Preliminary Report

Comparison of Resampling Methods for Bias-Reduced Estimation of Prediction Error: A Simulation Study Based on Real Datasets from Biomarker Discovery Studies

Keiji Kakumoto and Yoshiyuki Tochizawa
Research Institute of Tokushima, Otsuka Pharmaceutical Co., Ltd.
e-mail: Kakumoto.Keiji@otsuka.jp

Stepwise logistic regression is the traditional and most commonly used method for identifying biomarkers and evaluating the magnitude of their effects based on clinical data. Here, we evaluated the performance of the resampling methods leave-one-out cross-validation, 10-fold cross-validation, bootstrap, and .632+ bootstrap in terms of internal validation of prediction analysis using stepwise logistic regression. We conducted simulation studies to compare the ability of these methods to estimate prediction accuracy based on simulation settings (including statistical models) derived from two real biomarker discovery studies (Ogata et al., *Leukemia Research* 2012; 36: 1229–1236; Yoshimi et al., *Molecular Psychiatry* 2016; 21: 1504–1510). The simulation results revealed that leave-one-out cross-validation, 10-fold cross-validation, and .632+ bootstrap were comparable in terms of the root mean square error. We therefore recommend the application of these methods to similar biomarker discovery studies that involve approximately ten biomarkers with or without binary biomarkers (such as sex) and various degrees of correlation between the biomarkers.

*Key words:* 10-fold cross-validation; .632+ bootstrap; Bootstrap; Error rate; Leave-one-out cross-validation; Stepwise logistic regression.

1. Introduction

One of the key issues in the development of effective diagnostics and therapeutics is identifying biomarkers related to various disease statuses. Biomarkers are useful for the prevention, early detection, and treatment of many diseases. Recent biotechnological advances in the fields of genomics, proteomics, metabolomics, and flow cytometry have greatly accelerated the discovery of biomarkers.

Resampling methods, such as the cross-validation (CV) method, are used to perform internal validation of prediction analysis; they are useful for estimating prediction accuracy when applying a developed model to predict a particular disease status in future patients (Azuaje,
The established and useful resampling methods include leave-one-out CV (LOOCV), 10-fold CV, bootstrap, .632+ bootstrap (BT632+), and others (Hastie, Tibshirani, and Friedman, 2009).

Molinaro, Simon, and Pfeiffer (2005) compared the LOOCV, 10-fold CV, and BT632+ methods using simulated and real datasets in microarray experiments. They reported that although estimation bias decreased with increasing training sample size, the bias was generally low for the LOOCV method and was nearly uniformly low for the 10-fold CV method. The BT632+ method exhibited the lowest biases and low signal-to-noise ratios in moderate sample sizes; however, in other cases it demonstrated estimates with much higher biases. Hastie et al. (2009) also evaluated resampling methods based on error rates and reported similar results.

Davison and Hall (1992) compared the LOOCV and bootstrap methods in terms of asymptotic bias. LOOCV is less biased when the shift size is $n^{-1/2}$; as the size of the training sample, $n$, tends to infinity, the bias stays in the order $n^{-1}$. In contrast, the bias of the bootstrap method is exacerbated to the order $n^{-1/2}$. Abdi and Williams (2010) also showed that the error rate decreased for larger $n$.

In previous studies, we applied LOOCV for the analysis of specific biomarker discovery studies (Ogata et al., 2012; Yoshimi et al., 2016) in which the prediction variables are binary outcomes (see Sections 3.1 and 3.2 for further information regarding these studies). However, the relative performance of the resampling methods, including LOOCV, was not completely clear for the datasets from these studies (involving approximately 10 biomarkers with or without binary biomarkers and various correlations between the biomarkers). This is because the previous comparison studies assumed certain specific situations, such as those in the microarray experiments, for method comparison; moreover, they did not necessarily consider the implementation of the commonly used stepwise logistic regression.

In this study, we conducted a simulation study to evaluate the performance of the resampling methods in stepwise logistic regression analysis based on simulation settings (including statistical models) derived from two real biomarker discovery studies (Ogata et al., 2012; Yoshimi et al., 2016). In order to incorporate unequal prior relative frequencies between binary outcomes (e.g., disease and non-disease) we considered various cut-off values for the posterior probability in the fitted logistic models when classifying the samples.

An outline of the resampling methods is provided in Section 2, the simulation settings derived from the two real biomarker studies are described in Section 3, and the simulation results are detailed in Section 4. A concluding discussion is provided in Section 5.
2. Brief Outline of the Resampling Methods

2.1 Notation and Assumption

Suppose we study \( n \) subjects whose observed values are \( \{x_1 = (t_1, y_1), x_2 = (t_2, y_2), \ldots, x_n = (t_n, y_n)\} \), where \( y_i \) is a binary variable with 1 indicating that the \( i \)-th subject is ill, 0 indicating that the subject is not ill, and \( t_i \) is a \( r \)-dimensional row vector of the covariate describing various medical measurements of the \( i \)-th subject. These \( n \) subjects are referred to as the training sample in this study.

Let us assume that \( y_1, y_2, \ldots, y_n \) are independent Bernoulli variables with the probability \( \theta(t_i) = \Pr\{y_i = 1 \mid t_i\} \) given below:

\[
\logit(\theta(t_i)) = \beta_0 + t_i^\top \beta
\]

In this equation, \( \beta_0 \) is an unknown intercept parameter, \( \{t_1, t_2, \ldots, t_n\} \) are the observed values of independently and identically distributed random variables with an \( r \)-dimensional distribution (say \( F \)), and \( \beta \) is an unknown \( r \)-dimensional column vector of the covariate coefficient.

Let us predict that a subject with the covariate value \( t_0 \) will become ill if \( q(t_0) > c \), where \( q(t_0) \) is referred to as the prediction function and \( c \) is a cut-off value determined in advance. Note that the ROC curve is obtained by changing the cut-off value. The ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Our task for data analysis is to construct the best prediction function \( q(t_0) \) with the same form as \( \theta(t_0) = \Pr\{y_0 = 1 \mid t_0\} \) for a future variable \( y_0 \) corresponding to the value \( t_0 \) distributed with the same \( F \) distribution as \( t_1 \). For this task, we naturally estimate \( \beta_0 \) and \( \beta \) using an appropriate method such as the maximum likelihood method.

One important issue to be examined in detail when we apply our strategy, is how to estimate the expectation of the true error rate, say \( ER(t_0) \), for the realized prediction function. Additionally, the expectation of \( ER(t_0) \) with new random sampling is assumed as \( EER(t_0) \). \( ER(t_0) \) and \( EER(t_0) \) are described below (Molinaro et al., 2005):

\[
ER(t_0) = E_{x_0}[Q(y_0, q(t_0))] = \int Q(y_0, q(t_0))dx_0(t_0, y_0)
\]

\[
EER(t_0) = E_q[ER(t_0)] = \int ER(t_0)dq\{x_1, \ldots, x_n\},
\]

where \( q \) depends on sampling \( \{x_1 = (t_1, y_1), x_2 = (t_2, y_2), \ldots, x_n = (t_n, y_n)\} \). Therefore, \( ER(t_0) \) has variance following sampling and is not constant. \( EER(t_0) \) is an unbiased estimate of \( ER(t_0) \) and is constant.

Let the estimate of \( (\beta_0, \beta) \) be \( (b_0, b) \) and the realized prediction function be \( q(t) = b_0 + tb \). Then a natural estimate of \( ER(t_0) \) is \( R_{app} \), which signifies the apparent error rate (detailed below), because \( t_0 \) is independently distributed with the same distribution as \( t_1, t_2, \ldots, t_n \).
\[ R_{\text{app}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, q(t_i)), \]

where \( Q \) is an index function with the value 1 when \( y_i = 0 \) and \( q(t_i) \) is greater than \( c \) or when \( y_i = 1 \) and \( q(t_i) \) is less than or equal to \( c \), and has the value 0 otherwise.

However, the estimate \( R_{\text{app}} \) is known to be biased compared with \( ER(t_0) \) because of over-fitting and should therefore be adjusted to reduce the bias.

### 2.1.1 Leave-one-out cross-validation (LOOCV)

LOOCV is available for adjusting the estimate \( R_{\text{app}} \). Let the prediction function \( q(i)(t) \) constructed based on \( \{x_2, x_3, \ldots, x_n\} \), which is the \( x_1 \)-deleted training sample, similarly to \( q(t) \). Define \( q(i)(t) (i = 2, \ldots, n) \) likewise. \( R_{\text{loo cov}} \) (defined below), referred to as the cross estimate, is an estimate of the true error rate \( ER(t_0) \).

\[ R_{\text{loo cov}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, q(i)(t_i)) \]

### 2.1.2 Bootstrap

Bootstrap is available for adjusting the estimate \( R_{\text{app}} \). Let \( x_i^* \), \( (i = 1, 2, \ldots, n) \) be a sample randomly chosen from the training sample with replacement. The set \( \{x_1^*, x_2^*, \ldots, x_n^*\} \) is referred to as the bootstrap sample from the training sample. Let \( q^*(1)(t) \) be the prediction function constructed similarly to \( q(t) \) based on the bootstrap sample. \( R_{\text{boot}} \) (defined below), referred to as the cross estimate, is also an estimate of the true error rate \( ER(t_0) \).

\[ R_{\text{boot}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, q^*(1)(t_i)) \]

By repeating the sampling experiment \( B \) times to produce the bootstrap sample, we obtain \( B \)-tuple estimates \( \{R_{\text{boot}}^{(1)}(t_0), R_{\text{boot}}^{(2)}(t_0), \ldots, R_{\text{boot}}^{(B)}(t_0)\} \). \( R_{\text{boot}} \) (defined below), referred to as the bootstrap estimate, is also an estimate of the true error rate \( ER(t_0) \).

\[ R_{\text{boot}} = \frac{1}{B} \sum_{i=1}^{B} R_{\text{boot}}^{(i)} \]

\( R_{\text{boot}} \) is a better estimate than \( R_{\text{boot}}^{(i)} \) based on the individual bootstrap sample.

### 2.1.3 10-fold cross-validation (10-fold CV)

10-fold CV is available for adjusting the estimate \( R_{\text{app}} \). For this method, the data set is split into smaller sets. Let the prediction function \( q_k(i)(t) \) constructed based on \( 10 - 1 \) folds; the one-fold-deleted training sample that does not contain \( x_1 \), similarly to \( q(t) \). Define \( q_k(i)(t) (i = 2, \ldots, n) \) likewise. \( R_{\text{ten}} \) (defined below), referred to as the cross estimate, is an estimate of the true error rate \( ER(t_0) \).

\[ R_{\text{ten}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, q_k(i)(t_i)) \]
2.1.4 .632+ bootstrap (BT632+)

BT632+ is available for adjusting the estimate $R_{\text{app}}$. BT632+ was proposed by Efron and Tibshirani (1997) for reducing the upward bias of the Leave-one-out bootstrap (LOOBT). Let $x_{k*}$, $(k = 1, 2, \ldots, n)$ be a sample randomly chosen from the $x_i$-deleted $(i = 1, 2, \ldots, n)$ training sample with replacement by $j$-th $(j = 1, 2, \ldots, B)$ repeating. The set $\{x_{1*}, x_{2*}, \ldots, x_{n*}\}$ is referred to as the LOOBT sample from the training sample. Let $q_{(i)}^{*{(j)}}(t)$ be the prediction function constructed similarly to $q(t)$ based on the LOOBT sample, in which the training sample is the $j$-th bootstrap sample from the $x_i$-deleted dataset. $R_{\text{loobt}}$ (defined below), referred to as the LOOBT estimate, is also an estimate of the true error rate $ER(t_0)$.

$$R_{\text{loobt}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{B} \sum_{j=1}^{B} Q(y_i, q_{(i)}^{*{(j)}}(t_i))$$

Note,

$$R'_{\text{loobt}} = \min(R_{\text{loobt}}, \hat{\gamma}), \quad \hat{\gamma} = \sum_{i} \sum_{j} \frac{Q[y_i, q(t)]}{n^2}$$

$R_{\text{bt632+}}$ (defined below), referred to as the BT632+ estimate, is also an estimate of the true error rate $ER(t_0)$.

$$R_{\text{bt632+}} = (1 - w)R_{\text{app}} + wR'_{\text{loobt}}, \quad \text{where } w \text{ is between 0.632 and 1, constitutes the resubstitution estimate.}$$

$w$ has the form

$$w = \frac{0.632}{1 - 0.368R'}$$

$R'$ is given by

$$R = \frac{R'_{\text{loobt}} - R_{\text{app}}}{\hat{\gamma} - R_{\text{app}}}, \quad R' = \begin{cases} R(\cdot; R_{\text{loobt}}, \hat{\gamma} > R_{\text{app}}) \\ 0(\cdot; \text{thoersise}) \end{cases}$$

Taking gives the .632 bootstrap (BT632) originally proposed by Efron (1983). When the resubstitution error is zero, the BT632+ estimate becomes BT632, which results in systematic downward bias when there are no class differences (Breiman et al., 1984; Efron and Tibshirani, 1997). BT632+ aims to circumvent this problem by increasing the weight $w$ with respect to the growing level of overfitting.

2.2 Criterion for comparison

Because the objective of this study was to evaluate the use of $R_{\text{app}}, R_{\text{loocv}}, R_{\text{boot}}, R_{\text{ten}}$, and $R_{\text{bt632+}}$ as the bias-reduced estimate of $ER(t_0)$, it is logical to use the root mean squared error (RMSE) detailed below as the criterion for comparison:

$$RMSE = \sqrt{E[(R_M - ER(t_0))^2]},$$

where $R_M$ represents the estimate of each method, that is,
The expectation $E$ is taken over all distributions of the bootstrap sampling, $F$ on the covariate $t$, and Bernoulli variable $y$.

The RMSE can be separated into average bias and average variance components. $\text{BIAS}$ represents $\vert E(R_M) - E(ER(t_0)) \vert$ and $\text{VAR}$ represents $V(R_M - ER(t_0))$.

$$RMSE = \sqrt{(E(R_M) - E(ER(t_0)))^2 + V(R_M - ER(t_0))}$$

The significance level used in this study is 0.05.

3. Simulation

In order to simulate conditions in specific biomarker discovery studies, the simulation was based on previously published cases (Ogata et al., 2012; Yoshimi et al., 2016).

All analyses were performed using SAS® R9.3 (SAS Institute Inc., Cary, NC, USA) and R version 3.2.3 (R Core Team, 2015).

3.1 Myelodysplastic syndrome datasets

A multicenter, flow cytometry study was conducted to validate the differences in the immunophenotypes of myelodysplastic syndromes (MDS) disease types (Ogata et al., 2012). The data obtained from 115 patients were combined into three groups according to disease grade, i.e., low-grade MDS, refractory anemia with excess blasts, and acute leukemia transformed from MDS (AL-MDS). Data comparison showed that, with disease grade progression, CD34$^+$ myeloblast immunophenotypes were more immature, CD7 expression increased and CD15 expression decreased, the percentage of CD34$^+$ B-progenitors among the total CD34$^+$ cells decreased, and granulocyte granularity decreased. Stepwise logistic regression showed that, in addition to myeloblast percentages, the expression of CD7 and B7-H1 on myeloblasts was independently associated with 11 variables, including the binary data sex, of AL-MDS patients.

3.2 Cerebrospinal fluid metabolomics datasets

Although evidence of mitochondrial dysfunction in bipolar disorder (BD) pathogenesis has been reported, the precise biological basis remains unknown, thus hampering the identification of novel biomarkers (Yoshimi et al., 2016). In that study, metabolomics was performed on the cerebrospinal fluid (CSF) of male BD patients ($n = 54$) and age-matched male healthy controls ($n = 40$). Stepwise logistic regression revealed that isocitric acid (isocitrate) levels were significantly higher in the CSF of BD patients than of the healthy controls based on 16 variables comprised of the metabolomics data and body mass index (BMI).

3.3 Framework of simulation

In order to evaluate the seven estimates, $ER(t_0)$, $R_{\text{app}}$, $R_{\text{loocv}}$, $R_{\text{boot}}$, $R_{\text{ten}}$, $R_{\text{bt632+}}$, and $EER(t_0)$, we performed simulation experiments with 1000 replicates for each condition described below.
from previous studies (Ogata et al., 2012; Yoshimi et al., 2016) subsequent to normalization. The coefficient \( \beta \) was transposed from a reference paper to this study as follows: based on Ogata et al. (2012), \( \beta_1 = 2.73 \) to 2 or 1, \( \beta_2 = 0.60 \) to 1, and \( \beta_3 = 0.52 \) to 1; and based on Yoshimi et al. (2016), \( \beta_1 = 1.78 \) to 2 or 1. Each correlation coefficient value \( \tau \) of the correlation matrix \( \Sigma \) was transposed from the reference paper to this study as follows, \( -1 \leq \tau < -0.6 \) to \( \tau = -0.8 \), \( -0.6 \leq \tau < -0.4 \) to \( \tau = -0.5 \), \( -0.4 \leq \tau < -0.2 \) to \( \tau = -0.3 \), \( -0.2 \leq \tau < 0.2 \) to \( \tau = 0 \), \( 0.2 \leq \tau < 0.4 \) to \( \tau = 0.3 \), and \( 0.4 \leq \tau < 1 \) to \( \tau = 1 \).

The cut-off value \( c \) in the logistic model was set at 0.2, 0.5, and 0.8.

The sample size \( n \) was set as 20, 50, and 100.

The repetition of bootstrapping was \( B = 100 \) (Fu et al., 2012).

The assumption of \( F \) is described as follows: The covariates of \( F \) have a multi-dimensional normal distribution except for binary data. The coefficient \( \beta \) of \( F \) in the logistic distribution, the dimension \( r \) of the covariates, the covariance matrix \( \Sigma \) of \( F \), and the sample size \( n \) are set according to the cases detailed in Table 1. The intercept coefficient \( \beta_0 \) of \( F \) equals \(-1\) in the logistic distribution, similar to previous studies (Ogata et al., 2012; Yoshimi et al., 2016). Each independent Bernoulli variable \( y_i \) of illness follows the logistic distribution.

In the simulation based on Ogata et al. (2012) in which the dimension of the covariates is set as 11, the \( 9^{th} \) covariate is binary data set as sex. In the simulation based on Yoshimi et al. (2016), in which the dimension of the covariates is set as 16, the covariates do not contain binary data. The coefficient \( \beta \) and the covariance matrix \( \Sigma \) of \( F \) were assumed and were referred to as those from previous studies (Ogata et al., 2012; Yoshimi et al., 2016) subsequent to normalization. The \( \beta \) was transposed from a reference paper to this study as follows: based on Ogata et al. (2012), \( \beta_1 = 2.73 \) to 2 or 1, \( \beta_2 = 0.60 \) to 1, and \( \beta_3 = 0.52 \) to 1; and based on Yoshimi et al. (2016), \( \beta_1 = 1.78 \) to 2 or 1. Each correlation coefficient value \( \tau \) of the correlation matrix \( \Sigma \) was transposed from the reference paper to this study as follows, \( -1 \leq \tau < -0.6 \) to \( \tau = -0.8 \), \( -0.6 \leq \tau < -0.4 \) to \( \tau = -0.5 \), \( -0.4 \leq \tau < -0.2 \) to \( \tau = -0.3 \), \( -0.2 \leq \tau < 0.2 \) to \( \tau = 0 \), \( 0.2 \leq \tau < 0.4 \) to \( \tau = 0.3 \), and \( 0.4 \leq \tau < 1 \) to \( \tau = 1 \).

### Table 1. The simulation setting

| Case     | \( n \) | \( \beta \) | \( r \) | \( \Sigma \) |
|----------|--------|-------------|--------|-------------|
| Ogata211 | 20     | 1           | 11     | (1 -0.3 0 -0.3 0 0 0 0 0 0 0 0) |
|          | 50     | 0           | 0      | (0.5 0 0 0 0 0 0 0 0 0 0) |
|          | 100    | 0           | 0      | (0 0 0 0 0 0 0 0 0 0 0) |
| Ogata111 | 100    | 0           | 0      | (0 0 0 0 0 0 0 0 0 0 0) |
| Yoshimi2 | 20     | 0           | 0      | (0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5) |
|          | 50     | 0           | 0      | (0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5) |
|          | 100    | 0           | 0      | (0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5) |
| Yoshimi1 | 100    | 0           | 0      | (0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5) |

The sample size \( n \) was set as 20, 50, and 100.
0.4 ≤ τ < 0.6 to τ = 0.5, and 0.6 ≤ τ < 1 to τ = 0.8.

3.4 Simulation procedure

In this section, we explain the simulation procedure. The Monte Carlo method was performed according to Ounpraseuth et al. (2012).

**Step 1:** Generate the covariate \( t \) following distribution \( F \), using a large enough data volume (100,000) to perform the Monte Carlo method.

**Step 2:** Generate Bernoulli variable \( y \) based on the logistic model per case using the covariate \( t \) from step 1.

**Step 3:** Sample \( n \) from the data from step 2; repeat 1000 runs per case.

**Step 4:** Perform stepwise logistic regression for each sample from step 3.

**Step 5:** Apply each logistic model from step 4 to the original data from step 2; repeat 1000 runs per cut-off value \( c \).

**Step 6:** Perform both the CV and bootstrap methods; repeat 1000 runs per cut-off value \( c \).

**Step 7:** Calculate the RMSE per case based on steps 5 and 6.

4. Results

4.1 The myelodysplastic syndrome datasets case

Fig. 1 and Table 2 show the error rates of the methods. Because the BIAS is the same for all the methods based on the comparison results, the explanation of the comparison is detailed in the following paragraphs. Regarding \( ER(t_0) \): in Ogata211 (\( n = 20 \), \( c = 0.2 \) is inferior to \( c = 0.5 \) and 0.8; in the other cases, \( c = 0.5 \) is the best.

Fig. 2 and Table 3 show the RMSE of the methods, which evaluates the bias-reduction of the error rates in this study. In Ogata111 (\( n = 100 \), \( c = 0.8 \), \( R_{app} \) is comparable to \( R_{loocv} \), while in the other cases, \( R_{app} \) is inferior to \( R_{loocv} \). In Ogata211 (\( n = 20 \), \( c = 0.2 \); \( n = 50 \)) and Ogata111 (\( n = 100 \), \( c = 0.2 \), \( R_{boot} \) is inferior to \( R_{loocv} \), while in the other cases, \( R_{boot} \) is comparable to \( R_{loocv} \). In Ogata111 (\( n = 100 \), \( c = 0.2 \), \( EER(t_0) \) is comparable to \( R_{loocv} \), while in the other cases, \( EER(t_0) \) is superior to \( R_{loocv} \).

Fig. 3 and Table 4 present the BIAS of the methods. \( R_{app} \) and \( R_{boot} \) are inferior to \( R_{loocv} \). In Ogata211 (\( n = 20 \), \( c = 0.8 \); \( n = 50 \), \( c = 0.2 \) and 0.5; \( n = 100 \), \( c = 0.2 \) and 0.5), \( R_{ten} \) is superior to \( R_{loocv} \), while in the other cases \( R_{ten} \) is comparable to \( R_{loocv} \). In Ogata211 (\( n = 20 \), \( c = 0.5 \), \( R_{bt632+} \) is superior to \( R_{loocv} \), in Ogata211 (\( n = 20 \), \( c = 0.2 \)) and Ogata111 (\( n = 100 \), \( c = 0.5 \), \( R_{bt632+} \) is comparable to \( R_{loocv} \), and in the other cases \( R_{bt632+} \) is inferior to \( R_{loocv} \).

Fig. 4 and Table 5 show the \( \sqrt{VAR} \) of the methods. \( R_{app} \) is comparable to \( R_{loocv} \), \( R_{boot} \) is superior to \( R_{loocv} \), and \( R_{ten} \) is comparable to \( R_{loocv} \). In Ogata111 (\( n = 100 \), \( c = 0.8 \), \( R_{bt632+} \) is comparable to \( R_{loocv} \), while in the other cases \( R_{bt632+} \) is superior to \( R_{loocv} \). In Ogata211 (\( n = 50 \), \( c = 0.2 \)), \( EER(t_0) \) is comparable to \( R_{loocv} \), in Ogata111 (\( n = 100 \), \( c = 0.2 \)), \( EER(t_0) \) is inferior to \( R_{loocv} \), and in the other cases, \( EER(t_0) \) is superior to \( R_{loocv} \).
Fig. 1. The estimates and 95% confidence intervals (CIs) of error rates of the methods in the simulations based on Ogata’s paper.

Fig. 2. The estimates and 95% confidence intervals (CIs) of RMSEs of error rates of the methods in the simulations based on Ogata’s paper.
Fig. 3. The estimates and 95% confidence intervals (CIs) of BIASs of error rates of the methods in the simulations based on Ogata’s paper.

Fig. 4. The estimates and 95% confidence intervals (CIs) of $s$ of error rates of the methods in the simulations based on Ogata’s paper.
Table 2. The estimates and 95% confidence intervals (CIs) of the error rates of the methods in the simulations based on Ogata’s paper.

| Cases | ER($t_0$) | $R_{app}$ | $R_{loocv}$ | $R_{boot}$ | $R_{ten}$ | $R_{bt632+}$ | EER($t_0$) | Number of simulation |
|-------|-----------|-----------|-------------|------------|-----------|-------------|------------|---------------------|
|       | Estimate  | 95% CI    | P value     | Estimate   | 95% CI    | P value     | Estimate   | 95% CI    | P value     | Estimate   | 95% CI    | P value     | Estimate   | 95% CI    | P value     | Estimate   | 95% CI    | P value     |
| Ogata  |           |           |             |            |           |             |           |           |             |            |           |             |            |           |             |            |           |             |             |
| n=20  | 0.0017    | (0.0016, 0.0018) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) |
|       | 0.2079    | (0.2078, 0.2080) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) | 0.6000 | 0.2079 | (0.0016, 0.2081) |
|       |           |           |             |            |           |             |           |           |             |            |           |             |            |           |             |            |           |             |             |
|       |           |           |             |            |           |             |           |           |             |            |           |             |            |           |             |            |           |             |             |

P value shows the result that each $R_M$ and $R_{Loocv}$ are tested using paired t-test respectively. $R_M$ represents each of $ER(t_0)$, $R_{app}$, $R_{Loocv}$, $R_{boot}$, $R_{ten}$, $R_{bt632+}$, and $EER(t_0)$. Number of simulation shows the frequency of acceptance of the model in stepwise logistic regression.
Table 3. The estimates and 95% confidence intervals (CIs) of the RMSEs of the methods in the simulations based on Ogata's paper.

| Cut-off | $ER(t_0)$ | $R_{app}$ | $R_{loocv}$ | $R_{boot}$ | $R_{ten}$ | $R_{b632+}$ | $EER(t_0)$ |
|---------|------------|-----------|-------------|------------|-----------|-------------|------------|
| n=50    |            |           |             |            |           |             |            |
| 0.01-0.2| 0.0000 (0.0000, 0.0000) | 0.0166 (0.1631, 0.1761) | 0.1404 (0.1300, 0.1473) | 0.1518 (0.1500, 0.1535) | 0.1518 (0.1500, 0.1535) | 0.1336 (0.1271, 0.1401) | - |
| 0.03-0.5| 0.0000 (0.0000, 0.0000) | 0.1308 (0.1475, 0.1689) | 0.1303 (0.1371, 0.1447) | 0.1379 (0.1371, 0.1447) | 0.1379 (0.1371, 0.1447) | 0.1324 (0.1227, 0.1351) | - |
| 0.08-0.8| 0.0000 (0.0000, 0.0000) | 0.0974 (0.0922, 0.1026) | 0.1023 (0.0970, 0.1077) | 0.0985 (0.0914, 0.1058) | 0.0985 (0.0914, 0.1058) | 0.0896 (0.0847, 0.0944) | - |
| n=100  |            |           |             |            |           |             |            |
| 0.01-0.2| 0.0000 (0.0000, 0.0000) | 0.0795 (0.0670, 0.0923) | 0.0889 (0.0829, 0.0950) | 0.0889 (0.0829, 0.0950) | 0.0889 (0.0829, 0.0950) | 0.0877 (0.0745, 0.0955) | - |
| 0.03-0.5| 0.0000 (0.0000, 0.0000) | 0.0774 (0.0743, 0.0805) | 0.0874 (0.0843, 0.0905) | 0.0874 (0.0843, 0.0905) | 0.0874 (0.0843, 0.0905) | 0.0815 (0.0757, 0.0873) | - |
| 0.08-0.8| 0.0000 (0.0000, 0.0000) | 0.0695 (0.0620, 0.0770) | 0.0770 (0.0742, 0.0808) | 0.0770 (0.0742, 0.0808) | 0.0770 (0.0742, 0.0808) | 0.0667 (0.0663, 0.0691) | - |
| n=200  |            |           |             |            |           |             |            |
| 0.01-0.2| 0.0000 (0.0000, 0.0000) | 0.0236 (0.0156, 0.0292) | 0.0485 (0.0447, 0.0521) | 0.0485 (0.0447, 0.0521) | 0.0485 (0.0447, 0.0521) | 0.0475 (0.0427, 0.049) | - |
| 0.03-0.5| 0.0000 (0.0000, 0.0000) | 0.0656 (0.0535, 0.0769) | 0.0641 (0.0484, 0.0707) | 0.0641 (0.0484, 0.0707) | 0.0641 (0.0484, 0.0707) | 0.0654 (0.0516, 0.0699) | - |
| 0.08-0.8| 0.0000 (0.0000, 0.0000) | 0.0650 (0.0463, 0.0750) | 0.0661 (0.0444, 0.0721) | 0.0661 (0.0444, 0.0721) | 0.0661 (0.0444, 0.0721) | 0.0647 (0.0430, 0.0649) | - |
| n=500  |            |           |             |            |           |             |            |
| 0.01-0.2| 0.0000 (0.0000, 0.0000) | 0.0812 (0.0809, 0.0853) | 0.0825 (0.0805, 0.0845) | 0.0825 (0.0805, 0.0845) | 0.0825 (0.0805, 0.0845) | 0.0827 (0.0808, 0.0857) | - |
| 0.03-0.5| 0.0000 (0.0000, 0.0000) | 0.0612 (0.0456, 0.0711) | 0.0617 (0.0498, 0.0639) | 0.0617 (0.0498, 0.0639) | 0.0617 (0.0498, 0.0639) | 0.0612 (0.0491, 0.0631) | - |
| 0.08-0.8| 0.0000 (0.0000, 0.0000) | 0.0631 (0.0433, 0.0699) | 0.0676 (0.0486, 0.0767) | 0.0676 (0.0486, 0.0767) | 0.0676 (0.0486, 0.0767) | 0.0635 (0.0467, 0.0709) | - |

* means that each $R_M$ and $R_{loocv}$ are different regarding CIs. - means that each $R_M$ and $R_{loocv}$ are comparable regarding CIs. $R_M$ represents each of $ER(t_0)$, $R_{app}$, $R_{loocv}$, $R_{boot}$, $R_{ten}$, $R_{b632+}$, and $EER(t_0)$. 
Table 4. The estimates and 95% confidence intervals (CIs) of the BIAs of the methods in the simulations based on Ogata’s paper.

| Cases | O'cite | ER(t0) | Rapp | Rloocv | Rboot | Reten | Rbt632+ | EER(t0) |
|-------|--------|--------|-------|--------|--------|--------|---------|---------|
|       |        | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value |
| Ogata211 | n=20 | -0.139 (-0.193, 0.107) | 0.010 | -0.099 (-0.108, 0.091) | 0.010 | -0.139 (-0.194, 0.106) | 0.000 | 0.099 (-0.107, 0.097) | 0.771 | -0.016 (-0.116, 0.090) | 0.000 | 0.000 (-0.007, 0.006) | 0.000 |
|       | n=50 | -0.139 (-0.194, 0.106) | 0.001 | -0.099 (-0.108, 0.091) | 0.010 | -0.139 (-0.194, 0.106) | 0.000 | 0.099 (-0.107, 0.097) | 0.771 | -0.016 (-0.116, 0.090) | 0.000 | 0.000 (-0.007, 0.006) | 0.000 |
|       | n=100 | -0.139 (-0.194, 0.106) | 0.001 | -0.099 (-0.108, 0.091) | 0.010 | -0.139 (-0.194, 0.106) | 0.000 | 0.099 (-0.107, 0.097) | 0.771 | -0.016 (-0.116, 0.090) | 0.000 | 0.000 (-0.007, 0.006) | 0.000 |
|       | n=200 | -0.139 (-0.194, 0.106) | 0.001 | -0.099 (-0.108, 0.091) | 0.010 | -0.139 (-0.194, 0.106) | 0.000 | 0.099 (-0.107, 0.097) | 0.771 | -0.016 (-0.116, 0.090) | 0.000 | 0.000 (-0.007, 0.006) | 0.000 |
|       | n=500 | -0.139 (-0.194, 0.106) | 0.001 | -0.099 (-0.108, 0.091) | 0.010 | -0.139 (-0.194, 0.106) | 0.000 | 0.099 (-0.107, 0.097) | 0.771 | -0.016 (-0.116, 0.090) | 0.000 | 0.000 (-0.007, 0.006) | 0.000 |

P value shows the result that each $R_M$ and $R_{loocv}$ are tested using paired $t$-test respectively. $R_M$ represents each of $ER(t0)$, $R_{app}$, $R_{loocv}$, $R_{boot}$, $R_{ten}$, $R_{bt632+}$, and $EER(t0)$. 
Table 5. The estimates and 95% confidence intervals (CIs) of the $\sqrt{VAR}$s of the methods in the simulations based on Ogata's paper.

| Cases | Cut-off | $ER(t_0)$ | $R_{app}$ | $R_{loocv}$ | $R_{boot}$ | $R_{ten}$ | $R_{bt632+}$ | $EER(t_0)$ |
|-------|--------|-----------|-----------|-------------|------------|-----------|-------------|-----------|
|       |        | Estimate | 95% CI    | P-value     | Estimate   | 95% CI    | P-value     | Estimate   | 95% CI    | P-value     | Estimate   | 95% CI    | P-value     | Estimate   | 95% CI    | P-value     |
| Ogata211 | n=20  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0041  | (0.0097, 0.1094) | 0.218 | 0.0097  | (0.0097, 0.2053) | - | 0.0094  | (0.0020, 0.1050) | 0.039 | 0.0040  | (0.0009, 0.0056) | 0.003 | 0.0069  | (0.0013, 0.0127) | 0.005 |
|        | n=50  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0021  | (0.0068, 0.0469) | 0.464 | 0.0052  | (0.0097, 0.1053) | - | 0.0061  | (0.0019, 0.0719) | 0.245 | 0.0057  | (0.0010, 0.0068) | 0.002 |
|        | n=100 | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0027  | (0.0074, 0.0870) | 0.13 | 0.0048  | (0.0086, 0.1512) | - | 0.0099  | (0.0053, 0.1043) | 0.465 | 0.0019  | (0.0011, 0.0019) | 0.000 |
| Ogata211 | n=20  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0004  | (0.0063, 0.0659) | 0.959 | 0.0059  | (0.0054, 0.0637) | - | 0.0061  | (0.0055, 0.1013) | 0.275 | 0.0019  | (0.0002, 0.0250) | 0.000 |
|        | n=50  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0020  | (0.0043, 0.0687) | 0.76 | 0.0065  | (0.0045, 0.0685) | - | 0.0063  | (0.0016, 0.0513) | 0.034 | 0.0089  | (0.0013, 0.0067) | 0.000 |
|        | n=100 | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0046  | (0.0045, 0.0601) | 0.12 | 0.0048  | (0.0035, 0.0539) | 0.002 | 0.0040  | (0.0038, 0.0531) | 0.077 | 0.0192  | (0.0011, 0.0219) | 0.000 |
| Ogata111 | n=20  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0033  | (0.0042, 0.0666) | 0.396 | 0.0040  | (0.0047, 0.0648) | 0.002 | 0.0047  | (0.0069, 0.0647) | 0.004 | 0.0089  | (0.0031, 0.0065) | 0.000 |
|        | n=50  | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0027  | (0.0046, 0.0687) | 0.475 | 0.0043  | (0.0046, 0.0687) | 0.001 | 0.0043  | (0.0043, 0.0655) | 0.001 | 0.0044  | (0.0046, 0.0645) | 0.013 |
|        | n=100 | 0.0000  | (0.0000, 0.0000) | 0.0000 | 0.0046  | (0.0063, 0.0977) | 0.339 | 0.0040  | (0.0062, 0.0982) | 0.002 | 0.0035  | (0.0041, 0.0955) | 0.37 | 0.0043  | (0.0024, 0.0063) | 0.001 |

P value shows the result that each $R_M$ and $R_{loocv}$ are tested using paired $f$-test respectively. $R_M$ represents each of $ER(t_0)$, $R_{app}$, $R_{loocv}$, $R_{boot}$, $R_{ten}$, $R_{bt632+}$, and $EER(t_0)$. 
4.2 The cerebrospinal fluid metabolomics datasets case

Fig. 5 and Table 6 show the error rates of the methods. Because the $BIAS$ is the same for all the methods based on the comparison results, the explanation of the comparison is detailed in the following paragraphs. Regarding $ER(t_0)$, in Yoshimi2 ($n = 20$), $c = 0.2$ is inferior to $c = 0.5$ and 0.8, while in the other cases $c = 0.5$ is the best.

Fig. 6 and Table 7 present the RMSE of the methods, which evaluates the bias-reduction of error rates in this study. In Yoshimi2 ($n = 100$, $c = 0.2$ and 0.8) and Yoshimi1 ($n = 100$, $c = 0.8$), $R_{app}$ is comparable to $R_{loocv}$, while in the other cases, $R_{app}$ is inferior to $R_{loocv}$. $R_{boot}$, $R_{ten}$, and $R_{bt632+}$ are comparable to $R_{loocv}$. In Yoshimi1 ($n = 100$, $c = 0.2$), $EER(t_0)$ is comparable to $R_{loocv}$, while in the other cases $EER(t_0)$ is superior to $R_{loocv}$.

Fig. 7 and Table 8 show the $BIAS$ of the methods; $R_{app}$ and $R_{boot}$ are inferior to $R_{loocv}$. In Yoshimi2 ($n = 20$ and 50, $c = 0.8$), $R_{ten}$ is superior to $R_{loocv}$, while in the other cases $R_{ten}$ is comparable to $R_{loocv}$. In Yoshimi2 ($n = 20$, 50 and 100, $c = 0.2$ and 0.8) and Yoshimi1 ($n = 100$, $c = 0.2$ and 0.8), $R_{bt632+}$ is inferior to $R_{loocv}$, in Yoshimi2 ($n = 20$, $c = 0.5$), $R_{bt632+}$ is superior to $R_{loocv}$, and in the other cases, $R_{bt632+}$ is comparable to $R_{loocv}$.

Fig. 8 and Table 9 illustrate the $\sqrt{VAR}$ of the methods; $R_{app}$ is comparable to $R_{loocv}$. In Yoshimi2 ($n = 20$) and Yoshimi1 ($n = 100$, $c = 0.2$), $R_{boot}$ is superior to $R_{loocv}$, while in the other cases $R_{boot}$ is comparable to $R_{loocv}$. $R_{ten}$ is comparable to $R_{loocv}$. In Yoshimi2 ($n = 20$, $c = 0.2$ and 0.8; $n = 50$, $c = 0.2$) and Yoshimi1 ($n = 100$, $c = 0.2$), $R_{bt632+}$ is superior to $R_{loocv}$, while in the other cases $R_{bt632+}$ is comparable to $R_{loocv}$. In Yoshimi2 ($n = 20$, $c = 0.2$), $EER(t_0)$ is comparable to $R_{loocv}$, in Yoshimi1 ($n = 100$, $c = 0.2$), $EER(t_0)$ is inferior to $R_{loocv}$, and in the other cases $EER(t_0)$ is superior to $R_{loocv}$.

P value shows the result that each $M$ and $R_{loocv}$ are tested using paired $t$-test respectively. $R_M$ represents each of $ER(t_0)$, $R_{app}$, $R_{loocv}$, $R_{boot}$, $R_{ten}$, $R_{bt632+}$, and $EER(t_0)$.

4.3 Note

$BIAS$ of bootstrap is greater than LOOCV. Bootstrap estimates error rates lower than LOOCV. Because bootstrap estimates the error rate the same as the apparent error rate, medical and pharmaceutical workers tend to misinterpret bootstrap as more useful, in spite of its poor performance. In this report, in order to clearly correct this misunderstanding, we present the error rate as well as the $BIAS$.

5. Discussion and Conclusions

In this paper, we evaluated leave-one-out cross-validation, 10-fold cross-validation, bootstrap, and $.632+$ bootstrap using the root mean squared error via simulation by setting the three sample sizes. Additional considerations for the simulation settings included: the simulations were reflected by the unbiased estimates of actual cases, the true distributions of the
Fig. 5. The estimates and 95% confidence intervals (CIs) of error rates of the methods in the simulations based on Yoshimi’s paper.

Fig. 6. The estimates and 95% confidence intervals (CIs) of RMSEs of error rates of the methods in the simulations based on Yoshimi’s paper.

Jpn J Biomet Vol. 38, No. 1, 2017
Fig. 7. The estimates and 95% confidence intervals (CIs) of BIASs of error rates of the methods in the simulations based on Yoshimi's paper.

Fig. 8. The estimates and 95% confidence intervals (CIs) of $s$ of error rates of the methods in the simulations based on Yoshimi's paper.
Table 6. The estimates and 95% confidence intervals (CIs) of the error rates of the methods in the simulations based on Yoshimi’s paper.

| Cases | Estimate | 99% CI | P value | Estimate | 99% CI | P value | Estimate | 99% CI | P value | Estimate | 99% CI | P value | Estimate | 99% CI | P value | Number of simulation |
|-------|----------|--------|---------|----------|--------|---------|----------|--------|---------|----------|--------|---------|----------|--------|---------|---------------------|
| RM    | 0.0460  | (0.0370, 0.0547) | 0.000 | 0.2452  | (0.2379, 0.2525) | 0.000 | 0.3819  | (0.3740, 0.3906) | 0.000 | 0.3460  | (0.3400, 0.3520) | 0.000 | 0.2818  | (0.2735, 0.2900) | 0.000 | 0.2460  | (0.2400, 0.2520) | 0.000 | 539 |
| R_loocv | 0.2716  | (0.2646, 0.2786) | 0.000 | 0.1941  | (0.1860, 0.1965) | 0.000 | 0.2162  | (0.2105, 0.2219) | 0.000 | 0.1964  | (0.1897, 0.2032) | 0.000 | 0.2167  | (0.2105, 0.2226) | 0.000 | 0.2169  | (0.2105, 0.2229) | 0.000 | 0.2371  | (0.2371, 0.2371) | 0.000 | 539 |
| R_boot | 0.2943  | (0.2803, 0.2983) | 0.000 | 0.3241  | (0.3175, 0.3297) | 0.000 | 0.2970  | (0.2900, 0.3040) | 0.000 | 0.2863  | (0.2804, 0.2921) | 0.000 | 0.2770  | (0.2734, 0.2805) | 0.000 | 0.2770  | (0.2734, 0.2805) | 0.000 | 0.2746  | (0.2712, 0.2780) | 0.000 | 399 |
| R_ten | 0.2904  | (0.2862, 0.2946) | 0.000 | 0.2680  | (0.2644, 0.2716) | 0.000 | 0.2770  | (0.2735, 0.2806) | 0.000 | 0.2863  | (0.2804, 0.2921) | 0.000 | 0.2770  | (0.2734, 0.2805) | 0.000 | 0.2770  | (0.2734, 0.2805) | 0.000 | 0.2746  | (0.2712, 0.2780) | 0.000 | 399 |
| R_bt632+ | 0.3182  | (0.3173, 0.1950) | 0.000 | 0.1997  | (0.1997, 0.2032) | 0.000 | 0.1997  | (0.1997, 0.2032) | 0.000 | 0.2018  | (0.1997, 0.2049) | 0.000 | 0.1995  | (0.1990, 0.2032) | 0.000 | 0.2105  | (0.2105, 0.2152) | 0.000 | 0.2105  | (0.2105, 0.2152) | 0.000 | 1000 |
| EER(f0) | 0.3259  | (0.2575, 0.3950) | 0.000 | 0.2416  | (0.2380, 0.2454) | 0.000 | 0.2406  | (0.2407, 0.2528) | 0.000 | 0.2401  | (0.2374, 0.2438) | 0.000 | 0.2405  | (0.2407, 0.2520) | 0.000 | 0.2403  | (0.2380, 0.2449) | 0.000 | 0.2405  | (0.2380, 0.2449) | 0.000 | 1000 |
| RM    | 0.4343  | (0.4305, 0.4381) | 0.000 | 0.3791  | (0.3722, 0.4001) | 0.000 | 0.4112  | (0.4062, 0.4162) | 0.000 | 0.3956  | (0.3904, 0.4015) | 0.000 | 0.4110  | (0.4060, 0.4159) | 0.000 | 0.4002  | (0.4004, 0.4150) | 0.000 | 0.4406  | (0.4403, 0.4435) | 0.000 | 999 |
| R_loocv | 0.2925  | (0.2841, 0.2989) | 0.000 | 0.2667  | (0.2598, 0.2744) | 0.000 | 0.2335  | (0.2250, 0.2420) | 0.000 | 0.2616  | (0.2580, 0.2654) | 0.000 | 0.2616  | (0.2580, 0.2654) | 0.000 | 0.2616  | (0.2580, 0.2654) | 0.000 | 0.2616  | (0.2580, 0.2654) | 0.000 | 999 |
| R_boot | 0.2920  | (0.2851, 0.2989) | 0.000 | 0.2690  | (0.2622, 0.2758) | 0.000 | 0.2550  | (0.2489, 0.2616) | 0.000 | 0.2550  | (0.2489, 0.2616) | 0.000 | 0.2550  | (0.2489, 0.2616) | 0.000 | 0.2550  | (0.2489, 0.2616) | 0.000 | 0.2550  | (0.2489, 0.2616) | 0.000 | 999 |

P value shows the result that each $R_M$ and $R_{loocv}$ are tested using paired $t$-test respectively. $R_M$ represents each of $ER(f_0)$, $R_{app}$, $R_{loocv}$, $R_{boot}$, $R_{ten}$, $R_{bt632+}$, and $EER(f_0)$. Number of simulation shows the frequency of acceptance of the model in stepwise logistic regression.
Table 7. The estimates and 95% confidence intervals (CIs) of the RMSEs of the methods in the simulations based on Yoshimi’s paper.

| Cases | Cutoff | \( ER(t_0) \) | \( R_{app} \) | \( R_{loocv} \) | \( R_{boot} \) | \( R_{ten} \) | \( R_{632+} \) | \( EER(t_0) \) |
|-------|--------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|       |        | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| Yoshimi | \( n=20 \) | 0.0000 | (0.0000, 0.0000) | **0.1460** | (0.1387, 0.1534) | **0.1259** | (0.1109, 0.1320) | 0.1372 | (0.1304, 0.1441) | 0.1253 | (0.1186, 0.1321) | 0.1200 | (0.1137, 0.1262) | 0.0995 | (0.0895, 0.1094) | 0.0326 | (0.0263, 0.0390) |
|       | \( n=50 \) | 0.0000 | (0.0000, 0.0000) | **0.1490** | (0.1335, 0.1646) | **0.1444** | (0.1394, 0.1504) | 0.1184 | (0.1122, 0.1247) | 0.1141 | (0.1081, 0.1201) | 0.1065 | (0.1009, 0.1121) | 0.0836 | (0.0703, 0.0970) | 0.0864 | (0.0693, 0.0939) |
|       | \( n=100 \) | 0.0000 | (0.0000, 0.0000) | **0.1540** | (0.1389, 0.1692) | **0.1517** | (0.1360, 0.1676) | 0.1273 | (0.1193, 0.1355) | 0.1206 | (0.1126, 0.1286) | 0.1085 | (0.1005, 0.1164) | 0.0845 | (0.0696, 0.0993) | 0.0879 | (0.0722, 0.0936) |

* means that each \( R_M \) and \( R_{loocv} \) are different regarding CIs. - means that each \( R_M \) and \( R_{loocv} \) are comparable regarding CIs.

\( R_M \) represents each of \( ER(t_0) \), \( R_{app} \), \( R_{loocv} \), \( R_{boot} \), \( R_{ten} \), \( R_{632+} \), and \( EER(t_0) \).
Table 8. The estimates and 95% confidence intervals (CIs) of the BIASs of the methods in the simulations based on Yoshimi’s paper.

| Cases | ER(t₀) | R<sub>app</sub> | R<sub>loocv</sub> | R<sub>boot</sub> | R<sub>ten</sub> | R<sub>b<sub>632</sub>+</sub> | EER(t₀) |
|-------|--------|------------------|------------------|------------------|------------------|------------------|---------|
|       | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value |
| Yoshimi<sub>2</sub> | n=50 | 0.0000 (0.0000, 0.0000) | 0.0000 | -0.0011 (-0.2193, 0.2025) | 0.0000 | -0.0644 (-0.0795, -0.0553) | 0.0000 | -0.0551 (-0.0769, -0.0430) | 0.0000 | -0.0549 (-0.0791, -0.0307) | 0.0000 | -0.0195 (-0.0394, 0.0238) | 0.0000 |
|       | n=00 | 0.0000 (0.0000, 0.0000) | 0.0000 | -0.0011 (-0.0270, 0.0179) | 0.0000 | -0.0561 (-0.0475, 0.0618) | 0.0000 | -0.0566 (-0.0398, 0.0664) | 0.0000 | -0.0574 (-0.0381, 0.0565) | 0.0000 | -0.0195 (-0.0394, 0.0238) | 0.0000 |

P value shows the result that each \( R_M \) and \( R_{loocv} \) are tested using paired t-test respectively. \( R_M \) represents each of \( ER(t₀) \), \( R_{app} \), \( R_{loocv} \), \( R_{boot} \), \( R_{ten} \), \( R_{b<sub>632</sub>}+ \), and \( EER(t₀) \).
Table 9. The estimates and 95% confidence intervals (CIs) of the s of the methods in the simulations based on Ogata’s paper.

| Cases | Out of | \( ER(t_0) \) | \( R_{app} \) | \( R_{loocv} \) | \( R_{boot} \) | \( R_{ten} \) | \( R_{632+} \) | \( EER(t_0) \) |
|-------|--------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|
|       |        | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value | Estimate | 95% CI | P value |
| TestA | n=20   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.1053 | (0.0605, 0.1121) | 0.544 | 0.1081 | (0.1020, 0.1141) | 0.001 | 0.0884 | (0.0888, 0.1003) | 0.001 | 0.1071 | (0.1012, 0.1140) | 0.002 |
|       | n=50   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0960 | (0.0515, 0.1031) | 0.567 | 0.0992 | (0.0597, 0.1056) | 0.000 | 0.0895 | (0.0446, 0.0953) | 0.015 | 0.0890 | (0.0435, 0.0954) | 0.052 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0909 | (0.0456, 0.0967) | 0.619 | 0.0928 | (0.0497, 0.0989) | 0.001 | 0.0896 | (0.0471, 0.0951) | 0.000 | 0.0929 | (0.0472, 0.0990) | 0.002 |
| TestB | n=20   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0723 | (0.0490, 0.0759) | 0.604 | 0.0718 | (0.0485, 0.0751) | 0.079 | 0.0723 | (0.0459, 0.0756) | 0.023 | 0.0671 | (0.0463, 0.0781) | 0.034 |
|       | n=50   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0694 | (0.0466, 0.0727) | 0.609 | 0.0708 | (0.0475, 0.0740) | 0.029 | 0.0693 | (0.0467, 0.0723) | 0.041 | 0.0643 | (0.0437, 0.0677) | 0.059 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0556 | (0.0329, 0.0687) | 0.651 | 0.0544 | (0.0312, 0.0676) | 0.142 | 0.0532 | (0.0314, 0.0652) | 0.789 | 0.0611 | (0.0386, 0.0669) | 0.382 |
| TestC | n=20   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0471 | (0.0401, 0.0482) | 0.736 | 0.0466 | (0.0445, 0.0487) | 0.120 | 0.0468 | (0.0449, 0.0490) | 0.054 | 0.0443 | (0.0425, 0.0466) | 0.110 |
|       | n=50   | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0453 | (0.0434, 0.0494) | 0.406 | 0.0441 | (0.0423, 0.0462) | 0.092 | 0.0440 | (0.0427, 0.0466) | 0.762 | 0.0418 | (0.0400, 0.0437) | 0.382 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0426 | (0.0400, 0.0454) | 0.597 | 0.0425 | (0.0400, 0.0454) | 0.116 | 0.0430 | (0.0412, 0.0469) | 0.762 | 0.0427 | (0.0405, 0.0452) | 0.151 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0334 | (0.0275, 0.0393) | 0.113 | 0.0328 | (0.0269, 0.0383) | 0.197 | 0.0334 | (0.0275, 0.0391) | 0.779 | 0.0334 | (0.0274, 0.0397) | 0.041 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0485 | (0.0465, 0.0506) | 0.723 | 0.0480 | (0.0460, 0.0502) | 0.068 | 0.0488 | (0.0468, 0.0508) | 0.784 | 0.0454 | (0.0405, 0.0475) | 0.084 |
|       |        | 0.0000   | (0.0000, 0.0000) | 0.000 | 0.0422 | (0.0404, 0.0440) | 0.609 | 0.0422 | (0.0404, 0.0440) | 0.269 | 0.0443 | (0.0424, 0.0462) | 0.602 | 0.0427 | (0.0410, 0.0447) | 0.042 |

P value shows the result that each \( R_M \) and \( R_{loocv} \) are tested using paired \( f \)-test respectively. \( R_M \) represents each of \( ER(t_0) \), \( R_{app} \), \( R_{loocv} \), \( R_{boot} \), \( R_{ten} \), \( R_{632+} \), and \( EER(t_0) \).
samples were defined to evaluate the methods accurately, and the cut off values were set at three values for application of ROC. Our results indicate that the performances of leave-one-out cross-validation, 10-fold cross-validation, and .632+ bootstrap are comparable under previously encountered conditions.

In conclusion, we recommend the use of leave-one-out cross-validation, 10-fold cross-validation, or .632+ bootstrap in specific biomarker discovery studies involving binary outcome data, only approximately ten covariates with or without binomial covariates, such as sex, and low and/or high correlations between the covariates, thus eliminating misinterpretation due to cross-validation method selection.

Conflict of interest
The authors have no conflict of interest relevant to the content of this paper.

Acknowledgements
The authors wish to thank Dr. Isao Yoshimura for his helpful advice in preparing this manuscript. The authors are grateful to the two anonymous referees for their many constructive comments that substantially improved this manuscript. The authors also wish to thank Kiyoyuki Ogata M.D. PhD, Dr. Noriko Yoshimi, and the other members of these study groups for permission to use these valuable study data.

REFERENCES
Abdi, H. and Williams, L. J. (2010). Jackknife. In Encyclopedia of Research Design, 655–660. Thousand Oaks (CA): Sage.

Azuaje, F. (2003). Genomic data sampling and its effect on classification performance assessment. BMC Bioinformatics 4, 5.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. Wadsworth.

Davison, A. C. and Hall, P. (1992). On the bias and variability of bootstrap and CV estimates of error rate in discrimination problems. Biometrika 79, 279–284.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78, 316–331.

Efron, B. and Tibshirani, R. (1997). Improvement on cross-validation: the .632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning, Chapter 7.11. Springer.
Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307.

Ogata, K., Kakumoto, K., Matsuda, A., Tohyama, K., Tamura, H., Ueda, Y., et al. (2012). Differences in blast immunophenotypes among disease types in myelodysplastic syndromes: A multicenter validation study. *Leukemia Research* **36**, 1229–1236.

Ounpraseuth, S., Lensing, S. Y., Spencer, H. J., and Kodell, R. L. (2012). Estimating misclassification error: a closer look at CV based methods. *BMC Research Notes* **5**, 656.

Teh, S. Y., Othman, A. R., and Khoo, M. B. H. (2010). Dichotomous Logistic Regression with Leave-One-Out Validation. *World Academy of Science, Engineering and Technology* **62**, 994–1003.

Yoshimi, N., Futamura, T., Bergen, S. E., Iwayama, Y., Ishima, T., Sellgren, C., et al. (2016). Cerebrospinal fluid metabolomics identifies a key role of isocitrate dehydrogenase in bipolar disorder: evidence in support of mitochondrial dysfunction hypothesis. *Molecular Psychiatry* **21**, 1504–1510.