UW Biostatistics Working Paper Series

10-30-2017

# Using Multilevel Outcomes to Construct and Select Biomarker Combinations for Single-level Prediction

Allison Meisner
*University of Washington, Seattle*, meisnera@uw.edu

Chirag R. Parikh
*Program of Applied Translational Research, Department of Medicine, Yale School of Medicine, New Haven, CT; Department of Internal Medicine, Vetrans Affairs Medical Center, West Haven, CT*, chirag.parikh@yale.edu

Kathleen F. Kerr
*University of Washington*, katiek@u.washington.edu

# Using Multilevel Outcomes to Construct and Select Biomarker Combinations for Single-level Prediction

Allison Meisner[1], Chirag R. Parikh[2,3], and Kathleen F. Kerr[4]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,
Baltimore, Maryland, U.S.A.

[2]Program of Applied Translational Research, Department of Medicine,
Yale School of Medicine, New Haven, Connecticut, U.S.A.

[3]Department of Internal Medicine, Veterans Affairs Medical Center,
West Haven, Connecticut, U.S.A.

[4]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

## Abstract

Biomarker studies may involve a multilevel outcome, such as no, mild, or severe disease. There is often interest in predicting one particular level of the outcome due to its clinical significance. The standard approach to constructing biomarker combinations in this context involves dichotomizing the outcome and using a binary logistic regression model. We assessed whether information can be usefully gained from instead using more sophisticated regression methods. Furthermore, it is often necessary to select among several candidate biomarker combinations. One strategy involves selecting a combination on the basis of its ability to predict the outcome level of interest. We propose an algorithm that leverages the multilevel outcome to inform combination selection. We apply this algorithm to data from a study of acute kidney injury after cardiac surgery, where the kidney injury may be absent, mild, or severe. Using more sophisticated modeling approaches to construct combinations provided gains over the binary logistic regression approach in specific settings. In the examples considered, the proposed algorithm for combination selection tended to reduce the impact of bias due to selection and to provide combinations with improved performance. Methods that utilize the multilevel nature of the outcome in the construction and/or selection of biomarker combinations have the potential to yield better combinations.

**Keywords:** biomarker, combinations, multilevel

# 1 Background

In some clinical settings, a patient can experience one of several outcomes. For example, he can have no, mild, or severe disease. In the setting of cancer diagnosis, a patient can be disease-free, have a benign mass, or have a malignancy. However, it may be most important to be able predict one level of the outcome in particular, typically the level that poses the greatest threat in terms of morbidity and mortality. In the examples just given, this may be severe disease or the presence of a malignant tumor. Here, investigators are interested in “single-level prediction,” but a multilevel outcome is available. The question becomes whether and how the information from the multilevel outcome can be leveraged to improve prediction of the outcome level of interest.

It is becoming increasingly common for studies to measure several biomarkers in each participant. Such studies often seek to develop a combination of biomarkers that can be used in risk prediction. When a multilevel outcome is available, the development of such combinations becomes potentially more complicated; we consider two such complications.

The first complication relates to the construction of biomarker combinations, specifically, how the biomarkers should be combined when a multilevel outcome is available but there is interest in single-level prediction. The most common approach is to dichotomize the outcome and fit a binary logistic regression model. Of course, this discards some information available in the multilevel outcome. We evaluate the potential benefits of alternative regression methods that utilize the multilevel outcome.

The second complication concerns how a biomarker combination should be selected. In many studies, the number of candidate biomarker combinations is quite large. Investigators may consider, for example, all possible pairs of biomarkers. In a study with 20 biomarkers, there would be nearly 200 such candidate combinations. One strategy is to choose the combination with the best performance in terms of single-level prediction. As with combination construction, it may be possible to leverage the additional information in the multilevel outcome to aid in combination selection. We propose an algorithm for doing so, and provide examples to illustrate the benefits of this method.

We illustrate the application of our combination selection algorithm to data from the Translational Research Investigating Biomarker Endpoints in Acute Kidney Injury (TRIBE-AKI), a study of acute kidney injury (AKI) after cardiac surgery [1]. This study aims to use biomarkers measured immediately after surgery to provide an earlier diagnosis of AKI. Clinical definitions of AKI include both mild and severe types, though severe AKI is of primary clinical interest due to its impact on long-term morbidity and mortality [2]. As a result, there is interest in developing a biomarker combination to diagnose severe AKI.

## 1.1 Constructing Combinations

We consider a set of predictors $\mathbf{X}$ and an outcome $D$ with $K$ levels. We consider outcomes whose levels can be ordered by, for example, their clinical significance.

### 1.1.1 Models for Binary Outcomes

One set of regression-based approaches involves treating the outcome as binary by dichotomizing $D$ and/or subsetting the data and subsequently fitting one or more binary logistic regression models [7, 3, 4, 5, 6]. These include:

(i) **Standard.** One binary logistic model based on dichotomizing $D$ at some fixed level, $k'$: $\text{logit}\,\{P(D \leq k'|\mathbf{X} = \mathbf{x})\} = \alpha + \boldsymbol{\beta}^T\mathbf{x}$ [8, 9, 10, 11, 12].

(ii) **Each level vs. others.** $K$ binary logistic models comparing each level to the combination of the other levels: $\text{logit}\,\{P(D = k|\mathbf{X} = \mathbf{x})\} = \alpha_k + \boldsymbol{\beta}_k^T\mathbf{x}, \;\; k = 1, ..., K$ [14, 13].

(iii) **Each level vs. reference.** $(K - 1)$ binary logistic models comparing each level to a reference level $k'$: $\log\{P(D = k|\mathbf{X} = \mathbf{x})/P(D = k'|\mathbf{X} = \mathbf{x})\} = \alpha_k + \boldsymbol{\beta}_k^T\mathbf{x}, \;\; k \neq k'$ [15, 16].

(iv) **Sequential.** $(K-1)$ binary logistic models comparing each level to the combination of the levels above it: $\text{logit}\,\{P(D = 1|\mathbf{X} = \mathbf{x})\} = \alpha_1 + \boldsymbol{\beta}_1^T\mathbf{x}, \text{logit}\,\{P(D = 2|D \geq 2, \mathbf{X} = \mathbf{x})\} = \alpha_2 + \boldsymbol{\beta}_2^T\mathbf{x}$, etc. [14].

### 1.1.2 Models for Ordinal Outcomes

Several regression models are available that fully model $D$ (i.e., do not combine different levels of the outcome together) while accounting for the ordered, categorical nature of $D$ [18, 6]. Such ordinal methods do not assume equal spacing between the levels of $D$; they simply use the ordering of $D$ [19].

**Cumulative Logit Model** The cumulative logit model can be written as

$$\text{logit}\,\{P(D \leq k|\mathbf{X} = \mathbf{x})\} = \alpha_k + \boldsymbol{\beta}^T\mathbf{x}, \tag{1}$$

$k = 1, ..., K - 1$, where the $\alpha_k$ are ordered in $k$ [18]. Under model (1), the log cumulative odds ratio is proportional to the distance between the predictor values being compared and the proportionality constant does not depend on $k$ [18]:

$$\begin{aligned}\text{logit}\,\{P(D \leq k|\mathbf{X} = \mathbf{x}_1)\} - \text{logit}\,\{P(D \leq k|\mathbf{X} = \mathbf{x}_2)\}\\ = \boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2).\end{aligned}$$

As a result of this proportionality (sometimes referred to as the parallel slopes assumption), model (1) is also called the proportional odds model [18]. It is possible to include a separate

$\boldsymbol{\beta}$ vector, $\boldsymbol{\beta}_k$, for each value of $k$ [18, 21, 20, 3]. However, doing so may lead to crossing of the cumulative probability curves $P(D \leq k|\mathbf{X} = \mathbf{x})$ for some values of $\mathbf{X}$, violating the ordering of the cumulative probabilities [18, 21]. Importantly, the estimates provided by the cumulative logit model may be biased under case-control sampling [12, 22].

**Adjacent-Category Logit Model** The adjacent-category logit model can be written as

$$\begin{aligned}\text{logit}\left\{P(D = k|\mathbf{X} = \mathbf{x})/P(D = k+1|\mathbf{X} = \mathbf{x})\right\} \\ = \alpha_k + \boldsymbol{\beta}^T\mathbf{x},\end{aligned}$$

$k = 1, ..., K-1$ [18]. The adjacent-category logit model can be used with data from case-control studies [18].

**Continuation-Ratio Logit Model** The continuation-ratio logit model can be written as

$$\text{logit}\left\{P(D = k|D \geq k, \mathbf{X} = \mathbf{x})\right\} = \alpha_k + \boldsymbol{\beta}^T\mathbf{x},$$

$k = 1, ..., K-1$, where the $\alpha_k$ are ordered in $k$ [18]. Allowing a separate $\boldsymbol{\beta}$ vector for each $k$ gives the sequential approach described earlier [19, 3]. The estimates provided by the continuation-ratio logit model may be biased under case-control sampling [12].

**Stereotype Model** The stereotype model is a sort of compromise between models that incorporate the ordinality of the outcome and more flexible models (i.e., the baseline-category logit model defined below). The stereotype model actually includes a hierarchy of models that vary in flexibility, as defined by the dimension of the model, though the term "stereotype model" is generally reserved for the one-dimensional model and we focus on that model here [23]. This model can be written as

$$\log\left\{P(D = k|\mathbf{x})/P(D = K|\mathbf{x})\right\} = \alpha_k + \phi_k\boldsymbol{\beta}^T\mathbf{x}, \tag{2}$$

$k = 1, ..., K-1$ [23]. Essentially, this model allows some variation in the coefficient vector, but restricts $\boldsymbol{\beta}_k = \phi_k \boldsymbol{\beta}$ [23]. Identifiability constraints must be imposed on the $\phi_k$; typically, these are $\phi_1 = 0$ and $\phi_K = 1$ [23]. The definition of the stereotype model also typically includes the requirement that $\phi_1 < \phi_2 < ... < \phi_K$; when this holds, the one-dimensional model given in (2) is an ordered model [23]. However, it has been noted that this ordering does not need to be specified *a priori* and most statistical packages (e.g., R and Stata) do not impose such a restriction [24, 3]. Thus, in practice, a fairly flexible model is fit and used to assess whether the data suggest ordering [25, 23]. The stereotype model can be used with case-control data [12].

### 1.1.3 Models for Nominal Outcomes

The baseline-category logit model is a very flexible approach that considers the categorical nature of the outcome but does not incorporate the ordering [18, 26, 6]. The baseline-category logit model can be written as

$$\log \{P(D = k|\mathbf{x})/P(D = K|\mathbf{x})\} = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x},$$

$k = 1, ..., K-1$ [18]. Thus, the baseline-category logit model allows the effect of the predictors to vary with the level of the outcome [18]. The set of models specified by the each level vs. reference approach with $k' = K$ is parametrically equivalent to the baseline-category logit model. Furthermore, the adjacent-category logit model is a reparameterization of the baseline-category logit model with a common $\boldsymbol{\beta}$ [18].

### 1.1.4 Models for Continuous Outcomes

When the number of outcome levels $K$ is large, linear regression models may be appealing [6, 12, 27]. These models consider the ordering of the outcome but ignore the categorized nature of $D$ [6]. In other words, these models are only strictly valid if the intervals between consecutive outcome levels are considered equivalent [3]. Problems may arise in the use of these models with a multilevel outcome, including predictions beyond the reasonable range [27]. As a result of these issues, we do not consider this approach further.

### 1.1.5 Comparing Modeling Approaches

Some work has been done to compare the models described above, particularly in terms of efficiency of the model estimates. Briefly, gains in efficiency have been found for the baseline-category logit model relative to the each level vs. reference approach [15, 16], for the cumulative logit model relative to the standard approach [3, 9, 5], and for the stereotype model relative to the baseline-category logit model [28]. These results suggest that information can be gained from using all of the data in a single model, not dichotomizing $D$, and/or incorporating the ordinal nature of $D$.

Armstrong and Sloan conclude that in general, if the order of categories can be specified with confidence, models that incorporate this ordering are preferable to more flexible models [3]. In other words, it is reasonable to expect that when the outcome is ordinal, information is gained when this ordinality is used by the model [12]. Additionally, Harrell et al. note that models can exhibit lack of fit and yet still provide quite accurate predicted probabilities [29]. In comparing the cumulative logit model to the standard approach, Risselada et al. argued that the impact of “mild violations” of the proportional odds assumption is expected to be less severe than the loss of information resulting from dichotomizing $D$ [6]. On the other hand, others have noted that ordinal models become “increasingly unrealistic” as the number of outcome levels and/or predictors increases [28, 9].

### 1.1.6 Applications in Risk Prediction

Multilevel outcomes are frequently encountered in the risk prediction setting, and a common approach is to dichotomize the outcome and fit a binary logistic regression model (the standard approach defined above) [13, 7, 6, 17]. The literature on using multilevel outcomes for single-level prediction has largely focused on the area under the receiver operating characteristic (ROC) curve (AUC) as a measure of predictive capacity. Briefly, the AUC assesses the ability of a model to discriminate between individuals who have or will experience the outcome level of interest and those who do not have or will not experience the outcome level of interest; the AUC for a model that is able to perfectly separate these groups is 1, while the AUC for a useless model is 0.5 [30].

The previous work in this area has primarily involved using individual datasets to compare modeling strategies. Biesheuvel et al. compared the baseline-category logit model to the sequential approach and found fairly similar AUCs for both strategies [13]. Roukema et al. compared the baseline-category logit model, the sequential approach, and the each level vs. others strategy [14]. They found similar discriminatory power for all three strategies, though they employed variable selection procedures for all of the models, making comparisons difficult [14].

## 1.2 Combination Selection

Often, a number of candidate biomarker combinations are available, and some form of selection is required. When the goal is to use a biomarker combination for risk prediction, it seems appropriate to select combinations based on predictive capacity. For a binary outcome, one possibility is to use the AUC (e.g., [31]). That is, the AUC for each candidate combination is estimated, and the combination with the highest AUC is chosen.

Two challenges arise in utilizing this approach. The first is that when the same data are used to construct a biomarker combination and estimate the AUC (or other measure of performance) for that combination, the resulting AUC estimate will be optimistic relative to the AUC for the same fitted combination in independent data; we refer to this as "resubstitution bias" [32]. Methods such as bootstrapping can be used to correct the apparent AUC estimate [19].

An additional challenge applies to selection more generally. If many models are considered and a model is selected on the basis of some estimated measure of performance, that estimated measure of performance will be optimistically biased; we refer to this as "model selection bias" [32]. This idea has been explored in the bioinformatics/machine learning literature, where estimates of the classification error rate are often used to select a model. The estimated error rate for a model selected on the basis of its favorable error rate will be optimistic relative to the same model's error rate in independent data [33, 34, 35, 36, 37, 38, 39]. Cawley and Talbot call this issue "overfitting the model selection criterion" [35]. Although this issue has not yet been fully characterized in the clinical risk prediction setting, where the AUC is generally preferred over the classification error, the problem is expected to persist. In general, when some form of model selection is done and the performance of the chosen model is evaluated without accounting for the selection, that is, treating the selected model as though it were pre-specified, optimistic bias is expected [19, 40, 41, 42, 36].

# 2 Methods

We will suppose that for an outcome $D$ with $K$ levels, "single-level prediction" relates to predicting $D = K$.

## 2.1 Constructing Combinations

We have described several regression-based approaches to modeling a multilevel outcome. In particular, we can dichotomize the outcome or subset the data and use one of the four binary strategies, we can treat the outcome as ordered and use one of the four ordinal approaches, or we can use the more flexible baseline-category logit model. Using a binary strategy requires either combining several levels of the outcome together, or fitting several models to subsets of the data. Likewise, the ordinal models require restricting the nature of the relationship between the biomarkers and the outcome so as to achieve parsimony. The baseline-category logit model, on the other hand, imposes no such restrictions and includes all of the data in a single model; of course, this comes at the cost of having to estimate additional parameters. We use simulations to evaluate the impact of these modeling choices on the performance of the resulting estimated combinations. The key question is whether more sophisticated modeling approaches can offer improvements in performance in terms of single-level prediction over the standard approach, that is, a single binary logistic regression model.

Since we are interested in predicting $D = K$, we considered $k' = K - 1$ in the standard approach. Furthermore, for the purposes of predicting $D = K$, the standard approach and the each level vs. others strategy are identical, and so the latter was not considered further. Finally, as the baseline-category logit model is parametrically equivalent to the each level vs. reference approach with $k' = K$, and the former is generally more efficient than the latter, we did not include the each level vs. reference strategy in our investigation. Thus, we considered seven different modeling strategies: the standard approach ("Standard"), the sequential strategy ("Sequential"), the cumulative logit model ("CumLogit"), the adjacent-category logit model ("AdjCatLogit"), the continuation-ratio logit model ("ContRatLogit"), the stereotype model ("Stereo"), and the baseline-category logit model ("BaselineCat").

We considered two broad simulation scenarios. In the first scenario, the biomarkers were simulated such that the cumulative logit model with proportional odds did not hold; in the second scenario, the data were simulated under the cumulative logit model where the assumption of proportional odds held. In both scenarios, we considered two biomarkers, $\mathbf{X} = (X_1, X_2)$. We considered outcomes with either 3 or 5 levels, that is, $K = 3$ or $K = 5$. The combinations were constructed using training data with 200, 400, 800, or 1600 observations and evaluated in test data with $10^4$ observations. We simulated data such that $P(D = 1) = 0.1$ or 0.5 and $P(D = K) = 0.05$ or 0.3; when $K = 5$, $P(D = 2)$, $P(D = 3)$, and $P(D = 4)$ were approximately equal.

We used each of the modeling strategies to fit a linear combination of the biomarkers $\mathbf{X}$ in the training data, yielding estimates $\hat{\boldsymbol{\beta}}$. We then applied these estimates to the test data to determine $\hat{P}(D = K|\mathbf{X}, \hat{\boldsymbol{\beta}})$. Finally, we assessed the ability of $\hat{P}(D = K|\mathbf{X}, \hat{\boldsymbol{\beta}})$ to discriminate between $D = K$ and $D < K$ in the test data via the AUC.

In the simulations where the cumulative logit model with proportional odds did not hold (the first scenario mentioned above), the biomarkers had conditional bivariate normal distributions. In particular, for $K = 3$, we considered $(\mathbf{X}|D = 1) \sim N(\mathbf{0}, \Sigma)$, $(\mathbf{X}|D = 2) \sim N(\boldsymbol{\mu}, \Sigma)$, and $(\mathbf{X}|D = 3) \sim N(\mathbf{2}, \Sigma)$, and for $K = 5$, we considered $(\mathbf{X}|D = 1) \sim N(\mathbf{0}, \Sigma)$, $(\mathbf{X}|D = 2) \sim N(\mathbf{0.5}, \Sigma)$, $(\mathbf{X}|D = 3) \sim N(\mathbf{1}, \Sigma)$, $(\mathbf{X}|D = 4) \sim N(\boldsymbol{\mu}, \Sigma)$, and $(\mathbf{X}|D = 5) \sim N(\mathbf{2}, \Sigma)$. We used $\boldsymbol{\mu} \in \{\mathbf{-1}, \mathbf{0}, ..., \mathbf{2}, \mathbf{3}\}$ and $\Sigma = 2I_2$, where $I_2$ is a two-dimensional identity matrix. Other covariance matrices (including those with correlation between the biomarkers and unequal covariance matrices) were explored; details are given in Section S1.1 (Additional File 1).

To evaluate data generated by the cumulative logit model with proportional odds (second scenario), we simulated two independent normal biomarkers, both with mean 1 and variance 0.25. The outcome was then simulated as a multinomial random variable, where the success probabilities of the $K$ levels were determined by $\beta_{0i} + \boldsymbol{\beta}^\top \mathbf{X}$ such that the cumulative logit model held. Three sets of coefficients $\boldsymbol{\beta}$ were considered ($\boldsymbol{\beta} = (1, 2), (1, 1.5), (1, -1)$) and values of $\beta_{0i}$ were chosen such that the desired prevalences (given above) were achieved in a large dataset.

The simulations were repeated 1000 times.

## 2.2 Combination Selection

As above, we suppose that for an outcome $D$ with $K$ levels, "single-level prediction" relates to predicting $D = K$.

As with combination construction, the presence of a multilevel outcome requires that decisions about how to select a biomarker combination be made. One strategy is to simply estimate the AUC for $D = K$ vs. $D < K$ (including correcting this estimate for resubstitution bias due to any model fitting), and select the combination with the highest estimated

AUC. As discussed above, the estimated AUC for this selected combination will be optimistically biased relative to the AUC for the same fitted combination in independent data due to model selection bias. We propose an alternative strategy where combination selection is done on the basis of not only the AUC for $D = K$ vs. $D < K$, but also the AUC for $D = K - 1$ vs. $D < K - 1$, the AUC for $D = K - 2$ vs. $D < K - 2$, and so on. We anticipate that in some settings, the estimated AUC for $D = K$ vs. $D < K$ for the combination selected in this way will be less affected by model selection bias and so may be preferred. In particular, if some of the same biomarkers are associated with multiple levels of the outcome, our proposed method could offer improvements over the standard approach. Furthermore, we expect our approach to be useful when many biomarkers have modest associations with the outcome and the candidate combinations include subsets of these biomarkers.

More precisely, for $K = 3$, we define our algorithm (including constructing combinations and estimating their performance) as follows.

(1) In the training data, dichotomize $D$ at $D = 3$ vs. $D < 3$ and construct all candidate biomarker combinations using binary logistic regression.

(2) Based on the combinations fit in (1), estimate (i) the AUC for $D = 3$ vs. $D < 3$ and (ii) the AUC for $D = 2$ vs. $D = 1$ in the training data.

(3) Generate $B$ bootstrap samples from the training data.

  (a) In each bootstrap sample, dichotomize $D$ at $D = 3$ vs. $D < 3$ and construct all candidate biomarker combinations using binary logistic regression.

  (b) For each of the fitted combinations from (a), estimate (i) the AUC for $D = 3$ vs. $D < 3$ and (ii) the AUC for $D = 2$ vs. $D = 1$ in both the bootstrap sample and the training data.

  (c) Estimate the resubstitution bias as the average difference between the AUCs in the bootstrap sample and the training data across the $B$ samples.

(4) Correct the estimated AUCs from (2) using the estimated bias from (3c).

(5) Determine the ranks for each of the two sets of corrected AUCs from (4) across all fitted biomarker combinations. The "standard" approach involves choosing the combination with the best AUC for $D = 3$ vs. $D < 3$. The "new" approach involves choosing the combination with the best sum of ranks for the two AUCs.

(6) Apply the two chosen combinations to test data and estimate the AUC for $D = 3$ vs. $D < 3$ for each. The estimated model selection bias is the difference between the AUCs in the test data and the AUCs from (4).

In practice, test data may not be available, so it may not be possible to complete step (6). An `R` package including code to implement this method, `multiselect`, will be publicly available.

We used simulations to investigate the potential benefits of the proposed method. We considered five examples as a proof of concept; these are not intended to be exhaustive. In the first two examples, the cumulative logit model with proportional odds held, and in the other three, it did not. Throughout the simulations, there were $p = 30$ biomarkers and we considered the set of candidate combinations to be all possible pairs of these biomarkers, constructed via binary logistic regression. We used $B = 50$ bootstrap replicates, a training set of 400 observations, and a test set of $10^4$ observations. We repeated the simulations 500 times.

In Example 1, we had $\mathbf{X} \sim N(\mathbf{1}, 2\Gamma)$, where $\mathbf{X}$ was a vector of dimension 30 and $\Gamma$ was a $30 \times 30$ matrix where the diagonal elements were 1 and the off-diagonal elements were 0.3. The linear predictor was $\boldsymbol{\beta}^\top \mathbf{X}$, where $\beta_1 = 1, \beta_2 = 2, \beta_3 = ... = \beta_{16} = 0.5, \beta_{17} = ... = \beta_{30} = 0.1$. The outcome was simulated under the cumulative logit model such that $P(D = 1) = 0.6$, $P(D = 2) = 0.3$, and $P(D = 3) = 0.1$ in a large dataset. Example 2 was identical to Example 1, except that $P(D = 2) = 0.335$ and $P(D = 3) = 0.065$.

In Example 3, we had $P(D = 1) = 0.6$, $P(D = 2) = 0.335$, and $P(D = 3) = 0.065$. Additionally, $(\mathbf{X}|D = 1) \sim N(\mathbf{0}, 2\Gamma)$, $(\mathbf{X}|D = 2) \sim N(\boldsymbol{\beta}^{(2)}, 2\Gamma)$, and $(\mathbf{X}|D = 3) \sim N(\boldsymbol{\beta}^{(3)}, 2\Gamma)$

where $\mathbf{X}$ was a vector of dimension 30, $\Gamma$ was as defined above for Example 1, and $\beta_1^{(2)} = 1.5, \beta_2^{(2)} = 1, \beta_3^{(2)} = ... = \beta_{16}^{(2)} = 0.5, \beta_{17}^{(2)} = ... = \beta_{30}^{(2)} = 0.1, \beta_1^{(3)} = \beta_2^{(3)} = 2, \beta_3^{(3)} = ... = \beta_{16}^{(3)} = 0.8$, and $\beta_{17}^{(3)} = ... = \beta_{30}^{(3)} = 0.1$. Example 4 was identical to Example 3, except that $\beta_1^{(2)} = 1$ and $\beta_{17}^{(3)} = ... = \beta_{30}^{(3)} = 0.2$. Finally, Example 5 was identical to Example 3, except that $\beta_1^{(2)} = 1$, $\beta_{17}^{(2)} = ... = \beta_{30}^{(2)} = 0$, and $\beta_{17}^{(3)} = ... = \beta_{30}^{(3)} = 0.2$.

# 3 Results

## 3.1 Constructing Combinations

First we consider the scenario where the cumulative logit model with proportional odds did not hold. We present the results for a training set size of 400; the results for the other sample sizes were similar. Here, we focus on the results for $P(D = K) = 0.05$ and provide the full results in Section S1.1 (Additional File 1).

Table 1 presents the results for $K = 3$ and Table 2 presents the results for $K = 5$. We see that when $\boldsymbol{\mu} = -\mathbf{1}$, $\boldsymbol{\mu} = \mathbf{0}$, or $\boldsymbol{\mu} = \mathbf{1}$, the standard approach is comparable to or better than the other approaches. When $\boldsymbol{\mu} = \mathbf{2}$, the standard approach may do slightly worse than some of the ordinal approaches, particularly for $K = 3$. For $\boldsymbol{\mu} = \mathbf{3}$, the sequential approach, the stereotype model, and/or the baseline-category logit model offer some gains over the standard approach. In sum, when the cumulative logit model with proportional odds did not hold but there was some ordering in the outcome by the biomarkers (that is, $\boldsymbol{\mu}$ was not extreme), the standard approach did well, but when $\boldsymbol{\mu}$ was extreme, some of the alternative approaches demonstrated improved performance.

When the cumulative logit model with proportional odds held, the performance was comparable across the approaches considered for a training set size of 400 (Additional File 1: Section S1.2); similar patterns were seen for other sample sizes. Thus, even when the data were generated by an ordinal model, the standard approach did well in terms of the predictive capacity of the fitted combinations.

Table 1: Simulation results for $K = 3$, $n = 400$, and $P(D = 3) = 0.05$ when the cumulative logit model with proportional odds did not hold. The table presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy.

| Class | Model | $\mu$ = -1 | $\mu = 0$ | $\mu = 1$ | $\mu = 2$ | $\mu = 3$ |
|---|---|---|---|---|---|---|
| | | | | P(D=1) = 0.1 | | |
| Binary | Standard | 0.976 (0.974, 0.978) | 0.920 (0.915, 0.924) | 0.773 (0.764, 0.780) | 0.530 (0.508, 0.543) | 0.670 (0.642, 0.684) |
| | Sequential | 0.974 (0.971, 0.976) | 0.920 (0.915, 0.924) | 0.773 (0.764, 0.780) | 0.532 (0.519, 0.544) | 0.720 (0.708, 0.729) |
| Ordinal | CumLogit | 0.970 (0.946, 0.975) | 0.918 (0.912, 0.923) | 0.776 (0.769, 0.783) | 0.544 (0.536, 0.552) | 0.313 (0.306, 0.320) |
| | AdjCatLogit | 0.970 (0.952, 0.975) | 0.918 (0.912, 0.923) | 0.776 (0.769, 0.783) | 0.544 (0.536, 0.552) | 0.313 (0.306, 0.320) |
| | ContRatLogit | 0.971 (0.958, 0.976) | 0.918 (0.912, 0.923) | 0.776 (0.769, 0.783) | 0.544 (0.536, 0.552) | 0.313 (0.306, 0.320) |
| | Stereo | 0.976 (0.974, 0.978) | 0.920 (0.915, 0.924) | 0.776 (0.769, 0.783) | 0.535 (0.520, 0.547) | 0.724 (0.715, 0.732) |
| Nominal | BaselineCat | 0.976 (0.974, 0.978) | 0.920 (0.915, 0.924) | 0.773 (0.764, 0.780) | 0.532 (0.519, 0.544) | 0.720 (0.707, 0.728) |
| | | | | P(D=1) = 0.5 | | |
| Binary | Standard | 0.950 (0.946, 0.952) | 0.920 (0.915, 0.924) | 0.841 (0.834, 0.848) | 0.714 (0.705, 0.723) | 0.588 (0.571, 0.599) |
| | Sequential | 0.924 (0.911, 0.933) | 0.919 (0.915, 0.924) | 0.842 (0.834, 0.848) | 0.712 (0.701, 0.722) | 0.743 (0.733, 0.752) |
| Ordinal | CumLogit | 0.054 (0.050, 0.062) | 0.916 (0.907, 0.921) | 0.844 (0.838, 0.849) | 0.721 (0.715, 0.728) | 0.599 (0.593, 0.604) |
| | AdjCatLogit | 0.073 (0.054, 0.198) | 0.917 (0.911, 0.922) | 0.844 (0.838, 0.849) | 0.721 (0.715, 0.728) | 0.599 (0.593, 0.604) |
| | ContRatLogit | 0.094 (0.057, 0.409) | 0.917 (0.911, 0.922) | 0.844 (0.838, 0.849) | 0.721 (0.715, 0.728) | 0.599 (0.593, 0.604) |
| | Stereo | 0.950 (0.947, 0.953) | 0.920 (0.915, 0.924) | 0.844 (0.838, 0.849) | 0.718 (0.709, 0.725) | 0.749 (0.741, 0.756) |
| Nominal | BaselineCat | 0.950 (0.946, 0.952) | 0.920 (0.915, 0.924) | 0.841 (0.835, 0.848) | 0.712 (0.702, 0.722) | 0.743 (0.733, 0.752) |

Table 2: Simulation results for $K = 5$, $n = 400$, and $P(D = 5) = 0.05$ when the cumulative logit model with proportional odds did not hold. The table presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy.

| Class | Model | $\mu$ = -1 | $\mu = 0$ | $\mu = 1$ | $\mu = 2$ | $\mu = 3$ |
|---|---|---|---|---|---|---|
| | | | | P(D=1) = 0.1 | | |
| Binary | Standard | 0.870 (0.864, 0.875) | 0.851 (0.844, 0.856) | 0.802 (0.794, 0.810) | 0.721 (0.710, 0.730) | 0.636 (0.615, 0.647) |
| | Sequential | 0.732 (0.699, 0.760) | 0.836 (0.824, 0.845) | 0.802 (0.793, 0.810) | 0.720 (0.708, 0.729) | 0.693 (0.681, 0.703) |
| Ordinal | CumLogit | 0.134 (0.128, 0.144) | 0.804 (0.673, 0.843) | 0.804 (0.797, 0.811) | 0.728 (0.721, 0.735) | 0.650 (0.643, 0.656) |
| | AdjCatLogit | 0.140 (0.130, 0.161) | 0.831 (0.781, 0.847) | 0.804 (0.797, 0.811) | 0.728 (0.721, 0.735) | 0.650 (0.643, 0.656) |
| | ContRatLogit | 0.138 (0.130, 0.158) | 0.810 (0.690, 0.844) | 0.804 (0.796, 0.811) | 0.728 (0.721, 0.735) | 0.650 (0.643, 0.656) |
| | Stereo | 0.872 (0.867, 0.877) | 0.853 (0.847, 0.858) | 0.804 (0.797, 0.811) | 0.727 (0.718, 0.734) | 0.701 (0.692, 0.709) |
| Nominal | BaselineCat | 0.870 (0.864, 0.875) | 0.851 (0.844, 0.857) | 0.802 (0.794, 0.810) | 0.720 (0.709, 0.729) | 0.696 (0.684, 0.704) |
| | | | | P(D=1) = 0.5 | | |
| Binary | Standard | 0.893 (0.888, 0.898) | 0.883 (0.877, 0.888) | 0.856 (0.850, 0.862) | 0.814 (0.807, 0.821) | 0.769 (0.757, 0.777) |
| | Sequential | 0.791 (0.756, 0.824) | 0.878 (0.869, 0.884) | 0.856 (0.850, 0.862) | 0.814 (0.807, 0.820) | 0.790 (0.780, 0.798) |
| Ordinal | CumLogit | 0.878 (0.828, 0.891) | 0.883 (0.877, 0.888) | 0.858 (0.853, 0.864) | 0.818 (0.813, 0.824) | 0.777 (0.772, 0.782) |
| | AdjCatLogit | 0.866 (0.773, 0.889) | 0.883 (0.876, 0.888) | 0.858 (0.853, 0.864) | 0.819 (0.813, 0.824) | 0.777 (0.772, 0.782) |
| | ContRatLogit | 0.852 (0.720, 0.886) | 0.882 (0.875, 0.887) | 0.858 (0.853, 0.864) | 0.818 (0.813, 0.824) | 0.777 (0.772, 0.782) |
| | Stereo | 0.895 (0.890, 0.899) | 0.884 (0.879, 0.889) | 0.859 (0.853, 0.864) | 0.818 (0.812, 0.824) | 0.798 (0.790, 0.804) |
| Nominal | BaselineCat | 0.893 (0.888, 0.898) | 0.883 (0.877, 0.888) | 0.857 (0.851, 0.863) | 0.815 (0.808, 0.821) | 0.793 (0.786, 0.800) |

For small to moderate sample sizes, several of the approaches had issues with convergence. When the training set had 200 observations, the standard approach failed to converge in up to 3.1% of simulations, the sequential approach failed to converge in up to 38% of simulations, the stereotype model failed to converge in up to 2.6% of simulations, and the baseline-category logit model failed to converge in up to 1.4% of simulations. For training data with 400 observations, the sequential approach failed to converge in up to 7% of simulations. The proportion of convergence failures was below 0.2% for all methods for larger sample sizes.

## 3.2 Combination Selection

Table 3 presents the results for Examples 1 and 4 for the proposed combination selection method. The results for Examples 2, 3, and 5 show similar patterns; the full results are presented in Section S2 (Additional File 1). The results in Table 3 demonstrate some benefit to using the additional information available in the multilevel outcome to select a biomarker combination for single-level prediction, both in terms of the degree of model selection bias and the ability of the chosen combination to discriminate $D = 3$ from $D < 3$ in test data.

Table 3: Results for the proposed combination selection method for Examples 1 and 4. The table gives the median (interquartile range) of the estimated model selection bias and the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.

| | Method | Bias | AUC |
|---|---|---|---|
| Example 1 | Standard | 0.030 (0.020, 0.044) | 0.911 (0.905, 0.917) |
| | New | 0.014 (0.005, 0.026) | 0.916 (0.911, 0.923) |
| Example 4 | Standard | 0.042 (0.012, 0.068) | 0.794 (0.777, 0.834) |
| | New | 0.010 (-0.018, 0.037) | 0.831 (0.822, 0.838) |

There were no issues with the logistic regression model failing to converge in the Example 1 simulations, eight simulations (out of 500) had convergence issues in Example 2, and one simulation had convergence issues in each of Examples 3, 4, and 5.

### 3.2.1 Application to TRIBE-AKI

We applied our proposed method for combination selection to data from the TRIBE-AKI study. As noted above, the outcome in this study, AKI, is a multilevel outcome as patients may be diagnosed with no, mild, or severe AKI. Furthermore, of the biomarkers measured in the study, it is believed that only a subset are likely to be useful for early diagnosis. Thus, we considered all possible pairs of 14 biomarkers measured in the study.

The TRIBE-AKI study is a multicenter study, but we restricted attention to the largest center in order to avoid issues related to center differences. We used the biomarker measurements taken immediately after surgery, and removed observations missing any of these measurements. This left 465 observations (61 with mild AKI and 30 with severe AKI). We also log-transformed the biomarker measurements. As in the simulations, we applied our proposed method with 50 bootstrap replications.

The results for the ten best combinations in terms of the AUC for severe vs. no/mild AKI are given in Table 4. The combination with the highest AUC for severe vs. no/mild AKI, which would be selected by the standard approach, includes urine interleukin-18 (IL-18) and plasma N-terminal-pro-B-type natriuretic peptide (NT-proBNP). The estimated AUCs (corrected for resubstitution bias) for this combination were 0.8575 for severe vs. no/mild AKI and 0.6125 for mild vs. no AKI. The combination with the highest combined rank for the AUC for severe vs. no/mild AKI and the AUC for mild vs. no AKI, which would be selected by the proposed method, included plasma heart-type fatty acid binding protein (h-FABP) and plasma interleukin-6 (IL-6). The estimated AUCs (corrected for resubstitution bias) for this combination were 0.8365 for severe vs. no/mild AKI and 0.6757 for mild vs. no AKI. Thus, the AUC for severe vs. no/mild AKI for this second combination is slightly lower, but the AUC for mild vs. no AKI is substantially higher. It may be reasonable to expect that the estimated AUC for severe vs. no/mild AKI for the second combination (0.8365) is less affected by model selection bias than is the estimated AUC for severe vs. no/mild AKI for the first combination (0.8575), which may motivate choosing to validate the second combination instead of the first.

# 4 Discussion

When there is interest in using biomarker combinations for single-level prediction and a multilevel outcome is available, common practice is often to dichotomize the outcome for

Table 4: The ten best biomarker pairs in the TRIBE-AKI study. The table presents the ten pairs with the highest estimated AUC for severe vs. no/mild AKI. The estimated AUCs for severe vs. no/mild AKI and for mild vs. no AKI are presented. Both estimates are corrected for optimism due to resubstitution bias.

| Biomarkers | | AUC (Severe) | AUC (Mild) |
|---|---|---|---|
| Urine IL-18 | Plasma NT-proBNP | 0.8575 | 0.6125 |
| Plasma h-FABP | Urine IL-18 | 0.8495 | 0.6394 |
| Plasma h-FABP | Plasma BNP | 0.8464 | 0.6403 |
| Plasma h-FABP | Plasma NT-proBNP | 0.8459 | 0.6329 |
| Urine IL-18 | Plasma BNP | 0.8414 | 0.6168 |
| Plasma h-FABP | Urine KIM-1 | 0.8410 | 0.6400 |
| Plasma h-FABP | Plasma IL-6 | 0.8365 | 0.6757 |
| Plasma h-FABP | Plasma IL-10 | 0.8342 | 0.6405 |
| Plasma h-FABP | Plasma CKMB | 0.8271 | 0.6558 |
| Urine KIM-1 | Plasma TNTHS | 0.8253 | 0.6005 |

combination construction and selection. We have considered whether the information in a multilevel outcome could be more fully leveraged in the development of biomarker combinations for single-level prediction.

In the context of constructing biomarker combinations, we used simulations to compare seven regression-based approaches: two binary approaches, four ordinal approaches, and one nominal approach. We considered a variety of data-generating scenarios and found that when some separation in the biomarker distributions between $D = K$ and $D < K$ existed (i.e., $\boldsymbol{\mu} < \mathbf{2}$ in our first simulation scenario) or when the cumulative logit model with proportional odds held, the standard approach based on dichotomizing the outcome tended to work well in terms of the ability of the resulting combinations to predict $D = K$.

We have also proposed a method that utilizes the multilevel nature of the outcome in selecting a biomarker combination, as opposed to selecting a combination based solely on its ability to predict the targeted level. Simulations provide evidence that use of the proposed method may result in less model selection bias and could lead to selecting combinations with greater predictive capacity. We applied this method to data from the TRIBE-AKI study, where we demonstrated how the method could be used to select a combination in practice. This approach is expected to be most useful when there is some ordering in the biomarkers by the levels of $D$. It is important to study this method further in order to fully elucidate the settings in which it could be beneficial.

In using this method for selection, it is generally informative to look at the results for the candidate combinations, as we have done in Table 4 for the top ten pairs in the TRIBE-AKI study. If there is a clear "winner" in terms of the AUC for $D = 3$ vs. $D < 3$, that is, if this AUC is substantially higher for one candidate combination, it is probably reasonable to select that combination, regardless of the AUC for $D = 2$ vs. $D = 1$. This is because it is unlikely that such a markedly higher AUC estimate is due to model selection bias. On the other hand, if several combinations have fairly similar performance in terms of the AUC for $D = 3$ vs. $D < 3$, it may be worth using the AUC for $D = 2$ vs. $D = 1$ to aid in selection. One possible extension of this method could involve using a weighted average of ranks for the two AUCs, rather than the sum; additionally, using a weighted average of the AUC values themselves (as opposed to their ranks) could be considered.

# 5 Conclusions

When a multilevel outcome is available and there is interest in using biomarker combinations to predict a single level of the outcome, the common approach of dichotomizing the outcome necessarily discards some information. We have described when and how this information might be usefully recovered to advance the goal of single-level prediction, thereby providing insight into how best to use the data at hand.

# Supplementary Materials

The Supplementary Materials contain Sections S1 and S2 and are available with this paper. Section S1 contains results for simulations comparing methods for constructing combinations when the cumulative logit model with proportional odds did not hold (Section S1.1) and when the cumulative logit model with proportional odds held (Section S1.2). Section S2 contains results for simulations comparing methods for combination selection.

# Acknowledgements

This work was supported by the National Institutes of Health (F31 DK108356, R01 HL085757, K24 DK090203). The opinions, results, and conclusions reported in this article are those of the authors and are independent of the funding sources.

The authors wish to acknowledge the TRIBE-AKI study investigators: Steven G. Coca (Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, New York), Amit X. Garg (Institute for Clinical Evaluative Sciences Western, London, Ontario, Canada; Division of Nephrology, Department of Medicine, and Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada), Jay Koyner (Section of Nephrology, Department of Medicine, University of Chicago Pritzker School of Medicine, Chicago, Illinois), and Michael Shlipak (Kidney Health Research Collaborative, San Francisco Veterans Affairs Medical Center, University of California, San Francisco, San Francisco, California). Assays used in the TRIBE-AKI study were donated by Abbott Diagnostics, Randox Laboratories, and Beckman Coulter.

# References

[1] Parikh CR, Coca SG, Thiessen-Philbrook H, Shlipak MG, Koyner JL, Wang Z, et al. Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. J Am Soc Nephrol. 2011;22:1748–57.

[2] Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. Kidney Int. 2012;81:442–8.

[3] Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. Am J Epidemiol. 1989;129:191–204.

[4] Bartfay E, Donner A, Klar N. Testing the equality of twin correlations with multinomial outcomes. Am J Hum Genet. 1999;63:341–9.

[5] Maas AIR, Steyerberg EW, Marmarou A, McHugh GS, Lingsma HF, Butcher I, et al. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. Neurotherapeutics. 2010;7:127–34.

[6] Risselada R, Lingsma HF, Molyneux AJ, Kerr RSC, Yarnold J, Sneade M, et al. Prediction of two month modified Rankin Scale with an ordinal prediction model in patients with aneurysmal subarachnoid haemorrhage. BMC Med Res Methodol. 2010;10:86.

[7] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer Science & Business Media; 2008.

[8] Manor O, Matthews S, Power C. Dichotomous or categorical response? Analysing self-rated health and lifetime social class. Int J Epidemiol. 2000;29:149–57.

[9] McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. Clin Trials. 2010;7:44–57.

[10] Norris CM, Ghali WA, Saunders LD, Brant R, Galbraith D, Faris P, et al. Ordinal regression model and the linear regression model were superior to the logistic regression models. J Clin Epidemiol. 2006;59:448–56.

[11] Roozenbeek B, Lingsma HF, Perel P, Edwards P, Roberts I, Murray GD, et al. The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. Crit Care. 2011;15:3.

[12] Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. J Clin Epidemiol. 1997;50:45–55.

[13] Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KGM. Polytomous

logistic regression analysis could be applied more often in diagnostic research. J Clin Epidemiol. 2008;61:125–34.

[14] Roukema J, van Loenhout RB, Steyerberg EW, Moons KGM, Bleeker SE, Moll HE. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. J Clin Epidemiol. 2008;61:135–41.

[15] Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika. 1984;71:11–8.

[16] Bull SB, Donner A. A characterization of the efficiency of individualized logistic regressions. Can J Stat. 1993;21:71-8.

[17] Van Calster B, Valentin L, van Holsbeke C, Testa AC, Bourne T, van Huffel S, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol. 2010;10:96.

[18] Agresti A. Categorical data analysis. 3rd ed. John Wiley & Sons; 2013.

[19] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. 2nd ed. Springer Science & Business Media; 2015.

[20] Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol. 1997;26:1323–33.

[21] Liu I, Agresti A. The analysis of ordered categorical data: An overview and a survey of recent developments. Test. 2005;14:1–73.

[22] Strömberg U. Collapsing ordered outcome categories: a note of concern. Am J Epidemiol. 1996;144:421–4.

[23] Anderson JA. Regression and ordered categorical variables. J R Stat Soc Series B Stat Methodol. 1984;46:1–30.

[24] Lunt M. Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. Stat Med. 2005;24:1357–69.

[25] Feldmann U, Steudel I. Methods of ordinal classification applied to medical scoring systems. Stat Med. 2000;19:575–86.

[26] Bender R, Grouven U. Using binary logistic regression models for ordinal data with non-proportional odds. J Clin Epidemiol. 1998;51:809–16.

[27] Guisan A, Harrell FE. Ordinal response regression models in ecology. J Veg Sci. 2000;11:617–26.

[28] Campbell MK, Donner A. Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. J Am Stat Assoc. 1989;84:587–91.

[29] Harrell FE, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, et al. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. Stat Med. 1998;17:909–44.

[30] Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003.

[31] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 2006;22:e184-90.

[32] Kerr KF, Meisner A, Thiessen-Philbrook H, Coca SG, Parikh CR. RiGoR: reporting guidelines to address common sources of bias in risk model development. Biomark Res. 2015;3:2.

[33] Bernau C, Augustin T, Boulesteix A-L. Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. Biometrics. 2013;69:693–702.

[34] Boulesteix A-L, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009;9:85.

[35] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079–107.

[36] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A Stat Soc. 1995;158:419–66.

[37] Ding Y, Tang S, Liao SG, Jia J, Oesterreich S, Lin Y, et al. Bias correction for selecting the minimal-error classifier from many machine learning models. Bioinformatics. 2014;30:3152–8.

[38] Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix A-L. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26:1990–8.

[39] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.

[40] Lukacs PM, Burnham KP, Anderson DR. Model selection bias and Freedman's paradox. Ann Inst Stat Math. 2010;62:117–25.

[41] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56:441–7.

[42] Ye J. On measuring and correcting the effects of data mining and model selection. J Am Stat Assoc. 1998;93:120–31.

# Supplementary Material for "Using Multilevel Outcomes to Construct and Select Biomarker Combinations for Single-level Prediction"

Allison Meisner[1], Chirag R. Parikh[2,3], and Kathleen F. Kerr[4]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

[2]Program of Applied Translational Research, Department of Medicine, Yale School of Medicine, New Haven, Connecticut

[3]Department of Internal Medicine, Veterans Affairs Medical Center, West Haven, Connecticut

[4]Department of Biostatistics, University of Washington, Seattle, Washington

*ameisne1@jhu.edu*

# S1 Constructing Combinations

## S1.1 Cumulative Logit Model with Proportional Odds Did Not Hold

For the simulations where the cumulative logit model with proportional odds did not hold, the biomarkers had multivariate normal distributions conditional on $D$. We considered four different sets of covariance matrices for these distributions, and we call these sets $\Sigma_X$ where $\Sigma_X$ may be $\Sigma_1$, $\Sigma_2$, $\Sigma_3$, or $\Sigma_4$. These four sets of matrices are defined below.

For $K = 3$, we have:

- $\Sigma_1$
  - For $D = 1$, $D = 2$, and $D = 3$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
- $\Sigma_2$
  - For $D = 1$, $D = 2$, and $D = 3$: $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
- $\Sigma_3$
  - For $D = 1$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
  - For $D = 2$ and $D = 3$: $2\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$
- $\Sigma_4$

  - For $D = 1$ and $D = 2$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

  - For $D = 3$: $2\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

For $K = 5$, we have:

- $\Sigma_1$

  - For $D = 1$, $D = 2$, $D = 3$, $D = 4$, and $D = 5$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

- $\Sigma_2$

  - For $D = 1$, $D = 2$, $D = 3$, $D = 4$, and $D = 5$: $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

- $\Sigma_3$

  - For $D = 1$, $D = 2$, and $D = 3$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

  - For $D = 4$ and $D = 5$: $2\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

- $\Sigma_4$

  - For $D = 1$, $D = 2$, $D = 3$, and $D = 4$: $2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

- ◇ For $D = 5$: $2\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

The results for $\Sigma_X = \Sigma_1$ were presented in the paper.

### S1.1.1 $K = 3$

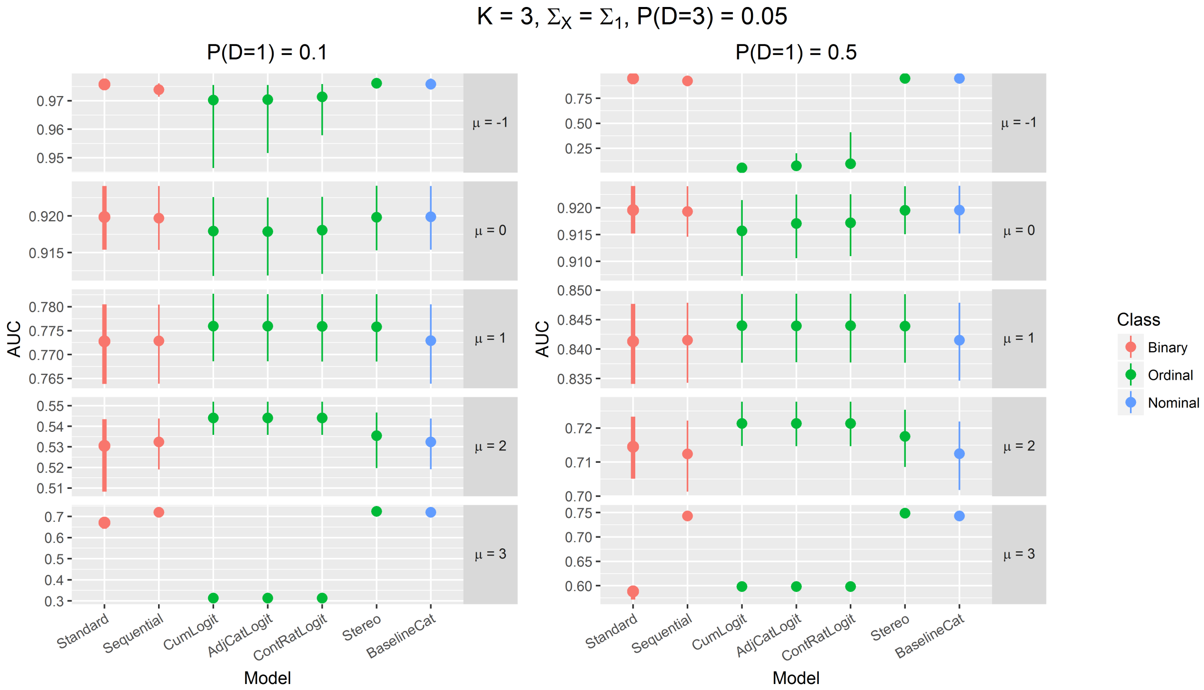

**Figure S1: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.05$, and $\Sigma_X = \Sigma_1$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

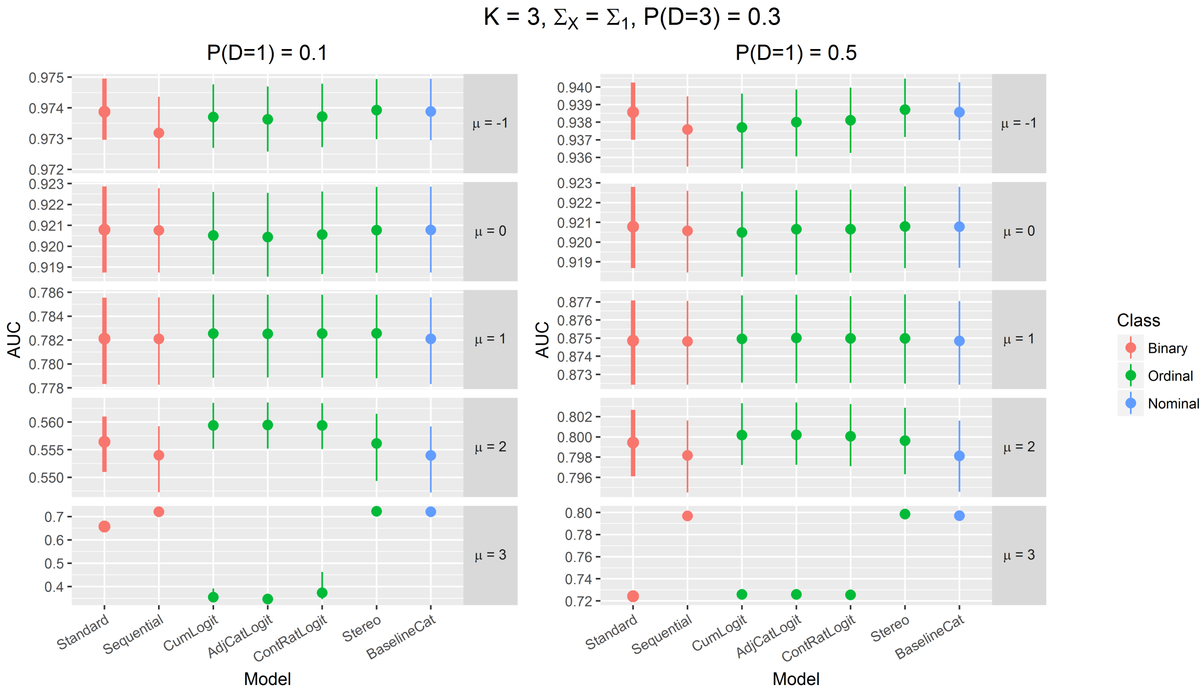

**Figure S2: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.3$, and $\Sigma_X = \Sigma_1$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

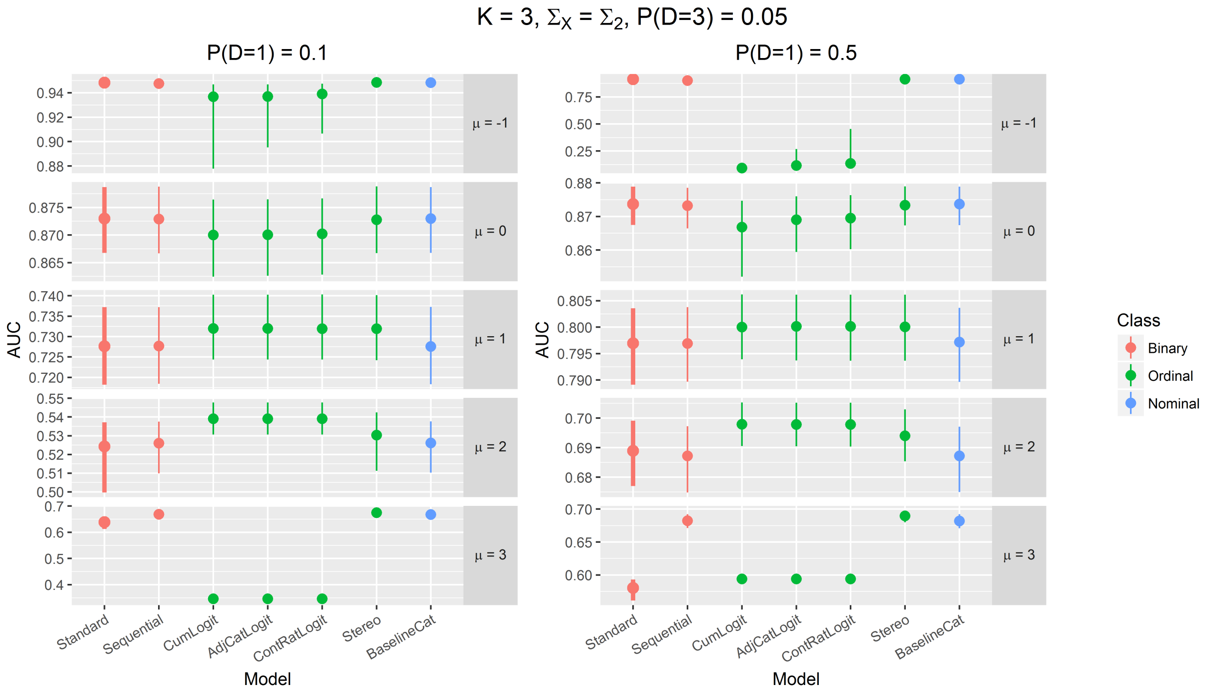

**Figure S3: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.05$, and $\Sigma_X = \Sigma_2$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

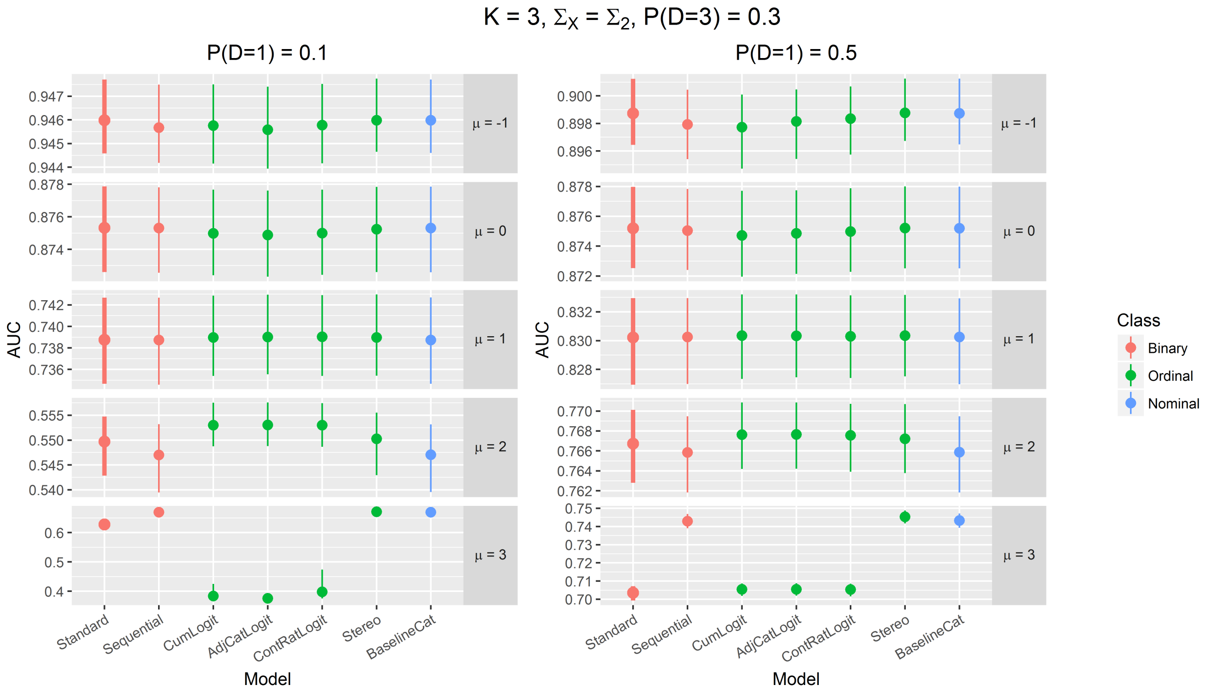

**Figure S4: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.3$, and $\Sigma_X = \Sigma_2$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

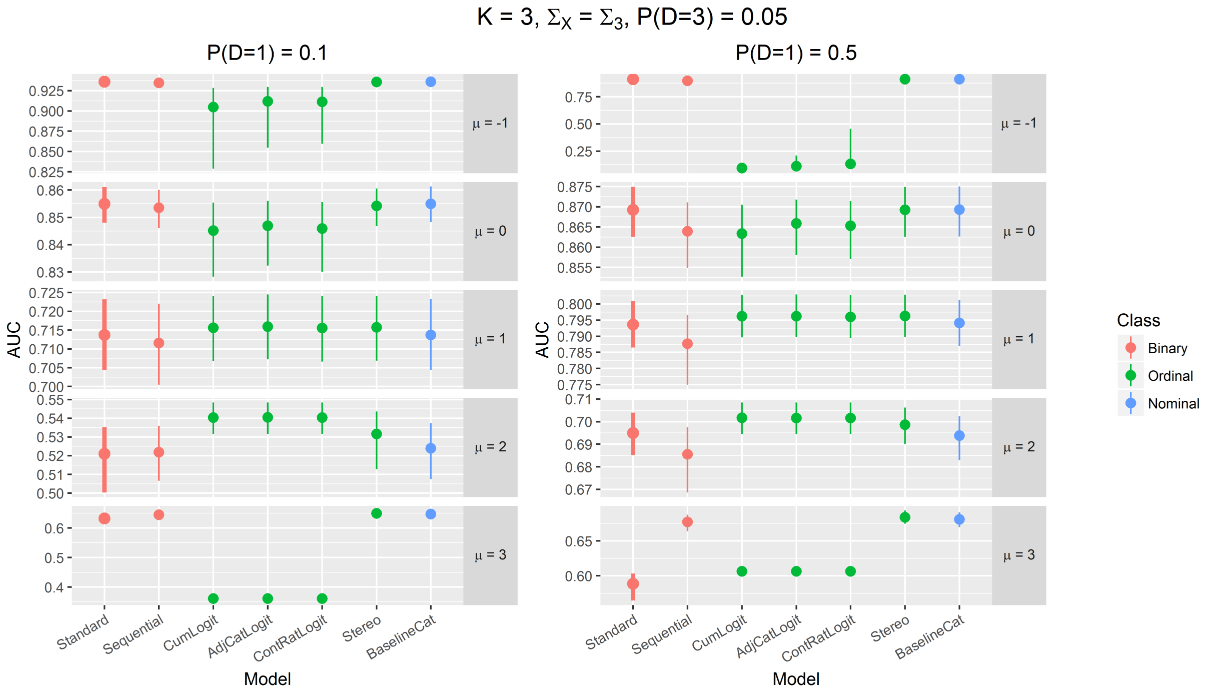

**Figure S5: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.05$, and $\Sigma_X = \Sigma_3$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

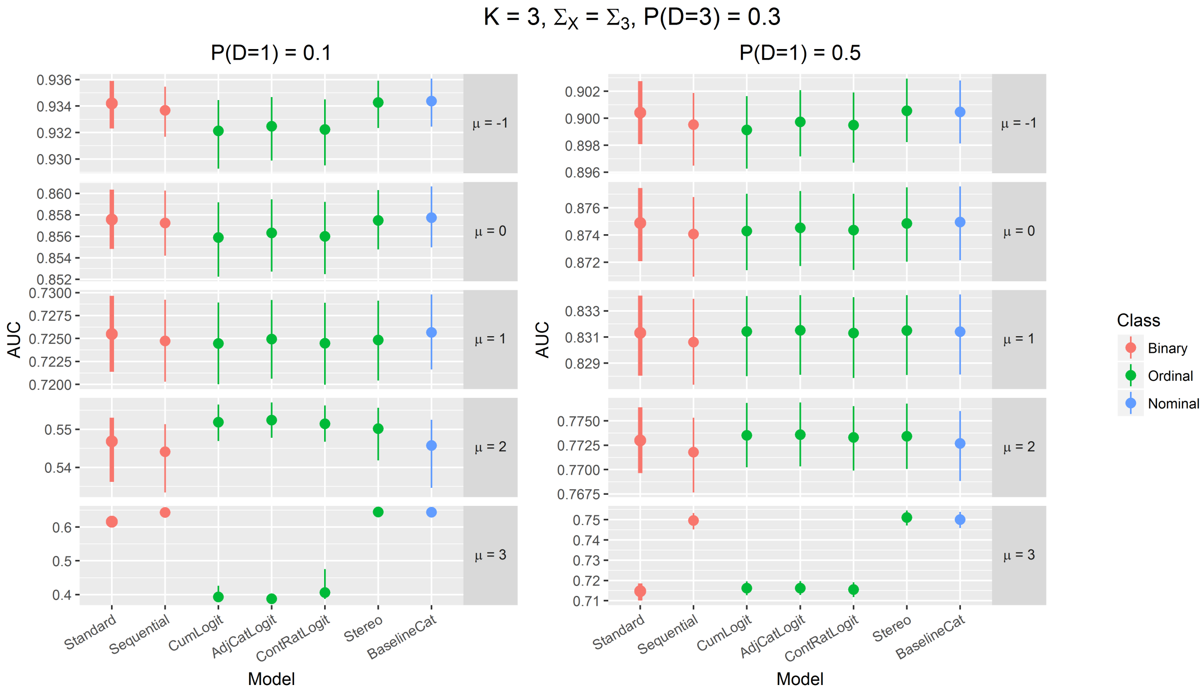

**Figure S6: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.3$, and $\Sigma_X = \Sigma_3$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

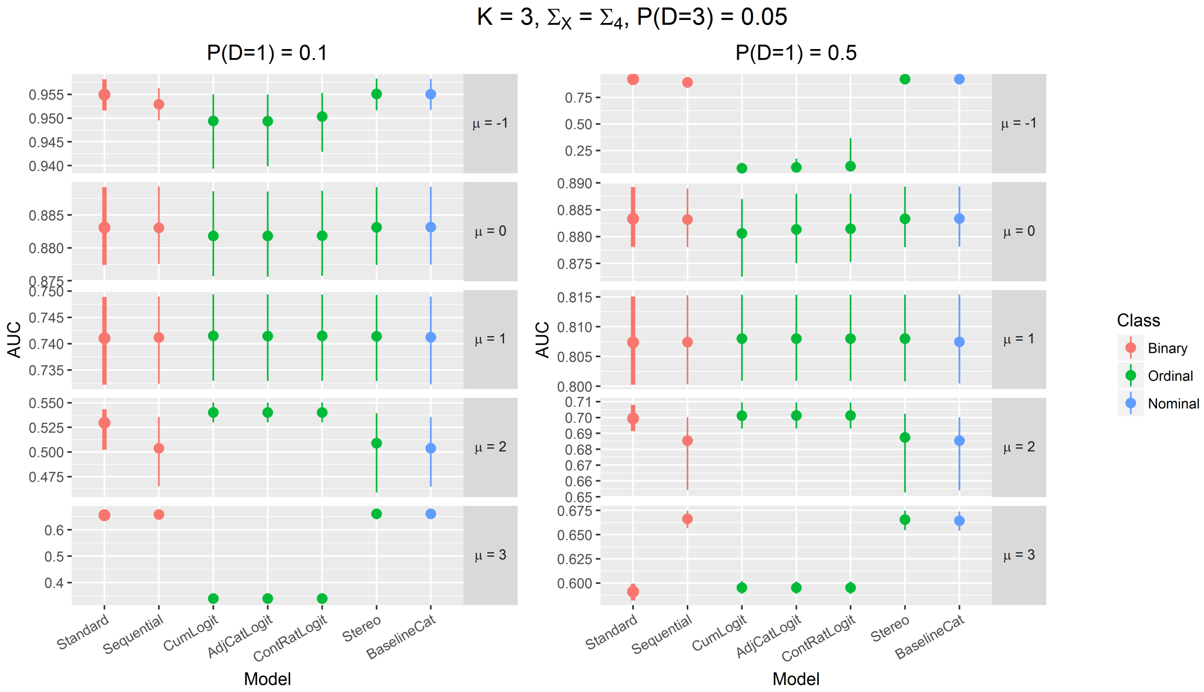

**Figure S7: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.05$, and $\Sigma_X = \Sigma_4$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

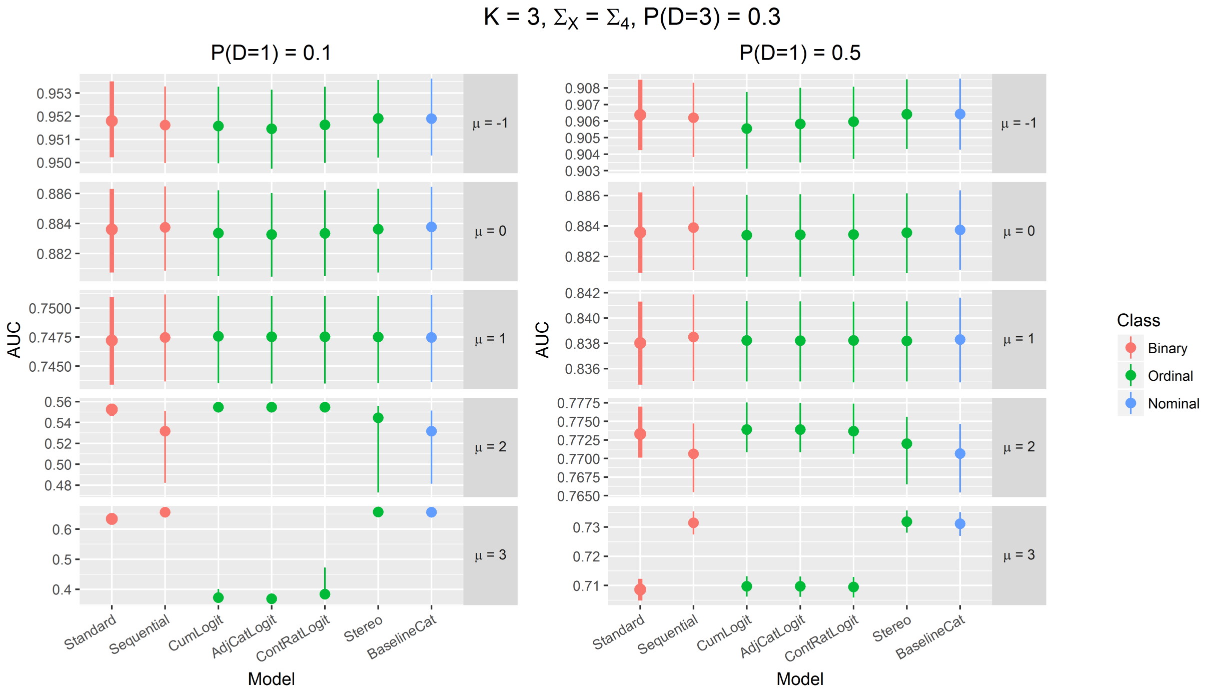

**Figure S8: Simulation results for $K = 3$, $n = 400$, $P(D = 3) = 0.3$, and $\Sigma_X = \Sigma_4$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

### S1.1.2 $K = 5$

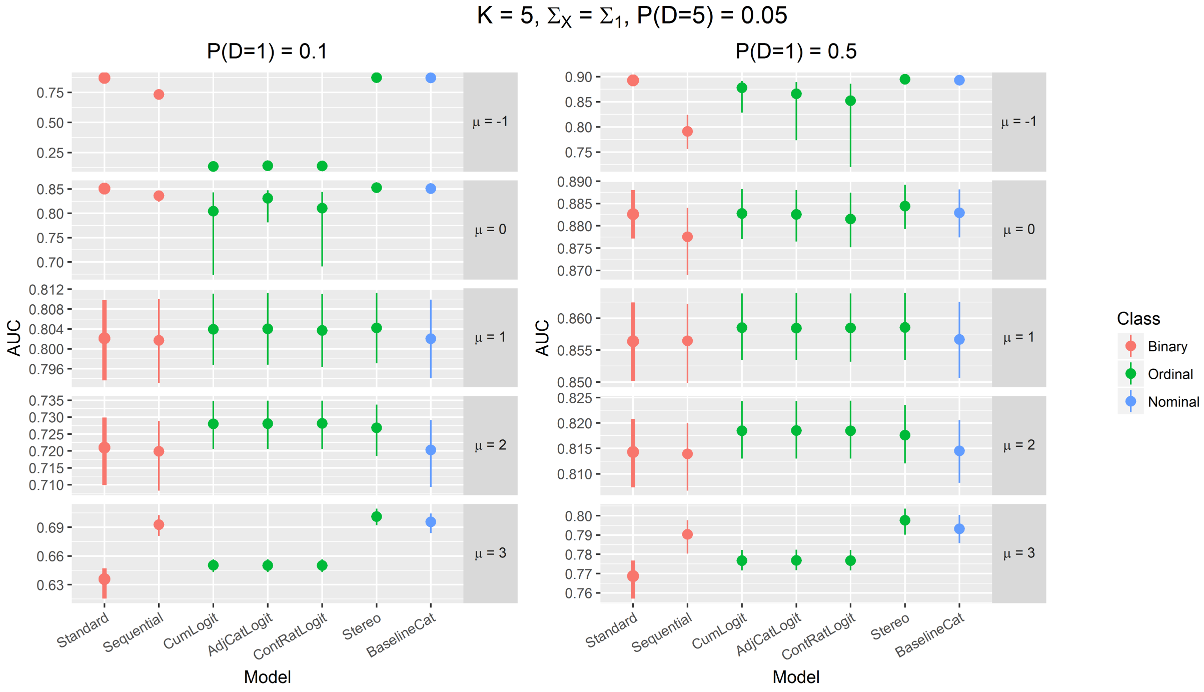

**Figure S9: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.05$, and $\Sigma_X = \Sigma_1$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

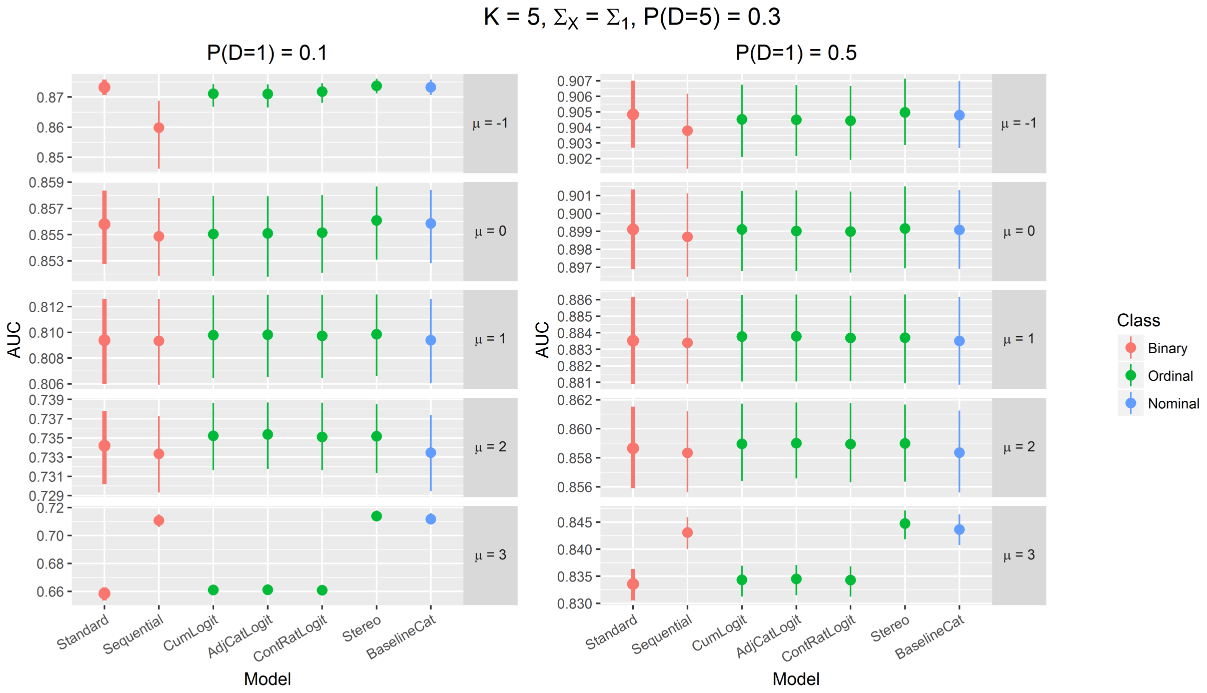

**Figure S10: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.3$, and $\Sigma_X = \Sigma_1$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

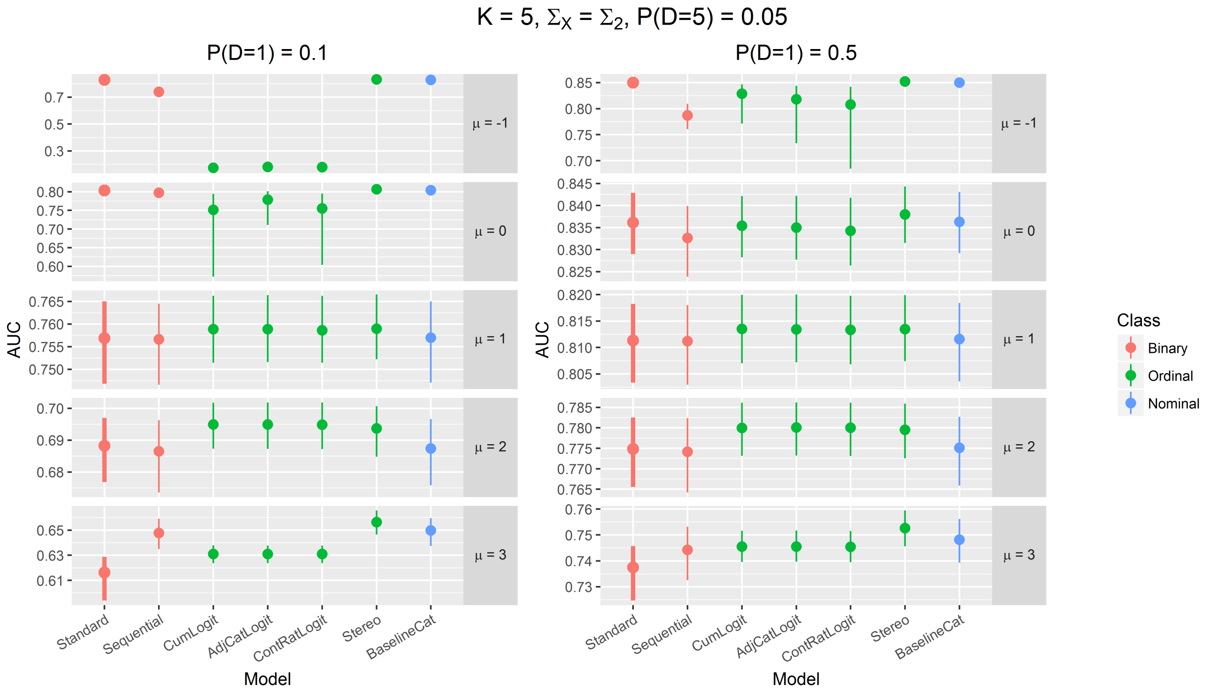

**Figure S11: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.05$, and $\Sigma_X = \Sigma_2$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

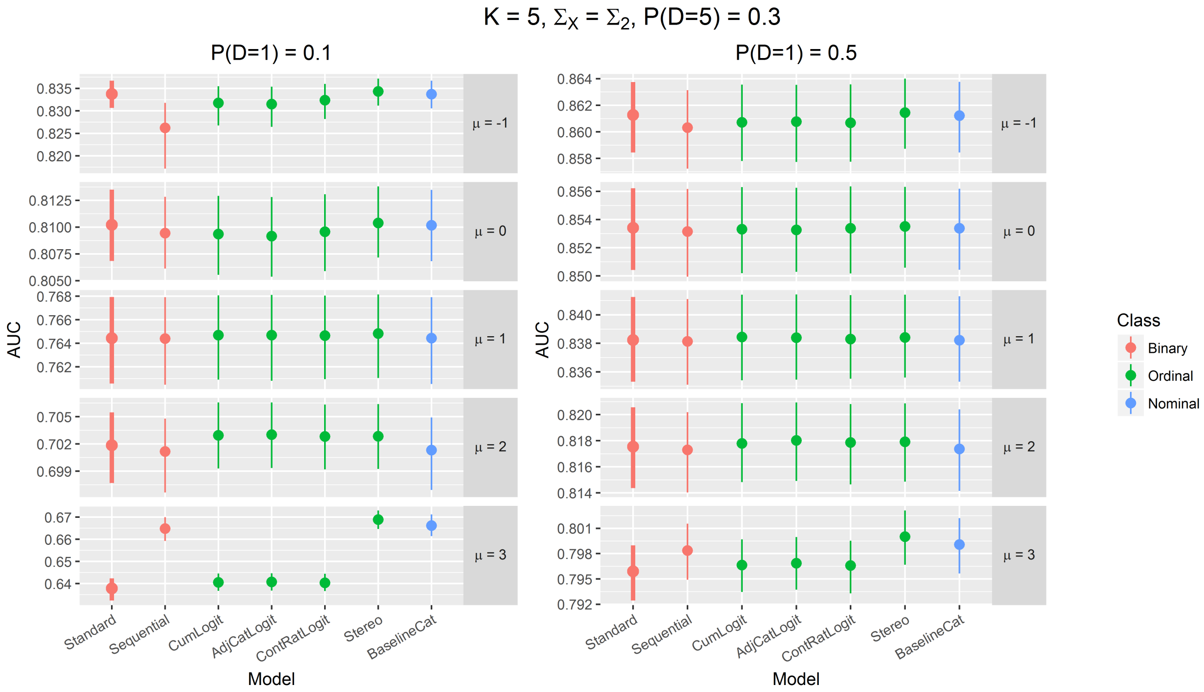

**Figure S12: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.3$, and $\Sigma_X = \Sigma_2$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

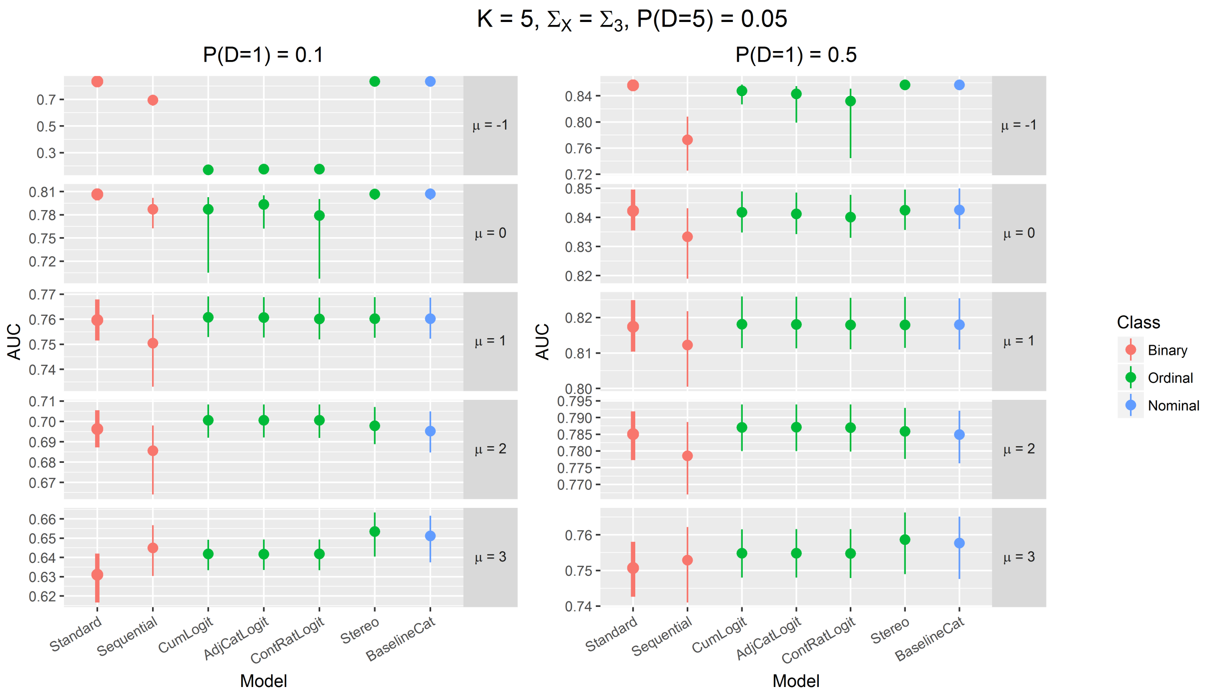

**Figure S13: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.05$, and $\Sigma_X = \Sigma_3$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

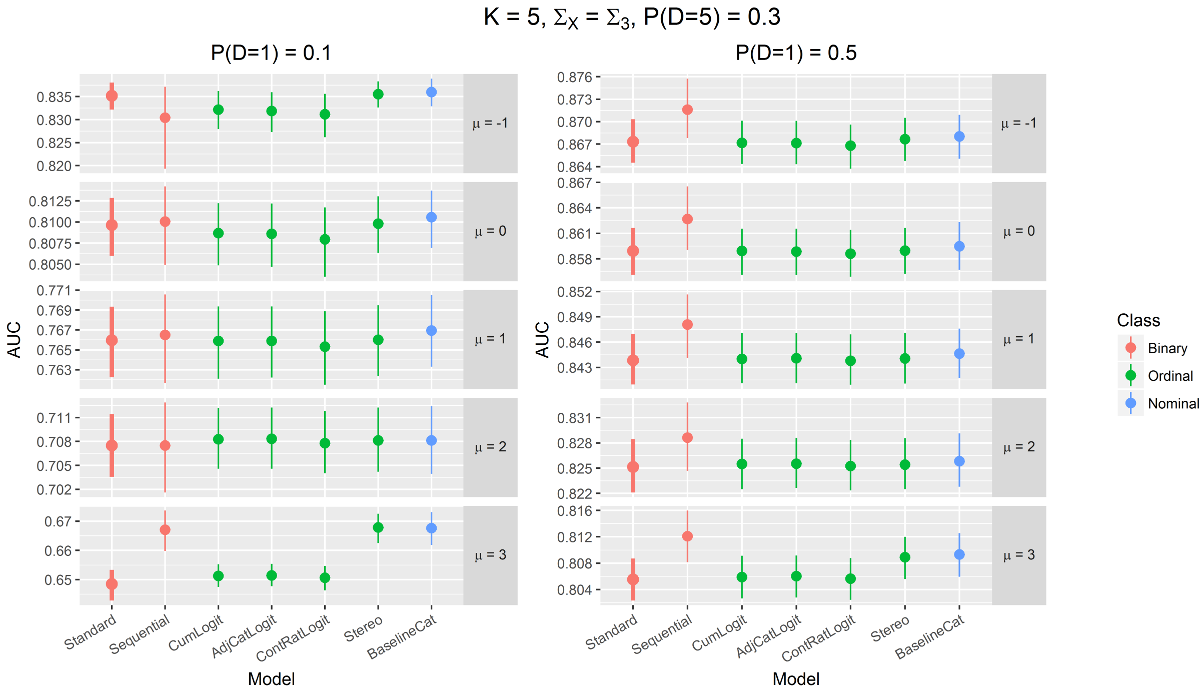

**Figure S14: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.3$, and $\Sigma_X = \Sigma_3$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

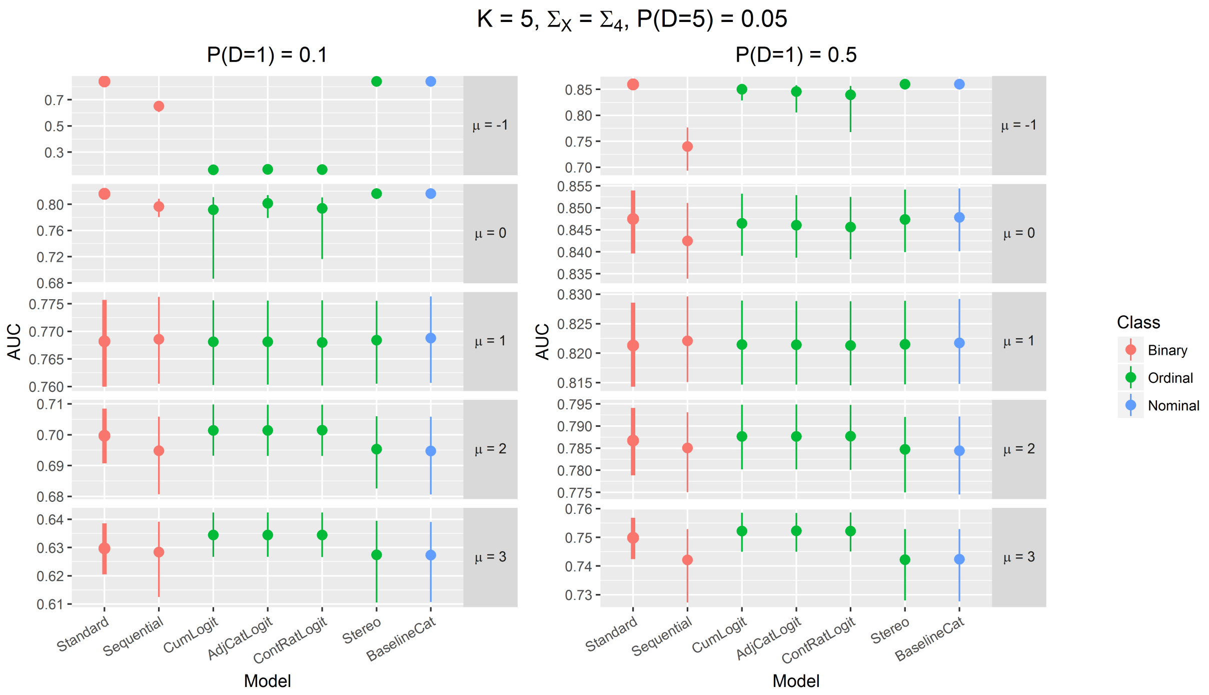

**Figure S15: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.05$, and $\Sigma_X = \Sigma_4$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

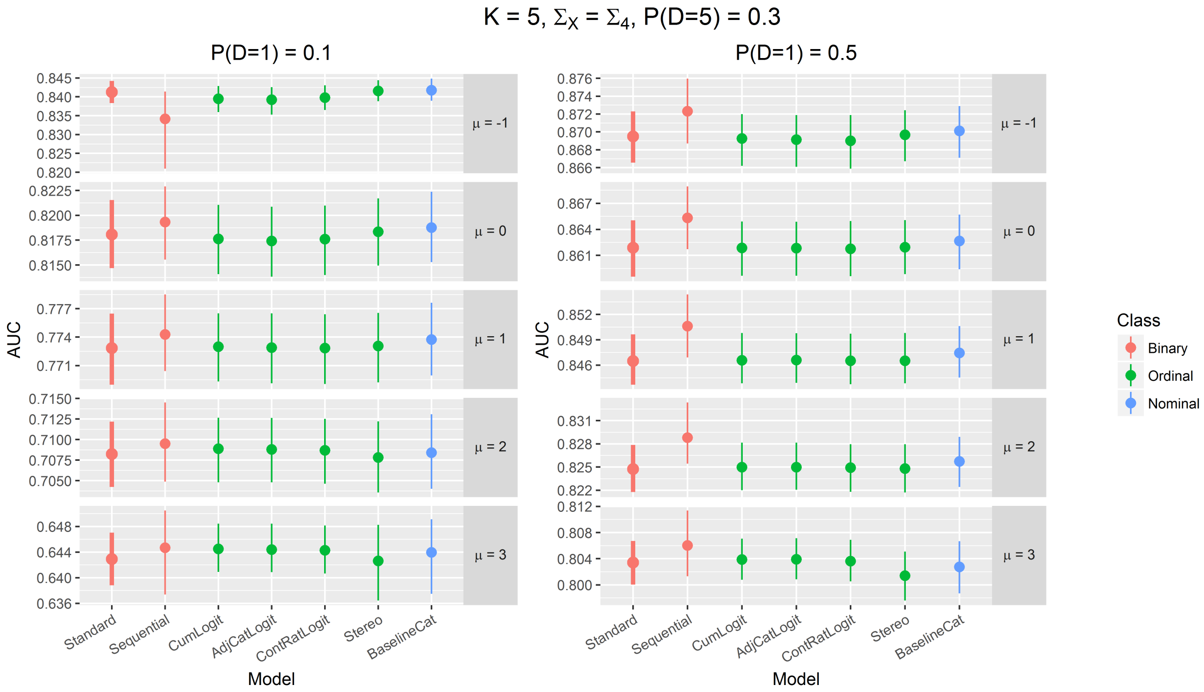

**Figure S16: Simulation results for $K = 5$, $n = 400$, $P(D = 5) = 0.3$, and $\Sigma_X = \Sigma_4$ when the cumulative logit model with proportional odds did not hold.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

## S1.2 Cumulative Logit Model with Proportional Odds Held

### S1.2.1 $K = 3$

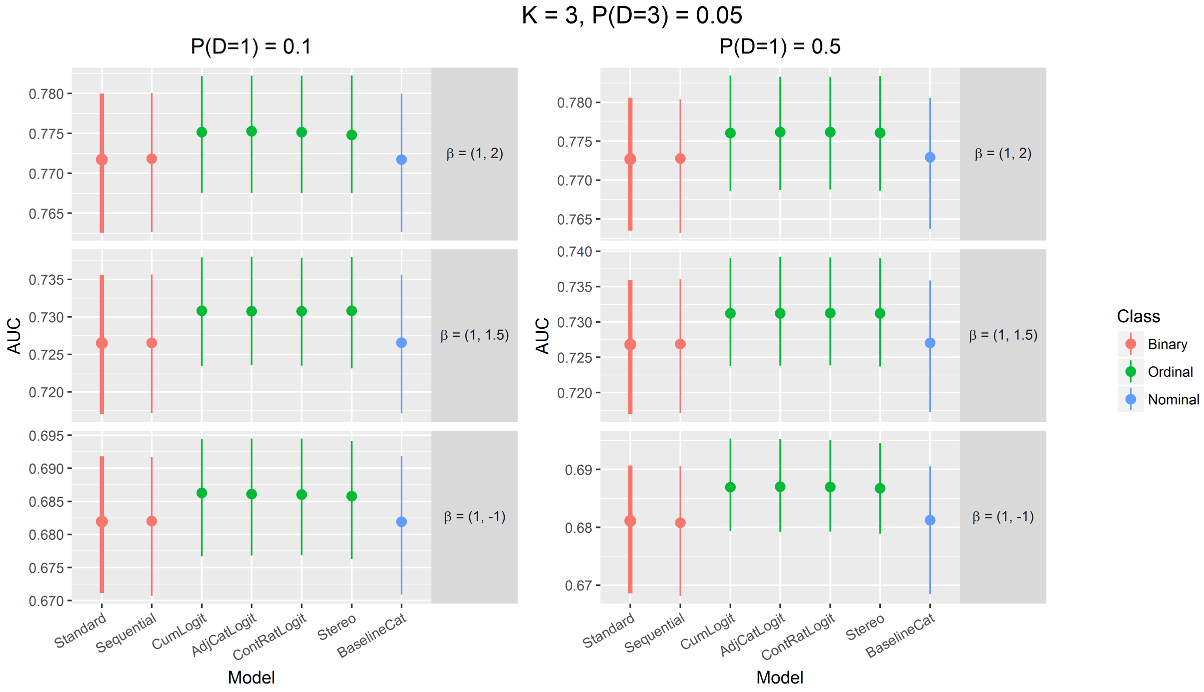

**Figure S17: Simulation results for $K = 3$, $n = 400$, and $P(D = 3) = 0.05$ when the cumulative logit model with proportional odds held.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

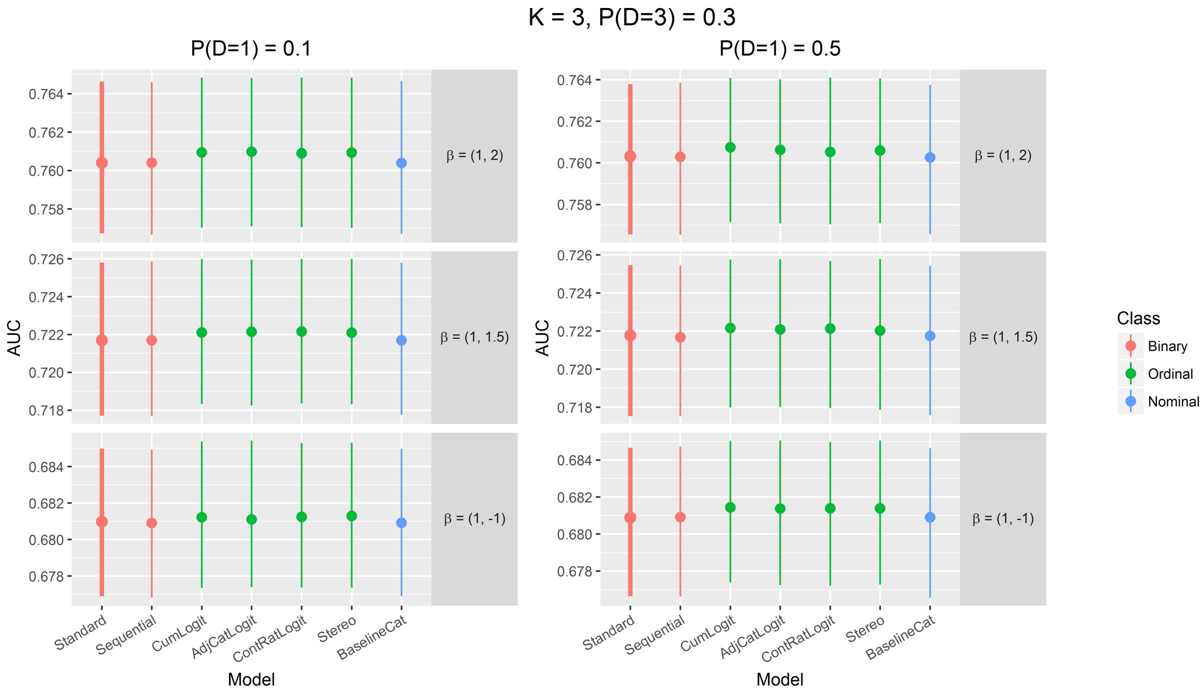

**Figure S18: Simulation results for $K = 3$, $n = 400$, and $P(D = 3) = 0.3$ when the cumulative logit model with proportional odds held.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

### S1.2.2 $K = 5$

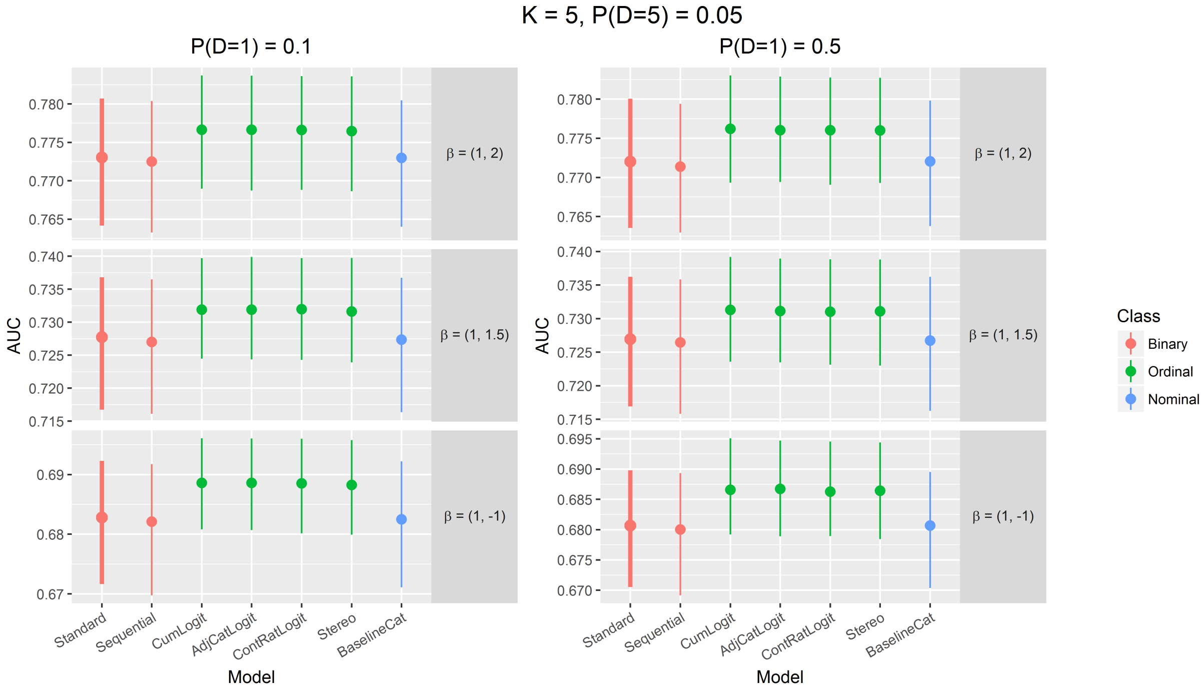

**Figure S19: Simulation results for $K = 5$, $n = 400$, and $P(D = 5) = 0.05$ when the cumulative logit model with proportional odds held.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

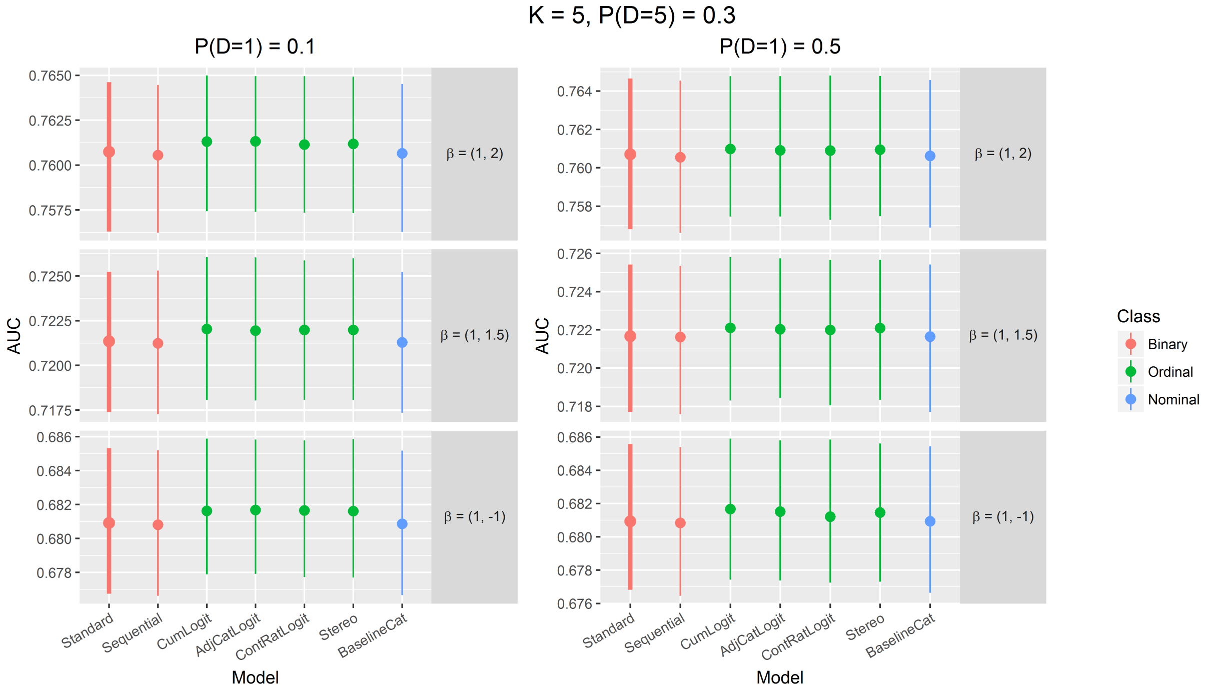

**Figure S20: Simulation results for $K = 5$, $n = 400$, and $P(D = 5) = 0.3$ when the cumulative logit model with proportional odds held.** Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each modeling strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The standard approach is indicated by a slightly thicker line and larger point.

# S2 Combination Selection

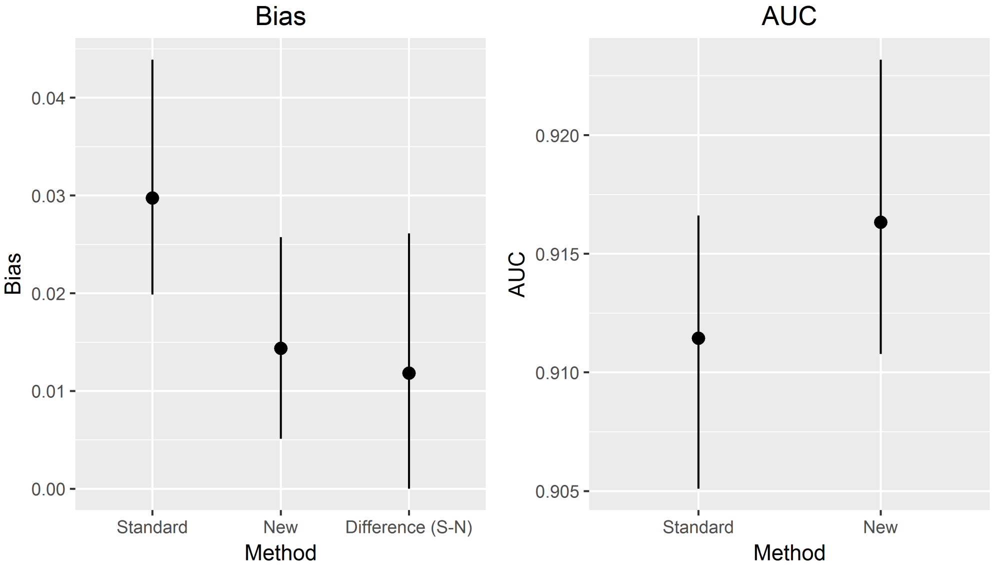

**Figure S21: Results for the proposed combination selection method for Example 1.** The plot on the left gives the median and interquartile range of the estimated model selection bias for the combinations selected by the two approaches (the standard approach and the new approach) and the difference in the estimated bias between the two approaches. The plot on the right gives the median and interquartile range of the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.

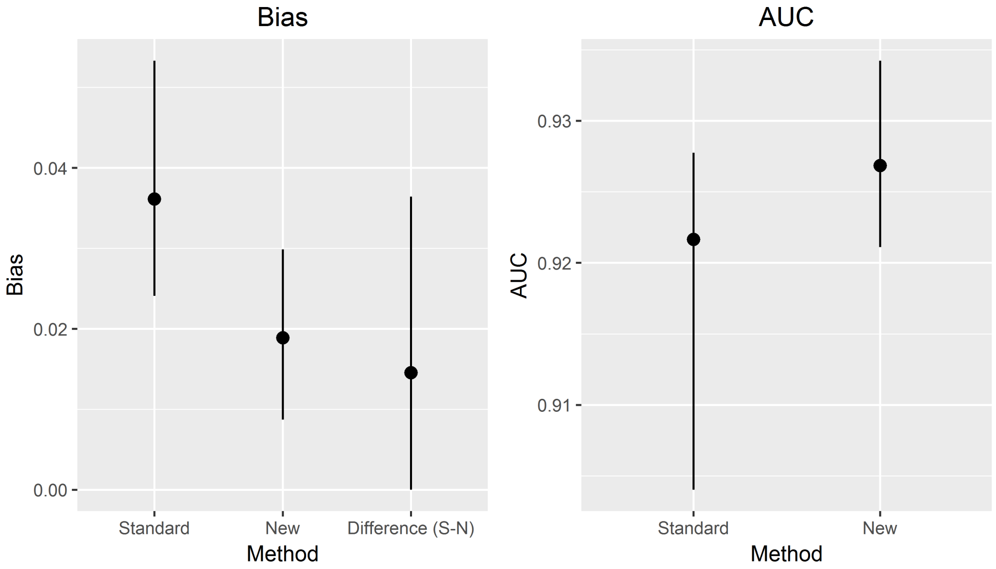

**Figure S22: Results for the proposed combination selection method for Example 2.** The plot on the left gives the median and interquartile range of the estimated model selection bias for the combinations selected by the two approaches (the standard approach and the new approach) and the difference in the estimated bias between the two approaches. The plot on the right gives the median and interquartile range of the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.

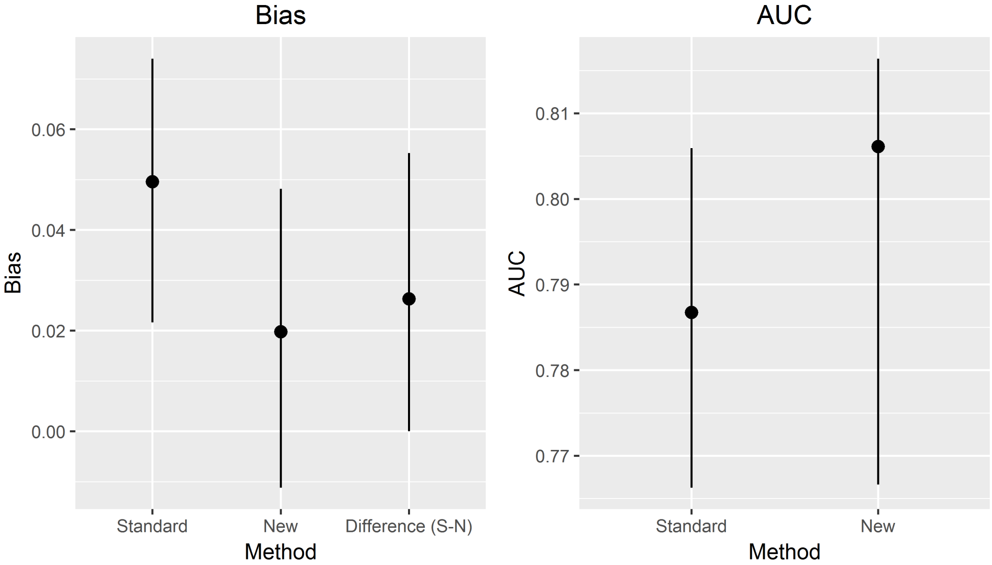

**Figure S23: Results for the proposed combination selection method for Example 3.** The plot on the left gives the median and interquartile range of the estimated model selection bias for the combinations selected by the two approaches (the standard approach and the new approach) and the difference in the estimated bias between the two approaches. The plot on the right gives the median and interquartile range of the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.

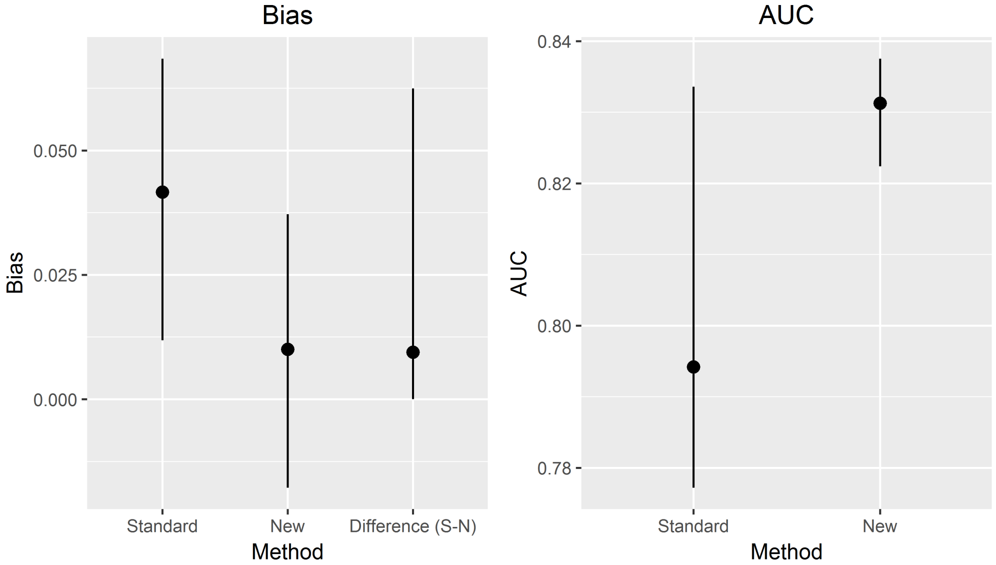

**Figure S24: Results for the proposed combination selection method for Example 4.** The plot on the left gives the median and interquartile range of the estimated model selection bias for the combinations selected by the two approaches (the standard approach and the new approach) and the difference in the estimated bias between the two approaches. The plot on the right gives the median and interquartile range of the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.

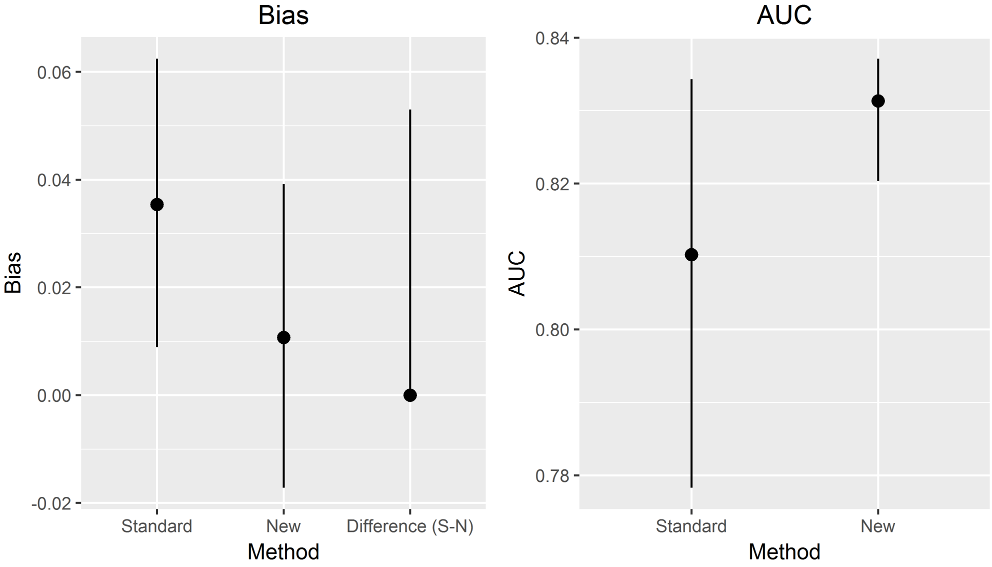

**Figure S25: Results for the proposed combination selection method for Example 5.** The plot on the left gives the median and interquartile range of the estimated model selection bias for the combinations selected by the two approaches (the standard approach and the new approach) and the difference in the estimated bias between the two approaches. The plot on the right gives the median and interquartile range of the AUC for $D = 3$ vs. $D < 3$ in test data for the combinations selected by the two approaches.