HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2017 March 15.

Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1017–1025. doi:10.1111/biom.12511.

Identifying Predictive Markers for Personalized Treatment Selection

Yuanyuan Shen[*] and **Tianxi Cai**[*]

[*]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Summary

It is now well recognized that the effectiveness and potential risk of a treatment often vary by patient subgroups. Although trial-and-error and one-size-fits-all approaches to treatment selection remains a common practice, much recent focus has been placed on individualized treatment selection based on patient information (La Thangue and Kerr, 2011; Ong et al., 2012). Genetic and molecular markers are becoming increasingly available to guide treatment selection for various diseases including HIV and breast cancer (Mallal et al., 2008; Zujewski and Kamin, 2008). In recent years, many statistical procedures for developing individualized treatment rules (ITRs) have been proposed. However, less focus has been given to efficient selection of predictive biomarkers for treatment selection. The standard Wald test for interactions between treatment and the set of markers of interest may not work well when the marker effects are non-linear. Furthermore, interaction based test is scale dependent and may fail to capture markers useful for predicting individualized treatment differences. In this paper, we propose to overcome these difficulties by developing a kernel machine (KM) score test that can efficiently identify markers predictive of treatment difference. Simulation studies show that our proposed KM based score test is more powerful than the Wald test when there is non-linear effect among the predictors and when the outcome is binary with non-linear link functions. Furthermore, when there is high-correlation among predictors and when the number of predictors is not small, our method also over-performs Wald test. The proposed method is illustrated with two randomized clinical trials.

Keywords

Treatment selection; Score test; Kernel machine; Kernel PCA; Perturbation

1. Introduction

The effectiveness and potential risk of a treatment often vary by patient subgroups (Duffy and Crown, 2008; La Thangue and Kerr, 2011). For instance, ER negative breast cancer patients benefit substantially from chemotherapy while ER positive patients do not benefit as compared to receiving tamoxifen alone (IBCSG, 2002). A gene-expression profile appears to

Correspondence to: Tianxi Cai.

Supplementary Materials

Web Appendix referenced in Section 2.2 and the source code for computation are available with this paper at the Biometrics website on Wiley Online Library.

be highly predictive of whether chemotherapy is beneficial for treating breast cancer patients and is now being further investigated by the TAILORx study (Zujewski and Kamin, 2008). The adverse risk of Abacavir for treating HIV infected patients is strongly associated with the presence of the HLA-B*5701 allele and thus Abacavir was recommended only for patients not carrying this allele (Mallal et al., 2008). Recently, the US Preventive Services Task Force issued new guidelines recommending against routine mammography screening for women under 50 (Nelson et al., 2009). On the other hand, such guidelines may not be appropriate for populations at increased risk and refinement of such recommendations warrants further research.

Many factors including genetics predisposition and environmental influences may play a role in a patient's treatment response. Incorporating information on clinical, biological and genomic markers into personalized prediction of treatment response holds great potential for identifying subgroups of patients who are most likely to benefit or are at high risk for toxicity from a particular therapy. Interventions can then be targeted to well-defined groups that are likely to benefit and at low risk of adverse event. In recent years, a wide range of statistical methods have been proposed for developing individualized treatment rules (ITRs) based on a set of baseline predictors (Qian and Murphy, 2011; Cai et al., 2011; Foster et al., 2011; Zhao et al., 2012; Zhang et al., 2012; Zhao et al., 2013). When the number of predictors for deriving ITRs is not small, it is important to only include informative markers since including a large number of unrelated markers may tamper the accuracy of the resulting ITR and lead to unnecessary cost associated with measuring the markers. Variable selection procedures have also been developed for both prediction and decision making (Gunter et al., 2011; Lu et al., 2013; Imai et al., 2013). However, in the high dimensional setting, variable selection procedures may not work well in identifying informative markers since many of such procedures are not consistent in variable selection and it is generally difficult to identify an appropriate tuning parameter to ensure selection consistency. For such settings, it would be desirable to perform testing on candidate markers and only develop ITRs using markers that are deemed predictive of treatment response.

Standard testing procedures for ITRs consider models that include interactions between the treatment group and the variables of interest and perform a Wald-type test on the interaction term. Rosenblum and van der Laan (2009) showed that even when the model is misspecified, the Wald test still obtains the correct size, if sandwich variance estimators are used. Despite the robustness property, such an approach suffers from two major limitations. First, the interaction term may not entirely capture markers' ability in predicting subject specific treatment effect (TE). When TE of interest is the treatment difference and the outcome Y is binary, the conditional TE given baseline predictor $\mathbf{X}$, $P(Y = 1 \mid T = 1, \mathbf{X}) - P(Y = 1 \mid T = 0, \mathbf{X})$, may depend on both the main effect and the interaction. For example, when $P(Y=1|T,\mathbf{X})=g(\alpha+\beta T+\boldsymbol{\gamma}_0^\top \mathbf{X}+T\boldsymbol{\gamma}_1^\top \mathbf{X})$ and $g(\cdot)$ is a distribution function, the conditional TE $g\{\alpha+\beta+(\boldsymbol{\gamma}_1+\boldsymbol{\gamma}_0)^\top \mathbf{X}\} - g(\alpha+\boldsymbol{\gamma}_0^\top \mathbf{X})$ is a function of both the main effect $\boldsymbol{\gamma}_0$ and the interaction effect $\boldsymbol{\gamma}_1$. Second, the standard Wald test restricts attention to linear marker effects. When the markers affect the outcome non-linearly or interactively, the Wald test may have little power in detecting the signal. In this paper, we propose a kernel machine (KM) based score test for identifying markers predictive of TE. The proposed KM testing

procedure can effectively incorporate non-linear effects and capture predictors that are predictive of treatment difference. We focus on the treatment difference scale because the value function of an ITR, $\mathscr{I}_{\mathbf{X}}:\mathbf{X} \to \{0,1\}$, in improving expected population outcome is directly captured by the treatment difference: $E\{\mathscr{I}_{\mathbf{X}}Y^{(1)}+(1-\mathscr{I}_{\mathbf{X}})Y^{(0)}\}=E\{\mathscr{I}_{\mathbf{X}}(Y^{(1)}-Y^{(0)})\}+E(Y^{(0)})$. The proposed testing procedures can be used to select important groups of baseline predictors that are predictive of treatment response. When a large number of potential predictors are available, biological or clinical knowledge can be used to group these predictors into meaningful subsets and the proposed testing procedures can be used to identify informative subsets. These subsets can then be used to form ITRs using existing methods such as those proposed in Zhao et al. (2012) and Zhang et al. (2012).

The rest of the paper is organized as follows. We introduce the KM test for ITR in section 2.1 and describe the resampling procedure for approximating the null distribution in section 2.2. Additional considerations including tuning parameter selection, dimension reduction via kernel principal component analysis (PCA), omnibus test incorporating kernel selection are given in section 2.3. In section 3.1, we present simulation results suggesting that the proposed procedures out-performs the traditional Wald test in various settings. The proposed procedures are applied to two randomized clinical trials in 3.2 and 3.3. We conclude with some remarks in section 4.

2. Treatment Selection Model

2.1 Score Statistic for Identifying Important Baseline Predictors for Treatment Selection

Suppose data for analysis comes from a randomized clinical trial (RCT), and consist of independent and identically distributed random variables $\{(Y_i, T_i, \mathbf{X}_i^\top)^\top, i=1,\ldots,n\}$, where Y is the disease outcome, T is a binary treatment indicator (1 for new treatment and 0 for standard treatment), and $\mathbf{X}$ represents baseline predictors. Let $Y^{(1)}$ and $Y^{(0)}$ be the counterfactual outcomes under the new and standard treatment, respectively.

To determine whether $\mathbf{X}$ is useful for guiding treatment selection, we quantify the TE for subjects with $\mathbf{X}$ based on the conditional treatment difference

$$\Delta(\mathbf{X})=\mu_1(\mathbf{X})-\mu_0(\mathbf{X}),$$

where $\mu_k(\mathbf{X}) = E(Y^{(k)}|\mathbf{X})$. Thus $\mathbf{X}$ is not informative for treatment selection if $\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$ is a constant. Thus, we aim to develop efficient testing procedures for the null hypothesis

$$H_0:\mu_1(\mathbf{X})-\mu_0(\mathbf{X})=\Delta_0, \quad (1)$$

where the constant $\Delta_0 = E\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\} = \mu_1 - \mu_0$ and $\mu_k = E(Y^{(k)})$. Under H_0,

$$\mathbf{R}_\psi=\mathrm{cov}\{Y^{(1)}-Y^{(0)},\psi(\mathbf{X})\}=\mathrm{E}\{(Y^{(1)}-Y^{(0)}-\Delta_0)\psi(\mathbf{X})\}=0, \text{ for any } \psi(\cdot).$$

and thus we propose to test (1) by constructing a test statistic summarizing the overall magnitude of $\mathbf{R}_{\boldsymbol{\psi}}$. To this end, we first obtain an empirical estimate of $\mathbf{R}_{\boldsymbol{\psi}}$ based on the observed RCT data. Specifically, by employing an inverse probability weighting (IPW) (Rotnitzky and Robins, 2005) estimator for the counterfactuals, we estimate $\mathbf{R}_{\boldsymbol{\psi}}$ as

$$\hat{\mathbf{R}}_{\psi}=n^{-1}\sum_{i=1}^{n}\hat{\delta}_i\boldsymbol{\psi}(\mathbf{X}_i)=n^{-1}\hat{\boldsymbol{\Delta}}^{\top}\boldsymbol{\Psi} \quad (2)$$

where $\boldsymbol{\Psi}=[\boldsymbol{\psi}(\mathbf{X}_1),\ldots,\boldsymbol{\psi}(\mathbf{X}_n)]^{\top}$, $\hat{\boldsymbol{\Delta}}=[\hat{\delta}_1,\ldots,\hat{\delta}_n]^{\top}$, $\overline{Y}_k=n_k^{-1}\sum_{\{T_i=k\}}Y_i, n_k=\sum_{i=1}^{n}I(T_i=k)$,

$$\hat{\delta}_i=\frac{(Y_i-\overline{Y}_1)I(T_i=1)}{\hat{\pi}_1}-\frac{(Y_i-\overline{Y}_0)I(T_i=0)}{\hat{\pi}_0},\ \text{and}\ \hat{\pi}_k=\frac{n_k}{n} \quad (3)$$

In order to test whether (2) is close to $\mathbf{0}$, the standard score-type test bstatistic takes the form of $\hat{\mathbf{R}}_{\psi}^{\top}\hat{\sum}_{\hat{\mathbf{R}}_{\psi}}^{-1}\hat{\mathbf{R}}_{\psi}$ and is approximately χ_q^2, where $\hat{\sum}_{\hat{\mathbf{R}}_{\psi}}$ is the covariance matrix estimate of $\hat{\mathbf{R}}_{\psi}$ and q is the dimension of $\boldsymbol{\psi}(\mathbf{X})$. However, such a test may suffer from power loss when $\boldsymbol{\psi}(\mathbf{X})$ are correlated and/or q is not small. In addition, the χ_q^2 distribution may not approximate the null distribution of $\hat{\mathbf{R}}_{\psi}^{\top}\hat{\sum}_{\hat{\mathbf{R}}_{\psi}}^{-1}\hat{\mathbf{R}}_{\psi}$ well especially when the covariance matrix is near singular. We instead summarize the overall effect of $\mathbf{X}$ based on the L_2 norm of (2) and propose the test statistic:

$$\hat{Q}_{\psi}=||n^{\frac{1}{2}}\hat{\mathbf{R}}_{\psi}||^2=n^{-1}\hat{\boldsymbol{\Delta}}^{\top}(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top})\hat{\boldsymbol{\Delta}} \quad (4)$$

This type of score test serves as a powerful alternative to the standard score test and can be viewed as a variance component test under various settings (Wu et al., 2010; Cai et al., 2011).

The choice of the basis functions $\boldsymbol{\psi}(\cdot)$ has a significant impact on the power of the resulting test. If the basis functions efficiently capture the non-linear characteristic of the data, one may achieve great power gain comparing to using the original data. However, in practice, it is often difficult to explicitly specify $\boldsymbol{\psi}(\cdot)$ to optimize power since prior knowledge of the underlying functional form is generally not available. We propose to overcome this difficulty by implicitly specifying the basis functions using the Reproducible Kernel Hilbert Space (RKHS). Let $\mathscr{H}_k$ be a RKHS generated by a given positive definite kernel function $k(\cdot,\cdot;\rho)$, and ρ is some tuning parameter associated with the kernel function (Cristianini and Shawe-Taylor, 2000), where the kernel function $k(\mathbf{x}_1,\mathbf{x}_2;\rho)$ measures the similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ and different choices of k lead to different RKHS. Some of the popular kernel functions

include the gaussian kernel $k(\mathbf{x}_1, \mathbf{x}_2;\rho)=\exp\{-0.5||\mathbf{x}_1 - \mathbf{x}_2||_2^2/\rho\}$ which can capture complex smooth non-linear effects; the linear kernel $k(\mathbf{x}_1, \mathbf{x}_2;\rho)=\rho+\mathbf{x}_1^\top \mathbf{x}_2$ which corresponds to $h(\mathbf{x})$ being linear in $\mathbf{x}$; and the quadratic kernel $k(\mathbf{x}_1, \mathbf{x}_2;\rho)=(\mathbf{x}_1^\top \mathbf{x}_2+\rho)^2$ which allows for 2-way interactive effects. By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), any $h(\mathbf{x}) \in \mathscr{H}_k$ has a *primal representation* with respect to the eigensystem of k. Specifically, under the probability measure of $\mathbf{x}$, k has eigenvalues $\{\lambda_l, l=1, \ldots, \mathscr{J}\}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_{\mathscr{J}}$ and eigenfunctions $\{\phi_l, l=1, \ldots, \mathscr{J}\}$ such that $k(\mathbf{x}_1, \mathbf{x}_2)=\sum_{l=1}^{\mathscr{J}} \lambda_l \phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, where $\mathscr{J}$ may be infinity and $\lambda_l > 0$ for any $l < \infty$. The basis functions, $\{\psi_l(\mathbf{x})= \sqrt{\lambda_l}\phi_l(\mathbf{x}), l=1, \ldots, \mathscr{J}\}$, span the RKHS $\mathscr{H}_k$. These basis functions can potentially be used in (2). The kernel functions may depend on the tuning parameter ρ. For the ease of presentation, we suppress ρ from k in remaining presentations although procedures for incorporating different choices of ρ in testing will be detailed in Section 2.3.

The basis functions $\{\psi_l(\cdot)\}$ inherently depend on the unknown distribution of $\mathbf{x}$, $\mathbb{P}(\mathbf{x}') = P(\mathbf{x} \leq \mathbf{x}')$, since $\phi_l(\mathbf{x}')$ is the solution to the integral equation $\int k(\mathbf{x}^*, \mathbf{x}')\phi_l(\mathbf{x}')\mathbb{P}(d\mathbf{x}') = \lambda_l \phi_l(\mathbf{x}^*)$. Thus, the basis functions are not directly available for inference. To estimate $\{\psi_l(\cdot)\}$, we apply a singular value decomposition to the observed kernel matrix $\mathbb{K}_n=[k(\mathbf{x}_i, \mathbf{x}_j)]_{n\times n}$:

$$\mathbb{K}_n=\hat{\mathbf{\Phi}}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Phi}}^\top=\hat{\mathbf{\Psi}}\hat{\mathbf{\Psi}}^\top, \text{ where } \hat{\mathbf{\Psi}}=\hat{\mathbf{\Phi}}\hat{\mathbf{\Lambda}}^{1/2}, \hat{\mathbf{\Lambda}}=\text{diag}\{a_1, \ldots, a_n\},$$

$a_1 \geqslant \ldots \geqslant a_n \geqslant 0$ are the eigenvalues of $\mathbb{K}_n$ and $\hat{\mathbf{\Phi}}=(\hat{\phi}_1, \ldots, \hat{\phi}_n)$ are the corresponding eigenvectors. It has been shown that $\hat{\mathbf{\Psi}}$ is effectively estimating the basis functions evaluated at the sample points, $\mathbf{\Psi} = [\psi_j(\mathbf{X}_i)]_{n\times n}$ (Koltchinskii and Giné, 2000; Braun et al., 2005). Replacing $\mathbf{\Psi}$ in (4) by $\hat{\mathbf{\Psi}}$, our KM score test statistic for ITR takes the form

$$\hat{Q}_\psi=\frac{1}{n}\hat{\mathbf{\Delta}}^\top\hat{\mathbf{\Psi}}\hat{\mathbf{\Psi}}^\top\hat{\mathbf{\Delta}}=\frac{1}{n}\hat{\mathbf{\Delta}}^\top\mathbb{K}_n\hat{\mathbf{\Delta}}. \quad (5)$$

We next detail procedures for approximating the null distribution of the statistic $\hat{Q}_\psi$.

2.2 Approximating the Null Distribution by Resampling Procedure

To approximate the distribution of (5) under H_0, we show in the Web Appendix that

$$\hat{Q}_\psi=n^{-1}\int\int k(\mathbf{x}, \mathbf{x}')d\hat{\Theta}(\mathbf{x})d\hat{\Theta}(\mathbf{x}') \quad (6)$$

and $\hat{\Theta}(\mathbf{x})=n^{-\frac{1}{2}}\sum_{i=1}^n \theta_i(\mathbf{x})+o_P(1)$, where $\hat{\Theta}(\mathbf{x})$ is defined in (2) in the Web Appendix

$$\theta_i(\mathbf{x})=\left\{\frac{(Y_i-\mu_1)I(T_i=1)}{\pi_1}-\frac{(Y_i-\mu_0)I(T_i=0)}{\pi_0}\right\}\{I(\mathbf{X}_i\le\mathbf{x})-\mathscr{F}(\mathbf{x})\} \quad (7)$$

$\pi_k = P(T = k)$, and $\mathscr{F}(\mathbf{x}) = P(\mathbf{X} \le \mathbf{x})$. We further show in the Web Appendix that $n^{-\frac{1}{2}}\hat{\Theta}(\mathbf{x})$ converges weakly to zero-mean Gaussian process $G(\mathbf{x})$ and hence

$$\hat{Q}_\psi=n^{-1}\int\int k(\mathbf{x},\mathbf{x}')d\hat{\Theta}(\mathbf{x})d\hat{\Theta}(\mathbf{x}')\to\int k(\mathbf{x},\mathbf{x}')dG(\mathbf{x})dG(\mathbf{x}'),\text{ in distribution.}$$

The limiting null distribution of $\hat{Q}_\psi$ takes a complex form, making explicit estimation infeasible. We propose to approximate the null distribution of $\hat{Q}_\psi$ via perturbation resampling, which has been used successfully in the literature to approximate the distribution of a wide range of regular estimators (Cai et al., 2005; Tian et al., 2007). Specifically, for a large number B, we generate independent standard normal random variables, $\{\mathbf{V}^{(b)}=(V_1^{(b)},\ldots,V_n^{(b)}),b=1,\ldots B\}$, independent of the observed data. For $b = 1, \ldots, B$, let the bth perturbed realization of $\hat{\Theta}(\mathbf{x})$ be

$$\hat{\Theta}^{(b)}(\mathbf{x})=n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{\theta}_i(\mathbf{x})V_i^{(b)},\ \text{ where }\hat{\theta}_i(\mathbf{x})=\hat{\delta}_i\{I(\mathbf{X}_i\le\mathbf{x}\}-\hat{\mathscr{F}}(\mathbf{x})\},$$

and $\hat{\mathscr{F}}(\mathbf{x})=n^{-1}\sum_{l=1}^{n}I(\mathbf{X}_l\le\mathbf{x})$. Subsequently, we obtain the perturbed counterpart of $\hat{Q}_\psi$ as

$$\hat{Q}_\psi^{(b)}=\int\int k(\mathbf{x},\mathbf{x}')d\hat{\Theta}^{(b)}(\mathbf{x})d\hat{\Theta}^{(b)}(\mathbf{x}')=(\hat{\mathbf{\Delta}}\odot\mathbf{V}^{(b)})^\top*(\hat{\mathbf{\Delta}}\odot\mathbf{V}^{(b)}) \quad (8)$$

where $*=-\mathbf{e}_n^\top-^\top\mathbf{e}_n+\mathbf{e}_n^\top\mathbf{e}_n,\mathbf{e}_n=n^{-1}\mathbf{1}_{n\times1}$ and $\odot$ denotes element-wise product. The null distribution of $\hat{Q}_\psi$ can be approximated by the empirical distribution of $\{\hat{Q}_\psi^{(b)},b=1,\ldots,B\}$. For an observed $\hat{Q}_\psi$, the p-value can be estimated as $\frac{1}{B}\sum_{b=1}^{B}I(\hat{Q}_\psi^{(b)}>\hat{Q}_\psi)$.

2.3 Additional Consideration: Scale Parameters, Kernel PCA and Kernel Selection

Kernels with Scale Parameters—Some kernels, such as the Gaussian kernel, involves a scale parameter ρ which has a great impact on the complexity of the resulting $\mathscr{H}_k$ and hence the power of the test. Unfortunately, the parameter ρ is not identifiable under H_0. To combine information from multiple choices of ρ, we take a similar approach as in Davies (1977) and consider the minimum p-value as the composite test statistic. Specifically, let $\{\rho_m, m = 1, \ldots, M\}$ be the list of candidate scale parameters. Let $\hat{Q}_{\psi,m}$ and $\hat{\mathbf{Q}}_{\psi,m}=\{\hat{Q}_{\psi,m}^{(1)},\cdots,\hat{Q}_{\psi,m}^{(B)}\}^\top$ denote the observed and perturbed test statistic corresponding to kernel $k(\cdot,\cdot;\rho_m)$, respectively, where the same set of perturbation variables

$\{\mathbf{V}^{(b)}=(V_1^{(b)},\ldots,V_n^{(b)}), b=1,\ldots B\}$ are used across all M scale parameters. Let $\hat{S}_m(\cdot)$ denote the empirical survival distribution of $\{\hat{Q}_{\psi,m}^{(b)}, b=1,\cdots,B\}$. Then we define minimum p-value across testing with M scale parameters as $\hat{p}_{min}=\min\{\hat{p}_m, m=1,\ldots,M\}$, where $\hat{p}_m=\hat{S}_m(\hat{Q}_{\psi,m})$. Although $\hat{p}_m$ is expected to be approximately uniform under H_0, the minimum p-value statistic $\hat{p}_{min}$ is no longer uniformly distributed. Nevertheless, the null distribution of $\hat{p}_{min}$ can be easily approximated using the perturbed realizations $\{\hat{\mathbf{Q}}_{\psi,m}, m=1,\ldots,M\}$. Specifically, the empirical distribution of $\{\hat{p}_{min}^{(b)}, b=1,\ldots,B\}$ can be used to approximate the null distribution of $\hat{p}_{min}$, where $\hat{p}_{min}^{(b)}=\min\{\hat{p}_m^{(b)}, m=1,\ldots,M\}$ and $\hat{p}_m^{(b)}=\hat{S}_m\{\hat{Q}_{\psi,m}^{(b)}\}$.

Kernel PCA—When the kernel space $\mathscr{H}_k$ is high dimensional, testing and estimation procedures based on such a space may not be efficient due to the high degrees of freedom (Braun et al., 2005). In addition, the null distribution of the test statistic tends to be more difficult to approximate in finite sample, leading to slightly inaccurate type I error (Cai et al., 2011). One approach to improving the power and maintaining proper size is to effectively reduce the dimensionality. When the eigenvalues of k decay quickly, $\mathscr{H}_k$ can be well approximated by the RKHS spanned by a truncated kernel $k^{(r_n)}(\mathbf{x}_1,\mathbf{x}_2)=\sum_{l=1}^{r_n}\lambda_l\phi_l(\mathbf{x}_1)\phi_l(\mathbf{x}_2)$, for some r_n such that $\sum_{l=r_n+1}^{\mathscr{J}}\lambda_l=o(\sum_{l=1}^{\mathscr{J}}\lambda_l)$. The error $\mathscr{E}_n=\|\mathbb{K}_n-\overset{(r_n)}{n}\|$ can be bounded by $O\{\lambda_r+\sum_{l=r_n+1}^{\infty}\lambda_l\}$, where $\overset{(r_n)}{n}$ is the kernel matrix constructed from kernel $k^{(r_n)}$ (Braun et al., 2005, Theorem 3.7). In many practical situations with fast decaying eigenvalues for k, r_n is typically fairly small and we can effectively approximate $\mathscr{H}_k$ by a finite dimensional space. Although $\overset{(r_n)}{n}$ is generally not attainable directly in practice, we may use kernel PCA to approximate $\overset{(r_n)}{n}$ as

$$\overset{(r_n)}{n}=[\hat{\boldsymbol{\phi}}_1,\ldots,\hat{\boldsymbol{\phi}}_{r_n}]\mathrm{diag}\{a_1,\ldots,a_{r_n}\}[\hat{\boldsymbol{\phi}}_1,\ldots,\hat{\boldsymbol{\phi}}_{r_n}]^\top=[\hat{\boldsymbol{\psi}}_1,\ldots,\hat{\boldsymbol{\psi}}_{r_n}][\hat{\boldsymbol{\psi}}_1,\ldots,\hat{\boldsymbol{\psi}}_{r_n}]^\top.$$

Replacing $\mathbb{K}_n$ by $\overset{(r_n)}{n}$ in (5), we obtain the kernel PCA approximated test b statistic

$$\hat{Q}_{PCA}=n^{-1}\hat{\boldsymbol{\Delta}}^\top\overset{(r_n)}{n}\hat{\boldsymbol{\Delta}}=\sum_{l=1}^{r_n}\|\hat{\boldsymbol{\Delta}}^\top\widehat{\psi_l}\|_2^2, \quad (9)$$

Obviously, $\hat{Q}_{PCA}$ reduces to $\hat{Q}_\psi$ when $r_n = n$.

Range of ρ—It is also important to choose the appropriate range of $\{\rho_m, m = 1, \ldots, M\}$, since the range will affect the size and power of the procedures. We use a data adaptive approach to select the range by taking into account the eigenvalues decay rate of the kernel for a given ρ, $\alpha(\rho)$, where we assume that $\lambda_j(\rho) = O\{j^{-\alpha(\rho)}\}$. We estimate the decay rate as the slope from fitting a robust linear regression $\log\{a_j(\rho)\} = \alpha \log(j) + \varepsilon$ with $j = 1, \ldots, r_n$. The range of ρ is chosen such that the corresponding estimated $\alpha(\rho)$ is between 1.2 and 2

and the vector $\{\rho_m, m = 1, \ldots, M\}$ is equally spaced on the logarithm scale within this range. When we select how many components to use in the singular value decomposition of $\mathbb{K}_n$, we choose r_n as the smallest r such that the estimated proportion of variation explained by the first r eigenfunctions, defined as $\{\sum_{l=1}^{r_n} a_l\}/\{\sum_{l=1}^{n} a_l\}$, is at least 0.99.

Omnibus Test with Kernel Selection—The choice of kernel k plays a critical role in the testing performance. It is generally difficult to decide a priori the optimal kernel to use for a particular dataset, since the underlying structure of the data is typically unknown. We propose an omnibus test that selects the kernel with the smallest p-value and account for the additional variability due to kernel selection through the perturbation procedure. Specifically, all candidate kernels, for example linear, Gaussian, and quadratic, will be applied to the dataset and the same resampling procedure as stated in section 2.2 will be carried out using the same set of perturbation variables $\{\mathbf{V}^{(b)}=(V_1^{(b)},\ldots,V_n^{(b)}), b=1,\ldots B\}$. Let $\{\hat{\mathscr{P}}_\kappa, \kappa=1,\ldots,\mathscr{K}\}$ denote the p-values from testing with $\mathscr{K}$ candidate kernels. Then we select kernel as the one with the smallest p-value. The final p-value for the omnibus test that accounts for kernel selection is obtained as $\hat{\mathscr{S}}\{\min(\hat{\mathrm{P}}_1,\ldots,\hat{\mathscr{P}}_{\mathscr{K}})\}$, where $\hat{\mathscr{S}}(\cdot)$ is the empirical survival distribution of $\{\min(\hat{\mathscr{P}}_1^{(b)},\ldots,\hat{\mathscr{P}}_{\mathscr{K}}^{(b)}), b=1,\ldots,B\}$ and $\hat{\mathscr{P}}_\kappa^{(b)}$ is the perturbed counterpart of $\hat{\mathscr{P}}_\kappa$ under H_0 constructed similarly as $\hat{p}_{min}^{(b)}$.

3. Numerical Studies

3.1 Simulation Study

We performed extensive simulation studies to compare the performances of our KM procedures to the Wald test with sandwich variance estimator, and the nonparametric test for treatment effect heterogeneity (Crump et al., 2008) (Crump). The Crump test, developed only for continuous Y, takes a similar form as the Wald test but uses a nonparametric power series estimation to estimate regression functions. The number of terms in the series is decided by cross-validation. We carried out our procedure with three kernels (i) linear (k_L), (ii) quadratic (k_Q) and, (iii) gaussian kernel (k_G), as well as the omnibus test that incorporates kernel selection. For conciseness, we only present results from the kernel PCA procedure where we select the first r_n eigenvectors that account for 99% of total variation. We studied both continuous and binary Y. The predictor $\mathbf{X}_{p\times 1}$ was generated from multivariate normal with mean zero, variance 4 and correlation ρ, where we let p = 5 and 20, and ρ = 0.2 and 0.5. We considered a total sample size of n = 500, 1000 and generate treatment indicator T from Bernoulli(0.5). To make fair comparisons, all procedures were applied to the entire vector $\mathbf{X}$. Results are summarized for target type I error of 0.05.

For continuous outcome, we generate Y from $Y = -35 + T + h_0(\mathbf{X}) + h_1(\mathbf{X})T + X_1X_2\varepsilon$, where ε follows a standard normal. Three different settings were considered for the predictor effect functions $h_0(\mathbf{X})$ and $h_1(\mathbf{X})$: (i) Null with $h_0(\mathbf{X}) = h_1(\mathbf{X}) = 0$; (ii) Linear effects with $h_0(\mathbf{X}) = X_3/2$ and $h_1(\mathbf{X}) = (X_1 + X_2 + X_5)/3$; and (iii) Non-linear effects $h_0(\mathbf{X}) = X_3/2$ and $h_1(\mathbf{X})=3(X_1^2+X_5^2+X_1X_5+X_1+X_5)/8$. For binary outcome, we generated Y from a logistic model logit $\{p(Y = 1|\mathbf{X}, T)\} = 0.3T + h_0(\mathbf{X}) + h_1(\mathbf{X})T$. Three settings were

considered for $h_0(\mathbf{X})$ and $h_1(\mathbf{X})$: (i) Null with $h_0(\mathbf{X}) = h_1(\mathbf{X}) = 0$; (ii) Linear effects with $h_0(\mathbf{X}) = X_3/2$ and $h_1(\mathbf{X}) = (X_1 + X_2 + X_5)/5$; and (iii) Non-linear effects $h_0(\mathbf{X}) = X_1$ and $h_1(\mathbf{X}) = 3(X_5^2/4 + X_3X_5/2 + \Phi(3X_3) \times X_5/3)/4$, where Φ is the distribution function of standard normal.

In Table 1, we present results for continuous Y. Under H_0, the empirical size of all procedures are reasonably close to the nominal level of 0.05, except for the Crump method which has somewhat inflated type I error, especially when $n = 500$. The proposed test with k_Q tend to be be slightly conservative when $p = 20$ and $\rho = 0.2$ due to the high dimensionality of the associated RKHS. Under the alternative with linear effects, our procedure with linear kernel has similar performance as the Wald test when the correlation among predictors is small and $p = 5$. However, as p and the correlation among predictors increase, our proposed procedure with k_L outperforms the Wald test. For example, when $\rho = 0.5$ and $p = 20$, the power at $n = 500$ and 1000 is {0.357, 0.562} for Wald test and {0.474, 0.706} for our proposed method with k_L. The power loss in the Wald can in part be attributed to the use of a p degree of freedom (DF) when the effective DF in the presence of high correlation could be much lower than p. On the contrary, our proposed test leverages the correlation, resulting a lower effective DF. When the effects are linear, our proposed test with k_Q suffers some power loss when $n = 500$ but has comparable power when $n = 1000$. On the other hand, our KM score test with k_G out performs all other tests even when the effects are linear. This is not surprising since when ρ is large, the $\mathcal{H}_{k_G}$ approximates the linear space (Cai et al., 2011) while allowing ρ to vary enables us to choose different basis functions to more efficiently capture the effects. The Crump method has power similar to or slightly higher than the Wald test across different p and ρ, but is slightly lower than our proposed procedures especially when $\rho = 0.5$. When the underlying effects are non-linear, both the Wald test and the KM score test with k_L perform poorly with low power, as expected, and the Crump method has similar performance as these two procedures. The KM score tests with both k_Q and k_G have substantially higher power across all settings. It is interesting to note that although the underlying effects are quadratic, the KM test with k_G has comparable or higher power when p is small. For the larger p of 20, the test with k_Q substantially outperforms k_G. One possible explanation is that the RKHS with k_G may not be an efficient approximation to capture $h_1(\mathbf{X}) - h_0(\mathbf{X})$ when compared to that based on k_Q. Across all scenarios, the omnibus test behaves close to the score test with the optimal kernel, which demonstrates the effectiveness of our kernel selection and perturbation procedure.

The results for binary outcome are presented in Table 2. All procedures maintain the type I error reasonably well although in this setting the Wald test has a slightly conservative size when $p = 20$. Unlike the setting with continuous outcome, our test is no longer expected to perform similarly to the Wald test even when the effects are linear since the two tests are capturing different aspects of the TE. When $p = 5$, the proposed test and the Wald test perform similarly. However, when $p = 20$, the KM score test with k_L substantially outperform the Wald test. For example, when $p = 20$, $\rho = 0.5$ and $n = 500$, the empirical power is 0.608 for the KM score test and only 0.229 for the Wald test. In this setting, the KM test with k_Q and k_G also perform quite comparably to the test with linear kernel, demonstrating the robustness of the test with non-linear kernels. When the underlying effects

are non-linear, the KM test with k_Q generally perform better than the tests assuming linear effects. Since the non-linear signals are mostly quadratic, the KM test with k_Q is generally more powerful than those from k_G although the procedures have similar performances when $p = 5$. Again, the power of the omnibus test is close to that of the test with optimal kernel for all scenarios.

3.2 Example: Predictors Useful for Individualized Treatment of HIV Infected Patients

We apply our methods along with the aforementioned existing methods to data from AIDS Clinical Trials Group Protocal 175 (ACTG175), which is a double-blind study that evaluated treatment with either a single nucleoside or two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1) (Hammer et al., 1996). The dataset contains 2139 HIV-infected subjects, where subjects were randomized to four different treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine and ddI monotherapy. Following the primary goal of the original study, we compare ZDV monotherapy ($T = 0$) to combination therapies ($T = 1$) and aim to identify baseline predictors that are associated with differential TE. We considered the long term immune response, defined as 96 (± 5) week CD4 counts, $CD4_{96}$, as the continuous outcome which was also used in Tsiatis et al. (2008). To test for predictors for ITR, we included 12 baseline covariates separated into 3 groups: (i) demographic information including age, weight, race and gender; (ii) risk factors including hemophilia status, homosexual activity, antiretroviral history, symptomatic status and history of intravenous drug use; and (iii) functional markers including Karnofsky score, baseline CD4 and baseline CD8 count. The goal is to test whether any group of covariates significantly affects the absolute risk reduction by different treatments, so the variables in the significant group can be used to guide treatment selection in the future. The results for the response being the continuous $CD4_{96}$ as defined are shown in Table 3(a). Our proposed method detected functional markers as being significantly predictive of treatment response with p-value about 0.01 and the demographic variables as being marginally significant with p-value 0.07 when the gaussian kernel is employed. On the other hand, the Wald test identified none of the predictor groups as significant. The omnibus procedure results in p-values close to the smallest p-value among the score tests with different kernels, while the Crump method achieves similar p-values to the Wald test.

3.3 Example: Predictors Useful for Treatment of Patients with Advanced Chronic Heart Failure

We also illustrate the proposed procedures using the Beta-Blocker Evaluation of Survival Trial (BEST), which is a randomized clinical trial to investigate if Bucindolol, a beta-blocker, would benefit patients with advanced chronic heart failure (CHF) (Beta-Blocker Evaluation of Survival Trial Investigators, 2001). The 2-year BEST study had 2708 participants randomized at 1:1 ratio to receive either Bucindolol or Placebo. We considered the Physician's Global Assessment (PGA) as the primary response. The PGA takes seven ordinal levels (1–3: different levels of worsening, 4:no change, 5–7: different levels of improvement) and we defined a binary outcome Y if the PGA $\geqslant 4$, reflecting some improvement. For baseline predictors, we considered four groups with grouping information provided in the original study database: (i) Ischemic CHF Etiology (ICE; 6 covariates), including prior myocardio infarction, stenosis, coronary artery disease etiology and so on;

(ii) Physical Exam (PE; 14 covariates), including heart rate, blood pressure, weight, height etc; (iii) Hematology Lab Test (HLT; 4 covariates): hematocrit, hemoglobin, platelet, and white blood count; (iv) Chemistry Lab Test (CLT; 19 covariates), including Clucose, Sodium, Calcium etc; and (v) Cardiac History (CH; 9 covariates), including Duration of CHF, Peripheral Vascular disease etc. The goal is to test whether any of these groups are significantly associated with treatment difference with respect to the binary outcome Y reflecting improvement in PGA.

Results given in Table 3(b) suggest that Physical Exam results are significantly associated with treatment difference with p-value 0.01 from the KM score test with k_G. Results of the KM test with k_L and k_Q are consistent with p-values 0.04 and 0.09, respectively. There is also suggestive evidence that Ischemic and CHF etiology may be associated with treatment response with a marginally significant p-value from the KM test with k_Q although the test is not significant for other kernels. Again, the Wald test failed to reject for any of the predictor groups.

4. Discussion

In this paper, we proposed a KM based score test to identify informative baseline predictors that can be useful for individualized treatment selection. Our method is robust due to the model-free construction of the statistic. Our proposed KM test is also generally more powerful than the existing Wald test. Numerical studies suggest that our proposed procedures could substantially outperform the Wald test, especially when testing for a moderate number of predictors that are correlated with each other and/or when the underlying effects are non-linear. Different kernel functions may be preferable for different types of signals. We also propose an omnibus test that combines information from multiple kernels. Simulation results suggest that the omnibus test performs well in selecting an optimal kernel for testing. When $\mathbf{x}$ consists of a mixture of discrete ($\mathbf{x}_D$) and continuous ($\mathbf{x}_C$) components with $\mathrm{x}=(\mathrm{x}_D^\top, \mathrm{x}_C^\top)^\top$, one may employ different kernels for $\mathbf{x}_D$ and $\mathbf{x}_C$ such as $k(\mathbf{x}_1, \mathbf{x}_2) = k_C(\mathbf{x}_{C1}, \mathbf{x}_{C2}) + k_D(\mathbf{x}_{D1}, \mathbf{x}_{D2})$, where k_C and k_D are different kernels. One may also account for heterogeneity in the covariate distribution by simply normalizing them to have equal variance and employ the same kernel. Examples of various kernels for different types of variables can be found in Hofmann et al. (2008).

The proposed procedure does not necessitate the estimation of the "main effects" of $\mathbf{X}$ on Y, which increases the robustness of our procedure since no model assumptions are required. On the other hand, inclusion of "main effects" properly may further increase the power of the test. Specifically, one may consider modifying $\widehat{\delta_i}$ as $\{Y_i - \overline{Y}_1 - \hat{h}(\mathbf{X}_i)\}I(T_i{=}1)/\hat{\pi}_1 - \{Y_i - \overline{Y}_0 - \hat{h}(\mathbf{X}_i)\}I(T_i{=}0)/\hat{\pi}_0$, for some $\hat{h}(\mathbf{X}_i)$ obtained via certain working models. Here, $\hat{h}(\mathbf{X})$ can be viewed as estimated "main effects" for continuous outcomes and proper choices of $\hat{h}(\cdot)$ may improve the power of the test. The power gain is achieved by leveraging the independence between treatment assignment and baseline covariates, similar to those augmentation procedures proposed in Tian et al. (2012)

and Zhang et al. (2008). Optimal choices of $\hat{h}(\cdot)$ to maximize the power gain warrant further research.

The proposed KM framework for testing can also be extended to estimate heterogenous treatment effects of interest with a given set of $\mathbf{X}$. If we assume the treatment effect is a linear function of the bases: $E(Y^{(1)} - Y^{(0)}|\mathbf{X}) = \beta^\top \overrightarrow{\psi(\mathbf{X})}$, then we may estimate the model parameter $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}$, the solution to $\sum_{i=1}^{n} \overrightarrow{\psi(\mathbf{X}_i)}[\hat{\delta}_i - \boldsymbol{\beta}^\top \overrightarrow{\psi(\mathbf{X}_i)}\}] = 0$ where $\overrightarrow{\psi(\mathbf{X}_i)} = \{1, \psi(\mathbf{X})^\top\}^\top$. Subsequently, one may use $\hat{\boldsymbol{\beta}}^\top \overrightarrow{\psi(\mathbf{X})}$ as basis for constructing ITRs. The performance of such an approach warrants further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grants R01 GM079330, R01 HL089778, and U54 HG007963.

References

Beta-Blocker Evaluation of Survival Trial Investigators. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. The New England journal of medicine. 2001; 344:1659–1667. [PubMed: 11386264]

Bilias Y, Gu M, Ying Z, et al. Towards a general asymptotic theory for cox model with staggered entry. The Annals of Statistics. 1997; 25:662–682.

Braun, ML., et al. PhD thesis. Friedrich-Wilhelms-Universität; Bonn: 2005. Spectral properties of the kernel matrix and their relation to kernel methods in machine learning.

Cai T, Tian L, Wei L. Semiparametric box–cox power transformation models for censored survival observations. Biometrika. 2005; 92:619–632.

Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12:270–282. [PubMed: 20876663]

Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics. 2011; 67:975–986. [PubMed: 21281275]

Cristianini, N., Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000.

Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics. 2008; 90:389–405.

Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika. 1977; 64:247–254.

Duffy MJ, Crown J. A personalized approach to cancer treatment: how biomarkers can help. Clinical chemistry. 2008; 54:1770–1779. [PubMed: 18801934]

Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in medicine. 2011; 30:2867–2880. [PubMed: 21815180]

Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions. Statistical methodology. 2011; 8:42–55.

Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine. 1996; 335:1081–1090. [PubMed: 8813038]

Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. The Annals of Statistics. 2008; 36:1171–1220.

Imai K, Ratkovic M, et al. Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics. 2013; 7:443–470.

Koltchinskii V, Giné E. Random matrix approximation of spectra of integral operators. Bernoulli. 2000:113–167.

La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. Nature reviews Clinical oncology. 2011; 8:587–596.

Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. Statistical methods in medical research. 2013; 22:493–504. [PubMed: 22116341]

Mallal S, Phillips E, Carosi G, Molina J, Workman C, Tomazic J, Jagel-Guedes E, Rugina S, Kozyrev O, Cid J, et al. HLA-B* 5701 screening for hypersensitivity to abacavir. New England Journal of Medicine. 2008; 358:568. [PubMed: 18256392]

IBCSG. (The International Breast Cancer Study Group) endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node negative breast cancer: A randomized trial. J Natl Cancer Inst. 2002; 94:1054–65. [PubMed: 12122096]

Nelson H, Tyne K, Naik A, Bougatsos C, Chan B, Humphrey L. Screening for breast cancer: an update for the US Preventive Services Task Force. Annals of Internal Medicine. 2009; 151:727. [PubMed: 19920273]

Ong F, Das K, Wang J, Vakil H, Kuo J, Blackwell W, Lim S, Goodarzi M, Bernstein K, Rotter J, et al. Personalized medicine and pharmacogenetic biomarkers: progress in molecular oncology testing. Expert review of molecular diagnostics. 2012; 12:593–602. [PubMed: 22845480]

Pollard, D. Empirical processes: theory and applications. Institute of Mathematical Statistics and American Statistical Association; 1990. NSF-CBMS Reginal Conference Series in Probability and Statistics 2

Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Annals of statistics. 2011; 39:1180. [PubMed: 21666835]

Rosenblum M, van der Laan MJ. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. Biometrics. 2009; 65:937–945. [PubMed: 19210739]

Rotnitzky A, Robins JM. Inverse probability weighting in survival analysis. Encyclopedia of Biostatistics. 2005

Tian L, Cai T, Goetghebeur E, Wei L. Model evaluation based on the sampling distribution of estimated absolute prediction error. Biometrika. 2007; 94:297–311.

Tian L, Cai T, Zhao L, Wei LJ. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. Biostatistics. 2012; 13:256–273. [PubMed: 22294672]

Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Statistics in medicine. 2008; 27:4658–4677. [PubMed: 17960577]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful snp-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics. 2010; 86:929–942. [PubMed: 20560208]

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012; 1:103–114. [PubMed: 23645940]

Zhang M, Tsiatis AA, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics. 2008; 64:707–715. [PubMed: 18190618]

Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association. 2013; 108:527–539. [PubMed: 24058223]

Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012; 107:1106–1118. [PubMed: 23630406]

Zujewski J, Kamin L. Trial assessing individualized options for treatment for breast cancer: the tailorx trial. Future oncology (London, England). 2008; 4:603–610.

Table 1

Empirical size and power for different methods, under various sample size, number of predictors, and correlation among predictors, with continuous outcome.

			size		nonlinear		linear	
		method	**n=500**	**n=1000**	**n=500**	**n=1000**	**n=500**	**n=1000**
ρ=0.2	p=5	Wald	0.050	0.050	0.317	0.519	0.543	0.791
		k_L	0.048	0.052	0.361	0.588	0.530	0.778
		k_Q	0.037	0.044	0.801	0.987	0.493	0.745
		k_G	0.033	0.039	0.994	1.000	0.611	0.909
		Omnibus	0.035	0.043	0.990	1.000	0.572	0.881
		Crump	0.069	0.063	0.361	0.559	0.582	0.795
	p=20	Wald	0.051	0.046	0.275	0.401	0.458	0.693
		k_L	0.044	0.047	0.318	0.541	0.498	0.745
		k_Q	0.028	0.031	0.653	0.933	0.425	0.679
		k_G	0.043	0.046	0.386	0.779	0.511	0.774
		Omnibus	0.032	0.037	0.603	0.921	0.461	0.736
		Crump	0.099	0.062	0.388	0.459	0.561	0.733
ρ=0.5	p=5	Wald	0.044	0.042	0.275	0.414	0.429	0.665
		k_L	0.054	0.051	0.316	0.480	0.466	0.709
		k_Q	0.041	0.051	0.893	0.985	0.371	0.613
		k_G	0.041	0.045	0.998	1.000	0.604	0.910
		Omnibus	0.046	0.048	0.995	1.000	0.530	0.866
		Crump	0.077	0.075	0.302	0.434	0.447	0.671
	p=20	Wald	0.039	0.041	0.215	0.319	0.357	0.562
		k_L	0.049	0.046	0.297	0.470	0.474	0.706
		k_Q	0.045	0.039	0.818	0.969	0.390	0.623
		k_G	0.047	0.046	0.706	0.990	0.517	0.808
		Omnibus	0.040	0.040	0.816	0.987	0.469	0.763

	size	nonlinear		linear	
method	n=500	n=1000	n=500	n=1000	n=500	n=1000
Crump	0.080	0.064	0.325	0.361	0.455	0.625

Table 2

Empirical size and power for different methods, under various sample size, number of predictors, and correlation among predictors, with binary outcome.

			size		nonlinear		linear	
		method	**n=500**	**n=1000**	**n=500**	**n=1000**	**n=500**	**n=1000**
ρ=0.2	p=5	Wald	0.046	0.053	0.217	0.440	0.549	0.891
		k_L	0.046	0.051	0.258	0.467	0.615	0.895
		k_Q	0.047	0.052	0.573	0.936	0.631	0.904
		k_G	0.051	0.051	0.560	0.943	0.615	0.894
		Omnibus	0.051	0.052	0.587	0.952	0.628	0.906
	p=20	Wald	0.025	0.033	0.095	0.209	0.220	0.590
		k_L	0.050	0.053	0.224	0.361	0.481	0.787
		k_Q	0.048	0.049	0.500	0.768	0.444	0.786
		k_G	0.051	0.055	0.338	0.576	0.486	0.796
		Omnibus	0.050	0.049	0.430	0.722	0.458	0.791
ρ=0.5	p=5	Wald	0.045	0.055	0.542	0.889	0.622	0.936
		k_L	0.053	0.051	0.784	0.966	0.683	0.931
		k_Q	0.056	0.053	0.933	1.000	0.645	0.911
		k_G	0.055	0.054	0.856	0.997	0.714	0.944
		Omnibus	0.057	0.052	0.910	0.999	0.693	0.935
	p=20	Wald	0.027	0.035	0.228	0.595	0.229	0.666
		k_L	0.050	0.052	0.720	0.951	0.608	0.883
		k_Q	0.053	0.055	0.813	0.985	0.561	0.851
		k_G	0.050	0.052	0.731	0.960	0.634	0.906
		Omnibus	0.051	0.053	0.793	0.982	0.599	0.882

Table 3

P-value for testing the overall effects of different groups of baseline predictors from the Wald test and the proposed KM score test with three kernels: k_L, k_Q and k_G.

(a) Treatment Effect on week 96 CD4 counts ACTG175

	demographic	risk factors	functional markers
Wald	0.18	0.96	0.27
k_L	0.24	0.99	0.02
k_Q	0.25	0.99	0.01
k_G	0.07	0.72	0.01
Omnibus	0.10	0.80	0.01
Crump	0.17	0.99	0.25

(b) Treatment Effect on PGA Response with BEST Study

	ICE	PE	HLT	CLT	CH
Wald	0.22	0.12	0.41	0.33	0.20
k_L	0.19	0.04	0.51	0.38	0.18
k_Q	0.05	0.09	0.13	0.06	0.23
k_G	0.18	0.01	0.08	0.19	0.23
Omnibus	0.08	0.04	0.10	0.09	0.20