Assessing discriminatory ability of random effects logistic models for clustered binary outcomes

M. Shafiqur Rahman*†, Gareth Ambler‡, and Rumana Z. Omar‡

Abstract

In multicentre studies patients are typically clustered within centres and are likely to be correlated. Typically, random effects logistic models are fitted to clustered binary outcomes. However, limited work has been done to assess the discriminatory ability of these models: the ability of the model to distinguish between low-and high-risk patients. The C-index has been used to assess discrimination in the standard logistic model. For clustered data, the naïve use of the standard C-index may lead to misleading conclusions regarding the model's discriminatory ability. This paper extends the standard C-index to use with random effects logistic models, resulting in an 'Overall' C-index and a Pooled cluster-specific C-index. Both indices have individual interpretation. The 'Overall' approach can produce two different values for the C-indices depending on type of predictions: conditional and marginal predictions. The methods are illustrated using real data on patients following heart valve surgery and their performances are investigated using simulation studies with several scenarios related to clustered data.

Keywords: clustered binary data, random effect model, discrimination, model validation.

1 Introduction

Random effects logistic models [1] are commonly used to analyse clustered binary data from multi-center studies. However, limited work has been done to assess the discriminatory ability of these models: the ability of the model to distinguish between low-and high-risk patients [2]. The C-index [3] is commonly used to assess the discriminatory ability of standard logistic models. For clustered data, the naïve use of the standard C-index may lead to misleading conclusions regarding the model's discriminatory ability. The naïve approach assesses the effects of the fixed predictors only, and the discriminatory ability may change if clustering effects are considered in addition to the effects of the fixed predictors. Furthermore, assessing the model's performance within each cluster may be of interest, particularly to identify outlying clusters. This paper extends the standard C-index to use with random effects logistic models. The paper begins with a brief description of the proposed C-index for independent binary outcomes, then discusses the estimation of these measures for clustered data. A simulation study is conducted to evaluate the performance of the new measures under various clustered data scenarios. The methods are illustrated using data on patients who had undergone heart valve surgery.

2 C-index for standard logistic regression

2.1 The model

Let Y_i $(i = 1, \ldots, N)$ be a binary outcome (0/1)from $Bernoulli(1, \pi_i)$ with $\pi_i = \Pr(Y_i = 1)$. The logistic regression model can be defined as

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

*Institute of Statistical Research and Training, University of Dhaka, Bangladesh
†Corresponding author's email: shafiq@isrt.ac.bd
‡Department of Statistical Science, University College London, UK

The term $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ is known as the 'prognostic index' (PI). The predictive form of this model can be written as

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

Predictions from the model depend on the estimate of $\boldsymbol{\beta}^T$, which is typically obtained by the method of maximum likelihood [4].

2.2 The C-index for logistic model

The C-index is numerically identical to the area under the receiver operating characteristic curve (AUC) [3]. It equals to the proportion of pairs in which the predicted event probability is higher for the subject who experienced the event of interest than that of the subject who did not experience the event. For a pair of subjects (i, j), where i and j correspond to those 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-index can be defined as

$$C = \Pr[\pi(\boldsymbol{\beta}|\boldsymbol{x}_i) > \pi(\boldsymbol{\beta}|\boldsymbol{x}_j)|Y_i = 1 \;\&\; Y_j = 0],$$

which is equivalent to

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

The above statistic can be estimated based on the Mann-Whitney U statistic [3].

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 PI derived by the model for subject i with event and for subject j without event, respectively. Further, let N_1 and N_0 be the number of events and non-events, respectively. Considering all pairs (i, j), the C-index can be estimated by analogy to the U statistic formulation [3] as

$$C = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} I\big(\eta_i^{(1)}, \eta_j^{(0)}\big), \tag{1}$$

where

$$I\big(\eta^{(1)}, \eta^{(0)}\big) = \begin{cases} 1 & \text{if } \eta^{(1)} > \eta^{(0)} \\ 0.5 & \text{if } \eta^{(1)} = \eta^{(0)} \\ 0 & \text{if } \eta^{(1)} < \eta^{(0)} \end{cases}.$$

The value of C ranges between 0.5 and 1: a value of 0.5 indicates that the model has no ability to discriminate between low and high risk subjects, whereas a value of 1 indicates that the model can perfectly discriminate between these two groups.

3 C-index for clustered data

We propose two approaches to calculate C-index in the clustered data setting, which results in an 'overall' and a 'pooled cluster-specific' indices. In the 'overall' approach, one calculates the C-index from a comparison of subjects within and between clusters, and the resulting C-index assesses the overall predictive ability of the model. In the 'pooled cluster-specific' approach, one calculates the validation measure for each cluster based on its original definition for standard logistic model along with a measure of precision. These measures are then pooled across clusters using the random-effects summary statistic method often used in meta analysis [5]. This approach yields a weighted average of the cluster-specific values, referred to as a

'pooled estimate'. The 'pooled cluster-specific' measure assesses the predictive ability of the predictors whose values vary within clusters.

3.1 The random effect logistic model for clustered data

Let Y_{ij} be a binary outcome (1/0) for the ith subject in the jth cluster of size n_j ($i = 1, \ldots, n_j; j = 1, \ldots, J$) and $\sum_{j=1}^{J} n_j = N$. It is assumed that $Y_{ij} \sim$ Bernoulli(π_{ij}), where $\pi_{ij} = \Pr(Y_{ij} = 1)$. The random-intercept logistic model is an extension of the standard logistic model with an additional cluster-specific random effect u_j. Typically u_js are $N(0, \sigma_u^2)$. The random-intercept logistic regression model is given by:

$$\text{logit}[\Pr(Y_{ij} = 1|u_j, \boldsymbol{x}_{ij})] = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_j,$$

The predictive form of the random effect logistic model for subject i in cluster j is given by

$$\pi(\boldsymbol{\beta}|u_j, \boldsymbol{x}_{ij}) = \frac{\exp[\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)]}{1 + \exp[\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)]},$$

where $\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)$=$\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_j$ is referred to as PI. Predictions from the model depend on the estimates of the model parameters $(\boldsymbol{\beta}^T, \sigma_u^2)$ and the random effect u_j. The model parameters can be estimated using adaptive Gaussian quadrature (AGQ) [6]. Using the estimates of the model parameters, the random effect u_j for the jth cluster can be obtained by empirical Bayes approach [7]. The above predictions based on u_j are called conditional predictions. Marginal predictions, π_{ij}(pa), can be made by integrating the conditional prediction $\pi(\boldsymbol{\beta}|u, \boldsymbol{x})$ over the (prior) random effects distribution.

3.2 C-index for random effect logistic model

For a pair of subjects (i, k) from clusters (j, l) respectively, where i and k correspond to subject who with event and those without event respectively, with event probability $\{\pi_{ij}(u), \pi_{kl}(u)\}$, the C-index can be defined as

$$C_{re(u)} = \Pr[\pi_{ij}(u) > \pi_{kl}(u)] \Leftrightarrow \Pr[\eta_{ij}(u) > \eta_{kl}(u)].$$

This applies to all possible pairs (i, k) in the data, where a pair may consist of subjects from the same cluster or from different clusters. If subjects are from different clusters, the cluster-specific random effect u values contribute in determining whether a pair is concordant and in $C_{re(u)}$, even if both subjects have the same predictor values. The random effects u however do not contribute in determining a concordant pair if both subjects are from the same cluster, as they share the same value of the random effect. Similarly, based on population average probabilities $\{\pi_{ij}(\text{pa}), \pi_{kl}(\text{pa})\}$, the C-index can be defined as

$$C_{PA} = \Pr[\pi_{ij}(\text{pa}) > \pi_{kl}(\text{pa})] \Leftrightarrow \Pr[\eta_{ij}(\text{pa}) > \eta_{kl}(\text{pa})].$$

3.2.1 The Overall C-index

Let $\eta_{ij}^{(1)}(u) = \eta_{ij}(u)|Y_{ij} = 1$ be PI for the ith subject with an event from the jth cluster, derived from $\pi_{ij}(u)$. Similarly, let $\eta_{kj}^{(0)}(u) = \eta_{kj}(u)|Y_{kj} = 0$ be PI for the kth subject without an event from the jth cluster. Let n_{1j} and n_{0j} be the number of subjects with an event and without an event respectively in the jth cluster. The total number of subjects with an event is $N_1 = \sum_j n_{1j}$, and the total number of subjects without an event is $N_0 = \sum_j n_{0j}$. Further,

let J_1 and J_0 be the total number of clusters with at least one subject with an event and one without an event, respectively. Note that $J \leq (J_1 + J_0) \leq 2J$.

Extending equation (1), the C-index for clustered data can be defined as

$$C_{re(u)} = \frac{1}{N_1 N_0} \sum_{j=1}^{J} \sum_{l=1}^{J} \sum_{i=1}^{n_j} \sum_{k=1}^{n_l} I\Big(\eta_{ij}^{(1)}(u), \eta_{kl}^{(0)}(u)\Big), \tag{2}$$

where $I(.)$ can be defined similarly as in eq(1). The C-index based on $\pi_{ij}(\text{pa})$ can be obtained using the same approach to that described in equation (2) but by replacing $\eta_{ij}^{(1)}(u)$ and $\eta_{kl}^{(0)}(u)$ by the corresponding prognostic indices derived from $\pi_{ij}(\text{pa})$. The resulting C-indices is denoted by C_{PA}.

3.2.2 Pooled cluster-specific C-index

Let $\hat{C}_j$ $(j = 1, \ldots, J)$ be the estimate of C-index for the jth cluster obtained using its standard definition, and $\hat{\sigma}_j^2$ be the corresponding estimated variance. The weighted average (pooled estimate) of the cluster specific estimates can be calculated as

$$\hat{C}_P = \bar{w}^{-1} \sum_{j=1}^{J} \hat{C}_j \hat{w}_j, \tag{3}$$

where $\hat{w}_j = 1/(\hat{\sigma}_j^2 + \hat{\tau}^2)$, $\bar{w} = \sum_{j=1}^{J} \hat{w}_j$, and $\hat{\tau}^2$ is the estimate of the between cluster variance and can be obtained as

$$\hat{\tau}^2 = \max\left\{0, \frac{\left[\sum_{j=1}^{J} \hat{a}_j (\hat{\theta}_j - \bar{\theta})^2\right] - (J-1)}{\sum_{j=1}^{J} \hat{a}_j - \sum_{j=1}^{J} \hat{a}_j^2 / \sum_{j=1}^{J} \hat{a}_j}\right\},$$

where $\hat{a}_j = 1/\hat{\sigma}_j^2$ and $\bar{C} = \sum_{j=1}^{J} \hat{a}_j \hat{C}_j / \sum_{j=1}^{J} \hat{a}_j$.

4 Simulation study

4.1 Simulation design

The properties of the C-indices such as bias and coverage were investigated by a simulation study. Both development and validation data were simulated from a true model based on random intercept logistic model. Prognostic models were developed using the simulated development data and then evaluated using the corresponding simulated validation data. The properties of the C-indices were investigated in a range of scenarios. A total of four ICC values such as 0%, 5%, 10%, and 20% were considered, to mimic scenario with different levels of clustering. Under each ICC value, development datasets each with 100 clusters of size 100 were generated. For each development set, validation datasets from several scenarios were generated. These include (i) 10 clusters of sizes 10 and 300, and (ii) 100 clusters of sizes 10, 30, and 100. For each of the development and validation scenarios, 500 datasets were generated.

For a sample of size N with J clusters, the predictor value x_{ij} for the ith subject in the jth cluster $(i = 1, \ldots, n_j; j = 1, \ldots, J)$ was generated from $\text{N}(0, 1)$, and the true random effects u_j were from $\text{N}(0, \sigma_u^2)$. Then the outcomes y_{ij} were generated from the Bernuolli distribution with probability calculated from the true random-intercept logistic model using

$$\pi(\beta_0, \beta_1 | x_{ij}, u_j) = \frac{\exp[\eta(\beta_0, \beta_1, x_{ij}, u_j)]}{1 + \exp[\eta(\beta_0, \beta_1, x_{ij}, u_j)]}. \tag{4}$$

As $X \sim \mathrm{N}(0,1)$, $\beta_1 X \sim \mathrm{N}(0, \beta_1^2)$, and therefore $\eta(\beta_0, \beta_1, x_{ij}, u_j)$ follows $\mathrm{N}(\beta_0, \beta_1^2+\sigma_u^2)$, assuming one subject per cluster. To simulate data under different ICC scenarios, the values of σ_u^2 were varied keeping the total predictive variability fixed to 1.4^2, and β_1 was determined from $\beta_1^2+\sigma_u^2 = 1.4^2$. In addition, β_0 was set to a fixed value of -1.8 to generate data with a prevalence of approximately 20% for each of the ICC scenarios.

Table I: Bias and coverage of Overall C-indices: true value=0.806

			Bias				Cov for 95% CI			
C-indices	# clusters	Size \ICC	0%	5%	10%	20%	0%	5%	10%	20%
$C_{re(u)}$	10	10	0.806	0.806	0.809	0.811	88	86	82	74
		300	0.806	0.806	0.806	0.806	90	90	89	91
	100	10	0.806	0.807	0.809	0.810	88	83	76	72
		30	0.806	0.806	0.807	0.808	89	88	86	84
		100	0.806	0.807	0.806	0.807	90	91	89	91
C_{PA}	10	10	0.806	0.805	0.794	0.786	88	83	78	72
		300	0.806	0.805	0.793	.785	90	86	80	73
	100	10	0.806	0.802	0.793	0.795	89	83	78	70
		30	0.806	0.804	0.795	0.787	89	85	76	72
		100	0.806	0.804	0.796	0.787	90	84	78	70

4.2 Results

When there was no clustering in the data (ICC=0%), the C-indices in general showed approximately unbiased estimates for all simulation scenarios (Table I). The C-index based on conditional prediction, $C_{re(u)}$, provided unbiased estimates when the clusters were large. The reason for bias in $C_{re(u)}$ when the clusters are small is possibly due to the poor estimation of the random effects. When the empirical Bayes estimates of the random effects were replaced by their true values in the calculation of $C_{re(u)}$ while still using the estimates of the fixed predictors, $C_{re(u)}$ showed a reasonably good performance even for the small clusters (result not shown). For all simulation scenarios, C_{PA} showed substantial negative bias (but of equal amount) in the presence of clustering, and the bias increased with increasing ICC values. In the presence of clustering (ICC $> 0\%$), C_{PA} showed substantial negative bias in the presence of clustering, and the bias increased with increasing ICC values. This is because C_{PA} ignore the actual contribution of the random effects and therefore underestimate the true value. It is seen from Table II, C_P were unbiased when clusters were large, but it showed large bias for small clusters. The possible reason for bias in C_P when the clusters are small is as follows. The prevalence of the outcome was set at 20% for the simulations. However, the number of events varied between the clusters for high values of the ICC. The minimum number of events required per cluster to calculate C-index is one. When calculating C-index based on small clusters, if

Table II: Bias and coverage of Pooled C-index C_P: true values=0.806, 0.785, 0.744, 0.722

			Bias				Cov for 95% CI			
C-indices	# clusters	Size \ICC	0%	5%	10%	20%	0%	5%	10%	20%
C_P	10	10	0.798	0.756	0.723	0.701	78	75	78	77
		300	0.806	0.785	0.743	.721	90	90	89	90
	100	10	0.795	0.778	0.723	0.702	76	73	75	72
		30	0.804	0.783	0.738	0.719	86	86	86	87
		100	0.806	0.785	0.743	0.722	90	90	89	91

the number of events for a cluster was too low, the cluster was ignored. Thus the calculation of the 'pooled estimate' was often based on a reduced number of clusters, resulting in bias.

5 Application

The new C-indices were illustrated using a dataset of patients who underwent heart valve surgery. The results showed that both the 'overall' and 'pooled cluster-specific' indices have a meaningful interpretation in a clustered data setting. Details of the data and results were not shown for space constraint.

6 Conclusions

This paper has described an adaptation of the C-index for use with models for clustered binary outcomes. Two approaches are proposed: an 'overall' and a 'pooled cluster-specific' indices. The 'Overall' approach produces two different values depending on the model predictions $\hat{\pi}_{ij}(u)$ and $\hat{\pi}_{ij}(\text{pa})$. The properties of the C-indices were evaluated by a simulation study in a range of clustered data scenarios. The simulation results showed that $C_{re(u)}$ showed reasonable performance when there was clustering in the data and the clusters were reasonably large, possibly due to the fact that the random effects were better estimated in larger clusters. C_{PA} performed poorly when there was a moderate level of clustering in the data, because they ignore the effect of clustering. The 'pooled cluster-specific' index, C_P, showed bias when the cluster sizes were small. This is because this approach ignores information from some of these clusters due to lack of events to calculate the index. In general, both the 'overall' and 'pooled cluster-specific' indices are recommended to use to assess the predictive ability of the cluster-data model. However, one needs to check whether the clusters are sufficiently large (for example, greater than 30) and each of these contains at least two events before using the 'pooled' measures.

References

[1] Stiratelli R, Laird NM, Ware JR. Random-effects models for serial observations with binary response. *Biometrics* 1984; **40**:961–71.

[2] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2009; **21(1)**:128–138.

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

[4] Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd edn. Wiley-Interscience Publication, 2000.

[5] DerSimonian R, Laird N. Meta analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.

[6] Pinheiro JC, Chao EC. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics* 2006; **15**:58–81.

[7] Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A* 2009; **172**(3):659–687.