# C-statistic for evaluating the added predictive ability in the binary risk models: An alternative to the existing approaches

Afrin Sadia Rumana*
Institution of Statistical Research and Training, Dhaka, Bangladesh - arumana@isrt.ac.bd

Mohammad Shafiqur Rahman
Institution of Statistical Research and Training, Dhaka, Bangladesh - shafiq@isrt.ac.bd

**Abstract**

Logistic regression models are frequently used in various settings of clinical research to predict the risk of a patient's future health status such as death or illness 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. Given their important role in clinical research, it is very essential to evaluate the predictive performance of the model e.g the ability of the model to distinguish between low-and high-risk patients-which is termed as 'discrimination'. Concordance statistic (C-statistic) is frequently used to quantify the discriminatory power of the logistic models. A particular problem of interest in this area is to quantify the added value of a set of influential predictors in the predictive performance of the model. Several proposals of C-statistics has been discussed in the literature, however none of these is able to translate the statistical significance of a new predictor into statistical significant improvement in the predictive performance of the model. More specifically, these C-statistics are not sensitive enough to the inclusion of additional predictors in the model, which leads to misleading conclusion on model's predictive/discriminatory performance. To address this problem, this paper proposed an alternative estimation for C-statistics. The new C-statistic is based on the concordance probability definition proposed by Gonen and Heller for Cox PH models, which quantifies the actual predictive value (or risk difference) added by the new predictor in the model. The method is illustrated by an application to low birth weight data. Further a simulation study was conducted to assess the performance of the new and existing C-statistics and compared the results. The results showed that the new C-statistic was more sensitive to quantify the added predictive value of the model.

**Keywords**:Logistic regression, C-index, ROC curve, Discrimination, Prediction, Predictive accuracy.

*Presenting author

# 1 Introduction

Logistic regression models are frequently used in various settings of clinical research to predict the risk of a patient's future health status such as death or illness 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. Given their important role in clinical research, it is very essential to evaluate the predictive performance of the model e.g the ability of the model to distinguish between low-and high-risk patients-which is termed as 'discrimination'. 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 of the logistic models. A particular problem of interest in this area is to quantify the added value of a set of influential predictors in the predictive/discriminatory performance of the model. The added predictive value of the new predictors in the model can be quantified using both parametric and non-parametric C-statistic. However, using the these C-statistics, statistical significance of a new predictor does not necessarily translate into statistical significant improvement in the predictive performance of the model [Chen et.al 2013, Pencina et al 2008, Delmer et al 2011]. More specifically, these C-statistics are not sensitive enough to the inclusion of additional predictors in the model, which leads to misleading conclusion on model's predictive/discriminatory performance. This is because the non-parametric C-statistic based on U-statistic is a ranked based measure and ranking the prognostic value(or log-odds) of two subjects rather than quantifying their actual risk differences.

Addressing this issue, this paper proposed an alternative estimation for C-statistics based on the concordance probability definition proposed by Gonen and Heller for Cox PH model [4]. Unlike the existing C-statistics, the new C-statistic quantifies the concordance between two subjects one with event and other without event by calculating the distance in prognostic values (or log-odds) derived from the model for these two subjects. Hence this new C-statistic is able to quantify the actual predictive value (or risk difference) added by the new predictor compared to those for the existing C-statistics.

# 2 Methodology

## 2.1 Binary logistic regression 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 is defined as

$$\text{logit}[\Pr(Y_i = 1|x_i)] = \log(\frac{\pi_i}{1-\pi_i}) = \eta_i = \beta^T x_i, \tag{1}$$

where $\beta^T$ is a vector of regression coefficients of length ($p$+1), and $x_i$ is the $i$th row vector of the predictor matrix $\mathbf{x}$ which has order $n \times (p+1)$. The term $\eta_i = \beta^T x_i$ is called as risk score or 'prognostic index (PI)'.

## 2.2 Concordance Statistics for Logistic Regression Model

Concordance statistic, identical to the area under the receiver operating characteristic curve (AUC), can be obtained by quantifying the concordance probability that the subject who experienced the event of interest had a higher predicted probability of experiencing the event than those who did not experience the event. For a pair of subjects $(i, j)$, the C-statistic can be defined as

$$\begin{aligned} C =& \Pr[\pi(\beta|x_i)|Y_i = 1 > \pi(\beta|x_i)|Y_i = 0] \\ =& \Pr[(\beta^T x_i|Y_i = 1) > (\beta^T x_j|Y_j = 0)]. \end{aligned} \tag{2}$$

The value of C-index ranges between 0.5 and 1 with a value of 0.5 indicates no discrimination and a value of 1 indicates perfect discrimination. Two approaches to calculate the concordance probability for binary data have been discussed: Non-parametric approach based on Mann-Whitney U statistic [5] and parametric approach based on binormal distribution assumption and method of maximum likelihood[6].

### 2.2.1 Non-parametric estimation

Let $\eta_i^{(1)} = \beta^T x_i|Y_i = 1$ and $\eta_j^{(0)} = \beta^T x_j|Y_j = 0$ be the prognostic index 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 as

$$C_U = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(\eta_i^{(1)}, \eta_j^{(0)}), \tag{3}$$

where

$$I(\eta_i^{(1)}, \eta_j^{(0)}) = \begin{cases} 1 & \eta^{(1)} > \eta^{(0)}, \\ \frac{1}{2} & \eta^{(1)} = \eta^{(0)}, \\ 0 & \eta^{(1)} < \eta^{(0)}. \end{cases}$$

### 2.2.2 Parametric estimation

Based on the central limit theorem, the prognostic index is likely to follow normal distribution as the dimension of the parameter vector $\beta$ increases . The estimation of the parametric C-index is as follows. Let us assume that $\eta_i^{(1)} = (\beta^T x_i|Y_i = 1) \sim N(\mu_1, \sigma^2)$ and $\eta_j^{(0)} = (\beta^T x_j|Y_j = 1) \sim N(\mu_0, \sigma^2)$. Therefore, $\eta_i^{(1)} - \eta_j^{(0)} \sim N(\mu_1 - \mu_0, 2\sigma^2)$. The parametric concordance statistic is

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

After standardising the term $\eta_i^{(1)} - \eta_j^{(0)}$ , $C^p$ can be obtained as

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

where $Z \sim (0, 1)$ and $\Phi(.)$ is the standard normal CDF. The estimate of $C_P$ can be obtained by replacing $\mu_1$, $\mu_0$ and and $\sigma_1^2$, $\sigma_0^2$ by their sample estimates $\bar{x_1}$, $\bar{x_0}$ and $s_1^2$, $s_2^2$, respectively.

### 2.2.3 Proposed New C-Statistic

The proposed a new estimator for C-statistic is based on the concordance probability definition of Gonen and Heller proposed for Cox PH model [4]. The Gonen and Heller's concordance statistic quantifies the concordance probability based on the predicted risk difference between the subjects of a pair. Similar to the non-parametric C-statistic for binary data the new C-statistic denoted by $K$ for binary data uses those pairs in which one subject developed the event and the other did not. However, $K$ quantifies the concordance probability by taking the differences in the risk predicted by the model for the subject with event and those without event. Let us consider a pair of subjects $(i, j)$ with event and without event who had predicted log-odds $\beta^T x_i$ and $\beta^T x_j$ respectively. The concordance probability for logistic model can be defined as

$$K = \Pr[(\beta^T x_i | Y_i = 1) \geq (\beta^T x_j | Y_j = 0)] = \frac{1}{1 + \exp(\beta^T x_j - \beta^T x_i)}$$

If $\beta^T x_i = \beta^T x_j$ the above formula produces probability 0.5 and if $\beta^T x_i > \beta^T x_j$ then it produces probability greater than 0.5. For all possible such pairs the above concordance probability can be estimated as

$$K = \frac{1}{\sum_i \sum_j I(\hat{\beta}^T x_i \geq \hat{\beta}^T x_j)} \sum_i \sum_j \frac{I(\hat{\beta}^T x_i \geq \hat{\beta}^T x_j)}{1 + \exp(\hat{\beta}^T x_j - \hat{\beta}^T x_i)} \tag{4}$$

# 3 Application

## 3.1 Low birth weight data

Low birth weight data [7] were collected on 189 women, 59 of which had babies with low birth weight and 130 of which had normal birth weight. 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/One)), history of hypertension (HT (Yes/no)), presence of uterine irritability (UI (Yes/No)), number of physician visits during the first trimester (FTV (None/One/Two)), and most important low birth weight (LOW (Birth Weight $\geq$ 2500g/Birth Weight $<$ 2500g)).

To assess the added predictive ability of risk models we develop several nested models with different predictive abilities-null model to full model. For each model C-statistics were calculated and compared their results to examine the ability of the C-statistics in quantifying the added predictive value. The results in Table 1 shows that the new C-statistic is more sensitive to inclusion of new predictor in the model, i.e. to quantify the added predictive ability of the model.

Table I: **Estimated Concordance statistics for the Low birth weight data**

| **Model** | **Variables** | $C_U$ | $C_P$ | $K$ |
|---|---|---|---|---|
| Model I | Null Model | .5 | .5 | .5 |
| Model II | Age | .55254 | .57493 | .57608 |
| Model III | Age+LWT | .62633 | .62060 | .63050 |
| Model IV | Age+LWT+RACE | .65325 | .65303 | .66759 |
| Model V | Age+LWT+RACE+SMOKE | .68370 | .69943 | .71449 |
| Model VI | Age+LWT+RACE+SMOKE+PTL | .71245 | .71666 | .73147 |
| Model VII | Age+LWT+RACE+SMOKE+PTL+HT | .74061 | .73969 | .76108 |
| Model VIII | Age+LWT+RACE+SMOKE+PTL+HT+UI | .74609 | .75058 | .77316 |

# 4 Simulation Study

A simulation study was conducted to evaluate the added predictive ability in the risk model using all types of C-statistics. First we simulate a continuous explanatory variable from a standard normal distribution. We determine a model as follows: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, where $\pi$ denotes the probability of occurring a binary condition. We considered several models with increasing predictive performances. We fixed the value of $\beta_0$ at -2.5 and increase the value of $\beta_1$ from 0.15 to 2.85 to form the models with increasing predictive performances. From each model, we then randomly generated binary response of size 200 from a Bernoulli distribution with subject-specific $\pi$ from the true model. For each simulation scenarios based on the values of $\beta_1$, we generation 500 data sets. We then fit a logistic regression model in the simulated dataset and estimated the C-statistics for the fitted model, which we refer to as the empirical C-statistic and takes average of 500 values. We investigate the subsequent improvement in the predictive performance by using C-statistics discussed due to increasing value of $\beta_1$ or log-odds ratio. The results in Figure 2 shows that the new C-statistic (K) is more sensitive to the inclusion of added predictive ability in the model than the other C-statistics.

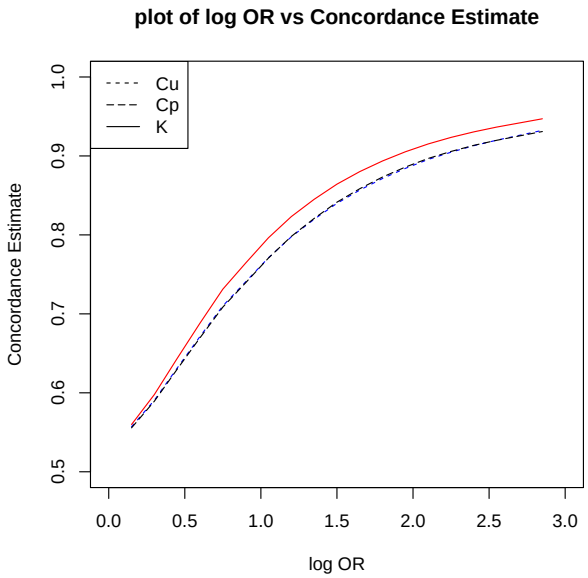


Figure 1: Estimates of the C-statistics against log OR

# References

[Chen et.al 2013] Chen et al. On the assessment of the added value of new predictive biomarkers. *BMC Medical Research Methodology* 2013; **13**:98.

[Pencina et al 2008] Pencina, MJ et al. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*; **27**: 157172.

[Delmer et al 2011] Demler et al. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality.*Statistics in Medicine* **30**: 14101418.

[4] Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92(4)**:1799-1809.

[5] Hanley JA, McNeil, BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982; **143**:29-36.

[6] Faraggi D, Reiser B (2004).Estimation of the area under the ROC curve. *Statistics in Medicine* 2002, **21**:3093-3106.

[7] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience Publication, 2nd edition, 2000.