Ranking Accuracy for Logistic-GEE Models

Ranking Accuracy for Logistic-GEE Models

Nasser Davarzani[1(✉)], Ralf Peeters[1], Evgueni Smirnov[1], Joël Karel[1],
and Hans-Peter Brunner-La Rocca[2]

[1] Department of Data Science and Knowledge Engineering, Maastricht University,
P.O.BOX 616, 6200 MD Maastricht, The Netherlands
{n.davarzani,ralf.peeters,smirnov,joel.karel}@maastrichtuniversity.nl
[2] Department of Cardiology, Maastricht University Medical Center,
Maastricht, The Netherlands
hp.brunnerlarocca@mumc.nl

Abstract. The logistic Generalized Estimating Equations (logistic-GEE) models have been extensively used for analyzing clustered binary data. However, assessing the goodness-of-fit and predictability of these models is problematic due to the fact that no likelihood is available and the observations can be correlated within a cluster. In this paper we propose a new measure for estimating the generalization performance of the logistic GEE models, namely ranking accuracy for models based on clustered data (RAMCD). We define RAMCD as the probability that a randomly selected positive observation is ranked higher than randomly selected negative observation *from another cluster*. We propose a computationally efficient algorithm for RAMCD. The algorithm can be applied for two cases: (1) when we estimate RAMCD as a goodness-of-fit criterion and (2) when we estimate RAMCD as a predictability criterion. This is experimentally shown on clustered data from a simulation study and a biomarkers' study.

Keywords: Clustered data · Generalized Estimating Equation · Goodness-of-fit · Predictability · Ranking accuracy

1 Introduction

Clustered data are common in biomedical, clinical, and social-science research [2,9,14]. They are defined as data with a clustered/grouped structure. A cluster (group) can consist of variable measurements of related subjects or repeated variable measurements for a single subject such that in either case the measurements may correlate.

To analyze clustered data, the correlation within clusters needs to be taken into account. To this end, Liang and Zeger [10] proposed an extension of the Generalized Linear Model (GLM) for clustered data with either dichotomous or continuous outcomes [16]. They introduced Generalized Estimating Equations (GEE) to estimate the parameters of the GLM model for dealing with correlated outcomes.

H. Boström et al. (Eds.): IDA 2016, LNCS 9897, pp. 14–25, 2016.
DOI: 10.1007/978-3-319-46349-0_2

The GEE models are widely used for analysis of clustered data, particularly if outcomes are binary (see e.g., [8]). However, due to the fact that no likelihood is available and the residuals (observed outcome minus expected terms) are correlated within a cluster, there is no consensus how to evaluate the GEE models.

This paper addresses the problem of evaluating logistic GEE models. The problem has been considered by several authors (see e.g., [6,7,13]). As a result, several criteria and tests have been proposed for assessing the goodness-of-fit of logistic GEE models. However, most of them have their own shortcomings making impossible having a commonly accepted criterion or test. Below we briefly describe relevant work and then propose our solution.

Barnhart and Williamson [4] proposed a model-based and robust goodness-of-fit test for logistic-GEE models. The method is based on partitioning the space of covariates into distinct regions. The main disadvantage of this method is that applying this method might be problematic when many continuous covariates contribute to the model, or sample sizes are small.

Williamson et al. [15] proposed a Kappa-like classification statistic to assess the model fit of GEE models with categorical outcomes. The disadvantage of the statistic is that for two-class imbalanced data it usually tends to be close to zero (i.e., it states that the model is poorly fitted). Moreover, since no distribution of the statistic is given, interpretation of the statistic is not obvious.

One of the well-established goodness-of-fit statistics for GEE is an *quasilikelihood under the independence model information criterion* (QIC) [12] which is the extension of Akiake's information criterion (AIC) [3]. As a goodness-of-fit and model-selection criterion, the model with smaller QIC is preferred. Since QIC is a function of both quasilikelihood (that depends on the size of the working dataset) and the number of estimated parameters in the GEE model, it indicates the quality of a model relative to other models, fitted with the same data set. That is why it might have different ranges for different data sets. Therefore, QIC is not an applicable criterion for comparing the goodness-of-fit of GEE models for different data sets.

If we generalize the aforementioned goodness-of-fit test statistics and criteria for logistic GEE models, we can derive the following shortcomings: (a) difficulty of interpretation, (b) a relative range of the criterion values (i.e., the range depends on the number of subjects and number of covariates in the model), (c) restriction on the number and types of covariates in the model being evaluated, (d) bias in case of two-class imbalanced data, and (e) inapplicability to indicate the predictability of the model being evaluated.

To propose a criteria that does not suffer from problems (a)–(e), we observe that: (1) logistic GEE models are models trained on clustered data, and (2) logistic GEE models output probabilities of being positive for test observations. The latter implies that logistic GEE models can induce an ordering over those observations. Thus, logistic GEE models actually solve the bipartite ranking task for clustered data [1]. The task is as follows: given labeled clustered data, find an ordering on test observations so that positive observations are ranked higher than negative ones. The standard measure for the quality of that ordering is ranking accuracy. However, it is not applicable for the logistic GEE models, since it does not take into account the within-cluster correlation that might be present, and thus it is not valid.

In this paper we extend the concept of ranking accuracy for clustered data. We propose a new measure that we call *ranking accuracy for models based on clustered data* (RAMCD). It is defined as a probability that a randomly selected positive observation is ranked higher than randomly selected negative observation *from another cluster*. By the definition RAMCD employs the within-cluster correlation in the data used. It focuses on estimating the generalization performance of the logistic GEE models when ranking uncorrelated observations.

We show that RAMCD can be used as a goodness-of-fit criterion and a predictability criterion (i.e., it can be used for estimating the generalization performance of the logistic GEE models beyond training data). For the latter we propose a modification to standard k-fold cross validation method applicable for clustered data.

When comparing RAMCD with the presented standard goodness-of-fit test statistics and criteria for logistic GEE models we observe that RAMCD does not suffer from any of problems (a) to (e) (given above). The main reasons are that: (1) RAMCD is a probability that is easy to interpreted; (2) RAMCD does not impose any restriction on the models being evaluated; (3) RAMCD is not biased for binary imbalanced data (since it indicates class separation); and (4) RAMCD can be used as a goodness-of-fit criterion and a predictability criterion.

The rest of the paper is organized as follows. Section 2 briefly formalizes the bipartite ranking task for clustered data and logistic-GEE model. RAMCD is introduced in Sect. 3. Section 4 provides the experiments and Sect. 5 concludes the paper.

2 Bipartite Ranking Task and Logistic GEE Models

The bipartite ranking task assumes that we have n subjects. The i-th subject is represented by a cluster of m_i observations such that the t^{th} observation, $t = 1, ..., m_i$ is given with p covariates $X_{it1}, ..., X_{itp}$ in $\mathbb{R}$ and a binary outcome variable Y_{it}. Hence, the i-th cluster is identified by $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$, where $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{im_i})'$ in which $\boldsymbol{X}_{it} = (X_{it1}, ..., X_{itp})$ is $1 \times p$ vector of covariates for observation t for subject i and $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_i})'$ is $m_i \times 1$ vector of binary outcomes. For any $i \neq j$ we assume that the correlation between $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ equals 0.0 while the components of each $\boldsymbol{Y}_i$ may be correlated and the covariates may be either fixed or changing at every cluster level. Given n clusters $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ for $i \in [1, n]$, the goal of bipartite ranking is to find a real-value ranking function that maps any observation $\boldsymbol{X}_{it}$ to real number. The ranking function can be used to induce ordering over the observations $\boldsymbol{X}_{it}$.

The logistic-GEE model solves the bipartite ranking task, since it is essentially a ranking function for clustered data. It describes the relationship between the covariates and outcome variables with the following equation:

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \boldsymbol{\beta}\boldsymbol{X_{it}'} \ , i = 1, ..., n,\ t = 1, ..., m_i, \tag{1}$$

where $\pi_{it} = \mathrm{E}(Y_{it}|\boldsymbol{X}_{it})$, β_0 is the population averaged intercept term and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is the vector of population averaged (or marginal) coefficients.

The logistic-GEE model can be obtained by estimating the unknown regression coefficient vector $\boldsymbol{\gamma} = (\beta_0, \boldsymbol{\beta})$. Estimating the coefficients can be done by solving the following generalized estimating equations [10]:

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i}{\partial \beta_h}\right)' \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\pi}_i) = 0,\ \ h = 0, ..., p, \tag{2}$$

where, for $i = 1, ..., n$, $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{im_i})'$, $\boldsymbol{V}_i = \boldsymbol{A}_i{}^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i{}^{1/2}$ is the working covariance matrix for $\boldsymbol{Y_i}$, $\boldsymbol{A_i}$, is a diagonal matrix diag$[\pi_{i1}(1-\pi_{i1}), ..., \pi_{im_i}(1-\pi_{im_i})]$, $\boldsymbol{\alpha}$ is an $m \times 1$ vector of unknown parameters, associated with the correlation between outcomes Y_{it} and Y_{is} of cluster i, $m = \max(m_1, ..., m_n)$, and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the working correlation matrix for $\boldsymbol{Y_i}$.

We note that the working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$, parameterized by $\boldsymbol{\alpha}$, might be defined in different ways depending on the nature of correlation between outcomes Y_{it} and Y_{is}. Zeger and Liang [16] proposed a method for estimating the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ in Eq. (3). The method operates by minimizing the weighted sum of squared residuals using IRLS, described in [11].

3 Ranking Accuracy for Models Based on Clustered Data

In this section we introduce the ranking accuracy for models based on clustered data (RAMCD). RAMCD is formally defined in Subsect. 3.1. The algorithm for computing RAMCD is provided in Subsect. 3.2 together with a complexity analysis. Subsect. 3.3 explains how the algorithm can be used for estimating RAMCD as a criterion of the model's goodness-of-fit and as a criterion for the model's predictability.

3.1 Definition

According Eq. (1) any logistic GEE model is essentially a scoring classifier. It outputs a score, a probability π_{it}, for any observation $\boldsymbol{X}_{it}$. Given a test data of n number of clusters $\langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$, the probabilities π_{it} induce an ordering over the observations from the clusters. To judge the quality of the probabilities π_{it}, we judge the quality of the ordering, they induce, and compare that ordering with the binary outcome variables Y_{it}. The standard measure for such a comparison is ranking accuracy [1]. It is defined as a probability that a randomly selected positive observation is ranked higher than randomly selected negative observation. However, as it might be seen from the definition, the ranking accuracy does not take into account the within-cluster correlation that might be present and thus it is not valid for clustered data. This calls for a new special ranking accuracy applicable for models based on clustered data.

We introduce the ranking accuracy for models based on clustered data (RAMCD) by analogy. Consider a set of observations $\boldsymbol{X}_{it}$ where each cluster is present with exactly one observation. The number of such sets equals $\sum_{i=1}^{n} m_i$. The probabilities π_{it} induce an ordering for each of these sets. To compare these orderings with the binary outcome variables Y_{it} we introduce RAMCD. RAMCD is defined as a probability that a randomly selected positive observation is ranked higher than randomly selected negative observation *from another cluster*. By the definition RAMCD employs the within-cluster correlation in the data used and focuses on estimating the generalization performance of the logistic GEE models when ranking uncorrelated observations.

RAMCD is easy to interpret, since it is a probability (i.e., it ranges between 0 and 1). The value of 1.0 indicates that the orderings imposed correspond completely to the binary outcome variables Y_{it} in the clustered data, and the value of 0.0 shows that the orderings are reversed to that with value of 1.0. The value of 0.5 is the worst case. It indicates bad orderings that do not correspond at all to the outcome variables. However, we note that RAMCD of 0.5 does not always imply a random logistic GEE model (e.g., when the data is class-imbalanced).

Below we introduce the exact formula for RAMCD. We first introduce statistics imposed by the binary outcome variables Y_{it}. Following the RAMCD definition we determine for any positive observation $\boldsymbol{X}_{it}$ the number P_{it} of negative observations from other clusters:

$$P_{it} = \sum_{j=1, j \neq i}^{n} \sum_{t=1}^{m_j} I\{Y_{jt} = 0\} \tag{3}$$

where I is the indicator function. The number P_{it} can be interpreted as the number of pairs that consist of positive observation $\boldsymbol{X}_{it}$ and negative observation from any other cluster. It is the same for any positive observation in cluster i. This implies that the number P_i of pairs for all the positive observations in cluster i is equal to:

$$P_i = \sum_{t=1}^{m_i} P_{it} I\{Y_{it} = 1\} \tag{4}$$

and the total number P of pairs of observations over all the clusters imposed by the binary outcome variables Y_{it} is equal to:

$$P = \sum_{i=1}^{n} P_i \tag{5}$$

Once the statistics imposed by the binary outcome variables Y_{it} have been defined, we introduce statistics for comparing the orderings imposed by probabilities π_{it}. We assume that for any observation $\boldsymbol{X}_{it}$ we have a probability estimate

π_{it} provided by a logistic GEE model. We rank the observations $\boldsymbol{X}_{it}$ according to π_{it}. To judge whether a particular positive observation $\boldsymbol{X}_{it}$ from cluster i is ranked properly we compute the number CP_{it} of correct pairs produced by the ranking through combining with all negative observations $\boldsymbol{X}_{jt}$ from all other clusters j such that $j \neq i$. The number CP_{it} is given by:

$$CP_{it} = \sum_{j=1, j \neq i}^{n} \sum_{t_j=1}^{m_j} (I\{\pi_{it} > \pi_{jt_j}\} + \frac{1}{2} I\{\pi_{it} = \pi_{jt_j}\}) I\{Y_{jt_j} = 0\} \quad (6)$$

Number CP_{it} does not stay the same for each positive observation in cluster i. Hence, the number CP_i of all correct pairs produced by combining all the positive observations $\boldsymbol{X}_{it}$ from cluster i with all the negative observations $\boldsymbol{X}_{jt}$ over all the clusters j given that $j \neq i$ is equal to:

$$CP_i = \sum_{t=1}^{m_i} CP_{it} I\{Y_{it} = 1\} \quad (7)$$

and the number CP of all the correct pairs produced by the ranking is:

$$CP = \sum_{i=1}^{n} CP_i \quad (8)$$

Thus, formally our RAMCD with respect to the ranking produced is defined equal to:

$$RAMCD = \frac{CP}{P} \quad (9)$$

3.2 Algorithm

Below in Fig. 1 we provide an algorithm for RAMCD. Given data with n number of clusters $\langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$, and a vector $\boldsymbol{\pi}_i$ of observation probabilities π_{it} for each cluster $\langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$, the algorithm computes RAMCD induced by the observation probabilities π_{it} w.r.t. outcome variable Y_{it}. The main steps are as follows. First, the algorithm computes the statistics imposed by the binary outcome variables Y_{it}: it computes number P_i for each cluster i (see formula (4)) and total number P (see formula (5)). Then, the algorithm computes statistics necessary for comparing the orderings imposed by probabilities π_{it}. For that purpose the observations $\boldsymbol{X}_{it}$ over all the clusters are sorted according to π_{it} in decreasing order of magnitude into list L_π. The algorithm scans the sorted list L_π to compute numbers CP_{it}, CP_i, and CP (initially set equal to 0). For list scanning it keeps a counter C_i for all the clusters $i \in [1, n]$ that represents the number of all correct pairs that start with a positive observation from cluster i and end with a negative observation from another cluster given that both observations have not been visited in list L_π. Therefore, C_i is initialized equal to $\frac{P_i}{m_i}$ which is the number of pairs derived by combining a positive observation from cluster i with all possible negative observations from other clusters.

```
Algorithm RAMCD
Input:    $n$ number of clusters $\langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$,
    Vector $\boldsymbol{\pi}_i$ of observation probabilities $\pi_{it}$ for each cluster $\langle \boldsymbol{X}_i, \boldsymbol{Y}_i \rangle$.
Output:
    RAMCD induced by the observation probabilities $\pi_{it}$ with respect to outcome variable $Y_{it}$.
     for $i := 1$ to $n$ do
          Compute $P_i$ using formula (4);
     Compute $P$ using formula (5);
     Sort all the observations $\boldsymbol{X}_{it}$ according to $\pi_{it}$ in decreasing order of magnitude into list $L_\pi$.
     for $i := 1$ to $n$ do
          $CP_i = 0$;
          $C_i = \frac{P_i}{m_i}$;
     $CP = 0$;
     for each observation $\boldsymbol{X}_{it}$ in $L_\pi$ do
          if $\boldsymbol{Y}_{it} = 0$ then
               for $j$ = 1 to $n$ do
                    if $j \neq i$ then
                         $C_j = C_j - 1$;
          else
               $CP_{it} = C_i$;
               $CP_i = CP_i + CP_{it}$;
     for $i := 1$ to $n$ do
               $CP = CP + CP_i$;
     return $\frac{CP}{P}$.
```

Fig. 1. Algorithm for computing ranking accuracy for models based on clustered data.

After the initialization the algorithm sequentially visits the observations $\boldsymbol{X}_{it}$ in the sorted list L_π. For each observation $\boldsymbol{X}_{it}$ the actions taken depends on the output variable Y_{it}. If the observation is negative ($Y_{it} = 0$), then the algorithm decrements the counter C_j for each cluster j different from cluster i. This is to indicate that all the positive observations $\boldsymbol{X}_{jt}$ with probability π_{jt} that is lower than π_{it} cannot form a correct pair with observation $\boldsymbol{X}_{it}$ according to the ordering imposed on L_π. If the observation is positive ($Y_{it} = 1$), then the algorithm assigns the counter value C_i to the number CP_{it} and then this number is added to number CP_i according to formula (7). Once all the numbers CP_i have been computed, the algorithm computes number CP (see formula (8)) and then outputs the RAMCD (see formula (9)).

The algorithm for RAMCD is computationally efficient. Its space complexity is $O(nm)$, where n is the number of clusters and m is the size of the clusters. This complexity is due to the sorted list L_π that has to be explicitly maintained by the algorithm. The time complexity is $O(nmlog_2(nm))$ and it coincides with the time complexity of the sorting algorithm[1]. We note that the time complexity of scanning list L_π is linear in the size of the list (nm) and that is why it does not influence the asymptotic time complexity.

[1] We assume the usage of efficient sorting algorithms like merge sort.

3.3 Goodness of Fit and Predictability

The ranking accuracy for models based on clustered data (RAMCD) can be used as a criterion of model's goodness-of-fit and as a criterion for model's predictability. If a logistic GEE model has been trained and tested on the same data, then RAMCD is a goodness-of-fit criterion. In this case RAMCD estimates how the logistic GEE model fits the data when only uncorrelated observations are taken into account.

If a logistic GEE model has been tested using k-fold cross validation on the data, then RAMCD is a predictability criterion. However, the randomization part of the cross validation has to be controlled such that observations within any cluster are being selected for only one folder. In this way we do not introduce additional bias when computing probabilities π_i due to the within-cluster correlation. This guarantees that the algorithm estimates RAMCD that indicates the predictability of the GEE model beyond training data when only uncorrelated observations are taken into account.

4 Experiments

In this section we present the experiments with RAMCD and QIC on simulated data and biomarker data. The experiments are employed to compare these two criteria.

4.1 Experiments with Simulated Data

This subsection presents two experiments with RAMCD of logistic GEE models on a simulated data. The simulated data is described by 30 time-dependent covariates (X_1, $X_2, \ldots, X_{30}$). It contains 500 clusters with maximum sizes of $m = 10$ and autoregressive working correlation structure of order 1 with correlation of 0.25.

The first experiment is in the context of the goodness-of-fit test. We compare RAMCD and QIC in a function of the GEE model complexity. For that purpose we add the covariates X_1 to X_{30} one by one into the GEE model and each time plot the RAMCD and QIC in Fig. 2. The Figure shows that the RAMCD and QIC follow similar trends in a function of the model complexity. There exists however some fluctuations of RAMCD when it is close to 0.5. In these cases GEE models are under-fitted and exhibit random performance which is not captured by QIC.

The second experiment is in the context of model selection: we employ RAMCD for forward feature selection when it is used as a goodness-of-fit criterion and when it is used as a predictability criterion. In the first case RAMCD is estimated on the simulated data and it is denoted as RAMCD. In the second case RAMCD is estimated using one-cluster-out cross validation on the simulated data and it is denoted as RAMCD-CV. In both cases we compare the results of the model selection with those obtained by QIC.

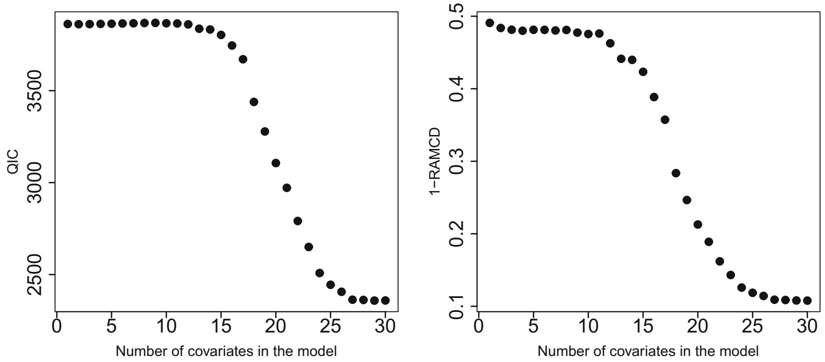

Fig. 2. QIC and RAMCD as functions of GEE model complexity.

The process of forward feature selection is sequential; i.e., the covariates are added one by one. It is guided by a hill-climbing search which for RAMCD (QIC) adds that covariate that maximizes (minimizes) the RAMCD (QIC) of the resulted GEE model. The process stops when further improvement is not possible.

Table 1. Forward feature selection for logistic-GEE model using RAMCD, RAMCD-CV, and QIC. Each box represents the selected covariate and the value of selection criterion (RAMCD, RAMCD-CV, or QIC). The bold variables are those that are not selected.

Step	RAMCD	RAMCD-CV	QIC	Step	RAMCD	RAMCD-CV	QIC
1	X_{18} (0.590242)	X_{18} (0.581824)	X_{18} (3791.772)	16	X_{25} (0.877261)	X_{25} (0.873539)	X_{25} (2464.293)
2	X_{17} (0.621768)	X_{17} (0.614062)	X_{17} (3732.281)	17	X_{27} (0.881865)	X_{27} (0.878116)	X_{27} (2425.637)
3	X_{16} (0.649819)	X_{16} (0.643470)	X_{16} (3658.207)	18	X_{26} (0.886264)	X_{26} (0.882337)	X_{26} (2387.744)
4	X_{19} (0.672773)	X_{19} (0.666522)	X_{19} (3587.769)	19	X_9 (0.889523)	X_9 (0.885488)	X_9 (2360.825)
5	X_{15} (0.696743)	X_{15} (0.690881)	X_{15} (3502.061)	20	X_{29} (0.890200)	X_{29} (0.885943)	X_{29} (2357.297)
6	X_{20} (0.718667)	X_{20} (0.713321)	X_{20} (3422.789)	21	X_2 (0.890596)	X_6 (0.886166)	X_6 (2355.108)
7	X_{13} (0.743073)	X_{13} (0.737916)	X_{13} (3325.778)	22	X_6 (0.890973)	X_{28} (0.886391)	X_{28} (2353.878)
8	X_{14} (0.759801)	X_{14} (0.754877)	X_{14} (3241.496)	23	X_{28} (0.891342)	X_2 (0.886536)	X_2 (2352.896)
9	X_{22} (0.778013)	X_{22} (0.773413)	X_{22} (3141.077)	24	X_5 (0.891657)	X_5 (0.886586)	X_5 (2352.300)
10	X_{21} (0.798329)	X_{21} (0.794059)	X_{21} (3028.665)	25	X_{30} (0.891823)	$\boldsymbol{X_4}$ (0.886572)	$\boldsymbol{X_7}$ (2352.430)
11	X_{12} (0.818305)	X_{12} (0.814269)	X_{12} (2905.000)	26	X_4 (0.891994)	$\boldsymbol{X_{30}}$ (0.886520)	$\boldsymbol{X_4}$ (2352.941)
12	X_{23} (0.834424)	X_{23} (0.830658)	X_{23} (2801.273)	27	X_7 (0.892142)	$\boldsymbol{X_7}$ (0.886462)	$\boldsymbol{X_{30}}$ (2353.607)
13	X_{24} (0.848698)	X_{24} (0.844904)	X_{24} (2698.514)	28	X_1 (0.892164)	$\boldsymbol{X_1}$ (0.886271)	$\boldsymbol{X_8}$ (2355.205)
14	X_{11} (0.861617)	X_{11} (0.858049)	X_{11} (2600.898)	29	X_8 (0.892181)	$\boldsymbol{X_8}$ (0.886086)	$\boldsymbol{X_1}$ (2356.971)
15	X_{10} (0.870470)	X_{10} (0.866909)	X_{10} (2527.606)	30	$\boldsymbol{X_3}$ (0.892134)	$\boldsymbol{X_3}$ (0.885836)	$\boldsymbol{X_3}$ (2359.079)

The results of model selection for RAMCD, RAMCD-CV, and QIC are provided in Table 1. The Table shows that RAMCD-CV and QIC are rather consistent: they lead to the same ordered set of covariates on the simulated data when the process of feature selection stops. This means that RAMCD-CV and

QIC result in the same GEE model. However, if we continue to add covariates after the stopping condition, the RAMCD-CV and QIC become less consistent. As expected, the values of RAMCDs are higher than those of RAMCD-CVs at each step which results in a bigger set of selected covariates. In this context we note that RAMCD is less consistent with RAMCD-CV and QIC than those two measures together.

4.2 Experiments with Biomarkers' Data

This subsection presents a model-selection (biomarker selection) process guided by RAMCD-CV on the data from the TIME-CHF study [5]. The TIME-CHF study (The Intensified versus standard Medical therapy in Elderly patients with Congestive Heart Failure) includes 499 patients aged 60 years or older, with left ventricular ejection fraction (LVEF) $< 45\,\%$ and NYHA II or more, from 15 centers in Switzerland and Germany. Patients were followed for 6 prespecified visits after baseline, 1^{st}, 3^{rd}, 6^{th}, 12^{th} and 18^{th} month. Six biomarkers, PREA (prealbumin), SST2 (soluble ST2), IL6 (Interleukin-6), hsCRP (high sensitivity C-reactive protein), GDF15 (growth differentiation factor 15), SFLT (soluble fms-like tyrosine kinase-1,) and BPsyst (Systolic blood pressure) and LVEF were measured at every visit and dosages of a heart failure (HF) drug Loop (Loop diuretics per se) were available on a daily basis. Patients were followed up for 19 months and the outcome variable for i^{th} patient at month t, Y_{it}, $i = 1, ..., 499$, $t = 1, ..., 19$, takes the value of one if the patient experienced HF hospitalization or death at the given month, otherwise zero. In this setup, more weight is given to the outcome death (weight 2 for death and 1 for the other observations).

The medication covariate Loop is down-sampled to monthly values by taking the average drug dosage during the previous month. Since the biomarkers, BPsyst and LVEF have been recorded just in six visits; obviously for these six measurements, the covariates gets the exact value, and between these six visits we used last observation carried forward method (LOCF) and put the value of the covariates of the previous visit. There exist eight fixed covariates that measured only at the baseline; Age, Gender (1 = male, 0 = female), Coronary artery disease (CAD), Kidney-disease, Diabetes, Anemia, Charlsonscore (Charlson comorbidity score) and Rales, where CAD, Kidney-disease and Diabetes are binary variable that indicates whether the patients are suffering from these diseases or not (1 = yes, 0 = no) and Rales (1 = abnormal lung sounds, 0 = normal lung sounds).

The goal of the study is to select the best subset of covariates (biomarkers) to explain the variation of the probability of HF hospitalization and death. To this end, we apply a forward feature selection process using the proposed RAMCD-CV and QIC as model-selection criteria to find the best GEE model. Table 2 shows the selected covariates for the GEE model at every step of forward selection process based on RAMCD-CV and QIC. Both criteria lead to the same selected subset of covariates (GDF15, SST2, CAD, Loop, hsCRP,

Table 2. Selected covariates at each step of forward selection method using RAMCD-CV and QIC as model selection criteria.

RAMCD-CV-Covariates	RAMCD-CV	p-values	QIC-Covariates	QIC	p-values
Intercept	0.5	0.000000	Intercept	1917.96	0.000000
GDF15	0.738816	0.000002	GDF15	1753.53	0.000002
SST2	0.767409	0.000001	SST2	1691.76	0.000001
CAD	0.775987	0.003211	Loop	1660.43	0.000060
Loop	0.784891	0.000060	CAD	1649.28	0.003211
hsCRP	0.790796	0.031393	BPsyst	1645.50	0.018600
Age	0.794360	0.020222	Age	1639.86	0.020222
BPsyst	0.795732	0.018600	hsCRP	1635.74	0.031393
Rales	0.796831	0.073626	Rales	1635.28	0.073626

Age, BPsyst and Rales), however, the selected subsets were obtained in different orders for each criterion. The estimated coefficients and corresponding *p*-values of selected covariates, when using RAMCD-CV as a model-selection, are presented in Table 2.

5 Conclusion

In this paper we proposed RAMCD as a new measure for estimating the generalization performance of logistic GEE models. RAMCD was defined as a probability that a randomly selected positive observation is ranked higher than randomly selected negative observation *from another cluster*. We showed that RAMCD focuses on estimating the generalization performance of the logistic GEE models when ranking uncorrelated observations. We proposed a computationally efficient algorithm for RAMCD and showed that it can be applied for two cases: (1) when we estimate RAMCD as a goodness-of-fit criterion and (2) when we estimate RAMCD as a predictability criterion. The algorithm was experimentally tested on clustered data from a simulation study and a biomarkers' study. The experiments showed that RAMCD is consistent with the QIC criterion.

We compared RAMCD with the standard goodness-of-fit test statistics and criteria for logistic GEE models: we observed that RAMCD does not suffer from any of their problems. The main reasons are that: (1) RAMCD is a probability that is easy to interpreted; (2) RAMCD does not impose any restriction on the models being evaluated; (3) RAMCD is not biased for binary imbalanced data (since it indicates class separation); and (4) RAMCD can be used as a goodness-of-fit criterion and a predictability criterion.

Finally, we note although RAMCD has been initially designed for the logistic GEE models, it is applicable to any model for bipartite ranking based on clustered data. This is due to the fact that RAMCD employs model's probabilities and data labels; i.e., it does not use any internal information from the model being tested. Thus, we conclude that RAMCD is a general measure for models for bipartite ranking based on clustered data.

References

1. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization bounds for the area under the ROC curve. J. Mach. Learn. Res. **6**, 393–425 (2005)
2. Ahsan, H., Chen, Y., Parvez, F., Zablotska, L., Argos, M., Hussain, I., Momotaj, H., Levy, D., Cheng, Z., Slavkovich, V., Van Geen, A.: Arsenic exposure from drinking water and risk of premalignant skin lesions in Bangladesh: baseline results from the health effects of arsenic longitudinal study. Am. J. Epidemiol. **163**(12), 1138–1148 (2006)
3. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
4. Barnhart, H.X., Williamson, J.M.: Goodness-of-fit tests for GEE modeling with binary responses. Biometrics **54**(2), 720–729 (1998)
5. Brunner–La Rocca, H.P., Buser, P.T., Schindler, R., Bernheim, A., Rickenbacher, P., Pfisterer, M., TIME-CHF-Investigators: Management of elderly patients with congestive heart failuredesign of the Trial of Intensified versus standard Medical therapy in Elderly patients with Congestive Heart Failure (TIME-CHF). Am. Heart J. **151**(5), 949–955 (2006)
6. Evans, S.R., Hosmer Jr., D.W.: Goodness of fit tests for logistic GEE models: simulation results. Commun. Stat. Simul. Comput. **33**(1), 247–258 (2004)
7. Evans, S., Li, L.: A comparison of goodness of fit tests for the logistic GEE model. Stat. Med. **24**(8), 1245–1261 (2005)
8. Hanley, J.A., Negassa, A., Forrester, J.E.: Statistical analysis of correlated data using generalized estimating equations: an orientation. Biometrics **157**(4), 364–375 (2003)
9. Lafata, J.E., Pladevall, M., Divine, G., Ayoub, M., Philbin, E.F.: Are there race/ethnicity differences in outpatient congestive heart failure management, hospital use, and mortality among an insured population? Med. Care **42**(7), 680–689 (2004)
10. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**(1), 13–22 (1986)
11. McCullagh, P.: Quasi-likelihood functions. Ann. Stat. **11**(1), 59–67 (1983)
12. Pan, W.: Akaike's information criterion in generalized estimating equations. Biometrics **57**(1), 120–125 (2001)
13. Pulkstenis, E., Robinson, T.J.: Two goodness-of-fit tests for logistic regression models with continuous covariates. Stat. Med. **21**(1), 79–93 (2002)
14. Titler, M.G., Jensen, G.A., Dochterman, J.M., Xie, X.J., Kanak, M., Reed, D., Shever, L.L.: Cost of hospital care for older adults with heart failure: medical, pharmaceutical, and nursing costs. Health Serv. Res. **43**(2), 635–655 (2008)
15. Williamson, J.M., Lin, H.M., Barnhart, H.X.: A classification statistic for GEE categorical response models. J. Data Sci. **1**, 149–165 (2003)
16. Zeger, S.L., Liang, K.Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics **42**(1), 121–130 (1986)