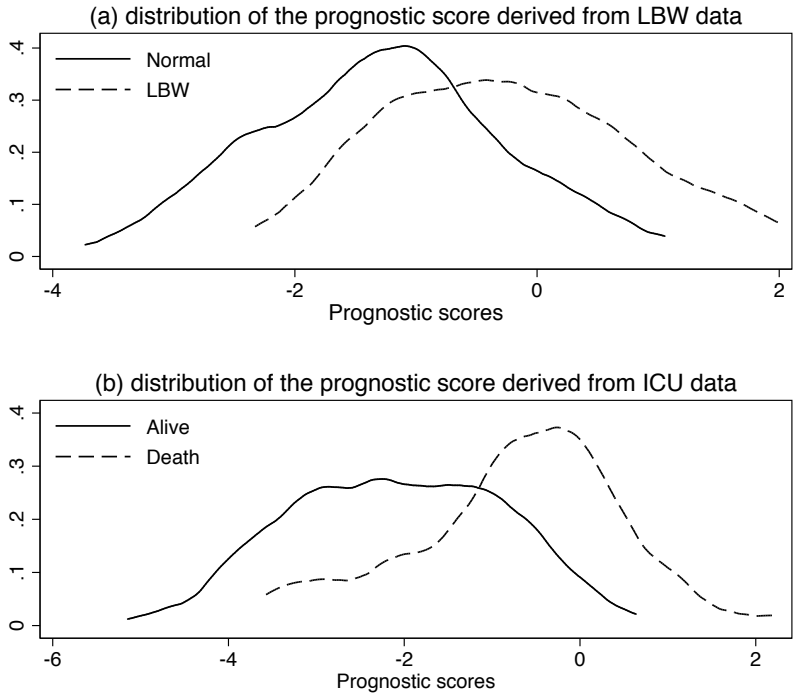

Figure 1: Visualization of the observed discrimination between event vs non-event derived from the respective data.

event were approximately normal. These findings motivated us to perform a simulation study to assess some statistical properties of a good estimator of the C-statistic and compare their results for practical recommendations.

4 Simulation Study

4.1 Simulation Design

The performance of different versions of the C-statistic was investigated using an extensive simulation study based on LBW data. More specifically, without generating all the risk factors (covariates) in the original data, we generated only binary outcome from Bernoulli distribution with probability ($\hat{\pi}$) derived from a true logistic model fitted using the LBW data. The model we developed for LBW data in section 3 was considered as true model. As distributional assumption of parametric approach of C-statistic is required, some simulation scenarios were created varying the shape of the distribution of the prognostic index as i)

bell shaped (normal), ii) right skewed, and iii) left skewed. Under each of the distributional scenarios, four scenarios were created by varying the prevalence as 45%, 30%, 20%, and 10% to mimic the scenarios with low to high prevalence of the binary outcome of interest. Under each of these scenarios, data were generated for different sample sizes such as 189, 100, 75, and 50, altogether for 16 scenarios. The sample size 189 represent the size of the original data and all other sample scenarios were created by taking a random sample of required size from 189. Under each simulation scenarios, a total of 1000 replicated datasets of the same size were generated. For each dataset, we developed a model and estimated each of the C-statistics (C_U, C_K, and C_P) under study. Finally bias, mean squared errors (MSE) and coverage for each estimator were reported. The bias was calculated as the difference between the estimate (average over 1000 replications) and the true C-statistic (derived for the true model). The MSE was calculated as mean of the squared differences between the estimated and true value over 1000 simulations and the coverage was calculated as the proportion of the 90% CIs out of 1000 containing the true C.

4.2 Simulation Results

The simulation results obtained considering the normal (bell-shaped) distribution of prognostic index associated with the true model were summarized in Table 2. The Table 2 suggest that, in general for all types of C-statistic, the amount of bias increase with the decreasing sample size, which is true for all scenarios with both high and low prevalence rate of the binary outcome. The amount of bias was greater for the parametric C_P compared to C_U and C_K, particularly for small sample situation. The MSE value, in general, increases with the decreasing sample size. When the sample size is small (50), lower MSE values are observed for C_U and C_K compared to C_P. When assessing against the varying prevalence rate of binary outcome, both the bias and MSE increased with decreasing prevalence rate, and the amount of bias and MSE is comparatively higher for low prevalence rate (Table 2). In terms of coverage of 90% CI for true C, both C_U and C_K showed better performance than C_P particularly when sample size is small. Both non-parametric and smooth estimators (C_U, C_K) showed comparable results in all simulation scenarios under the bell-shaped distribution of the prognostic index.

When assessing the performance of the C-statistics under the right skewed distribution of the prognostic index derived from the model, the results suggest that both bias and MSE increase with decreasing sample size, which is true for all types of the C-statistic (Table 3). Of them, C_P showed greater amount of bias compared to those associated with the C_U, C_K for all scenarios under the right skewed distribution. Although the MSE for C_P is comparatively lower than those associated with the other C-statistics under study when sample size is large but higher when sample size is small (50). In terms of coverage, both the C_U and C_K showed better performance than that associated with C_P. Similar results can be observed for all simulation scenarios under the left skewed distribution of the prognostic index (Table 4). The amount of bias for all types of C-statistic for all scenarios under skewed (both left and right skewed) prognostic index are slightly larger than those associated with

Table 2: Empirical comparison of concordance statistics when the distribution of prognostic score is normal distribution

Prev.	Sample size	C-statistic	Estimate	Bias	MSE	Coverage
45%	189	C_U	0.74639	0.00039	0.00108	0.918
		C_K	0.74639	0.00039	0.00108	0.918
		C_P	0.74702	0.00102	0.00105	0.922
	100	C_U	0.74744	0.00144	0.00216	0.906
		C_K	0.74744	0.00144	0.00215	0.906
		C_P	0.74747	0.00147	0.00192	0.924
	75	C_U	0.76083	0.01483	0.00332	0.868
		C_K	0.76084	0.01484	0.00332	0.870
		C_P	0.76199	0.01599	0.00294	0.918
	50	C_U	0.72038	-0.02563	0.00698	0.912
		C_K	0.72033	-0.02567	0.00699	0.910
		C_P	0.64758	-0.09843	0.01005	0.994
30%	189	C_U	0.74853	-0.00147	0.00130	0.906
		C_K	0.74852	-0.00148	0.00130	0.906
		C_P	0.74969	-0.00031	0.00125	0.904
	100	C_U	0.75167	0.00167	0.00236	0.910
		C_K	0.75168	0.00168	0.00236	0.910
		C_P	0.75184	0.00184	0.00215	0.930
	75	C_U	0.76591	0.01591	0.00317	0.876
		C_K	0.76594	0.01594	0.00318	0.876
		C_P	0.76729	0.01729	0.00299	0.906
	50	C_U	0.72143	-0.02857	0.00762	0.902
		C_K	0.72139	-0.02861	0.00761	0.902
		C_P	0.64694	-0.10307	0.01098	0.994
20%	189	C_U	0.75527	0.00227	0.00143	0.918
		C_K	0.75526	0.00226	0.00143	0.920
		C_P	0.75615	0.00315	0.00133	0.932
	100	C_U	0.75030	0.00269	0.00328	0.892
		C_K	0.75032	0.00267	0.00327	0.890
		C_P	0.74911	-0.00389	0.00305	0.908
	75	C_U	0.77109	0.01809	0.00468	0.856
		C_K	0.77102	0.01802	0.00467	0.858
		C_P	0.77129	0.01829	0.00434	0.894
	50	C_U	0.72313	-0.02987	0.00769	0.912
		C_K	0.72301	-0.02999	0.00766	0.918
		C_P	0.65809	-0.09492	0.00945	1.000
10%	189	C_U	0.75736	-0.00264	0.00319	0.904
		C_K	0.75733	-0.00267	0.00319	0.908
		C_P	0.75721	-0.00279	0.00321	0.890
	100	C_U	0.75661	-0.00339	0.00475	0.894
		C_K	0.75661	-0.00339	0.00475	0.894
		C_P	0.75559	-0.00441	0.00419	0.918
	75	C_U	0.78224	0.02224	0.00714	0.830
		C_K	0.78231	0.02231	0.00710	0.830
		C_P	0.78199	0.02299	0.00692	0.872
	50	C_U	0.79238	0.03238	0.01738	0.844
		C_K	0.76010	0.00010	0.00927	0.900
		C_P	0.71708	0.04292	0.00299	0.994

the scenarios under the bell-shaped (normally distributed) prognostic index. Both C_U and C_K showed comparable results for all scenarios under the skewed distribution.

Table 3: Empirical comparison of concordance statistic for right skewed distribution of the prognostic score

Prev.	Sample size	C-statistics	Estimate	Bias	MSE	Coverage
45%	189	C_U	0.70204	0.00304	0.00139	0.908
		C_K	0.70205	0.00305	0.00139	0.908
		C_P	0.69381	-0.00519	0.00039	0.998
	100	C_U	0.70226	0.00326	0.00235	0.916
		C_K	0.70222	0.00322	0.00235	0.918
		C_P	0.68074	-0.01826	0.00086	0.980
	75	C_U	0.68681	-0.01219	0.00380	0.906
		C_K	0.68679	-0.01221	0.00380	0.908
		C_P	0.69078	-0.00823	0.00071	0.990
	50	C_U	0.63857	-0.06043	0.01022	0.844
		C_K	0.63852	-0.06048	0.01023	0.848
		C_P	0.56863	-0.13037	0.01701	0.014
10%	189	C_U	0.54562	0.00862	0.00187	0.884
		C_K	0.54562	0.00862	0.00187	0.884
		C_P	0.55712	0.02012	0.00192	0.876
	100	C_U	0.55334	0.01634	0.00294	0.922
		C_K	0.55333	0.01633	0.00294	0.922
		C_P	0.56856	0.03156	0.00305	0.902
	75	C_U	0.56627	0.02927	0.00479	0.882
		C_K	0.56628	0.02928	0.00479	0.884
		C_P	0.59015	0.05315	0.00616	0.802
	50	C_U	0.52794	-0.00907	0.00649	0.916
		C_K	0.52791	-0.00909	0.00649	0.916
		C_P	0.57752	0.04052	0.00167	0.990

5 Discussion

The concordance statistic (C-statistic) is frequently used to assess the discriminatory ability of a risk model developed in the logistic regression framework for binary data. Given the several approaches of the C-statistics in the literature, this paper evaluated some commonly used C-statistics by an extensive simulation study and illustrating them using two datasets of different prognostic abilities and compared their performance in order to make practical

Table 4: Empirical comparison of concordance statistic for left skewed distribution of the prognostic score

Prev.	Sample size	C-statistics	Estimate	Bias	MSE	Coverage
45%	189	C_U	0.79052	0.00552	0.00088	0.922
		C_K	0.79055	0.00556	0.00087	0.922
		C_P	0.77477	-0.01023	0.00030	0.994
	100	C_U	0.79399	0.00899	0.00164	0.904
		C_K	0.79396	0.00896	0.00164	0.902
		C_P	0.75776	-0.02724	0.00098	0.988
	75	C_U	0.79698	0.01198	0.00251	0.878
		C_K	0.79689	0.01189	0.00249	0.880
		C_P	0.77182	-0.01318	0.00043	0.990
	50	C_U	0.73734	-0.04766	0.00693	0.878
		C_K	0.73734	-0.04766	0.00687	0.874
		C_P	0.56838	-0.21662	0.04693	0.010
10%	189	C_U	0.95115	0.00415	0.00037	0.842
		C_K	0.95115	0.00415	0.00364	0.846
		C_P	0.90344	-0.04357	0.00202	0.061
	100	C_U	0.96152	0.01453	0.00053	0.658
		C_K	0.96165	0.01465	0.00053	0.656
		C_P	0.91825	-0.02875	0.00094	0.326
	75	C_U	0.95302	0.00602	0.00077	0.820
		C_K	0.95257	0.00557	0.00074	0.813
		C_P	0.89346	-0.05354	0.00311	0.049
	50	C_U	0.57601	-0.37098	0.17672	0.978
		C_K	0.42374	-0.52326	0.30323	0.233
		C_P	0.28923	-0.65778	0.43382	0.072

recommendations. Illustration using two datasets of different prognostic abilities suggest that there are some differences in the estimates between the C-statistics under study probably due to the difference in the prevalence of the binary outcome, sample size, and the distribution of the prognostic score derived from the model. Further, the simulation studies based on low-birth weight data suggest that all the C-statistics (C_U, C_K, and C_P) provide comparable results when the distribution of the prognostic index derived from the model is normal and sample size is large. However, all of them showed increasing bias and MSE with decreasing sample size and the proportion of binary outcome, and greater bias and MSE for the skewed distribution of the prognostic index. Of them, the parametric C-statistic, C_P, showed worst performance by providing largest amount of bias and MSE and poor coverage probability when sample is small or prevalence of binary outcome is low or distribution of the prognostic index is skewed or any combination of these scenarios. Both the C_U and C_K

showed comparable results in most simulation scenarios.

The reason for the poor performance of the parametric C-statistic (C_P) when sample size is small or prevalence of the binary outcome is low is violation of the normality assumption for the prognostic index derived form the model. Whereas the non-parametric C-statistics (C_U and C_K) do not require such assumption and hence showed better performance. However, the only advantage of parametric approach is the simplicity of the estimation of the confidence interval using well known delta method (Reiser, 2000; Faraggi, 2000) over the non-parametric approaches suggested by DeLong et al. (1988). With respect to computational aspects, all types of C-statistics, particularly C_P, are easier to implement in the standard statistical software. In addition, several packages and functions are available in the commonly used statistical software such as R, Stata, and SAS. For example, the packages `lroc`, and `roctab` for non-parametric C_U, `rocfit` for parametric C_P are available in Stata; `pAUC`, `PRROC` and `ROCit` for C_U in R; and `PROC LOGISTIC` for both C_U and C_P in SAS.

This study restricts the illustration of the methods using only two practical datasets and a simulation study based on one of these datasets, where simulation scenario were created considering maximum sample size of 189, which is equal to the sample size of the original dataset. Further simulation could be possible by creating simulation scenarios of very large sample size (larger than 189) to check if there is any difference in the results between different types of C-statistics under study. However, for the scenario with sample size 189, all the methods under study showed comparable results, suggesting similar results will be found for the simulation with large sample size. We actually tried here to create some simulation scenarios that reflect the characteristics of the practical data that we have. This does not mean that the methods have limited use to other datasets. We therefore discuss some guidelines, based on the findings of the study and software availability, for practical users of the C-statistics as follows.

If the sample size is large and distribution of the prognostic index derived from the model is normal, one can apply one of the C-statistics under study as all they perform equally for those situations. However, the parametric C_P could be a reasonable option as it easier to implement in most statistical software. For any other conditions, that is, if the distribution of the prognostic index is skewed or sample size is small or prevalence of outcome is low or any combination of these conditions exists, one may consider either of the non-parametric method, C_U, or Kernel based method, C_K, as both perform equally. Between these two C-statistics, the C_K requires appropriate choice of bandwidth selection and self-written program for implementation because of limited availability of programs in statistical software. However, the C_U overcomes all of these constraints and hence recommended to use in practice. Finally, we recommend to check the sample size, the prevalence of the binary outcome, and the distribution of the prognostic index derived from the risk model before selecting an appropriate estimator of the C-statistic for assessing the discriminatory power of the risk model developed in the logistic regression framework.

References

Altman, D. G. and Royston, P. (2000), "What do we mean by validating a prognostic model?" *Statistics in Medicine*, 19, 453–473.

Antolini, L., Nam, B.-H., and D'Agostino, R. B. (2004), "Inference on correlated discrimination measures in survival analysis: a nonparametric approach," *Communications in Statistics-Theory and Methods*, 33, 2117–2135.

Austin, P. C. and Steyerberg, E. W. (2012), "Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable," *BMC Medical Research Methodology*, 12, 82.

Choodari-Oskooei, B., Royston, P., and Parmar, M. K. (2012), "A simulation study of predictive ability measures in a survival model I: explained variation measures," *Statistics in Medicine*, 31, 2627–2643.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988), "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, 44, 837–845.

Faraggi, D. (2000), "The effect of random measurement error on receiver operating characteristic (ROC) curves," *Statistics in Medicine*, 19, 61–70.

Faraggi, D. and Reiser, B. (2002), "Estimation of the area under the ROC curve," *Statistics in Medicine*, 21, 3093–3106.

Hall, P. G. and Hyndman, R. J. (2003), "Improved methods for bandwidth selection when estimating ROC curves," *Statistics & Probability Letters*, 64, 181 – 189.

Hanley, J. A. and McNeil, B. J. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, 143, 29–36.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996), "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, 15, 361–387.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013), *Applied logistic regression*, vol. 398, John Wiley & Sons.

Lloyd, C. J. (1998), "Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems," *Journal of the American Statistical Association*, 93, 1356–1364.

Mann, H. B. and Whitney, D. R. (1947), "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, 50–60.

Metz, C. E. (1978), "Basic principles of ROC analysis," in *Seminars in nuclear medicine*, Elsevier, vol. 8, pp. 283–298.

Molodianovitch, K., Faraggi, D., and Reiser, B. (2006), "Comparing the Areas Under Two Correlated ROC Curves: Parametric and Non-Parametric Approaches," *Biometrical Journal*, 48, 745–757.

Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., and Altman, D. G. (2009), "Prognosis and prognostic research: what, why, and how?" *Bmj*, 338, b375.

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.

Pencina, M. J., Fine, J. P., and D'Agostino Sr., R. B. (2017), "Discrimination slope and integrated discrimination improvement – properties, relationships and impact of calibration," *Statistics in Medicine*, 36, 4482–4490.

Pulit, M. (2016), "A new method of kernel-smoothing estimation of the ROC curve," *Metrika*, 79, 603–634.

Reiser, B. (2000), "Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves," *Statistics in Medicine*, 19, 2115–2129.

Royston, P. and Altman, D. G. (2010), "Visualizing and assessing discrimination in the logistic regression model," *Statistics in Medicine*, 29, 2508–2520.

Steyerberg, E. W. (2008), *Clinical prediction models: a practical approach to development, validation, and updating*, Springer Science & Business Media.

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 some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, 21, 128.

Wyatt, J. C. and Altman, D. G. (1995), "Commentary: Prognostic models: clinically useful or quickly forgotten?" *Bmj*, 311, 1539–1541.

Zhou, X.-H. and Harezlak, J. (2002), "Comparison of bandwidth selection methods for kernel smoothing of ROC curves," *Statistics in Medicine*, 21, 2045–2055.

Received: May 31, 2019

Accepted: July 27, 2019