Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods

Hisashi Noma1 | Tomohiro Shinozaki2 | Katsuhiro Iba3,4 | Satoshi Teramukai5 | Toshi A. Furukawa6,7

Abstract
In assessing prediction accuracy of multivariable prediction models, optimism corrections are essential for preventing biased results. However, in most published papers of clinical prediction models, the point estimates of the prediction accuracy measures are corrected by adequate bootstrap-based correction methods, but their confidence intervals are not corrected, for example, the DeLong’s confidence interval is usually used for assessing the C-statistic. These naïve methods do not adjust for the optimism bias and do not account for statistical variability in the estimation of parameters in the prediction models. Therefore, their coverage probabilities of the true value of the prediction accuracy measure can be seriously below the nominal level (eg, 95%). In this article, we provide two generic bootstrap methods, namely, (1) location-shifted bootstrap confidence intervals and (2) two-stage bootstrap confidence intervals, that can be generally applied to the bootstrap-based optimism correction methods, that is, the Harrell’s bias correction, 0.632, and 0.632+ methods. In addition, they can be widely applied to various methods for prediction model development involving modern shrinkage methods such as the ridge and lasso regressions. Through numerical evaluations by simulations, the proposed confidence intervals showed favorable coverage performances. Besides, the current standard practices based on the optimism-uncorrected methods showed serious undercoverage properties. To avoid erroneous results, the optimism-uncorrected confidence intervals should not be used in practice, and the adjusted methods are recommended instead. We also developed the R package predboot for implementing these methods (https://github.com/nomahi/predboot). The effectiveness of the proposed methods are illustrated via applications to the GUSTO-I clinical trial.

Keywords
bootstrap, confidence interval, discrimination and calibration, multivariable prediction model, optimism
In the development of clinical prediction models, multivariable prediction models have been essential statistical tools for incorporating multiple predictive factors to construct diagnostic and prognostic algorithms. A multivariable prediction model is usually constructed by an adequate regression model (eg, a logistic regression model for a binary outcome) based on a series of representative patients from the source population, but their “apparent” predictive performances such as discrimination and calibration measures are biased from their actual performances for external populations. This bias is known as “optimism” in prediction models. Practical guidelines (eg, the TRIPOD statements) recommend adopting principled optimism adjustment methods for internal validations, and these are currently the standard statistical analysis methods in practice. In particular, the bootstrap-based correction methods, that is, the Harrell’s bootstrapping bias correction, and methods, have been recommended.

However, the optimism corrections are mainly applied only to point estimates of the prediction performance measures in current practice. Even in recent leading medical journals, although many papers provided optimism-corrected estimates, their confidence intervals were usually not optimism-corrected, for example, for C-statistics, and many papers provided solely the conventional DeLong’s confidence interval. Since the point estimate of the naïve, uncorrected predictive accuracy measures is biased, the actual coverage rates of their confidence intervals can be seriously below the nominal level (eg, 95%). The reported predictive performance estimates can directly influence clinical guidelines or medical practices, so assuring the validity of their inferences is a relevant problem.

In previous methodological studies, resampling-based confidence intervals have been discussed for CV. However, there have been no effective techniques to construct valid confidence intervals based on the bootstrap-based optimism correction methods. In this article, we propose effective methods to construct the confidence intervals, particularly we provide two generic bootstrap algorithms, namely, (1) location-shifted bootstrap confidence intervals and (2) two-stage bootstrap confidence intervals, that can be widely applied to various approaches to prediction model development involving modern shrinkage methods such as the ridge regressions. In numerical evaluations using simulations, the proposed confidence intervals based on the three optimism correction methods showed favourable coverage performances. However, the current standard practices based on the optimism-uncorrected methods showed marked undercoverage properties. We recommend against adopting the naïve optimism-uncorrected confidence intervals, and instead propose using the improved methods in practice. We also illustrate the effectiveness of the proposed methods via simulation studies and applications to real-world data from the GUSTO-I trial.

2 | ESTIMATION OF PREDICTION ACCURACY MEASURES

2.1 | Logistic regression model for clinical predictions

First, we briefly introduce the fundamental methods for multivariable prediction models and their prediction accuracy measures. In this article, we consider to construct a logistic regression prediction model for a binary outcome, but the proposed methods can similarly be applied to other types of prediction models, for example, the Cox regression for censored time-to-event outcomes. Let \( y_i \) (\( i = 1, 2, \ldots, n \)) be a binary outcome variable (1: event occurring, or 0: not occurring) and let \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \) (\( i = 1, 2, \ldots, n \)) be \( p \) predictor variables for an individual \( i \). Based on the logistic regression model, the probability of event occurrence \( \pi_i = \Pr(y_i = 1|x_i) \) is modeled as

\[
\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})},
\]

where \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) are the regression coefficients containing the intercept. Plugging an appropriate estimate \( \hat{\beta} \) into the above model for \( \beta \), the estimated probability \( \hat{\pi}_i \) (\( i = 1, 2, \ldots, n \)) is defined as the risk score of individual patients, and this score is used as the criterion score to determine the predicted outcome.

For estimating the regression coefficients \( \beta \), the maximum likelihood (ML) estimation is a standard approach that has favorable theoretical properties such as asymptotic efficiency. Shrinkage regression methods that use penalized likelihood functions can also be adopted, such as the ridge and lasso regressions. Through the penalizations, the resultant regression coefficient estimates are shrunk towards zero and thereby can reduce overfitting. Other prediction algorithms involving machine learning methods have also been used, and the following proposed methods can generally be applied to these methods.
2.2 | Prediction accuracy measures

For assessing the predictive performances of the developed multivariable models, several accuracy measures are considered regarding their discrimination and calibration abilities.\(^2\) For example, discrimination refers to a predictor’s ability to separate patients with different responses, and the most commonly used measure is the \(C\)-statistic, which is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant.\(^1\) The \(C\)-statistic also corresponds to the area under the curve (AUC) of the empirical receiver operating characteristic (ROC) curve for the risk score.\(^2\) The \(C\)-statistic ranges from 0.5 to 1.0, with larger values corresponding to superior discriminant performance. All of the prediction accuracy measures may have biases from their actual performances for external populations if they are assessed for the derivation dataset itself,\(^3,4\) and this bias is known as optimism in prediction models. To appropriately assess these accuracy measures, adequate optimism adjustments are needed, and the bootstrap-based methods are widely adopted.\(^3,4,7\) In Sections 2.3 to 2.5, we briefly review the three bootstrap-based methods, namely, the Harrell’s bias correction, 0.632, and 0.632+ methods.

2.3 | Harrell’s bias correction

Currently, the most widely applied bootstrap-based correction method in practice is Harrell’s bias correction,\(^1\) which is based on the conventional bootstrap bias correction.\(^20,21\) The algorithm is summarized as follows:

1. Let \(\hat{\theta}_{\text{app}}\) be the apparent predictive performance estimate for the original population.
2. Generate \(B\) bootstrap samples by resampling with replacement from the original population.
3. Construct a prediction model for each bootstrap sample, and calculate the predictive performance estimate for it. Note that the same modeling steps used to develop the final model are performed. We denote the \(B\) bootstrap estimates as \(\hat{\theta}_{1,\text{boot}}^\prime, \hat{\theta}_{2,\text{boot}}^\prime, \ldots, \hat{\theta}_{B,\text{boot}}^\prime\).
4. Using the \(B\) prediction models constructed from the bootstrap samples, calculate the predictive performance estimates for the original population: \(\hat{\theta}_{1,\text{orig}}, \hat{\theta}_{2,\text{orig}}, \ldots, \hat{\theta}_{B,\text{orig}}\).
5. The bootstrap estimate of optimism is obtained as

\[
\hat{\Lambda} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{b,\text{boot}} - \hat{\theta}_{b,\text{orig}}).
\]

Subtracting the estimate of optimism from the apparent performance, the bias-corrected predictive performance estimate is obtained as \(\hat{\theta}_{\text{app}} - \hat{\Lambda}\).

The bias-corrected estimate is calculable by a relatively simple algorithm, and some simulation-based numerical evidence has shown that it has favorable properties under realistic situations.\(^7,22\) However, a certain proportion of patients in the original population (approximately 63.2%, on average) should be overlapped in the bootstrap sample. The overlap may cause overestimation of the predictive performance,\(^19\) and several alternative methods have therefore been proposed.

2.4 | The 0.632 method

The 0.632 method\(^5\) was proposed as another bias correction technique to address the overlapping problem. In each bootstrapping, we formally regard the “external” subjects who are not included in the bootstrap sample as a “test” dataset for the prediction model developed in the bootstrap sample. Then, we calculate the predictive performance estimates for the external samples by the developed \(B\) prediction models \(\hat{\theta}_{1,\text{out}}, \hat{\theta}_{2,\text{out}}, \ldots, \hat{\theta}_{B,\text{out}}\), and we denote the average measure as \(\hat{\theta}_{\text{out}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{b,\text{out}} / B\). Thereafter, the 0.632 estimator is defined as a weighted average of the predictive performance estimate in the original sample \(\hat{\theta}_{\text{app}}\) and the external sample estimate \(\hat{\theta}_{\text{out}}\):

\[
\hat{\theta}_{0.632} = 0.368 \times \hat{\theta}_{\text{app}} + 0.632 \times \hat{\theta}_{\text{out}}
\]

The weight 0.632 is derived from the approximate proportion of subjects included in a bootstrap sample. Since the subjects that are included in a bootstrap sample are independent from those that are not, the 0.632 estimator can be interpreted as an extension of CV. However, the 0.632 estimator is associated with overestimation bias under highly overfit situations, when the apparent estimator \(\hat{\theta}_{\text{app}}\) has a large bias.\(^6\)
2.5 | The 0.632+ method

Efron and Tibshirani\(^6\) proposed the 0.632+ method to address the problem of the 0.632 estimator by taking into account the amount of overfitting. They define the relative overfitting rate \( R \) as

\[
R = \frac{\hat{\theta}_{\text{out}} - \hat{\theta}_{\text{app}}}{\gamma - \hat{\theta}_{\text{app}}}.
\]

\( \gamma \) corresponds to “no information performance,” which is the predictive performance measure for the original population when the outcomes are randomly permuted. \( \gamma \) can usually be computed as an empirical value from the permuted datasets, and we can use theoretically valid values for some predictive measures, for example, we can adopt \( \gamma = 0.50 \) for the case of the C-statistic.\(^7\) The overfitting rate \( R \) approaches 0 when there is no overfitting (\( \hat{\theta}_{\text{out}} = \hat{\theta}_{\text{app}} \)), and approaches 1 when the degree of overfitting is large. Then, the 0.632+ estimator is defined as

\[
\hat{\theta}_{0.632+} = (1 - w) \times \hat{\theta}_{\text{app}} + w \times \hat{\theta}_{\text{out}},
\]

\[
w = \frac{0.632}{1 - 0.368 \times R}.
\]

Note that the weight \( w \) ranges from 0.632 (\( R = 0 \)) to 1 (\( R = 1 \)). Hence, the 0.632+ estimator approaches the 0.632 estimator when there is little overfitting and approaches the external sample estimate \( \hat{\theta}_{\text{out}} \) when there is marked overfitting.

3 | CONFIDENCE INTERVALS OF THE PREDICTION ACCURACY MEASURES

3.1 | Location-shifted bootstrap confidence interval

The bootstrap-based optimism correction methods are essential for bias corrections of the prediction measures, but currently there are no effective methods based on the bootstrap-based corrections to construct confidence intervals that consider the optimism. Conventional analytical approaches for apparent measures (e.g., the DeLong’s confidence interval for C-statistic\(^8\)) and the naïve bootstrap confidence interval should provide invalid confidence intervals under realistic situations due to the biases.

The first approach we propose here is the location-shifted bootstrap confidence interval. This approach is simple. Based on the asymptotic theory for the bootstrap,\(^20,21\) the naïve bootstrap confidence interval for the apparent measures can adequately evaluate the approximate statistical variability of the predictive measures under large sample settings. However, its location should be shifted upwardly by the bias of the point estimate. Thus, the invalidity of the naïve bootstrap confidence interval is expected to be addressed if the bias of the location is adjusted when the sample size is sufficiently large. The proposal here is to correct the “location” of the naïve bootstrap confidence interval by the estimated bias. The algorithm to calculate the confidence limits is provided as follows.

**Algorithm 1.** Location-shifted bootstrap confidence interval

1. For a multivariable prediction model, let \( \hat{\theta}_{\text{app}} \) be the apparent predictive measure for the derivation population and let \( \hat{\theta} \) be the optimism-corrected predictive measure obtained from the Harrell’s bias correction, 0.632, or 0.632+ method.
2. In the computational processes of \( \hat{\theta} \), we can obtain a bootstrap estimate of the sampling distribution of \( \hat{\theta}_{\text{app}} \) from the \( B \) bootstrap samples. Compute the bootstrap confidence interval of \( \hat{\theta}_{\text{app}} \) from the bootstrap distribution, \( (\hat{\theta}_{\text{app}, L}, \hat{\theta}_{\text{app}, U}) \); for the 95% confidence interval, they are typically calculated by the 2.5th and 97.5th percentiles of the bootstrap distribution.
3. Calculate the bias estimate by optimism, \( \hat{\delta} = \hat{\theta}_{\text{app}} - \hat{\theta} \).
4. Then, the location-shifted bootstrap confidence interval is computed as \( (\hat{\theta}_{\text{app}, L} - \hat{\delta}, \hat{\theta}_{\text{app}, U} - \hat{\delta}) \).
Note that the advantage of this method is the simplicity and low computational burden of calculating the confidence limits. It only requires the bootstrap confidence interval by $\hat{\theta}_{\text{app}}$ and the bias estimate $\hat{\delta}$. These quantities can be obtained within the bootstrap processes of the optimism correction methods, and additional burdensome computations are not needed. Adjusting the location of the apparent bootstrap confidence interval is straightforward, but the location-shifted confidence interval is justified by the large-sample theory, because the apparent bootstrap confidence interval becomes a valid confidence interval asymptotically, and the bias also converges to 0.

However, the apparent bootstrap confidence interval only quantifies the variability of the apparent predictive measure $\hat{\theta}_{\text{app}}$. The optimism-corrected measure $\hat{\theta}$ generally has a larger variability related to the variability of the optimism correction quantity $\hat{\delta}$ and the correlation between $\hat{\theta}_{\text{app}}$ and $\hat{\delta}$. Therefore, it can underestimate the statistical variability and have undercoverage properties under moderate sample settings.

### 3.2 | Two-stage bootstrap confidence interval

To address the undercoverage properties, the variability of the optimism measures $\hat{\delta}$ and the correlations between $\hat{\theta}_{\text{app}}$ and $\hat{\delta}$ should be adequately considered. However, the correlations between their constituent components are quite complicated and are difficult to assess adequately. Thus, numerical approaches are also effective for assessing their variabilities simultaneously. We propose the following two-stage bootstrap approach that aims to directly obtain bootstrap distributions of the optimism-corrected statistics.

**Algorithm 2.** Two-stage bootstrap confidence interval

1. Generate $B$ bootstrap samples by resampling with replacement from the original population.
2. Develop a multivariable prediction model for each bootstrap sample, and calculate the optimism-corrected predictive measures for the $B$ bootstrap samples $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$, using the Harrell's bias correction, 0.632, or 0.632+ method, by performing bootstrap resampling from the bootstrap sample. We can obtain the bootstrap distributions of the optimism-corrected predictive measures by $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$.
3. Then, compute the bootstrap confidence interval from the bootstrap samples $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$ for the 95% confidence interval, they are typically calculated by the 2.5th and 97.5th percentiles of the bootstrap distribution.

The two-stage bootstrap confidence interval adequately addresses the variabilities of the optimism-corrected measures themselves that involve the correlations mentioned above. In the numerical evaluations in Section 4, the two-stage bootstrap methods provided wider confidence intervals in general and showed better coverage properties compared with the apparent bootstrap confidence interval. In addition, the two-stage bootstrap directly corresponds to the formal bootstrap confidence intervals of the Harrell's bias correction, 0.632, and 0.632+ methods, thus their asymptotic validities are assured theoretically.

However, a difficulty of this approach is the computational burden. Many studies recommend performing more than 1000 resamplings for calculating bootstrap confidence intervals, so if we conduct 2000 bootstrap resamplings for both of the two-stage bootstraps, $2000 \times 2000 = 4,000,000$ iterative computations are needed to construct the multivariable prediction models, requiring enormous computational time. However, the performance of computers is continually increasing, and therefore these tasks might not be difficult in the near future. At present, an appropriate solution is to collaborate with statisticians who have access to high-performance computer systems. If requested, our group is willing to engage in collaborations on valuable research projects.

The R package *predboot*, which can conduct the two proposed methods for logistic regression models with the ML, ridge, and lasso estimations, is available at [https://github.com/nomahi/predboot](https://github.com/nomahi/predboot). Also, examples of *predboot* codes for implementing the estimations are available in e-Appendix A.

### 4 | APPLICATIONS

To illustrate the effectiveness of the proposed methods in practice, we present their applications to a real-world dataset, namely that of the GUSTO-I trial. The GUSTO-I dataset has been adopted by many performance evaluation studies of multivariable prediction models, and here we specifically used the West region dataset. GUSTO-I was a comparative
TABLE 1 C-statistics and 95% confidence intervals (C.I.) for the GUSTO-I trial dataset for the 8-variable models

|                          | ML estimation | Ridge estimation | Lasso estimation |
|--------------------------|---------------|------------------|------------------|
| **Apparent C-statistic** |               |                  |                  |
| DeLong's C.I.            | 0.819 (0.783, 0.854) | 0.819 (0.784, 0.855) | 0.819 (0.787, 0.857) |
| Apparent bootstrap C.I.  | 0.819 (0.788, 0.858) | 0.819 (0.787, 0.858) | 0.819 (0.787, 0.857) |
| **Harrell's bias correction** |            |                  |                  |
| Location-shifted bootstrap C.I. | 0.810 (0.779, 0.849) | 0.811 (0.779, 0.850) | 0.811 (0.779, 0.849) |
| Two-stage bootstrap C.I. | 0.810 (0.777, 0.850) | 0.811 (0.787, 0.857) | 0.811 (0.784, 0.839) |
| **0.632 estimator**      |               |                  |                  |
| Location-shifted bootstrap C.I. | 0.810 (0.779, 0.849) | 0.812 (0.780, 0.851) | 0.811 (0.779, 0.849) |
| Two-stage bootstrap C.I. | 0.810 (0.777, 0.850) | 0.812 (0.788, 0.857) | 0.811 (0.784, 0.840) |
| **0.632+ estimator**     |               |                  |                  |
| Location-shifted bootstrap C.I. | 0.810 (0.779, 0.849) | 0.812 (0.780, 0.851) | 0.811 (0.779, 0.849) |
| Two-stage bootstrap C.I. | 0.810 (0.777, 0.850) | 0.812 (0.788, 0.857) | 0.811 (0.784, 0.840) |

clinical trial that assessed four treatment strategies for acute myocardial infarction. Here we adopted death within 30 days as the outcome variable. There were 17 covariates: two variables (height and weight) are continuous, one (smoking) is ordinal, and the remaining 14 (age, gender, diabetes, hypotension, tachycardia, high risk, shock, no relief of chest pain, previous myocardial infarction, hypertension history, hypercholesterolemia, previous angina pectoris, family history of myocardial infarction, and ST elevation in >4 leads) are binary; age was dichotomized at 65 years old. For smoking, which is a three-category variable (current smokers, ex-smokers, and never-smokers), we generated two dummy variables (ex-smokers vs never-smokers, and current smokers vs never-smokers) and used them in the analyses. The clinical trial dataset can be downloaded from http://www.clinicalpredictionmodels.org.

We considered two modeling strategies: (1) 8-predictor models (age > 65 years, female gender, diabetes, hypotension, tachycardia, high risk, shock, and no relief of chest pain), which were adopted in several previous studies, and (2) 17-predictor models, which included all the variables mentioned above. The EPVs for these models were 16.9 and 7.5, respectively. For the two modeling strategies, we constructed multivariable logistic prediction models by ML estimation and the two shrinkage penalized regression approaches, the ridge and lasso regressions. We applied the proposed methods to these prediction models, and calculated the C-statistics and the bootstrap-based confidence intervals. The number of bootstrap resamplings was consistently set to 2000; for the two-stage bootstrap methods, the total resampling number was $2000 \times 2000 = 4 000 000$. We assessed the Monte Carlo errors of the two-stage bootstrap confidence intervals by varying the resampling numbers, and the results are presented in e-Appendix B.

For the 8-predictor models, the results are presented in Table 1. The corrected optimisms from the apparent C-statistics were 0.009 for ML estimation and 0.007 to 0.008 for the ridge and lasso regressions. The 95% confidence intervals by DeLong's method and apparent bootstrap were located around the optimism-uncorrected C-statistic, and was a little influenced by the biases. For the ML estimation, the two proposed methods provided optimism-corrected confidence intervals, as expected. Further, for the ridge estimation, the locations of the 95% confidence intervals moved upward, that is, the bootstrap distributions of the optimized-corrected C-statistics moved upward. For the lasso regression, the lower confidence limits moved upward compared with the location-shifted bootstrap confidence intervals. On the other hand, the upper confidence limits moved downward, that is, the bootstrap distributions were limited to narrower ranges. The results were not very different between the three optimism correction methods, that is, the Harrell's bias correction, 0.632, and 0.632+ methods.

For the 17-predictor models, the results are presented in Table 2. The corrected optimisms from the apparent C-statistics were generally larger than those from the 8-predictor models, and became 0.021 to 0.022 for ML estimation and 0.018 to 0.019 for ridge and lasso regressions. The DeLong's and apparent bootstrap confidence intervals could be influenced by biases. The location-shifted bootstrap confidence intervals would correct the relatively large biases. For the two-stage bootstrap confidence intervals, the overall results were similar to those obtained with the ML estimation; they provided slightly wider confidence intervals compared with the location-shifted bootstrap confidence intervals. For the ridge estimation, the locations of the lower confidence limits moved upward, as did the bootstrap distributions of the
optimism-corrected C-statistics. These results indicate that the AUCs became larger. For the lasso regression, the lower confidence limits were very different from those of the location-shifted bootstrap confidence intervals, but the upper confidence limits moved downward. The confidence intervals indicate that the standard errors of the C-statistics became smaller but the prediction performances became worse as a result of the strong shrinkage with the lasso regressions. For the 17-predictor models, the results were also not very different among the three optimism correction methods, namely the Harrell, 0.632, and 0.632+ methods.

5 | SIMULATION EXPERIMENTS

5.1 | Simulation settings

To evaluate the performances of the proposed methods, we conducted simulation studies based on the GUSTO-I trial.13,14 We explain the settings of simulation studies following the ADEMP structure of Morris et al26 as follows.

Aims: The aims of the simulation studies were to assess the validities of the proposed methods and to compare their performances with those of the conventional standard methods.

Data-generating mechanisms: We considered a certain range of conditions with various factors that can affect predictive performance: the events per variable (EPV) (1, 3, 5, 7, 10, 20, and 40), the expected event fraction (0.125 and 0.0625), the number of candidate predictors (8 and 17 variables), and the regression coefficients of the predictor variables (two scenarios, as explained below). A total of 56 scenarios covering all combinations of these settings were investigated. The settings of the EPV and the event fraction were based on those used in previous studies.7,27 For the regression coefficients of the predictor variables (except for the intercept ϒ0), we considered two scenarios: one fixed to the ML estimate for the GUSTO-I dataset (coefficient type 1) and the other fixed to the lasso estimate for the same dataset (coefficient type 2). With coefficient type 1, all the predictors had some effect on the risk of events, while with coefficient type 2, some of the predictor effects were null and the others were relatively small compared with scenario 1; the two scenarios are considered as plausible settings estimated from the real world dataset using valid estimation methods. The intercept ϒ0 was set to properly adjust the event fractions. We also conducted additional simulations, shown in e-Appendix C, that varied the expected event fractions to 0.25 and 0.50. Also, the sample size of the derivation cohort n was determined as follows: (the number of candidate predictor variables × EPV)/(expected event fraction). The predictor variables were generated as random numbers based on the parameters estimated from the GUSTO-I dataset; for the details of the covariate information, see Section 4. Three continuous variables (height, weight, and age) were generated from a multivariate normal distribution with the same mean vector and covariance matrix as in the GUSTO-I data. For smoking, an ordinal variable, random numbers were generated from a multinomial distribution using the same proportions as in the GUSTO-I data; this variable was converted to two dummy variables before being incorporated into the prediction models.
The remaining binary variables were generated from a multivariate binomial distribution using the same marginal probabilities and correlation coefficients estimated from the GUSTO-I dataset. We used the \texttt{mipfp} package to generate the correlated binomial variables. We also conducted additional simulations, shown in e-Appendix C, that used resamples of the covariates from the original GUSTO-I dataset and that completely preserved the correlation structure. The event occurrence probability \( \pi_i \) \((i = 1, 2, \ldots, n)\) was determined based on the generated predictor variables \( x_i \) and the logistic regression model \( \pi_i = \frac{1}{1 + \exp(-\beta^T x_i)} \). The outcome variable \( y_i \) was generated from a Bernoulli distribution with a success probability \( \pi_i \).

**Estimands:** We evaluated the prediction performance of the \( C \)-statistic, which is the most popular discriminant measure for clinical prediction models. The estimand was set to the empirical AUC of the ROC curve for 500 000 independently generated external test observations. We estimated the AUC using the apparent \( C \)-statistic and the bootstrap-based optimism-corrected \( C \)-statistic.

**Methods:** We evaluated the coverage performances and precisions of 95\% confidence intervals for the AUC by (1) DeLong’s confidence interval, (2) the apparent bootstrap confidence interval, (3-5) the location-shifted bootstrap confidence intervals by the Harrell, 0.632, and 0.632+ methods, and (6-8) the two-stage bootstrap confidence intervals by the Harrell, 0.632, and 0.632+ methods. The number of bootstrap resamplings was consistently set to 1000; for the two-stage bootstrap methods, the total resampling number was 1000 \( \times 1000 = 1000000 \). The multivariable prediction model was constructed by ML estimation.

**Performance measures:** The coverage rates of the AUC and the expected widths were adopted as evaluation measures. We conducted 1000 simulations under each scenario, and empirical measures of these quantities were evaluated.

5.2 | **Results**

The results of the simulations are presented in figs. 1 to 3 using the nested-loop plots of Rücker and Schwarzer. Specifically, the results of the coverage rates of 95\% confidence intervals are shown in Figure 1, and the expected widths
FIGURE 2  The expected widths of 95% confidence intervals in the simulation studies by the DeLong method, the apparent bootstrap confidence interval, and the two-stage bootstrap (TSB) confidence intervals for the Harrell, 0.632, and 0.632+ methods; the expected widths of the location-shifted confidence intervals are consistent with those of the apparent bootstrap confidence interval [Colour figure can be viewed at wileyonlinelibrary.com]

of these confidence intervals in Figure 2. First, under most of the 56 scenarios, the apparent bootstrap confidence interval showed marked undercoverage properties. The coverage rates were especially small when EPV was small and/or the number of predictor variables was large. The DeLong’s confidence interval had similar trends and showed undercoverage properties, but the actual coverage rate was relatively large compared with the apparent bootstrap confidence interval. These results indicate that these naive methods lacking optimism corrections usually misestimate the statistical variability of prediction accuracy measures.

The proposed methods showed clearly more favorable coverage performances. For the location-shifted bootstrap confidence intervals, all three methods performed well and their coverage rates were around the nominal level (95%) when EPV was relatively large (≥10). However, they showed minor undercoverage properties in general, and the trend was fairly strong under relatively small EPV settings. The undercoverage properties were pronounced under especially small EPV settings (EPV = 1, 3). The undercoverage properties would be caused by the fact that the location-shifted bootstrap confidence intervals only quantify the statistical errors by the apparent bootstrap confidence interval, and may underestimate the total statistical variabilities as mentioned in Section 3. However, these results show that the estimated variation with the apparent bootstrap confidence interval certainly quantifies the actual statistical variabilities under relatively large sample sizes, although they can be easily computed by the bootstrap outputs using the ordinary optimism correction methods.

In addition, for the two-stage bootstrap confidence intervals, the coverage rates were superior to the location-shifted bootstrap confidence intervals in general. Under all the scenarios except for the settings with EPV = 1, the realized coverage rates were around the nominal level (95%), and their expected widths were slightly larger than the apparent bootstrap confidence intervals (ie, equal to those of the location-shifted confidence intervals). The differences of coverage rates with the location-shifted confidence intervals were markedly under those of the EPV ≤ 10 settings. However, under especially small EPV settings (EPV = 1), the coverage performances were not very good, and the confidence interval based
on Harrell's method showed undercoverage properties. That based on the 0.632 estimator also demonstrated undercoverage under a few scenarios. In contrast, the confidence interval using the 0.632+ estimator showed valid coverage rates; this estimator is therefore recommended under small EPV settings. Comparing coefficient types 1 and 2, the general performances of the three methods were comparable, but the coverage probabilities were generally larger under the former settings. In summary, the two-stage bootstrap methods provided values that were nearly identical to the nominal level under most of these settings. This indicates that the two-stage bootstrap can adequately assess the statistical variabilities of the optimism-corrected predictive measures.

6 | DISCUSSION

In developing multivariable prediction models, bootstrapping methods for internal validations of discriminant and calibration measures have been increasingly used in recent clinical studies. Although most published papers have presented the optimism-uncorrected confidence intervals (eg, DeLong’s method for the C-statistic), they sometimes provide inaccurate and misleading evidence. In this article, our simulations clearly showed the inadequacy of the naïve methods that do not consider optimism, and their use is not recommended in practice. Adequate alternative methods should be used instead.

In this article, we proposed two effective methods to construct the confidence intervals to address this important issue. The most highly recommended approach is to use the two-stage bootstrap methods. These adequately reflect the statistical variability of the optimism-corrected prediction measures, and they provide appropriate confidence intervals as shown in the simulation studies. However, one difficulty is that they have a heavy computational burden. Their application in practice will be difficult if analysts cannot access a high-performance computer that can conduct parallel computations. However, the prices of these high-performance machines have been gradually becoming more reasonable, and they will not be considered to be special tools in the future. A current solution is to collaborate with statisticians who have access to high-performance computer systems. We are willing to collaborate on valuable research projects upon request. Also, the computations can be easily implemented using the R package predboot (https://github.com/nomahi/predboot).

An alternative approach is the use of location-shifted bootstrap confidence intervals. Although these may underestimate statistical variabilities, as mentioned in Section 3.1, the actual coverage rates of true predictive accuracy measures approach the nominal level when the sample sizes are reasonably large. In our simulation studies, the coverage performances were favorable under EPV ≥ 20, and would be acceptable even under EPV = 10. They were certainly better than the Delong’s and apparent bootstrap confidence intervals. In addition, the location-shifted bootstrap confidence intervals can be computed by the outputs of bootstrap algorithms for calculating the optimism-corrected prediction measures, and a pragmatic advantage is that no additional computational burdens are needed.

In future methodological studies, alternative effective computational methods might be developed, such as those of LeDell et al., which combine analytical and Monte Carlo approaches. These are relevant issues for further research. Also, the proposed methods can be applied to other optimism correction methods by bootstrap, for example, the Efron-Gong method. Further theoretical discussions comparing the optimism correction methods, especially the Efron-Gong and 0.632 methods, would be future relevant issues. In summary, for the evaluations of predictive accuracies of multivariable prediction models, the naïve confidence intervals are no longer recommended, and appropriate methods should be adopted in practice. The proposed methods in this article certainly provide a pragmatic solution and can serve as effective practical tools.

ACKNOWLEDGEMENTS

This study was supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant number: JP19H04074).

DATA AVAILABILITY STATEMENT

The GUSTO-I West region dataset is available at http://www.clinicalpredictionmodels.org.

ORCID

Hisashi Noma https://orcid.org/0000-0002-2520-9949
REFERENCES

1. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.

2. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. New York, NY: Springer; 2019.

3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

4. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.

5. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc*. 1983;78(382):316-331.

6. Efron B, Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. *J Am Stat Assoc*. 1997;92(438):548-560.

7. Steyerberg EW, Harrell FE Jr, Borsboom GI, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54(8):774-781.

8. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

9. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat*. 2015;9(1):1583-1607.

10. Jiang W, Varma S, Simon R. Calculating confidence intervals for prediction error in microarray classification using resampling. *Stat Appl Genet Mol Biol*. 2008;7(1):8.

11. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Dent Tech*. 1970;12(1):55-67.

12. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996;58(1):267-288.

13. GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med*. 1993;329(10):673-682.

14. Lee KL, Woodlief LH, Topol EJ, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. *Circulation*. 1995;91(6):1659-1668.

15. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. *JAMA*. 2017;317(10):1068-1069.

16. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105-1117.

17. Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley; 2013.

18. Lee AH, Silvapulle MJ. Ridge estimation in logistic regression. *Commun Stat Simul Comput*. 1988;17(4):1231-1257.

19. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.

20. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York, NY: CRC Press; 1994.

21. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge, MA: Cambridge University Press; 1997.

22. Iba K, Shinozaki T, Maruo K, Noma H. Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Med Res Methodol*. 2021;21(1):9.

23. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*. 1999;52(10):935-942.

24. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000;19(8):1059-1079.

25. Steyerberg EW, Eijkemans MJ, Habbema JD. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl*. 2001;55(1):76-88.

26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.

27. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474.

28. Dai B, Ding S, Wahba G. Multivariate Bernoulli distribution. *Ther Ber*. 2013;19(4):1465-1483.

29. Barthélemey J, Susse T. mifp: an R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *J Stat Softw*. 2018;86:Code Snippet 2.

30. Rücker G, Schwarzer G. Presenting simulation results in a nested loop plot. *BMC Med Res Methodol*. 2014;14:129.

31. Pencina MJ, D’Agostino RB. Evaluating discrimination of risk prediction models: the C statistic. *JAMA*. 2015;314(10):1063-1064.

32. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat*. 1983;37(1):36-48.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.