Distribution Agreement

Signature:

______________________________ _______________

Zhiwei Zhao Date

Assessment of diagnostic accuracy after biomarker combination in the same study: The issue of over-optimism and potential solutions

By

Zhiwei Zhao
Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____
Yijian(Eugene) Huang, PhD
Committee Chair

____
Yuan Liu, PhD
Committee Member

Assessment of diagnostic accuracy after biomarker combination in the same study: The issue of over-optimism and potential solutions

By

Zhiwei Zhao

B.S.
Xiamen University
2017

Thesis Committee Chair: Yijian(Eugene) Huang, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2019

Abstract

Assessment of diagnostic accuracy after biomarker combination in the same study:
The issue of over-optimism and potential solutions

By Zhiwei Zhao

In disease diagnosis, biomarker combination is an important method in disease diagnosis since it is usually not enough to consider only a single marker. There are a few studies focusing on the biomarker combination rules. In practice, it will be ideal if researchers have independent training and validation datasets. However, it is usually not the case in real word. In fact, it is well-known that using single dataset for both development and evaluation of a combination rule could produce over optimism problem. In this thesis, we are trying to address this problem. We used logistic regression to generate the combination rule. Then, area under the ROC curve (AUC) was used as the assessment method. The k-fold cross-validation was used in order to solve the over optimism. To reduce the bias, we proposed and introduced a two-sample jackknife bias-reduced approach as well as bootstrap bias-reduced approach. As for inference, bootstrap was introduced to estimate the standard error and a double bootstrap was proposed to improve the estimate for bootstrap bias-reduced estimators. A prostate cancer data was used as an illustration of the aforementioned methods in real word application.

Assessment of diagnostic accuracy after biomarker combination in the same study: The issue of over-optimism and potential solutions

By

Zhiwei Zhao

B.S.
Xiamen University
2017

Thesis Committee Chair: Yijian(Eugene) Huang, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2019

Contents

1 Introduction

Biomarker is a portmanteau of "biological marker", which refers to a subcategory of medical signs (Strimbu and Tavel, 2010)[1]. It is the most quantifiable and objective one among medical signs. The importance of biomarker study is increasing in the areas of streamlining drug research, diagnostic disease, medical personalization and clinical endpoints surrogation. The disease diagnosis with biomarkers is the topic of our special interest. But a single biomarker is usually not sufficient enough to get a precise diagnostic result as most biomarkers reveal complementary information. For example, in prostate cancer we usually have more than one subtypes and for each subtype, some biomarkers could be really informative whereas at the same time, they could be totally non-informative for other markers (Kornberg et al, 2018)[2]. In this situation, only using a single biomarker as diagnostic rule can misdiagnose some subtypes as no-disease if the marker used in the study happened to be the one only informative for a specific subtype (Chan et al, 2013; Sanda et al, 2017)[3, 4]. Thus, biomarker combination is an important topic in medical research as it is supposed to have a better performance in diagnosis. Several other studies which are related to atherosclerotic coronary heart disease, osteoarthritis (OA) and Alzheimer's disease (AD) has also emphasized the necessity of biomarker combinations (review Liu et al, 2005; Frolich et al, 2017 and Williams, 2009 for more details)[5, 6, 7]. Also, it has been shown that combinations of biomarkers may lead to more sensitive screening rules of cancer detection (Etzioni et al. 2003) [8]. These increasing applications of biomarker combinations to facilitate the diseases diagnosis highlight the need of careful assessment of the performance of the diagnostic rules (i.e. biomarker combinations in our study).

Plenty of studies have been conducted about the way of biomarker combinations. Neyman-Pearson Lemma has been used to diagnosis which led to the insights about combination of multiple markers when the joint distribution of biomarkers is known for disease and controls (Green and Swets, 1966; Egan, 1975; Baker, 2000)[9, 10, 11]. The nonparametric method (Baker, 2000) [11] , logistic regression and boosting (Qu et al, 2002 and Yasui et al, 2003) [12, 13] has been used when the joint distributions of biomarkers are unknown for disease and control groups (see Feng and Yasui, 2004 for more reviews of these methods)[14]. In addition, it has been proved that logistic regression could produce an optimal rules for combinations when it holds(McIntosh and Pepe, 2002)[15]. And another useful way for combination of biomarkers is defined by a linear discriminant function which maximizes the area under the operating characteristic (ROC) curve (Su and Liu, 1993) [16], but this method is limited to the multivariate normally distributed biomarkers and some extra works has been done to relax this assumption (Pepe and Thompson, 2000) [17].

There are many methods used as the assessments of diagnostic rules. The measure of area under the ROC curve (AUC) is a widely used index of the assessment of diagnostic accuracy. It can be interpreted as an index which measures the distance between the distributions of scores for diseased and disease-free subjects, in a distribution-free sense (Pepe and Thompson, 2000) [17]. Here ”score” is a combination of biomarkers which is used in diagnosis. There are other measures of accuracy such as true-positive rate (TPR) and false-negative rate (FPR). These are methods which depend on the setting of implemented technology with varied optimal threshold (Dodd and Pepe, 2003) [18].

In practice, the intention of getting an accurate diagnostic rule also requires less over-optimism issues as well as low biases. One potential and the most prominent source of bias among the diagnostic tests is the interpretation of the "gold-standard" test and diagnostic test results should be independent and blinded (Sahpiro 1999; Begg 1991; Begg and McNeil 1998) [19, 20, 21]. Research should be especially careful when using the same dataset to both develop and evaluate a diagnostic rule. It is well known that over-optimism may arise if the performance is evaluated with the same derivation data. However, these kind of discussions are more focusing on the diagnostic models instead of a specific consideration of the statistics used for accuracy evaluation such as AUC (Schutte et.al, 2011)[22]. Nevertheless, with the risk of over-optimism, researchers continue using the same dataset to develop and assess the rules especially when the sample size is limited.

In this thesis, we used AUC to evaluate the biomarker combinations and focused on the over-optimism issues as well as the high biases when only single dataset is available. We reviewed the existing methods for estimation and inference of AUC as an assessment of combination rules (logistic regression was used in our special case). We introduced new methods to solve problems of over-optimism and bias reduction, conducted simulations to compare these methods and applied to a prostate cancer study. In simulations, one general setup was analyzed, but a special situation was also considered. In the illustrated case study, a study based on the research from Sanda (2017)[4] was used with 514 observations, in which three important biomarkers were combined and the performance was evaluated by the proposed methods mentioned in simulation study. All simulations and analysis were conducted in R (version: 3.5.2).

2 Methods

2.1 Notations

To describe the technical procedures, $\mathbf{Y} = (Y_1, ..., Y_p)$ is considered as predictors (or biomarkers). The diagnostic rule then will be written as $L(\mathbf{Y}) = \alpha_1 Y_1 + \cdots + \alpha_p Y_p$ and will be called as "score". Here, $\alpha_1, \ldots, \alpha_p$ is the linear combination coefficients. If logistic regression holds, they are the coefficients for logistic regression. Otherwise, if the logistic regression violates, they will be the limits of estimated coefficients under large sample size. And the prediction rule will be invariant even if multiplying a constant to the score function.

Particularly, in our study, we need to consider situations with cases and controls. Thus, we define $\{\underset{\sim}{Y}_{D1}, ..., \underset{\sim}{Y}_{Dn_D}\}$ and $\{\underset{\sim}{Y}_{\bar{D}1}, ..., \underset{\sim}{Y}_{\bar{D}n_{\bar{D}}}\}$ as observed biomarker vectors for cases and controls respectively, where $\underset{\sim}{Y}_{Di} = (Y_{D,i1}, \ldots, Y_{D,ip})$ represents the observed biomarkers for the i-th subject in cases and $\underset{\sim}{Y}_{\bar{D}i} = (Y_{\bar{D},i1}, \ldots, Y_{\bar{D},ip})$ represents those for the i-th subject in controls. Furthermore, the estimated coefficients in logistic regression will be defined as $\{\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_p\}$ and the estimated score function will be defined as $\hat{L}(\mathbf{Y}) = \hat{\alpha}_1 Y_1 + \cdots + \hat{\alpha}_p Y_p$.

2.2 Point Eatimation

2.2.1 Standard Approach

In diagnostic studies, when the number of biomarkers is finite, the combination rules will be asymptotically approximate the true results if observations are infinite. However, it is usually not true. The simplest and most common method to evaluate the combination accuracy is using a single dataset for diagnostic rule generation as well as evaluation. Such an approach provides a very fast analysis as well as a very simple operation. However, it ignores the correlation between the training dataset and testing dataset and hence lead to over-optimism and high-biased results. We used this approach as a reference (or standard) approach in our study and compared the performances of other methods with it.

Specifically, we ran the logistic regression on the whole dataset and then calculated the AUC for assessment. Theoretically, AUC is the probability of scores for cases greater than the probability of scores for controls. That is, $AUC = Pr(L(\mathbf{Y}_{cases}) > L(\mathbf{Y}_{controls}))$, where $\mathbf{Y}_{cases}$ and $\mathbf{Y}_{controls}$ are the biomarker vectors for case and control.

In fact, the AUC can be estimated by the frequencies of scores for cases that exceed scores for controls plus 0.5 multiplied by the frequencies of equal scores as an adjustment, which is:

$$\widehat{AUC} = \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \{I[L(\underset{\sim}{Y}_{Di}) > L(\underset{\sim}{Y}_{\bar{D}j})] + \frac{1}{2} I[L(\underset{\sim}{Y}_{Di}) = L(\underset{\sim}{Y}_{\bar{D}j})]\},$$

where n_D and $n_{\bar{D}}$ are numbers of observations for cases and controls and the definition for $\underset{\sim}{Y}_{Di}$ and $\underset{\sim}{Y}_{\bar{D}j}$ are the same as in notation part.

2.2.2 Cross-Validation

In an ideal situation, there should exist two independent datasets so that an independent validation dataset is available to evaluate the diagnostic rule. However, in practice, it is often not the case and hence there usually exists over-optimism issues. In this situation, the use of cross-validation (CV) is a typical method to solve this problem since the procedure of CV will assess the result of a statistical analysis on an independent dataset and the use of an independent dataset during the development phase are supposed to reduce over-optimism (Boulesteix, 2015)[23]. The general procedure for CV is to partition the data into subsets for training and testing and there are several ways to do it such as leave-one-out cross validation (LOOCV), generalized cross validation (GCV), leave-K-out cross validation (LKOCV) and k-fold cross validation (see Syed, 2011 as a review of these methods)[24].

A k-fold cross-validation method was used in our study. This method randomly divides the data into k parts, one part will be used as the testing and the rest as the training. K times iterations are executed, and the average of the results is the final result. So, the estimator function will be:

$$\widehat{AUC}_{CV} = \frac{1}{k}\sum_{i=1}^{k} \widehat{AUC}_{CV,i},$$

where $\widehat{AUC}_{CV,i}$ is the cross-validation estimator using i-th sub-dataset as testing. This kind of result from k-fold cross-validation tends to be conservative which produces negative bias estimators, but the effect of bias reduction for AUC here is disputed (more details could be referred in simulation and discussion parts).

2.2.3 Jackknife Bias Correction

Jackknife has been proposed initially for bias reduction (Quenouille, 1949) and then been developed as a tool to estimate variance. In most situation, this jackknife approach is used for one-sample case where only single sample is considered. But it is less applicable in our study as we usually have two independent samples of cases and controls. Schechtman and Wang (2004) [25] has proposed a two-sample jackknife bias-reduced estimator, but it has a heavy calculation load. Therefore, we proposed a new jackknife estimator based on the logic of the one-sample situation but is more reasonable in a two-sample situation.

Similar as the way to develop the one-sample jackknife bias-reduced estimator, this jackknife estimator is derived based on the bias estimator from a two-sample estimating expectation. To derive this method, we consider:

$$E(\widehat{AUC}) \approx AUC_0 + \frac{\alpha}{n_D} + \frac{\beta}{n_{\bar{D}}} \tag{1}$$

where α and β are unknown constant which we want to estimate, and AUC_0 is the theoretical AUC under a specific combination rule.

Then, consider two estimators based on the average of delete-one estimators for cases and controls, respectively. It means we will estimate the AUC based on delete one sample from cases each time and iterated for the number of cases times. Then these estimators will be the delete-one estimators for cases. And do similar things for controls to get the delete-one estimators for controls. The expectation for these two estimators will be:

$$E(\widetilde{AUC}_{D(\cdot)}) \approx AUC_0 + \frac{\alpha}{n_D - 1} + \frac{\beta}{n_{\bar{D}}}$$

and

$$E(\widetilde{AUC}_{\bar{D}(\cdot)}) \approx AUC_0 + \frac{\alpha}{n_D} + \frac{\beta}{n_{\bar{D}} - 1}$$

where $\widetilde{AUC}_{D(\cdot)}$ is the average of delete-one AUC estimators for cases and $\widetilde{AUC}_{\bar{D}(\cdot)}$ is that of controls. Hence, the two unknown constant could be estimated by:

$$\hat{\alpha} \approx -(\widehat{AUC} - \widetilde{AUC}_{D(\cdot)})n_D(n_D - 1) \tag{2}$$

$$\hat{\beta} \approx -(\widehat{AUC} - \widetilde{AUC}_{\bar{D}(\cdot)})n_{\bar{D}}(n_{\bar{D}} - 1) \tag{3}$$

By subtracting the bias terms in equation (1) using the estimated results from equation (2) and (3) (i.e. subtract $\frac{\hat{\alpha}}{n_D}$ and $\frac{\hat{\beta}}{n_{\bar{D}}}$) from the estimated $\widehat{AUC}$, the estimator equation is:

$$\widehat{AUC}_{jack} = (n_D + n_{\bar{D}} - 1)\widehat{AUC} - (n_D - 1)\widetilde{AUC}_{D(\cdot)} - (n_{\bar{D}} - 1)\widetilde{AUC}_{\bar{D}(\cdot)}$$

2.2.4 Bootstrap Bias Correction

Bootstrap is a common method to estimate statistics such as mean, variance or quantiles. Efron and Tibshirani (1993) [26] mentioned a bootstrap bias-reduced estimator and its improvement, but they are hardly used in biomarker studies. The purpose of bias-reduced bootstrap estimator is also to reduce a bias term. Then the estimator function is,

$$\begin{aligned}\widehat{AUC}_{boot} &= \widehat{AUC} - (\frac{1}{B}\sum_{k=1}^{B}\widehat{AUC}_k^{\star} - \widehat{AUC}) \\ &= 2\widehat{AUC} - \frac{1}{B}\sum_{k=1}^{B}\widehat{AUC}_k^{\star}\end{aligned}$$

where $\widehat{AUC}$ is also the standardized evaluation method, $\widehat{AUC}_k^{\star}$ is the estimator in each bootstrap step and B is the iteration times for bootstrap procedure.

In addition, Efron and Tibshirani (1993) [26] had also proposed an improvement estimator based on this method. This improvement is supposed to have a better convergence rate than the previous one. In other words, the number of iteration times to a stable value for bootstrap should be smaller for the improvement than the old estimator.

Based on their concepts,

$$\widehat{AUC}_{bootimprove} = \widehat{AUC} - (\frac{1}{B}\sum_{k=1}^{B}\widehat{AUC}_k^{\star} - \widetilde{AUC}) \tag{4}$$

where $\widetilde{AUC}$ is a new estimator based on the average of bootstrap distributions. In this study, $\widetilde{AUC}$ was estimated by adding a weight within logistic and AUC calculation procedures. And the weight is calculated based on the average of observation frequencies in each bootstrap iteration.

2.3 Variance Estimation

In large sample study, bootstrap is a good method to estimate statistics like estimating variance or mean. In our study, we will use it to estimate the variance. But the situation is complicated since we need to think about the randomness in both the risk score (or score function) estimator (which comes from the coefficients estimators in logistic regression) and the AUC estimator. Specifically, in each iteration, the dataset from bootstrap ran a logistic regression and then AUC was calculated. The variance then would be the variance of the estimators based on the bootstrap samples. We used these estimated variances to construct the 95% confidence intervals for all of the methods above by re-centering the Wald CI's for each approach. Here, re-centering means to calculate the Wald CI based on estimators of each method.

For standard approach, this construction of confidence interval was proper since these bootstrap variances were the estimated variance for standard estimators. In large sample study, these variances were supposed to asymptotically approximate the variances for cross-validation estimators, jackknife bias-reduced estimators or bootstrap bias-reduced estimators. However, in our study with relatively small sample size, using the variances estimated based on the standardized approach was not so appropriate. With this concern, a double-bootstrap for the bootstrap bias-reduced approach was also discussed in our study in order to improve the inference accuracy.

The way to do that was to resample the data by bootstrap and then applied the bootstrap bias-reduced estimator method on this resampled data. So, there would be an outside bootstrap procedure which resampled the data and would also be an inside bootstrap procedure which estimated the bootstrap bias-reduced estimator. Then, the variance was the variance of the estimators for the outside bootstrap samples. Also, we considered the coverage probabilities of the confidence intervals in simulation study and we will talk more about this later.

Additionally, a log-transformation of these confidence interval construction was also considered. Log-transformation is a spontaneous method people will do for statistics lay between 0 and 1. Here we will take the log-transformation of AUC estimators and construct the confidence interval by considering the variance (or standard deviation) of the log transform.

3 Simulations

3.1 Situation with only one informative biomarker

In a basic simulation, supposing independent case and control groups, each group contained the same number of observations. Biomarkers in both case and control groups were supposed to be multivariate normally distributed and uncorrelated. For convenience, the identity matrix were used as the covariance matrix. In the basic simulation study, only one biomarker in case group was considered to be informative (i.e. having a mean not equal to 0). In this case, unknown parameter mean for this informative biomarker could be derived with $\mu = \sqrt{2}\Phi^{-1}(AUC_0)$, where Φ^{-1} was the cumulative distribution function of standard normal distribution and AUC_0 was the theoretical area under the curve which was self-defined. The other biomarkers were considered to be non-informative and having 0 means. For the bias estimator, $Bias(\widehat{AUC}) = E(\widehat{AUC}) - AUC_0$ was applied. In order to get the expectation, $1,000$ datasets were generated by simulation, for each dataset, estimated $AUC's$ were calculated using the aforementioned methods. Then, the average of the $1,000$ estimators could approximate the expectation $AUC's$ and the biases could be calculated.

In order to explore a coverage probability of the confidence intervals for each iteration, we used the bootstrap method as we mentioned in the method part to estimate the variance of each estimator and to construct the 95% Wald CI's. Then the coverage probability was the percentage of intervals which contained the true AUC_0. Also, aforementioned double bootstrap was applied to bootstrap bias-reduced estimators in order to get more reasonable coverage probabilities.

3.2 Situation with subtypes

At the same time, a special setup was also under our interests as in real world dataset would be more complex and biomarkers could be more complicatedly distributed. One typical situation in real life is there are subtypes of disease and some biomarkers are informative for a specific subtype but are non-informative for other subtypes. In this case, the previous simulation seemed to be provided less information. In our study, to simplify the simulation, we considered 3 subtypes of disease, with each subtype having one informative biomarker. In other words, 3 biomarkers were simulated, but each marker would be informative only for one subtype. Also, equal numbers of subtypes were sampled, since with this balanced dataset, the final biomarker combination could take the average of the combinations for each subtype due to the symmetric distribution of markers in our simulation. Then, the mean for the informative biomarker fo each subtype would be $\mu = \sqrt{6}\Phi^{-1}(AUC_0)$, and the covariate matrix would be a diagonal matrix with $\frac{1}{\sqrt{3}}$ diagonals. Again, $1,000$ datasets were simulated, and the biases calculation as well as the coverage probabilities construction were as the same as previous.

3.3 Results

Table 1 and 2 shows the main results of our assessments for the basic structure including the bias, standard deviation and coverage probability. Four basic setups were considered, with the AUC_0 be either 0.6 or 0.8 and the number of biomarker as well as the number of observations differed. We can observe the bias to standard deviation ratios tend to be larger for standard approach and those for bias-reduction method (jackknife and bootstrap) are much smaller. The ratios for cross-validation method is at the middle between standard and bias-reduction methods. The estimated results from cross-validation (either 3-fold or 10-fold) are conservative, with negative ratios exist which represented negative biases. The standard deviation within the 1,000 iterations for standard method is close to the standard error estimated by bootstrap method. And comparing to the standard deviation for the standard approach, all of the rest results have higher standard deviations and jackknife performs the worst.

The coverage probabilities are all less than expected value of 95%. The worst coverage probabilities are from jackknife estimators with only nearly 50% $\sim$ 70% confidence intervals contained the true AUC_0. Those two cross-validation methods also have lower coverage probabilities with most of the setups having 70% $\sim$ 80% coverage. Bootstrap methods have good coverage probabilities with most of them are around 90%. Log-transformation is not really helpful with respect to improving the coverage probabilities, on the contrary, the results for log-transform are somehow poorer.

The results for improved bootstrap bias-reduced estimator is not included in these tables as it performed similar to that of bootstrap bias-reduction and the number of bootstrap iteration times is less considerable since it could be calculation burdensome if we want to include a large iteration number in our simulations.

Table 3 compares different method for calculation of coverage probabilities for bootstrap bias-reduced estimators. We compared the probabilities calculated by recentered confidence interval (which is based on the estimated standard error for standard approach) and those calculated by double bootstrap method. The coverage probabilities after doing double bootstrap is closer to the true confidence level (95%) than that of re-centered method, the estimated standard errors are higher for double bootstrap and the log-transformation helps with improvement of coverage probabilities for double bootstrap method. Only one setup in our simulations was applied with double bootstrap method since the heavy calculation load of this method. But the results here are encouraging.

Table 4 shows the results for setup with three subtypes and with one biomarker informative for each subtype. We only included the results for $AUC_0 = 0.6$ due to the intensive calculation time but most of the results were pretty similar to what we have in situation with only one informative biomarker. The bias to standard deviation ratio is still large for standardized method. But it can also be noticed that, the ratio for 3-fold cross-validation is poorly performed. Actually, the results for 3-fold CV are in general the worst with respect to all the statistics we estimated but also the 3-fold results in table 2 are not obviously performed so badly. For the rest methods, jackknife has the highest standard deviation as well as the poorest coverage probability and standard method has the lowest SD and the highest coverage.

Table 1: Main Results: Bias and SD for standard, 3-forld and 10-fold cross-validation, jackknife and bootstrap bias-reduced estimators

	No. of	Standard				10-fold Cross-Validation			3-fold Cross-Validation			Jackknife(two-sample)			Bootstrap		
Setups	Biomarkers	AUC	Bias/SD	SD	SE	AUC	Bias/SD	SD	AUC	Bias/SD	SD	AUC	Bias/SD	SD	AUC	Bias/SD	SD
AUC_0	2	0.618	0.355	0.050	0.051	0.580	-0.258	0.077	0.573	-0.328	0.081	0.590	-0.062	0.158	0.599	-0.008	0.064
$= 0.6$	3	0.631	0.670	0.047	0.051	0.572	-0.368	0.076	0.564	-0.442	0.081	0.601	0.004	0.156	0.605	0.076	0.0612
$n_{\bar{D}} = 50$	4	0.643	0.889	0.048	0.050	0.562	-0.454	0.084	0.556	-0.534	0.082	0.605	0.029	0.175	0.608	0.127	0.0594
$n_D = 50$	5	0.655	1.158	0.048	0.049	0.559	-0.477	0.086	0.555	-0.526	0.087	0.603	0.019	0.175	0.609	0.147	0.061
AUC_0	2	0.608	0.205	0.039	0.038	0.590	-0.203	0.050	0.586	-0.266	0.052	0.595	-0.045	0.120	0.599	-0.023	0.043
$= 0.6$	3	0.615	0.415	0.037	0.037	0.583	-0.327	0.052	0.578	-0.398	0.056	0.604	0.037	0.116	0.598	-0.038	0.0434
$n_{\bar{D}} = 100$	4	0.624	0.024	0.036	0.037	0.580	-0.376	0.053	0.574	-0.459	0.056	0.6004	0.003	0.124	0.602	0.045	0.045
$n_D = 100$	5	0.629	0.029	0.035	0.036	0.575	-0.464	0.054	0.569	-0.542	0.058	0.599	-0.006	0.135	0.607	0.152	0.044
AUC_0	2	0.805	0.125	0.044	0.043	0.796	-0.085	0.048	0.794	-0.132	0.049	0.8003	0.005	0.072	0.780	-0.006	0.046
$= 0.8$	3	0.809	0.216	0.042	0.042	0.789	-0.218	0.049	0.786	-0.280	0.051	0.795	-0.065	0.081	0.801	0.018	0.045
$n_{\bar{D}} = 50$	4	0.812	0.278	0.045	0.041	0.783	-0.320	0.054	0.777	-0.413	0.057	0.798	-0.028	0.0890	0.8001	0.002	0.045
$n_D = 50$	5	0.818	0.415	0.043	0.040	0.779	-0.383	0.055	0.773	-0.462	0.058	0.799	-0.008	0.092	0.799	-0.028	0.047
AUC_0	2	0.802	0.065	0.032	0.030	0.798	-0.064	0.033	0.797	-0.107	0.033	0.8004	0.007	0.049	0.7999	-0.004	0.032
$= 0.8$	3	0.804	0.144	0.030	0.030	0.795	-0.160	0.033	0.793	-0.215	0.033	0.797	-0.046	0.056	0.798	-0.053	0.032
$n_{\bar{D}} = 100$	4	0.808	0.258	0.030	0.030	0.794	-0.185	0.034	0.791	-0.271	0.034	0.801	0.009	0.063	0.799	-0.027	0.033
$n_D = 100$	5	0.809	0.322	0.029	0.029	0.790	-0.295	0.033	0.787	-0.388	0.034	0.801	0.014	0.062	0.802	0.064	0.032

Table 2: Main Results: Coverage Probabilities for standard, 3-forld and 10-fold CV, jackknife and bootstrap bias-reduced estimators

	No. of	Standard		10-fold Cross-Validation		3-fold Cross-Validation		Jackknife(two-sample)		Bootstrap	
Setups	Biomarkers	Coverage Probability	Prob after log-trans	Coverage Probability	Prob after log-trans	Coverage Probability	Prob after log-trans	Coverage Probability	Prob after log-trans	Coverage Probability	Prob after log-trans
AUC_0	2	92.5%	90.1%	83.2%	80.5%	77.2%	75.3%	55.4%	–	86.9%	86.0%
$= 0.6$	3	89.1%	85.7%	80.2%	75.8%	75.8%	72.2%	48.2%	–	88.4%	85.1%
$n_D = 50$	4	86.1%	83.7%	76.5%	69.4%	72%	66.7%	46.9%	–	89.7%	85.0%
$n_{\bar{D}} = 50$	5	82.1%	76.8%	73%	64.3%	70.9%	64.3%	44.6%	–	87.7%	82.8%
AUC_0	2	91.4%	90.4%	87%	85.7%	83.5%	81.5%	54.5%	–	90.0%	89.4%
$= 0.6$	3	91.8%	90.4%	82.8%	80.2%	78.4%	75.7%	45.7%	–	88.0%	85.9%
$n_{\bar{D}} = 100$	4	91.3%	88.3%	81.6%	77.6%	78.1%	73.7%	45.7%	–	88.0%	85.9%
$n_D = 100$	5	85.6%	82.4%	78.4%	75.2%	75.3%	71.2%	43.4%	–	87.4%	83.4%
AUC_0	2	92.1%	91.8%	90.9%	90.0%	90.8%	90.4%	75.1%	–	91.3%	90.9%
$= 0.8$	3	91.4%	90.6%	89.1%	87.2%	88.5%	87.2%	69.7%	–	92.3%	91.4%
$n_{\bar{D}} = 50$	4	88.3%	87.7%	84.7%	82.8%	83.0%	80.7%	63.9%	–	92.0%	90.6%
$n_D = 50$	5	87.3%	86.5%	83.5%	81.0%	81.9%	78.7%	62.7%	–	89.7%	88.8%
AUC_0	2	91.5%	91.6%	92.7%	92.1%	92.8%	92.4%	77.5%	–	93.0%	93.0%
$= 0.8$	3	93.4%	93.2%	92.3%	90.8%	90.9%	89.9%	71.6%	–	93.5%	92.1%
$n_{\bar{D}} = 100$	4	92.3%	91.4%	90.1%	89.6%	89.5%	87.6%	65.0%	–	91.7%	90.6%
$n_D = 100$	5	88.4%	87.2%	91.3%	89.7%	87.7%	86.5%	64.8%	–	92.6%	92.1%

Table 3: Coverage Probabilities Comparison for Bootstrap Bias-Reduced Estimators

setups	No. of Biomarkers	Re-centered Method			Double Bootstrap		
		coverage	log-trans Prob	SE	coverage	log-trans Prob	SE
$AUC_0 = 0.6$	2	86.9%	86.0%	0.0508	92.0%	94.3%	0.0590
$n_{\bar{D}} = 50$	3	88.4%	85.1%	0.0503	91.9%	95.5%	0.0569
$n_D = 50$	4	89.7%	85.0%	0.4970	92.6%	95.3%	0.0556
	5	87.7%	82.8%	0.0493	90.7%	95.2%	0.0557

Table 4: Results for setups with only one informative biomarker for each subtype, 3 subtypes

3 biomarkers	AUC	Bias/SD	SD	Coverage Probabilities
$AUC_0 = 0.6$,$n_D = 50$,$n_{\bar{D}} = 50$				
Standardized Method	0.6443	0.8735	0.0507	86.4%
Jackknife Method	0.6208	0.1304	0.1593	48.3%
Bootstrap Method	0.6205	0.3109	0.0660	84.9%
Cross Validation				
3-fold	0.4962	-0.9825	0.1057	50.1%
5-fold	0.5589	-0.4616	0.0891	74.4%
10-fold	0.5761	-0.2900	0.0826	80.2%
$AUC_0 = 0.6$,$n_D = 100$,$n_{\bar{D}} = 100$				
Standardized Method	0.6321	0.8350	0.0373	88.8%
Jackknife Method	0.6176	0.1598	0.1099	51.5%
Bootstrap Method	0.6192	0.4393	0.0437	88.5%
Cross Validation				
3-fold	0.4989	-1.2879	0.0785	38.9%
5-fold	0.5706	-0.4765	0.0617	77.5%
10-fold	0.5895	-0.1908	0.0552	83.9%

4 Real Data Application

For illustration, we applied the various estimation and inference methods to a prostate cancer study. The data for this illustration were taken from a cohort study investigating the effect of combined T2:ERG and PCA3 on detection of aggressive prostate cancer(Sanda et al,2017)[4]. In their study, they want to evaluate the priori primary hypothesis that combined measurement of PCA3 and T2:ERG RNA in the urine after digital rectal examination would improve specificity over measurement of prostate-specific antigen alone for detecting prostate cancer. As a result, the use of combining testing of these two biomarkers improved the specificity twice (from 18% to 39%), with respect to predicting aggressive prostate cancer at initial biopsy. There are 514 eligible participants (156 cases and 358 controls) from 748 previous prospective cohort participants. T2:ERG and PCA3 as well as pre-biopsy prostate-specific antigen (PSA) were the three biomarkers of interest in our study, other covariates including age, family prostate cancer history, race and digital rectal exam (DRE) were also been considered. All of the above methods for accuracy assessment were applied in this dataset, bootstrap method was also used to estimate the estimated variance and to construct the 95% confidence interval.

Table 5 shows the results for real data analysis. We can see the estimated AUC are really close to each other comparing all the methods, but it could still be noticed that the standard method had the slightly higher AUC. The log-transformed confidence intervals tend to be wider. The number of bootstraps for the bootstrap bias-reduction method is 2,000. The reason we used a larger iteration time is the results tend to be more stable around 2,000 compared to a small iteration times such as 100.

Table 5: Results for real data analysis

	Model with Biomarkers Only			With Covariates[1]		
	AUC	CI	log-CI	AUC	CI	log-CI
Standard	0.7588	(0.713,0.805)	(0.698,0.819)	0.7580	(0.712,0.804)	(0.697,0.819)
Jackknife	0.7585	(0.708,0.799)	(0.693,0.814)	0.7578	(0.673,0.765)	(0.658,0.780)
Bootstrap[2]	0.7580	(0.712,0.804)	(0.698,0.819)	0.7578	(0.711,0.804)	(0.696,0.818)
C-V						
3-fold	0.7569	(0.711,0.803)	(0.696,0.818)	0.7533	(0.707,0.780)	(0.692,0.814)
5-fold	0.7552	(0.709,0.801)	(0.695,0.816)	0.7573	(0.709,0.801)	(0.694,0.816)
10-fold	0.7581	(0.712,0.804)	(0.698,0.819)	0.7548	(0.711,0.803)	(0.696,0.818)

[1] Adjusted for covariates of family prostate cancer history, age, race and digital rectal exam.

[2] Number of bootstrap iterations is 2,000.

5 Discussion

In this study, we investigated the potential over-optimism problem when using the same dataset for both developing and evaluating the methods in biomarker studies. As a result, the over-optimism problem did exist especially when the true area under the curve was relatively small or most of the biomarkers were non-informative. As expected, the k-fold cross-validation worked on solve the over-optimism problem and tended to have conservative results. However, what is less known, as demonstrated in our study, this method may also have considerable bias although in the other direction. The 3-fold cross-validation, not surprisingly, performed worse especially in the setup with 3 subtypes. The reason is when doing 3-fold CV, the training dataset will be relatively smaller and provide less information for development of combination rules. But we should also notice that the bad performances of 3-fold were not so bad in the basic setups. It could be explained by the fact that in basic setups, the biomarkers are all informative or non-informative across the disease so the information in training and testing should be balanced no matter how to divide the data, but in the special setup with 3 subtypes, the biomarker are not all informative across disease. Therefore, when the data was randomly divided, the unbalanced subtypes in training and testing can lead to bad results since the information in training and testing could be also unbalanced. By comparing the standard deviations and standard errors for standardized method, we can see the poor performance of the coverage probabilities was caused by the bias since SD and SE are close to each other.

The jackknife bias-reduced estimator reduced the bias well, but it seemed to produce high variance estimators. It is reasonable to have an increasing variance in bias-reduced estimator as there are trade-offs between bias and variance when we are doing evaluations. However, the trade-off here for the jackknife method seemed to be extremely high with the coverage probabilities were every poorly performed. One reasonable explain is the jackknife turned to perform worse for estimators that are not smooth functions of the sample data, such as median. As our estimator function of AUC here is not a smooth function, this method could perform poorly. This issue merits further investigation as the two-sample jackknife variance estimator of the AUC was proved to perform better than other closed-form variance estimator (Bandos, Guo and Gur, 2017)[27]. In other words, the jackknife methods are good tool for variance estimators for AUC, but it's less useful in the context of bias reduction as the variance went too large.

The bootstrap bias-reduced estimator performed well. It has less bias than the standardized method and with lower variance trade-off compared to jackknife estimators. The improved bootstrap method is considered to have a better convergence rate, but the comparison of the iteration times will need a much higher calculation load, which we have not discussed. Also, we could notice that all of the coverage probabilities seemed not performed well, the log transformation was not helpful at all and even produced worse results. One possible reason is the abuse of the estimated variance based only on the standardized method. To explore this, we performed a double bootstrap in the bootstrap bias-reduced method. The estimated standard errors increased which is expected as the trade-off between bias and standard error exist. There's an obviously promotion of the probabilities and the log-transformation had a better coverage which is around 95%. However, we only conducted this method with one setup and one method. The calculation will be burdensome if we want to explore more details about double bootstrap, parallel calculation or other calculation method might be necessary for further consideration.

In the more complicated setup simulation, all of the method seemed to be consistent as the conclusion we got from the simple setups. However, other setups are also deserved to be studied. Situations when the logistic regression is violated are desirable. One way to do this is to performed bootstrap in the real data and kind of simulated based on it. Again, the calculation will be burdensome, so we did not discuss it.

In the real data analysis, the results for AUC estimator are similar. It could be the reason that we have a relatively larger sample size than what we have used in our simulations. One should notice that a fluctuated estimator for bootstrap or cross-validation is reasonable since both cross-validation and bootstrap methods have a randomized steps which make the result unstable. Therefore, situations could be much more complex in real world, and some more precise and complicated methods or models need to be considered.

References

[1] Kyle Strimbu and Jorge A. Tavel M.D. What are biomarkers? Curr Opin HIV AIDS, (5(6)):463–466, 2010.

[2] Zachary Kornberg et al. Genomic biomarkers in prostate cancer. Transl Androl Urol, (7(3)):459–471, 2018.

[3] Sam W. Chan et all. Early detection of clinically significant prostate cancer at diagnosis: a prospective study using a novel panel of tmpprss2: Ets fusion gene markers. Cancer Med, (2(1)):63–75, 2013.

[4] Martin G. Sanda et al. Association between combined tmprss2:erg and pca3 rna urinary testing and detection of aggressive prostate cancer. JAMA Oncol., (3(8)):1085–1093, 2017.

[5] Enrique F. Schisterman Aiyi Liu and Yan Zhu. On linear combinarions of biomarkers to improve diagnostic accuracy. Statistics in Medicine, (24):37–47, 2005.

[6] Pliver Peters et al Lutz Frolich. Incremental value of biomarker combinations to predict progression of mild cognitive impairment to alzheimer's dementia. Alzheimer's Research and Therapy, (9):84, 2017.

[7] Frances MK Williams. Biomarkers: in combination they may do better. Arthritis Res Ther, (11(5)):130, 2009.

[8] Pepe M Smith R. Etzioni R., Kooperberg C. and Gann P.H. Combining biomarkers to detect disease with application to prostate cancer. Biostatistics, (4):523–538, 2003.

[9] D.M. Green and J/A/ Swets. Signal detection theory and psychophysics. New York: Wiley, 1996.

[10] J.P. Egan. Signal detection theory and roc analysis. New York: Academic Press, 1975.

[11] S.G Baker. Identifying combinations of cancer markers for further study as triggers of early intervention. Biometrics, (56):1082–1087, 2000.

[12] B. Adam Y. Qu and et al. Boosted decision tree analysis of seldi mass spectral serum profiles discriminates prostate cancer from non-cancer patients. Clinical Chemistry, (48):1853–1843, 2002.

[13] M. Pepe Yutaka Yasui and et al. A data-analytic strategy for protein-biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biometrics, (4):449–463, 2003.

[14] Ziding Feng and Yutaka Yasui. Statistical considerations in combining biomarkers for disease classification. Disease Markers, (20):45–51, 2004.

[15] Martin W. McIntosh and Margaret Sullivan Pepe. Combining several screening tests: Optimality of the risk score. Biometrics, (58):657–664, 2002.

[16] Su J.Q. and Liu J.S. Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association, (88):1350–1355, 1993.

[17] MS Pepe and ML Thompson. Combining diagnostic test results to increase accuracy. Biostatistics, (1(2)):123–140, 2000.

[18] Lori E Dodd and Margaret Sullivan Pepe. Semiparametric regression for the area under the receiver operating characteristic curve. Journal of the American Statistical Association, (98):462,409–417, 2003.

[19] Colin B. Begg. Advances in statistical methodology for diagnostic medicine in the 1980?s. Statistics in Medicine, (VOL. 10):1887–1895, 1991.

[20] David E shapiro. The interpretation of diagnostic tests. Statistical Methods in Medical Research, (8):113–134, 1999.

[21] Colin B. Begg and Barbara J. McNeil. Assessment of radiologic tests: control of bias and other design considerations. Radiology, (167):565–69, 1988.

[22] Bradley Axelrod Dr. Christian Schutte, Scott Millis and Sarah VanDyke. Derivation of a composite measure of embedded symptom validity indices. The Clinical Nueropsychologist, (25(3)):454–462, 2011.

[23] Anne-Laure Boulesteix. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLos comput Biol, (11(4)):e1004191, 2015.

[24] Ali R. Syed. A review of cross validation and adaptive model selection. ScholarWorks at Georgia State University, 2011.

[25] Edna Schechtman and Suojin Wang. Jackknifing two-sample statistics. Journal of Statistical Planning and Inference, (119(2)):329–340, 2004.

[26] Tibshirani R.J. Efron, B. An introduction to the bootstrap. New York, NY: Chapman and Hall, 1993.

[27] ben Guo Andriy I. Bandos and David Gur. Jackknife variance of the partial area under the empirical roc curve. Stat Methods Med Res., (26(2)):528–541, 2017.