Statistics in Biosciences (2020) 12:353–375
https://doi.org/10.1007/s12561-020-09275-2

Methods for Feature Selection in Down-Selection of Vaccine Regimens Based on Multivariate Immune Response Endpoints

Ying Huang[1] · Aliasghar Tarkhan[2]

Received: 5 May 2019 / Revised: 3 December 2019 / Accepted: 19 March 2020 / Published online: 15 April 2020

Abstract

In clinical trials, it is often of interest to compare and order several candidate regimens based on multiple endpoints. For example, in HIV vaccine development, immune response profiles induced by vaccination are key for selecting vaccine regimens to advance to efficacy evaluation. Motivated by the need to rank and choose a few vaccine regimens based on their immunogenicity in phase I trials, Huang et al. (Biostatistics 18(2):230–243, 2017) proposed a ranking/filtering/selection algorithm that down-selects vaccine regimens to satisfy the superiority and non-redundancy criteria, based on multiple immune response endpoints. In practice, many candidate immune response endpoints can be correlated with each other. An important question that remains to be addressed is how to choose a parsimonious set of the available immune response endpoints to effectively compare regimens. In this paper, we propose novel algorithms for selecting immune response endpoints to be used in regimen down-selection, based on importance weights assigned to individual endpoints and their correlation structure. We show through extensive simulation studies that pre-selection of endpoints can substantially improve performance of the subsequent regimen down-selection process. The application of the proposed method is demonstrated using a real example in HIV vaccine research, although the methods are also applicable in general to clinical research for dimension reduction when comparing regimens based on multiple candidate endpoints.

Keywords Correlation · Down-selection · Feature selection · Importance weight · Measurement error · Vaccine trial

Ying Huang and Aliasghar Tarkhan have contributed equally.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12561-020-09275-2) contains supplementary material, which is available to authorized users.

✉ Ying Huang
yhuang@fredhutch.org

Extended author information available on the last page of the article

1 Introduction

When developing a vaccine against a rare disease such as HIV/AIDS, efficacy trials are typically large and operationally challenging to conduct, making it critical to select and rank candidate vaccine regimens based on their immunogenicity in phase I studies before the regimens can be advanced to efficacy evaluation. We and others in the HIV Vaccine Trials Network (HVTN) have been developing statistical approaches and frameworks for this process. For example, out of 15 vaccine regimens studied in Phase I trials by the HVTN, combining 5 unique prime-boost types and 3 Env dose × adjuvant types, Huang et al. [13] described statistical approaches for selecting up to 3 regimens to advance for concurrent testing in a later, multi-regimen HIV vaccine efficacy trial. For maximum operational efficiency, it is desirable to have a strong statistical framework for selecting the most promising regimens in such a process, as the maximum number of regimens allowed to be selected is typically pre-determined based on the budget limit. The design of one such multi-regimen phase IIb HIV vaccine efficacy trial in Southern Africa was laid out by Gilbert et al. [8], to evaluate one or more qualifying prime-boost vaccine regimens for efficacy against a shared placebo arm. In phase I trials designed to inform down-selection of vaccine regimens, the immunogenicity of a given vaccine (as characterized by multiple immune response endpoints such as T-cell or antibody responses) is an essential criterion in regimen selection. Moreover, vaccine-induced multivariate immune response biomarkers play a key role in vaccine development as potential correlates of a vaccine's protective effect in preventing HIV infection; that is, a vaccine's efficacy can be predicted based on the magnitude and breadth of the immune responses elicited by the vaccine [6, 7].

Huang et al. [14] investigated the problem of how to rank and down-select a small number of vaccine regimens based on a given set of immune response endpoints. While others have also studied this type of "pick-the-winner" problem, previous work typically focused either on the selection of a single best regimen based on a few endpoints [19–21] or on the comparison of two regimens with respect to univariate or multivariate endpoints [1, 2, 4, 16, 18, 22, 23]. The particular problem of selecting the best several regimens based on multivariate endpoints has unique challenges. Huang et al. [14] addressed this down-selection problem through the development of formal superiority and non-redundancy criteria for selecting regimens. The formal superiority criterion states that regimens with superior immunogenicity are preferred. The non-redundancy criterion states that when more than one regimen can be advanced, it is desirable to advance regimens with diverse immune profiles such that different vaccines (acting via potentially different mechanisms) for HIV prevention can be evaluated in the efficacy trial. A rank/filtering/selection (RFS) algorithm based on ranking and hypothesis testing was developed [14] to select regimens satisfying these two criteria, where a pre-determined set of immune response endpoints was used for comparison between regimens. Multi-test adjustment was proposed to account for the multiple pairs of regimens for comparison and the multivariate endpoints considered in order to control the probability of selecting regimens with redundant immune profiles into the final set.

In practice, however, phase I immunogenicity studies often yield a large number of candidate immune response endpoints, evaluated over multiple laboratories. These immune response endpoints can involve different immune classes, such as cellular or humoral responses, as well as different antigens. Moreover, these immune response endpoints can be correlated with each other and differ in their strength as predictors of a vaccine's protective effect. Entering all possible immune response endpoints into the down-selection algorithm would create an unnecessary measurement burden and also have a negative impact on the down-selection process. A numeric example in Huang et al. [14] shows that when multivariate immune response endpoints are highly correlated with each other, the multi-test adjustment implemented in the RFS algorithm can be too conservative and have reduced power to detect differences between regimens. How to choose a parsimonious subset of immune response endpoints from a larger number of candidate endpoints to enter into down-selection is an important open question that needs to be addressed. In this paper, we aim to fill this gap. We investigate and propose algorithms for selecting immune response endpoints to enter into down-selection, taking into account information about the importance of individual endpoints as correlates of a vaccine's protective effect as well as the correlation structure among these endpoints. While this research is motivated by the down-selection problem in the HIV vaccine trial setting, it also has more general applications for feature selection when a large number of candidate endpoints are available for comparison between intervention groups.

In Sect. 2, we describe the problem setting. After briefly reviewing the down-selection criteria and the RFS down-selection algorithm based on multivariate endpoints, we propose our feature selection algorithms for selecting endpoints to enter the RFS algorithm, based on importance weights assigned to individual endpoints and their correlation structure. In addition, we introduce a special risk model setting to (1) help interested readers understand the connection between down-selection based on immune response data and the ranking of vaccine efficacy and (2) provide guidance on the practical choice of importance weights for individual endpoints. In Sect. 3, we conduct extensive numerical studies to compare the performance of the proposed feature selection algorithms combined with RFS to that of naive RFS without feature selection under various settings. In Sect. 4, we apply the proposed methods to an immune response data example from HIV vaccine trials. We then complete the paper with concluding remarks.

2 Methods

We consider a phase I randomized immunogenicity trial with m vaccine arms. Suppose a set of vaccine-induced immune response endpoints $\boldsymbol{X}$ of dimension p can be measured at a single time point for every vaccine recipient. Let $j = 1, \dots, m$ be the regimen indicator, with n_j indicating the sample size for the jth regimen. Let k be the participant indicator, which takes values $1, \dots, n_j$ for regimen j. Let $i = 1, \dots, p$ be the immune response endpoint indicator. We use X_{ijk} to indicate the value for the ith

immune response endpoint from participant k from regimen j and let μ_{ij} indicate the mean of X_{ijk}.

The goal of the down-selection process is to select up to $Q(1 \leq Q < M)$ regimens that are superior among the set of candidate regimens and that have unique immune profiles. Huang et al. [14] developed superiority and non-redundancy criteria for the down-selection process, as well as a rank/filtering/selection (RFS) algorithm for selecting regimens based on multivariate immune response endpoints to meet these criteria. Below, we briefly review the down-selection criteria and the RFS algorithm; further details can be found in Huang et al. [14].

2.1 Review of the Down-Selection Criteria and Algorithm

Without loss of generality, assume that for each individual immune response endpoint considered, a larger immune response is associated with a better protective effect of the vaccine and thus is desirable. In practice, it is possible for some immune response measures to have a "harmful" effect, in the sense that a higher immune response is associated with lower vaccine efficacy. In such a situation, one can use the negative value of the original immune response measurement as the endpoint of interest in the framework we will describe next. Suppose the comparison between regimens with respect to individual endpoints is based on mean values. That is, a regimen A is considered *superior*, *inferior*, or *equivalent* to another regimen B with respect to a particular endpoint i if the mean of the endpoint in A is larger ($\mu_{Ai} > \mu_{Bi}$), smaller ($\mu_{Ai} < \mu_{Bi}$), or the same ($\mu_{Ai} = \mu_{Bi}$) compared to that in B. A regimen A is defined to be *superior* to regimen B (or equivalently B inferior to A) with respect to their immune profiles if A is superior to B with respect to at least one endpoint and is not inferior to B with respect to any endpoint. Two regimens A and B are said to have *equivalent* immune profiles if they are equivalent with respect to each endpoint considered. When a set of vaccine regimens is selected, the superiority criterion [13] is fully satisfied if no selected regimen is inferior to any other regimen that enters down-selection. The non-redundancy criterion [14] argues that with limited resources for conducting the efficacy trial, it is desirable to select vaccine regimens with diverse immune profiles, so that diverse potential mechanisms that mediate a vaccine's protective effects can be investigated in the efficacy trial. A vaccine regimen A is defined to be *non-redundant* to regimen B if A is not superior, not inferior, and not equivalent to B with respect to its immune profile. Non-redundancy is satisfied for a set of selected regimens if it is satisfied for any regimen pair in the set. An example illustrating the two criteria is presented in Web Supplementary Fig. 1.

Targeting these two criteria, Huang et al. [14] proposed a "ranking, filtering, and selection" (RFS) algorithm to down-select regimens. In RFS, all regimens are first ranked according to a univariate summary score across individual endpoints (such as the weighted mean $\sum_{i=1}^{p} w_i\mu_{ij}$), with a pre-determined positive weight w_i for each individual endpoint reflecting its importance or usefulness in predicting vaccine efficacy. The top-ranked regimen is selected first, after which the rest of the regimens are evaluated sequentially according to their rank. In each step, a new regimen is compared with each regimen in the set selected earlier. Through hypothesis testing, if relative to any regimen already selected, one fails to declare that the new regimen is superior or non-redundant, the new regimen will not be selected into the set; otherwise, the new regimen is selected and all regimens selected earlier that fail to show non-redundancy to the newly selected regimen will be filtered out. This process is repeated until Q regimens have been selected or until all regimens have been evaluated once. As HIV vaccine efficacy trials are large and operationally challenging, satisfaction of the non-redundancy criterion is of paramount importance to avoid wasting valuable resources by advancing redundant regimens to efficacy evaluation. The RFS algorithm controls the probability of selecting non-redundant regimens into the final set through Bonferroni correction and uses the Holm procedure [10] to account for the comparison between multiple regimen pairs and the testing of multiple immune response endpoints. Details of the RFS algorithms are included in Web Supplementary Appendix A.

When the vaccine-induced immune responses are uncorrelated or have only small correlations between each other, the RFS algorithm performs well in selecting the desired regimens and excluding undesired regimens. However, as also demonstrated in Huang et al. [14], the presence of highly correlated immune response endpoints can decrease performance of the down-selection process. In particular, the multi-test adjustment to declare a significant difference between regimens can be too conservative when the endpoints are highly correlated, leading to sub-optimal power for detecting a difference between the regimens. Moreover, in vaccine immunogenicity studies, many candidate immune response endpoints are typically available; some of them might be useful surrogate endpoints for predicting a vaccine's protective effect (albeit with inherent measurement error), whereas others might have limited or no utility as surrogate endpoints, as they are not actually related to differential vaccine efficacy across regimens. Importantly, including all available candidate endpoints could potentially reduce performance of the down-selection process and lead to unnecessarily high measurement cost in phase I down-selection trials.

The objective of this paper is thus to develop dimension reduction algorithms for pre-selecting a subset of immune response endpoints most useful for down-selection to feed into the RFS algorithm. Next we propose two different algorithms for feature selection based on the correlation structure and/or relative importance of immune response endpoints. We also present special model settings for infection risk conditional on vaccine-induced immune response endpoints to help interested readers understand the connection between down-selection based on immune response data and the ranking of vaccine efficacy and the rational for practical choices of importance weights.

2.2 Penalization-Based Feature Selection Method

In this section, we propose a penalization-based feature selection method based on optimizing the reward of adding more endpoints for use in regimen down-selection. One way to choose a set of useful and not highly correlated endpoints is to consider the importance of adding immune response endpoints as a gain and the correlation among immune response endpoints as a cost. We propose to write a reward function as

$$R(\mathbf{w}_S, \boldsymbol{\Sigma}_S) = g(\boldsymbol{w}_S) - c(\boldsymbol{\Sigma}_S). \tag{1}$$

In (1), S is a subset of chosen immune response endpoints; $\mathbf{w}_S = \{w_i : i \in S\}$ (with $w_i > 0$) and $\boldsymbol{\Sigma}_S$ are the vector of importance scores and the correlation matrix of immune response endpoints in subset S; $g(.)$ and $c(.)$ are the gain and the cost functions corresponding to adding more immune response endpoints. Note that both $g(\mathbf{w}_S)$ and $c(\boldsymbol{\Sigma}_S)$ are increasing functions of $|S|$, i.e., the cardinality of set S. Thus, there should be an optimum subset of immune response endpoints that maximizes the reward function in (1) as

$$S^* = \underset{S \subseteq \{1,\ldots,p\}}{\operatorname{argmax}} R(\mathbf{w}_S, \boldsymbol{\Sigma}_S). \tag{2}$$

For general functions $g(.)$ and $c(.)$, we need to search all possible subsets of immune response endpoints S. For instance, with p as the number of immune response endpoints, we need to search for all $2^p - 1$ (empty set is excluded) possible subsets. It is thus desirable to choose $g(.)$ and $c(.)$ such as to minimize the number of searches. We propose a simple and efficient way of defining $g(.)$ and $c(.)$ as additive functions of $\mathbf{w}_S$ and $\boldsymbol{\Sigma}_S$ as

$$R(\mathbf{w}_S, \Sigma_S) = \sum_{i \in S} w_i - \lambda \sum_{i,i' \in S, i \neq i'} |\rho_{ii'}| \tag{3}$$

where λ controls the cost due to correlation among the chosen immune response endpoints in set S, and needs to be chosen carefully. If we choose a very small λ, we allow highly correlated immune response endpoints to enter the RFS algorithm, which can result in a more conservative decision regarding the comparison between regimens. If we choose a very large λ, we might discard some important immune response endpoints, which can result in sub-optimal performance for selecting the desired regimens. Here, we use a simple approach to alleviate this issue by normalizing the importance vector as,

$$w_i^{norm} = \frac{w_j}{\max_{j \in \{1,\ldots,p\}} w_j}. \tag{4}$$

Now we can find S^* as

$$S^* = \underset{S \subseteq \{1,\ldots,p\}}{\operatorname{argmax}} \left(\sum_{i \in S} w_i^{norm} - b \sum_{i,i' \in S, i < i'} |\rho_{ii'}| \right), \tag{5}$$

where b is a new tuning parameter establishing a trade-off between reward and penalty of adding more immune response endpoints; we use constraint $i < i'$ to choose each correlation coefficient only one time (note that the correlation matrix is symmetric). Assume that we have found a subset of immune response endpoints S so far, the reward of adding a new immune response endpoint $i' \notin S$ to S is

$$
\begin{aligned}
r_{i'} &= R(\mathbf{w}^{norm}_{S\cup\{i'\}}, \Sigma_{S\cup\{i'\}}) - R(\mathbf{w}^{norm}_{S}, \Sigma_S) \\
&= w^{norm}_{i'} - b \sum_{i\in S} |\rho_{ii'}|.
\end{aligned} \tag{6}
$$

Therefore, adding a new endpoint i' to the chosen subset S has a reward of $w^{norm}_{i'} - b\sum_{i\in S}|\rho_{ii'}|$. This observation guides us to two rules: (1) choose endpoints with highest rewards first; and (2) exclude endpoints with negative rewards. Therefore, we can propose a heuristic sequential algorithm, based on these two rules, that significantly reduces search complexity. We start with one immune response endpoint with highest w^{norm} (or highest importance) and choose the second immune response endpoint that gives the highest reward, i.e, highest $r_{i'}$ in (6). We continue adding new immune response endpoints until there is no reward, i.e., $r_{i'}$ is negative for all remaining $i' \notin S$. Fig. 1 summarizes our proposed heuristic algorithm, the property of which is stated in Theorem 1 below.

Theorem 1 *Assume that the proposed heuristic algorithm stops with the set* S. *There is no further increase in* $R(\mathbf{w}^{norm}_S, \boldsymbol{\Sigma}_S)$ *by adding any more endpoints, i.e.,* S *gives a local maximum of the reward function in* (3).

Proof The heuristic algorithm stops with the chosen set S, meaning that for any single-added endpoint $k \notin S$ we have $r_k < 0$. Therefore, there is no improvement in the overall reward by adding just one endpoint. However, the possibility remains that adding a set of endpoints to the available set S might improve the overall reward. We show that this is not possible. Assume that after stopping with set S by our heuristic algorithm, there exists S^* such that $R(w^{norm}_{S\cup S^*}, \boldsymbol{\Sigma}_{S\cup S^*}) > R(w^{norm}_S, \boldsymbol{\Sigma}_S)$. We can then find S' and k' such that $S^* = S' \cup k'$ and $R(w^{norm}_{S\cup S'}, \boldsymbol{\Sigma}_{S\cup S'}) \leq R(w^{norm}_S, \boldsymbol{\Sigma}_S)$. If not, we can replace S^* with any S' which has one less endpoint than S^*, and continue such decomposition. A desired decomposition is guaranteed at least when $S^\star$ contains only two endpointS, according to the definition of S.

Note that we already showed that for $S' = \emptyset$ (i.e., adding only one endpoint), the situation described above can not happen. Then assume that $S' \neq 0$ and we aim to add more than one endpoint. To be able to add endpoint k' along with set S' (i.e., at least two endpoints), we need to have

$$
\begin{aligned}
r_{k'} &= w^{norm}_{k'} - b \sum_{i\in S\cup S'} |\rho_{ik'}| \\
&= w^{norm}_{k'} - b \sum_{i\in S} |\rho_{ik'}| - b \sum_{i\in S'} |\rho_{ik'}| > 0.
\end{aligned} \tag{7}
$$

Since $|\rho_{ik'}| \geq 0$ for any pair (i, k') and $b \geq 0$, (7) at least requires that $w^{norm}_{k'} - b\sum_{i\in S}|\rho_{ik'}| > 0$, which means adding a single endpoint like k' to set S can improve the overall reward. This is in contradiction with the stopping criterion of our proposed heuristic algorithm, because it was stopped at set S with no further improvement in the overall reward by adding a single endpoint. Thus, we conclude that there is not such a set S' and that $R(\mathbf{w}^{norm}_S, \boldsymbol{\Sigma}_S)$ cannot be further improved after it is stopped with set S. □

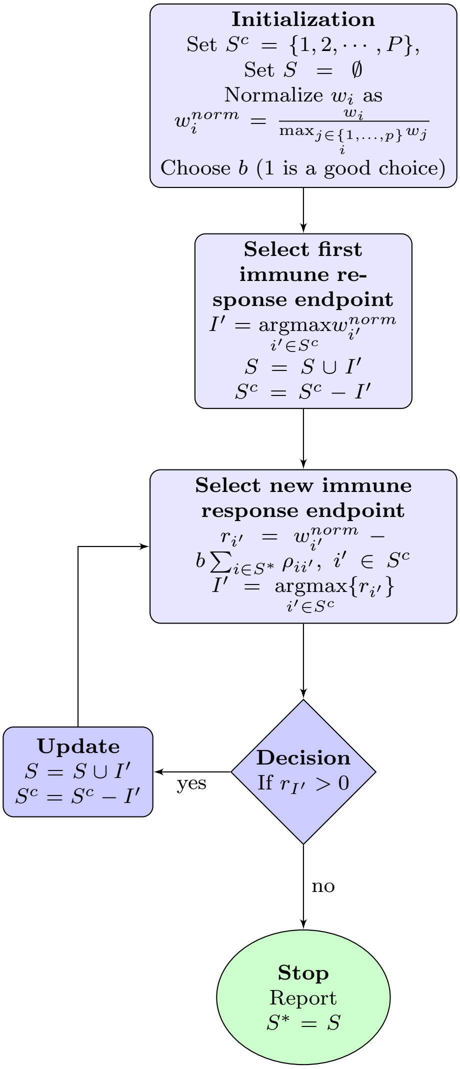
Initialization
Set $S^c = \{1, 2, \cdots, P\}$,
Set $S = \emptyset$
Normalize w_i as
$w_i^{norm} = \frac{w_i}{\max_{j \in \{1,\ldots,p\}} w_j}$
Choose b (1 is a good choice)
Select first immune response endpoint
$I' = \underset{i' \in S^c}{\text{argmax}} w_{i'}^{norm}$
$S = S \cup I'$
$S^c = S^c - I'$
Select new immune response endpoint
$r_{i'} = w_{i'}^{norm} - b \sum_{i \in S^*} \rho_{ii'}, \; i' \in S^c$
$I' = \underset{i' \in S^c}{\text{argmax}} \{r_{i'}\}$
Decision
If $r_{I'} > 0$
yes
Update
$S = S \cup I'$
$S^c = S^c - I'$
no
Stop
Report
$S^* = S$

Fig. 1 System flowchart of the proposed penalization-based algorithm. Here, w_i is the importance weight for endpoint i. At each step, let S indicate the set of endpoints already selected, let S^c indicate the set of endpoints to be selected from, and let I' indicate the endpoint that will move from S^c to S. Here $r_{i'}$ is the reward for moving endpoint i' from S^c to S, and S^* is the final set of selected immune response endpoints that will be used in regimen down-selection

Note that $w_i^{norm} \in [0, 1]$ (thanks to normalization) and $|\rho_{ii'}| \in [0, 1]$, and simulation results for different settings suggest that $b = 1$ can be a legitimate choice.

2.3 Bayesian Information Criterion (BIC)-Based Feature Selection Method

In the previous section, we dealt with an unsupervised learning problem in (5) and came up with a legitimate choice of the tuning parameter after re-parameterization of our reward function. As an alternative way to tackle this problem, we adopt clustering methods to select an appropriate subset of the immune response endpoints based on the observed values in different regimens. We assume that the correlation matrix $\boldsymbol{\Sigma}$ among multivariate immune response endpoints is available and can be used for clustering. We adopt the BIC-based clustering method presented by Fraley and Raftery [5] and use the "NbClust" R package [3] for clustering. By clustering, the most correlated immune response endpoints nest in the same cluster. When there is stronger correlation among immune response endpoints, there will be fewer clusters and vice versa. Now we need to select the desired subset of immune response endpoints based on the clustering results. We propose to choose one endpoint per cluster that has the highest importance value (i.e., w) in that cluster. Note that our proposed BIC-based selection method tends to choose fewer immune response endpoints when immune response endpoints are highly correlated, since there are fewer clusters from which to choose one endpoint. Figure 2 shows the flowchart of the BIC-based selection method.

2.4 Importance Weights for Immune Response Endpoints

To implement the proposed feature selection methods, we need an importance weight w_i reflecting the importance of each immune response endpoint i to a vaccine's protective effect. The importance weight is also used later in the RFS algorithm to compute a weighted average across endpoints for ranking vaccine regimens. In practice, the assignment of these importance weights would require biological knowledge and/or data-driven hypotheses about the importance of individual immune response endpoints; sensitivity analyses based on a range of choices for weights are always helpful to evaluate the robustness of feature selection to different weight choices. For example, expert opinions could be elicited from lab scientists regarding the putative predictive clinical importance to vaccine efficacy of each immune response endpoint. When there are preliminary vaccine trial data available for modeling the risk of HIV as a function of vaccination

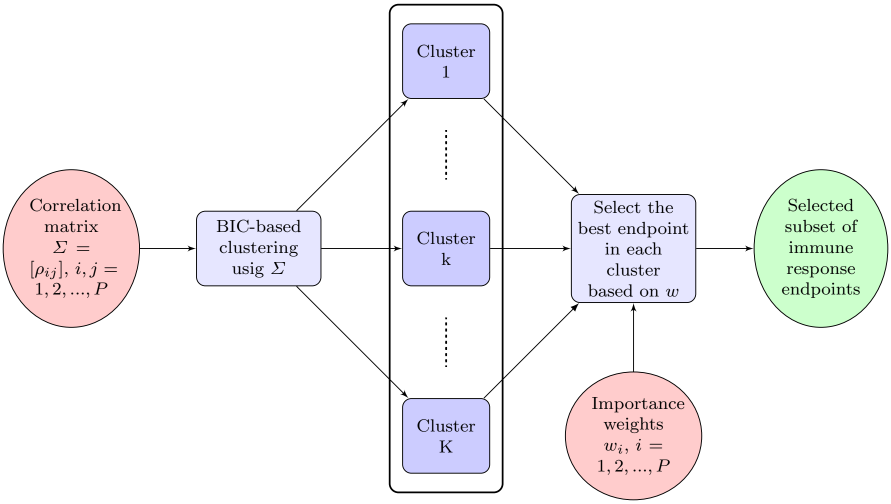
Correlation matrix $\Sigma = [\rho_{ij}], i,j = 1,2,...,P$
BIC-based clustering usig Σ
Cluster 1
Cluster k
Cluster K
Select the best endpoint in each cluster based on w
Importance weights w_i, $i = 1,2,...,P$
Selected subset of immune response endpoints

Fig. 2 System flowchart of the weighted BIC-based feature selection method

status, the potential outcome of an immune response endpoint if receiving a vaccine, and their interaction, the magnitude of this interaction term can serve as a practical weight for the endpoint as it characterizes the capacity of the endpoint for modifying vaccine's protective effect. Next, we show the connection between these interaction terms and the optimal weights for combining multiple endpoints in regimen ranking, using a special risk modeling setting.

We present a special model for disease risk conditional on *potential* multivariate immune responses if receiving a vaccine regimen. Based on this model, one can analytically derive the optimal weights for individual immune response endpoints such that ranking based on the weighted average of immune response endpoints is equivalent to the ranking of the corresponding efficacies of the vaccine regimens. Details can be found in Web Supplementary Appendix B. Modeling disease risk conditional on potential immune responses if receiving a particular vaccine regimen (instead of actual observed immune responses in a trial participant) is appealing for HIV vaccine down-selection for the following reasons: (1) HIV vaccine trials typically enroll healthy individuals who have not been previously exposed to HIV vaccine antigens and thus their HIV-specific immune responses if receiving placebo would be zero; (2) modeling disease risk conditional on potential vaccine-induced immune responses allows prediction and ranking of vaccine efficacy based on immunogenicity induced by candidate vaccine regimens observed in phase I/II studies. In particular, let T be the treatment indicator with $T = 0$ indicating placebo receipt, and $T = j$ for $j = 1, \dots, m$ indicating receiving the jth regimen out of m different vaccine regimens. Further, let Y be the binary indicator for disease (HIV infection). We consider the following models for vaccine-induced immune responses and disease risk.

(I) For $j \in 1, \dots, m$, let $\boldsymbol{X}_j$ be the potential outcome of a set of immune response endpoints of dimension p for a participant if receiving vaccine regimen j, i.e., the immune responses induced by vaccine regimen j. Note that this is just the set of endpoints considered in down-selection, which may not necessarily include all immune response endpoints important to vaccine efficacy. The potential outcome $\boldsymbol{X}_j$ does not depend on the actually assigned T; among an individual receiving placebo, $\boldsymbol{X}_j$ is essentially a counterfactual outcome that is not observed. Suppose $\boldsymbol{X}_j = \{X_{1j}, \dots, X_{pj}\}$ are multivariate normal among vaccine group j with common variance across vaccine regimens. Without loss of generality, we assume that the vaccine-induced immune response endpoints are standardized in a way such that each immune response endpoint is centered at zero with unit variance when $j = 1$. That is, we have

$$\mathbf{X} | T = j \sim \mathcal{N}\left[\begin{bmatrix} \delta_{1j} \\ \delta_{2j} \\ \dots \\ \delta_{pj} \end{bmatrix}, \boldsymbol{\Sigma}\right] \tag{8}$$

where $\delta_{i1} = 0$ by definition. Note that we essentially only assume a constant correlation structure for immune responses induced by different vaccine regimens. That is, if the original immune assay measures have different variability across different vaccines, one can scale the measure by its corresponding standard deviation first to achieve the common variance of X_j across vaccine j.

(II) Suppose the following risk model holds for comparing vaccine regimen j with placebo conditional on $\mathbf{X}_j$ (immune response elicited by jth vaccine):

$$P(Y=1|T,\mathbf{X}_j)=\Phi\Big(\gamma_0+\gamma_1 I(T>0)+\boldsymbol{\gamma}_2^T(\mathbf{X}_j-\boldsymbol{\delta}_j)+\boldsymbol{\gamma}_3^T\mathbf{X}_j I(T>0)\Big) \tag{9}$$

for $T=0$ or j , where $\boldsymbol{\delta}_j=\{\delta_{1j},\dots,\delta_{pj}\}$ and Φ is the cumulative distribution function of a standard normal distribution. Under model (9), we have $P(Y=1|T=0,\mathbf{X}_j)=\Phi\Big(\gamma_0+\boldsymbol{\gamma}_2^T(\mathbf{X}_j-\boldsymbol{\delta}_j)\Big)$. The disease prevalence among the placebo group $P(Y=1|T=0)$ can then be derived by integrating $P(Y=1|T=0,\mathbf{X}_j)$ with respect to distribution of $\mathbf{X}_j$; it is straightforward to see that the location-shift $\boldsymbol{\delta}_j$ in vaccine-induced immune response endpoints $\mathbf{X}_j$ across vaccine regimens as in (8) ensures the derivation of a common prevalence of disease among placebo recipients based on model (9) for any $j\in\{1,\dots,m\}$. As show in Web Supplementary Appendix B1, based on the (8) and (9), the disease prevalence if receiving the jth vaccine regimen is

$$\begin{aligned}P(Y=1|T=j)&=\int P(Y=1|T=j,\mathbf{X}_j)dF_{\mathbf{X}_j}(\mathbf{X}_j)\\&=\Phi\left(\frac{\gamma_0+\gamma_1+\boldsymbol{\gamma}_3^T\boldsymbol{\delta}_j}{\sqrt{1+(\boldsymbol{\gamma}_2+\boldsymbol{\gamma}_3)^T\Sigma(\boldsymbol{\gamma}_2+\boldsymbol{\gamma}_3)}}\right),\end{aligned} \tag{10}$$

which is a monotone increasing function of $\boldsymbol{\gamma}_3^T\boldsymbol{\delta}_j$. In other words, the efficacy of a vaccine in preventing infection is a monotone increasing function of the weighted mean of the immune response endpoint induced by the vaccine with weight equal to $-\boldsymbol{\gamma}_3$. Therefore, the ranking of the efficacy of different vaccine regimens is the same as the ranking of the vaccine regimens' immunogenicity based on this weighted mean, with weights proportional to the interaction between immune response endpoints and treatment status in the multivariate risk model (9).

In practice, however, it is rarely the case that the immune response endpoints measured from a phase I immunogenicity study for down-selection have all been measured together in a previous efficacy trial to allow for the risk modeling in (9), which is based on multivariate immune response endpoints to estimate the optimal weight. It is more likely that prior information is available for modeling disease risk as a function of a univariate immune response endpoint. One practical way to accommodate this situation is to weight each immune response endpoint by the magnitude of its interaction with treatment in the risk model conditional on the specific immune response endpoint. For the ith immune response endpoint, the marginal risk model comparing vaccine regimen j vs placebo conditional on X_{ij} can be derived from (8) and (9) as

$$P(Y=1|T,X_{ij})=\Phi(\beta_{0i}+\beta_{1i}I(T>0)+\beta_{2i}(X_{ij}-\delta_{ij})+\beta_{3i}X_{ij}I(T>0)) \tag{11}$$

for $T = 0, j$. If one has prior information about model (11), then one can use $-\beta_{3i}$ as the weight for immune response endpoint i. That is, one gives weight to each immune response endpoint based on its interaction with the treatment in a disease risk model based on potential outcome of univariate immune response endpoint. As mentioned in Sect. 2.1, here we assume that for each immune response endpoint, a larger immune response is associated with higher vaccine efficacy so $-\beta_3 > 0$. For immune responses that are associated with decreased vaccine efficacy, one can take the negative value of the assay measurement as the endpoint so this condition is satisfied. More details about the analytical form of β_{3i} under models (8) and (9) can be found in Web Supplementary Appendix B2. In general, even without assuming models (8) and (9), the interaction between individual immune response endpoint and vaccination under a generalized linear risk model like in (11) is still a meaningful measure of effect modification and can thus be used to derive the importance weight of the endpoint. Various approaches have been developed in the literature for estimating the risk model (11) conditional on the vaccine-induced immune response [7, 12, 14, 15]. In scenarios where a partial set of multivariate immune responses is available from existing efficacy trial data, we can also estimate importance weights for each endpoint in the partial set by fitting a risk model conditional on the partial set of immune responses together [11].

In practice, any immune response measured in the laboratory is accompanied by a certain degree of inherent measurement error. That is, if X_{ij} described above is the *ith* true underlying immune response endpoint induced by the *jth* vaccine regimen, in real life we might observe K_i responses $X_{ijk} = X_{ij} + e_{ijk}$ that characterize X_{ij} but with additional measurement error $e_{ijk} \sim N(0, \sigma^2_{ijk})$. For example, to characterize an underlying binding antibody response, several immune response endpoints might be obtained using different antigens. Importance weights can be similarly assigned to these observed immune response endpoints based on their interaction with treatment status in the disease risk model conditional on the vaccine-induced immune response measured with error. Detailed derivations of the corresponding risk model under models (8) and (9) are provided in Web Supplementary Appendix B3.

3 Simulation Studies

In this section, we conduct numerical studies to assess the performance of our proposed methods for feature selection in down-selection of vaccine regimens. We consider several settings where $p = 30$ observed immune response endpoints subject to measurement error from $n = 50$ individuals within each of 9 regimen arms are simulated from a multivariate normal distribution with variance 1 and regimen-specific mean, and varying correlation structure between these immune response endpoints. Here we assume the correlation structure among immune response endpoints is available from a pilot study and that this information can be used to guide the decision about which immune response endpoints to collect when designing a down-selection study, although the correlation structure could alternatively be estimated at the time of data analysis.

We apply the RFS algorithm developed in Huang et al. [14] for regimen down-selection, using the weighted average as the criterion for ranking regimens based on pre-specified weights of immune response endpoints. The weights adopted are proportional to the interaction between univariate endpoint and treatment status in the model of disease risk conditional on corresponding endpoint (further details can be found in Web Supplementary Appendix C). We investigate the performance of our feature selection algorithms (penalization-based & BIC-based) combined with the RFS, and compare their performance with that of the naive RFS using the whole sets of immune response endpoints. We define regimens that should be selected as those that satisfy both the superiority and non-redundancy criteria, as well as those that should be ranked among the top Q; the remaining regimens should not be selected. Without loss of generality, here we evaluate settings where there are no other regimens in the whole set that have identical immune response profiles as the top Q regimens that should be selected. With a maximum allowable $Q = 3$, we compare various approaches with respect to the following performance criteria: (i) the average total number of regimens selected, (ii) the percentage of regimens that should be selected among the selected set, namely the ‘positive predictive value’ (PPV), (iii) the probability that a regimen that should be selected is actually selected, i.e. the ’true positive rate’ (TPR), and (iv) the probability that a regimen that should not be selected is mistakenly selected, i.e. the ’false positive rate’ (FPR). Performance results presented are averaged over 1,000 Monte-Carlo simulations.

For the penalization-proposed feature selection algorithm, we adopt $b = 1$ in our simulations. We find that the results are fairly robust to an interval around $b = 1$ (an example is given in Web Supplementary Appendix C, Setting IV). For the BIC-based feature selection algorithm, we use the NbClust R package for clustering the observed immune response endpoints.

In simulation setting I, we assume that the $p = 30$ observed immune response endpoints can be grouped into two clusters of 15 endpoints each. This clustering can be attributed to two underlying true immune response endpoints that are associated with vaccine efficacy. The two underlying immune response endpoints have correlation ρ. Within each cluster, the 15 measured immune response endpoints are equal to the underlying true immune response endpoints plus some measurement error, such that they have the same mean as the true underlying endpoints. In particular, the immune response endpoints within the first cluster of regimen 1 are empirically measured versions (i.e., with inherent measurement error and different noise levels) of the first true underlying immune response endpoint with mean δ and variance 1, while the immune response endpoints within the second cluster of the regimen 1 are empirically measured versions of the second underlying endpoint with zero-mean and unit-variance. We have an opposite data generation mechanism for regimen 2. For regimen 3, we assume both true points have mean $\delta/2$ and variance 1 and all immune response endpoints measured are empirically measured versions of the underlying true endpoints. Based on the means of the 30 immune response endpoints measured with error, regimens 1, 2, and 3 are non-redundant to each other and superior to the other regimens, and are thus the desired regimens to be selected. The first three immune response endpoints in each cluster have smaller measurement error (with measurement error standard deviation = 0.1) than the rest of the

12 immune response endpoints (with measurement error standard deviation = 1). Therefore, the first three endpoints in each cluster have higher importance weights and thus better chances of being chosen during feature selection. The mean structures of the individual endpoints for each regimen and the weights of the individual endpoints for varying ρ's in simulation setting I are presented in Web Supplementary Table 1.

The performance with respect to various operational criteria as a function of δ (effect size, see Appendix C in Web Supplementary materials) for setting I is presented in Fig. 3 for $\rho = 0.3$ and in Supplementary Figs. 2 and 3 for $\rho = 0.1$ and

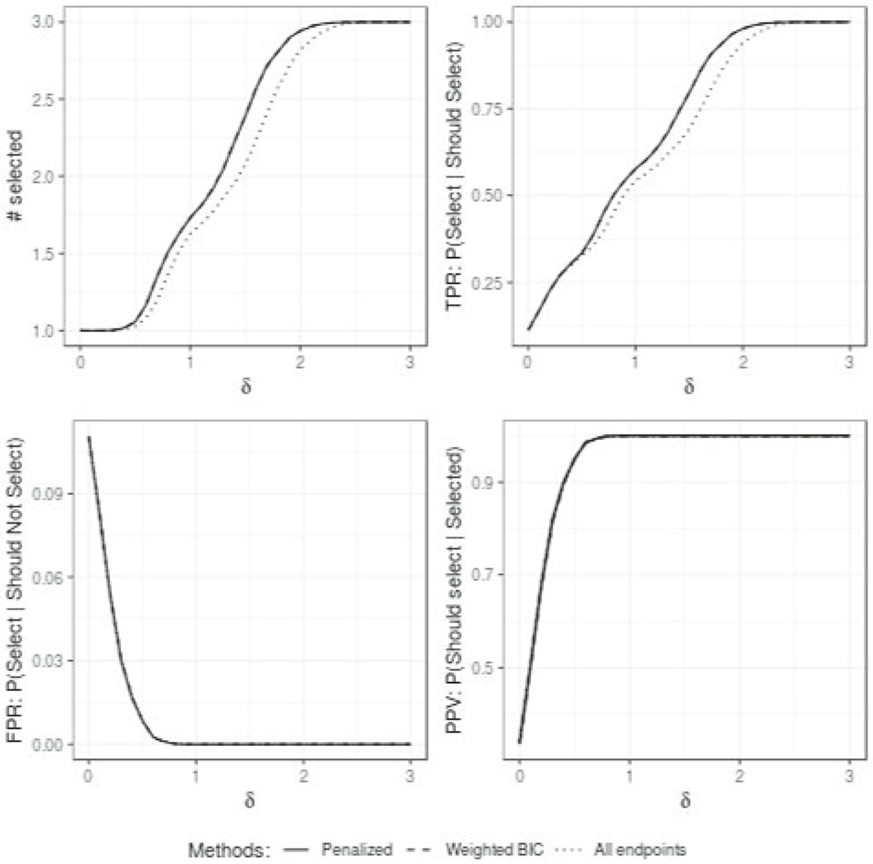

Fig. 3 Comparison of the proposed penalization-based ($b = 1$), BIC-based, and naive selection methods with respect to the average number of selected regimens, TPR, FPR, and PPV over 1000 Monte Carlo simulations. The naive method includes all immune response endpoints in down-selection. We consider two clusters, each containing 15 immune response endpoint measurements. These endpoint measurements are empirically measured versions of the underlying true endpoints in each cluster; some of the measurements are noisier than others. The correlation between the two true underlying endpoints is $\rho = 0.3$ (medium-correlation scenario). For more details, see setting I in Appendix C of Supplementary materials

0.5, respectively. When effect size δ increases, all approaches tend to select the correct number of regimens; their PPV and TPR approach 1 with small FPRs. Both the proposed penalization-based method and the BIC-based method select around 6.7% (one endpoint from each cluster) of the total immune response endpoints to enter regimen down-selection. This results in a lower computational burden in the RFS stage and also better performance compared to the naive methods, which enter all measured immune response endpoints into the RFS. By selecting immune response endpoints with smaller measurement error and thus larger importance weights, our proposed feature-selection methods outperform the naive method , with an apparent increase in TPR for medium effect size (δ from 1.0 to 2.0).

Simulation setting II is a modified version of setting I: within each cluster, only the first five immune response endpoints have non-zero means and are empirically measured versions of the true underlying immune response endpoint. The rest of the 10 immune response endpoints measured have mean zero and thus are not useful surrogate endpoints, since they are not associated with differential vaccine efficacy across different vaccine regimens. For simulation setting II, the mean structure of the individual immune response endpoints for each regimen and the weights of individual endpoints for varying ρ's are presented in Web Supplementary Table 2.

The performance with respect to various operational criteria as a function of δ for setting II is presented in Fig. 4 for $\rho = 0.3$ and in Web Supplementary Figs. 4 and 5 for $\rho = 0.1$ and 0.5, respectively. Like in setting I, both the penalization-based and the BIC-based selection methods select around 6.7% (one endpoint from each cluster) of the total immune response endpoints to enter the RFS algorithm due to the same correlation structure between the two settings. Substantial improvements in performance using the feature-selection methods are observed in setting II compared to the naive method without any feature selection, with increased TPR for various effect sizes, and decreased FPR and increased PPV for small effect size ($\delta < 1.0$). Feature selection is particularly important in this setting relative to setting I, since the redundant "pure noise" endpoints utilized by the naive method will only increase the multi-test adjustment burden without contributing to the differentiation between regimens. Meanwhile, inclusion of "pure noise" variables tends to select nonoptimal regimens and make FPR and PPV much worse for the naive method without feature selection compared to methods with feature selection. This is in contrast to setting I, where each immune response endpoint is useful to some extent for differentiating regimens. As a result, when more endpoints are included in setting I, the test for differentiating regimens is in general more conservative, but the impact on FPR and PPV is minimal.

In simulation setting III, we consider a modified version of setting II such that only the first 10 immune response endpoints in each cluster have non-zero means and are empirically measured versions of the true underlying immune response endpoint in each cluster. The rest of the five immune response endpoints measured in each cluster have mean zero and thus are not useful surrogate endpoints. Within each cluster, the first five immune response endpoints have measurement error standard deviation 0.1, the second five endpoints have measurement error standard deviation 0.2, and the last five endpoints have measurement error standard deviation 1. As a result, the first five endpoints in each cluster have higher importance weights and

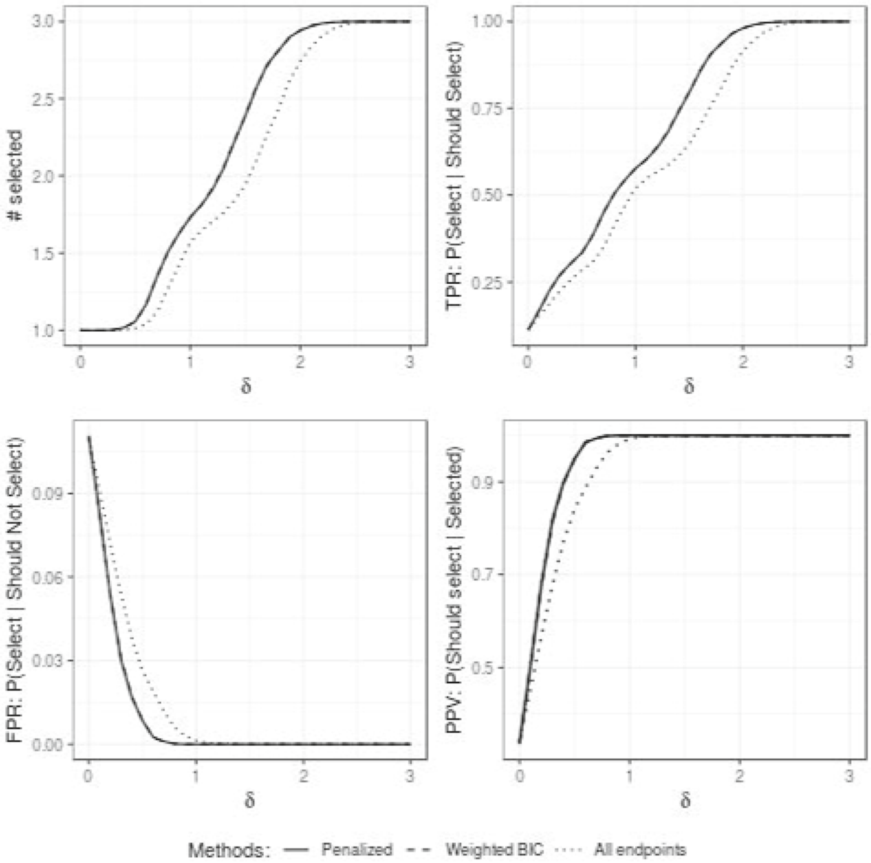

Fig. 4 Comparison of the proposed penalization-based ($b = 1$), BIC-based, and naive selection methods with respect to the average number of selected regimens, TPR, FPR, and PPV over 1000 Monte Carlo simulations. The naive method includes all immune response endpoints in down-selection. We consider two clusters, each containing 15 immune response endpoint measurements: 5 are non-zero mean noisy measurements of the underlying true endpoint in each cluster and the remaining 10 are “pure noise” variables with zero mean that are not associated with the differential vaccine efficacy across regimens. The correlation between the two true underlying endpoints is $\rho = 0.3$ (medium-correlation scenario). For more details, see setting II in Appendix C of Supplementary materials

thus better chances of being chosen during feature selection. For simulation setting III, the mean structure of the individual immune response endpoints for each regimen and the weights for individual endpoints for varying ρ's are presented in Web Supplementary Table 3.

The performance with respect to various operational criteria as a function of δ for setting III is presented in Fig. 5 for $\rho = 0.3$ and in Web Supplementary Figs. 6 and 7 for $\rho = 0.1$ and 0.5, respectively. In this setting, the penalization-based and BIC-based selection methods select around 6.7% (one endpoint from each cluster) and 13.3% (two endpoints from each cluster) of the total immune response

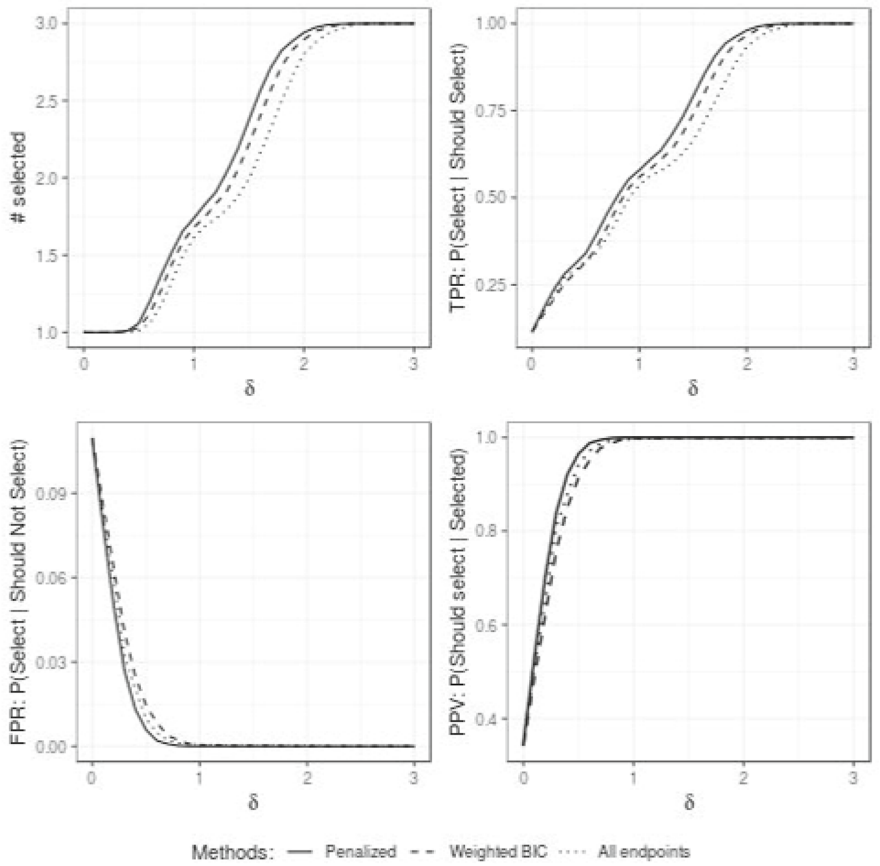

Fig. 5 Comparison of the proposed penalization-based ($b = 1$), BIC-based, and naive selection methods with respect to the average number of selected regimens, TPR, FPR, and PPV over 1000 Monte Carlo simulations. The naive method includes all immune response endpoints in down-selection. We consider two clusters, each containing 15 immune response endpoint measurements: 10 are non-zero mean noisy measurements of the underlying true endpoint in each cluster and the remaining 5 are "pure noise" variables with zero mean that are not associated with the differential vaccine efficacy across regimens. The correlation between the two true underlying endpoints is $\rho = 0.3$ (medium-correlation scenario). For more details, see setting III in Appendix C of Supplementary materials

endpoints to enter the RFS algorithm, respectively. The BIC-based method tends to estimate more clusters than needed due to the increased heterogeneity in measurement error, leading to a larger number of immune response endpoints selected. Substantial improvements in performance using penalization-based feature-selection methods are observed compared to the naive method without any feature selection, with increased TPR for various effect sizes, and decreased FPR and increased PPV for small effect size ($\delta < 1.0$). The penalization-based method also outperforms the BIC-based method in this setting. Compared to the naive method without any feature selection, the BIC-based feature selection method improves TPR but also has larger FPR and smaller PPV for small effect size.

We also studied two other settings (IV and V) with three clusters (dictated by three underlying true immune response endpoints) and 10 immune response endpoints measured within each cluster. Details of these settings are presented in Web Supplementary Appendix C and Supplementary Tables 4 and 5. The corresponding simulation results are presented in Web Supplementary Figs. 8–13. Moreover, we investigated performance based on mixed binary and continuous endpoints in another setting VI, which is derived from setting II by discretizing 40% of the continuous endpoints into binary endpoints based on comparison of each value with the corresponding mean value (See Web Supplementary Appendix C). The corresponding simulation results are presented in Web Supplementary Figs. 14–16. A similar pattern comparing the proposed feature selection methods with the naive method without feature selection is observed in these settings.

As we discussed in Sect. 2, $b = 1$ is a legitimate choice after reformulating our problem based on normalized $\boldsymbol{w}$ (i.e., $\boldsymbol{w}^*$), based on various exploratory studies. An example is shown in Web Supplementary Figs. 17–18, where we present results for simulation setting IV with different choices of b, demonstrating robust performance achieved with $b = 1$.

4 Data Example

In this section, we illustrate the application of our proposed methods using a real data example that was used in Huang et al. [13] for down-selection of vaccine regimens in HIV vaccine trials. This example includes immune response data from five different vaccine regimens. The first regimen is the partially efficacious vaccine regimen used in the RV144 Thai trial [17], which tested an ALVAC-HIV prime with a gp120 AIDSVAX B/E boost. In the RV144 trial, 16,395 participants were recruited and randomized (1:1) to vaccine or placebo; of these, 8,197 participants were assigned to receive vaccine injections at weeks 0, 4, 12, and 24. Among the vaccine recipients who were uninfected at week 26, immune responses were measured from 41 cases (i.e. infected before month 42) and 205 controls (i.e. free of infection over 42 months) selected in a 5:1 ratio to cases among strata defined by gender, number of vaccine injections received, and per-protocol status [9]. Immune response endpoints measured at week 26 for the 205 uninfected vacinees were included in this data example (RV144T). Also included are data from four other regimens in a recently completed HVTN phase I trial (HVTN 096): NYVAC prime plus NYVAC + AIDSVAX B/E boosts (T1), NYVAC + AIDSVAX B/E prime plus NYVAC + AIDSVAX B/E boosts (T2), DNA prime plus NYVAC + AIDSVAX B/E boosts (T3), and DNA + AIDSVAX B/E prime plus NYVAC + AIDSVAX B/E boosts (T4). Participants in each of the four vaccine arms of HVTN 096 received vaccine injections at weeks 0, 4, 12, and 24; immune response data measured from 19, 18, 17, and 19 vaccine recipients in T1, T2, T3, and T4, respectively, are included in this data example. Eight immune response endpoints from each regimen are included in the analysis dataset: one neutralizing antibody (NAb) response endpoint measured using the TZM-bl assay, one CD4+ T-cell response endpoint measured using intracellular cytokine staining (ICS), and six IgG binding antibody response endpoints measured using the binding antibody multiplex assay (BAMA) with different HIV antigens, including AE.A244 V1V2 Tags/293F (B1), gp70_B.CaseA2 V1V2/169K (B2), gp70_B.CaseA_V1_V2(B3), A244 gp120 gDneg/293F/mon (B4), vaccine insert (B5) and Con S gp140 CFI(B6).

Figure 6 shows a heatmap of the correlations between the eight immune response endpoints along with their hierarchical clustering based on the correlation matrix. The ICS readouts have the lowest correlations with all other immune response endpoints. Correlations between NAb and other immune response endpoints are also low. In contrast, strong correlations are observed among BAMA readouts B1, B4, B5 and B6 and between B2, and B3.

We first entered all eight immune response endpoints into the RFS down-selection process as in Huang et al. [13]. The down-selection process ranked the vaccine regimens in the following order: 096T3, 096T1, 096T4, 096T2, and RV144T; 096T3 was selected as the only superior vaccine regimen. We then applied the proposed feature selection algorithms to the candidate immune response endpoints before entering the regimen down-selection. Two different sets of weights for individual immune response endpoints were considered. In the first set, equal weights were assigned to each endpoint. Based on this weighting strategy, the penalization-based algorithm chose three endpoints: ICS, B1, and NAb. Since all endpoints have the same importance wights, we chose ICS first because it has the lowest correlation with the other endpoints. We next chose B1 since it has the lowest correlation with

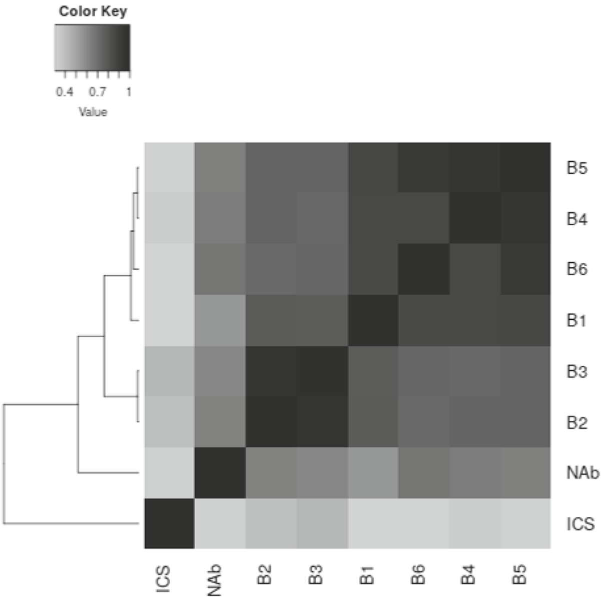

Fig. 6 Heatmap of spearman correlations between different immune response endpoints in the data example

ICS. NAb was the third endpoint chosen because it has low correlations with ICS and B1. The other five endpoints B2 - B6 were discarded because of their relatively high correlations with B1. Based on the three selected endpoints, the RFS algorithm ranked the vaccine regimens in the order of 096T3, 096T4, 096T1, RV144T, and 096T2; 096T3 was the only vaccine regimen selected. Therefore, with only 37.5% of immune response endpoints, the vaccine regimen selected is the same as when we used all immune response endpoints for down-selection. The BIC-based feature selection algorithm assigned B2 and B3 in the same clusters and assigned the remaining six endpoints into six individual clusters. It selected seven (B1, B2, B4, B5, B6, ICS, and NAb) out of eight endpoints. Based on these selected endpoints, the RFS algorithm ranked the vaccine regimens in the order of 096T3, 096T1, 096T4, 096T2, and RV144T. Again, 096T3 was the only regimen selected, this time with 87.5% of the immune response endpoints.

We also investigated our feature selection algorithms using a second set of weights that correspond to the magnitude of the log-transformed univariate odds ratio (OR) (0.7 for B1–B6 and 1.08 for ICS and NAb) for the association of each individual immune response endpoint with infection risk, as reported in Haynes et al. [9]. Note that if one assumes the infection risk among placebo receipts does not depend on the vaccine-induced immune response, then this log(OR) among vaccine recipients is the same as the interaction coefficient β_3 presented in (11). Based on this weighting strategy, the penalization-based algorithm chose two immune response endpoints, B2 and B6. In particular, among endpoints B1-B6 that have high importance weights, B2 has the lowest correlation with other endpoints and was chosen first. After that, B6 was chosen because of its low correlation with B2. Other endpoints were discarded by the algorithm due to either high correlations with B2 and B6 or low importance weights. Based on the two selected endpoints (25% out of eight), the RFS algorithm ranked the vaccine regimens in the order of 096T3, 096T1, 096T4, 096T2 and RV144T; 096T3 was the only vaccine regimen selected. The BIC-based feature selection algorithm under the second set of weights selected the same seven endpoints as the equal weights setting; again, 096T3 was the only regimen selected.

5 Discussion

In HIV vaccine development, effectively screening and appropriately down-selecting candidate regimens based on their immunogenicity before advancing to future efficacy trials is important for saving both time and financial resources. Identifying a parsimonious set of immune response endpoints most relevant for a vaccine's protective effect is essential in this down-selection practice, because the identification of such a set can lead to better down-selection performance and also significant resource savings in terms of lab assay measurements. Moreover, the resulting reduction in the number of immune response endpoints entering regimen down-selection can reduce the computational complexity in down-selection, which increases exponentially as the number of endpoints increases.

Motivated by the application of down-selecting HIV vaccine regimens based on their immunogenicity, here we developed new algorithms for dimension reduction when multiple candidate endpoints are available to rank and select regimens. The proposed algorithms combine two pieces of information: (1) a weight reflecting an individual immune response endpoint's relevance for predicting a vaccine's protective effect, and (2) the correlation structure between immune response endpoints. We also demonstrated through extensive numerical studies that prior feature selection combined with a subsequent down-selection algorithm can achieve better performance with respect to selection of the desired regimens.

Immune response endpoints from HIV phase-I immunogenicity studies can involve multiple immune classes. When constructing the list of candidate immune response endpoints for regimen down-selection, we typically start with endpoints that span various different immune classes. The penalization-based or clustering-based approach for feature selection can then be applied to the list of candidates, ignoring the immune class information. As endpoints within the same class tend to be more correlated with each other, redundant endpoints within a class are likely to be excluded during feature selection.

A caveat of the proposed feature selection algorithms is the need to input a relative importance weight of an individual immune response endpoint with respect to its relevance to the vaccine's protective effect. This weight is important for selecting desired endpoints. Consider a situation in which we have two assays, both measuring the same underlying true response, but with different levels of noise. The endpoint with smaller measurement error would have larger importance weight (or larger interaction with treatment estimated from pilot data). The endpoint with larger measurement error would have smaller importance weight. As a result, immune endpoints with smaller signal-noise ratios are more likely to be excluded in the penalized-based feature selection process, due to their smaller importance weights. In practice, one can also consider first screening out assays with low signal-noise ratios. Usually the relative importance weight can be estimated by eliciting expert input (e.g. regarding whether an assay is more tied to the mechanism of protection of a given vaccine) or can be substituted with the interaction between a vaccine-induced immune response endpoint and vaccine status in infection risk, estimated based on existing data. When weights are estimated with pilot data, the uncertainty associated with the estimated weights would affect the performance of feature selection. Therefore, during the process of HIV vaccine development, it is important for researchers to continue updating the importance weights as new data become available, in order to reduce the uncertainty in weight estimation. Meanwhile, since the choice of importance weights depends on untestable assumptions about the risk prediction model conditional on new vaccine regimens and vaccine-induced immune responses, it is important to perform sensitivity analysis in practice to evaluate the robustness of feature selection and regimen down-selection under a range of plausible importance weights.

Acknowledgements The authors thank the HVTN lab program for generating the ICS data (Dr. Julie McElrath's lab), BAMA data (Dr. Georgia Tomaras's lab), and neutralization data (Dr. David Montefiori's lab) for the HVTN 096 trial and Dr. Gepi Pantaleo for chairing the HVTN 096 trial. The authors thank the participants, investigators, and sponsors of the HVTN 096 trial and the RV144 trial. Research reported in this publication was supported by the National Institutes of Health (Grants 2R37AI05465-10, UM1AI068635, and R01 GM106177-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Bloch DA, Lai TL, Tubert-Bitter P (2001) One-sided tests in clinical trials with multiple endpoints. Biometrics 57(4):1039–1047
2. Bloch DA, Lai TL, Su Z, Tubert-Bitter P (2007) A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. Stat Med 26(6):1193–1207
3. Charrad M, Ghazzali N, Boiteau V, Niknafs A (2015) NbClust: determining the best number of clusters in a data Se. CRAN
4. Follmann D (1996) A simple multivariate test for one-sided alternatives. J Am Stat Assoc 91(434):854–861
5. Fraley C, Raftery DE (1998) How many clusters? which clustering method? Answers via model-based cluster analysis. Comput J 41:578–588
6. Gilbert PB, Huang Y (2016) Predicting overall vaccine efficacy in a new setting by re-calibrating baseline covariate and intermediate response endpoint effect modifiers of type-specific vaccine efficacy. Epidemiol Methods 5(1):93–112
7. Gilbert PB, Hudgens MG (2008) Evaluating candidate principal surrogate endpoints. Biometrics 64(4):1146–1154
8. Gilbert PB, Grove D, Gabriel E, Huang Y, Gray G, Hammer SM, Buchbinder SP, Kublin J, Corey L, Self SG (2011) A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple hiv vaccine regimens. Stat Commun Infect Dis 3(1):4
9. Haynes Barton F, Gilbert Peter B, Juliana MM, Zolla-Pazner S et al (2012) Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N Engl J Med 366(14):1275–1286 PMCID:PMC3371689
10. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
11. Huang Y, Gilbert PB (2011) Comparing biomarkers as principal surrogate endpoints. Biometrics 67(4):1442–1451
12. Huang Y, Gilbert PB, Wolfson J (2013) Design and estimation for evaluating principal surrogate markers in vaccine trials. Biometrics 69(2):301–309
13. Huang Y, DiazGranados C, Janes H, Huang Y, Metch B, Grant S, Sanchez B, Phogat S, Koutsoukos M, Kanesa-Thasan N (2016) Selection of hiv vaccine candidates for concurrent testing in an efficacy trial. Curr Opin Virol 17:57–65
14. Huang Y, Gilbert PB, Fu R, Janes H (2017) Statistical methods for down-selection of treatment regimens based on multiple endpoints, with application to hiv vaccine trials. Biostatistics 18(2):230–243
15. Juraska M, Huang Y, Gilbert PB (2018) Inference on treatment effect modification by biomarker response in a three-phase sampling design. Biostatistics
16. Perlman MD, Wu L (2004) A note on one-sided tests with multiple endpoints. Biometrics 60(1):276–280
17. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S et al (2009) Vaccination with ALVAC and AIDS-VAX to prevent HIV-1 infection in Thailand. N Engl J Med 361(23):2209–2220
18. Röhmel J, Gerlinger C, Benda N, Läuter J (2006) On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. Biometrical J 48(6):916–933
19. Sargent DJ, Goldberg RM (2001) A flexible design for multiple armed screening trials. Stat Med 20(7):1051–1060
20. Simon R, Wittes RE, Ellenberg SS (1985) Randomized phase II clinical trials. Cancer Treat Rep 69(12):1375–1381
21. Steinberg SM, Venzon DJ (2002) Early selection in a randomized phase II clinical trial. Stat Med 21(12):1711–1726
22. Tamhane AC, Logan BR (2004) A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. Biometrika 91(3):715–727
23. Tang D-I (1994) Uniformly more powerful tests in a one-sided multivariate problem. J Am Stat Assoc 89(427):1006–1011

Affiliations

Ying Huang[1] · **Aliasghar Tarkhan**[2]

[1] Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[2] Department of Biostatistics, University of Washington, Seattle, WA 98109, USA