Received: 4 May 2019 | Accepted: 7 September 2020

DOI: 10.1111/biom.13373

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Net benefit index: Assessing the influence of a biomarker for individualized treatment rules

Yiwang Zhou[1] | **Peter X.K. Song**[1] | **Haoda Fu**[2]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

[2] Eli Lilly and Company, Lilly Corperate Center, Indianapolis, Indiana

Correspondence
Peter X.K. Song, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109.
Email: pxsong@umich.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: DMS 1811734; National Institute of Environmental Health Sciences, Grant/Award Number: R01ES024732

Abstract

One central task in precision medicine is to establish individualized treatment rules (ITRs) for patients with heterogeneous responses to different therapies. Motivated from a randomized clinical trial for Type 2 diabetic patients on a comparison of two drugs, that is, pioglitazone and gliclazide, we consider a problem: utilizing promising candidate biomarkers to improve an existing ITR. This calls for a biomarker evaluation procedure that enables to gauge added values of individual biomarkers. We propose an assessment analytic, termed as *net benefit index (NBI)*, that quantifies a contrast between the resulting gain and loss of treatment benefits when a biomarker enters ITR to reallocate patients in treatments. We optimize reallocation schemes via outcome weighted learning (OWL), from which the optimal treatment group labels are generated by weighted support vector machine (SVM). To account for sampling uncertainty in assessing a biomarker, we propose an NBI-based test for a significant improvement over the existing ITR, where the empirical null distribution is constructed via the method of stratified permutation by treatment arms. Applying NBI to the motivating diabetes trial, we found that baseline fasting insulin is an important biomarker that leads to an improvement over an existing ITR based only on patient's baseline fasting plasma glucose (FPG), age, and body mass index (BMI) to reduce FPG over a period of 52 weeks.

KEYWORDS
biomarker, bootstrap null distribution, clinical trial, O-learning, personalized medicine

1 | INTRODUCTION

Utility of the newly discovered biomarkers, such as omics-markers, from basic sciences to facilitate better and more cost-effective clinical practice is of critical importance in translational medicine. In connection to the emerging field of personalized medicine, one central task is to update the existing individualized treatment rules (ITRs) using new biomarkers with the aim to receive better clinical benefit. A noticeable shortcoming in the current statistical literature of personalized medicine is a lack of such methods to evaluate the significance of individual biomarkers in improving the existing ITRs. Perhaps this has been regarded as a small mathematical problem, but such methodological need is not easy to be addressed appropriately given many practical constraints involved, such as clinical benefit and medical cost associated with the inclusion of such biomarkers in daily clinical practice. This paper is intended to fill in this technical gap with a specific objective of developing a new statistical procedure to assess the usefulness of a biomarker in the context of personalized medicine.

Motivated from a randomized clinical trial comparing two drugs, that is, pioglitazone and gliclazide, on treating Type 2 diabetic patients, we propose, examine, and

illustrate a new statistical framework, termed as *net benefit index (NBI)*, to analytically and numerically quantify added values of candidate biomarkers when they are used to update an existing ITR. We consider an existing ITR involving age, body mass index (BMI), and baseline fasting plasma glucose (FPG) to maximize the reduction of FPG after 52 weeks of treatment. With several new variables like Hemoglobin A1c (HbA1c), fasting insulin, and so forth, we want to evaluate their added values, and decide which one or ones can significantly improve the existing ITR. Added value of a promising biomarker is gauged by a contrast between the resulting gain and loss of clinical benefits from reallocations of treatments through a revised ITR with the utility of this new biomarker. Reallocation is optimized by outcome weighted learning (OWL) (Zhao *et al.*, 2012). In addition, we consider an NBI-based test for significance of a certain added value in which the empirical null distribution is generated via stratified permutation by treatment arms. Through this procedure, biomarkers that significantly improve an existing ITR are sequentially selected to revise current allocation rules, where the biomarker selection is controlled under false discovery rate (FDR) to avoid the issue of overfitting. Note that overfitting in terms of the number of biomarkers is practically unattractive due to higher costs and longer time spent on collecting samples that essentially produce redundant information, which may undermine the accuracy and interpretation of an estimated ITR.

While NBI allows to evaluate the contribution of new biomarkers under controlled FDR, it avoids some other problems known in existing biomarker selection techniques in the context of personalized medicine. In the field of decision making, variable selection should target primarily at prescriptive variables that help prescribe the optimal action, instead of the predictive variables that reduce the variability and increase the accuracy of an estimator. A prescriptive variable has to have a qualitative interaction with the treatment (Gunter *et al.*, 2011). A variable is said to qualitatively interact with the treatment if there exists at least two distinct, nonempty sets within the space of the variable, for which the treatment arms that maximize the expected clinical benefit are distinct (Gunter *et al.*, 2011). Qian and Murphy (2011) propose a two-stage Q-learning (Q denoting "quality") (Watkins, 1989) procedure that employs the l_1-penalty for variable selection to estimate an optimal ITR. Lu *et al.* (2013) develop a penalized regression framework, known as a kind of A-learning (A denoting "advantage") (Murphy, 2003) that allows to simultaneously estimate an optimal ITR and to select an important variable. However, neither of these two methods specifically targets at the selection of prescriptive variables. Gunter *et al.* (2011) propose two variable-ranking criteria, U-score and S-score, for variable selection via quantitative interactions. But one limitation of these criteria is the ignorance of the correlations between the variables. To overcome this issue, Fan *et al.* (2016) develop a sequential advantage selection (SAS) method based on a modified version of S-score. SAS sequentially evaluates additional values of new variables via qualitative interactions, so that it can avoid identifying any variables marginally important but jointly unimportant. However, SAS lacks its direct relevance in clinical practice as it does not directly optimize treatment benefit objective for ITR, instead building models with sequentially added interaction terms under a statistical criterion of mean squared error. Different from these existing methods, NBI has the following advantages: (a) NBI directly optimizes treatment grouping labels to maximize the expected clinical benefit; (b) NBI selects important prescriptive variables beneficial for treatment allocation; (c) NBI is naturally applicable for nonlinear decision rules due to the invocation of support vector machine (SVM); (d) NBI sequantially selects biomarkers into an existing ITR through FDR control.

The remainder of the paper is organized as follows. Section 2 introduces the motivating diabetes clinical trial, followed by the framework of NBI for biomarker assessment in Section 3. Section 4 evaluates the proposed NBI test through simulation experiments. The NBI method is illustrated by the motivating clinical trial in Section 5. Section 6 contains some concluding remarks. Some additional results are available in the Supporting Information.

2 | APPLICATION: A DIABETES CLINICAL TRIAL

This is a randomized control and double-blind trial that aims to compare the therapeutic effects of pioglitazone and gliclazide in treating patients with Type 2 diabetes. Pioglitazone and gliclazide are two common oral medications with different therapeutic mechanisms for the treatment of Type 2 diabetic patients. A total of 1270 patients with Type 2 diabetes were recruited into the trial. All the eligible patients were randomized to a 52-week treatment period. The outcomes of interest is the change of FPG between the last posttreatment measurement and baseline. FPG was measured repeatedly at baseline and at weeks 4, 8, 12 up to 52. Other variables measured at baseline included age, BMI, HbA1c, fasting insulin, high-density lipoproteins (HDL), low-density lipoproteins (LDL), aspartate transferases (AST), alanine transferases (ALT), total cholesterol, triglycerides, creatinine, and gamma-glutamyl transferase (GGT). After deleting subjects with missing data, a sample of 830 patients remains available for analysis, with 424 assigned to pioglitazone and 406 assigned to gliclazide. Due to loss of follow-up, some of the patients had the last

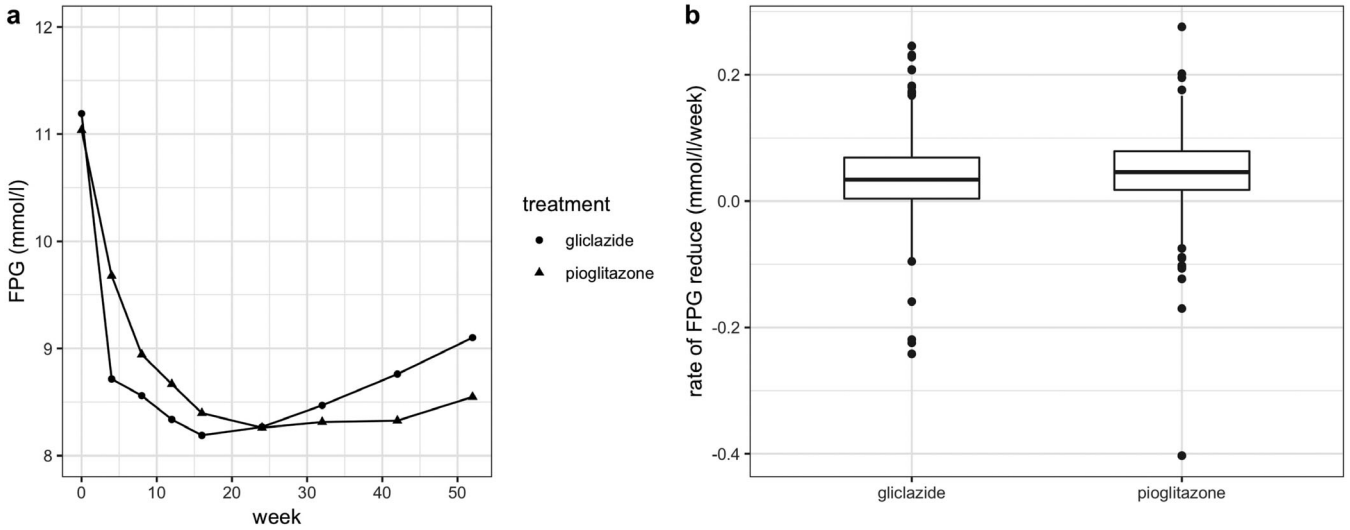

FIGURE 1 (a) Effects of pioglitazone and gliclazide on reducing fasting plasma glucose (FPG) during the 52-week period; (b) Average reduction rate of FPG by pioglitazone and gliclazide in the 52-week period

posttreatment measurement taken at week 32 or 42, resulting in a shorter period of treatment. Charbonnel *et al.* (2005) perform a noninferiority test for the differential treatment effects between these two drugs on the reduction of FPG and illustrate a significantly greater mean reduction of FPG by pioglitazone (2.4 mmol/L) than by gliclazide (2.0 mmol/L), with a treatment difference of 0.4 mmol/L in favor of pioglitazone (95% CI 0.1 to 0.7 mmol/L). The comparison of FPG reduction rate illustrates that pioglitazone leads to a more effective FPG reduction (0.049 mmol/L/week) than gliclazide (0.038 mmol/L/week), with a treatment difference of 0.011 mmol/L/week (95% CI 0.003 to 0.019 mmol/L/week) (Figure 1).

The previous analysis finds a significant treatment difference on population average between pioglitazone and gliclazide in reducing FPG. However, for individual patients, some taking pioglitazone may receive little benefit, while some taking gliclazide may receive significant benefit. Given such heterogeneous responses to these drugs, an optimal ITR is deemed necessary to increase the benefit by shuffling patients in a systematic way to assign each patient to the "right" drug. This purpose of reallocation may be formulated and achieved with the aim to maximize the expected FPG reduction via a revised drug allocation rule. Consider a simple preliminary ITR that involves only baseline FPG, age, and BMI, denoted by ITR(b.FPG, age, BMI). Age and BMI are two commonly used demographics for treatment assignment, while baseline FPG is a key clinical factor representing a personal reference level for the target endpoint. Among the available additional candidate biomarkers, we want to determine which ones may provide significant added values to improve ITR(b.FPG, age, BMI); if there are some, the expanded ITR is expected to provide higher treatment benefit than that given by the preliminary ITR.

3 | FORMULATION

3.1 | OWL and optimal ITR

Consider a two-armed randomized clinical trial where each patient is randomly assigned a treatment $A \in \mathcal{A} = \{-1, 1\}$. $A = -1$ is the traditional treatment, say *gliclazide*. $A = 1$ is the new treatmnet, say *pioglitazone*. Complete randomization implies that the treatment allocation scheme is independent of patients' prognostic variables, denoted as $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathcal{X} \subseteq \mathbb{R}^d$. Potential clinical benefit $B^*(A)$ is the outcome that would result if a patient were assigned to A. Since each patient takes only one treatment, the observed clinical benefit is given by $B = I(A = 1)B^*(1) + I(A = -1)B^*(-1)$, which is determined by A. The other potential clinical benefit is latent with no data captured. Suppose that B is bounded with a larger value of B being clinically more desirable. The primary aim of personalized medicine is to establish a decision rule $D(\boldsymbol{X})$, a mapping $\mathcal{X} \to \mathcal{A}$, that maximizes the expectation of the clinical benefit. The following three assumptions are typically required for computing the expectation of the clinical benefit: (a) consistency assumption: $B = I(A = 1)B^*(1) + I(A = -1)B^*(-1)$; (b) no unmeasured confounders assumption: $A \perp \{B^*(a)\}_{a\in\mathcal{A}}|\boldsymbol{X}$; (c) positivity assumption: $P\{P(A = a|\boldsymbol{X}) > 0\} = 1, \forall a \in \mathcal{A}$ (Robins, 1997).

In this paper, the optimal ITR refers to the decision rule $D^*(\boldsymbol{X})$ that maximizes the expected clinical benefit

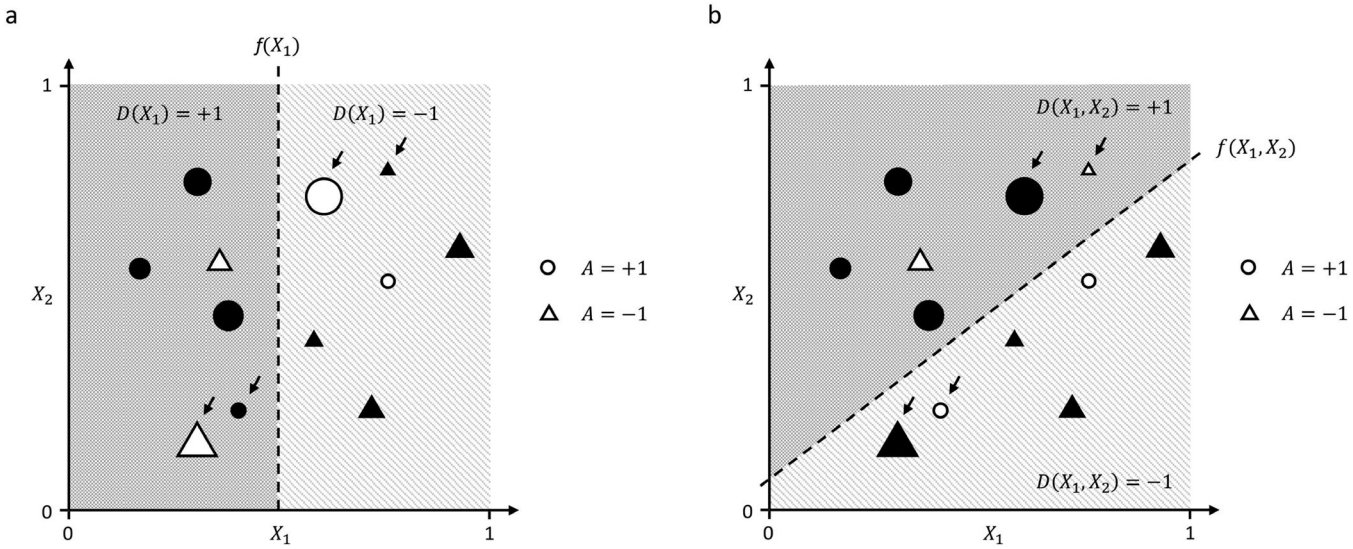

FIGURE 2 Changes of the subjects included in the calculation of $V(D)$ when new biomarker X_2 is included in the estimation of the decision function. Sizes of the circles and triangles reflect the magnitude of the clinical benefit B. Black circles and triangles are the subjects included in the calculation of $V(D)$ since they have $D(\boldsymbol{X}) = A$. Subjects in group "gain" and "loss" are pointed by arrows. (a) Decision function $f(X_1)$ estimated only on the existing biomarker X_1; (b) decision function $f(X_1, X_2)$ estimated on X_1 and X_2

$V(D) = E\{\frac{I(A=D(\boldsymbol{X}))}{P(A|\boldsymbol{X})}B\}$, which according to Zhao *et al.* (2012) may be formulated as a weighted classification problem that can be solved by SVM in the context of OWL (see details in the Supporting Information). The optimization problem of OWL has very relevant interpretations to our definition of NBI. For patients with large observed benefits, the optimality encourages to allocate the same treatment type as the one previously assigned. Conversely, for those receiving small observed benefits, the optimality tends to assign the alternative treatment type. Note that in the diabetes trial, the actually implemented allocation is complete randomization, that is, $P(A|\boldsymbol{X}) = P(A) = 1/2$. But this might not give the best personalized drug allocation rule as randomization primarily aims to control confounding not to maximize treatment benefit. Clearly, the optimal decision rule $D^*(\boldsymbol{X})$ will give a higher overall clinical benefit in comparing to completely randomized trial.

3.2 | Treatment reallocation

Let $D(\boldsymbol{X})$ be a decision rule based on features $\boldsymbol{X}$. The optimization for $D(\boldsymbol{X})$ imposed by OWL encourages concordant treatment assignment on patient who receives clearly treatment benefit. In other words, with the invocation of OWL, patients who are assigned $D(\boldsymbol{X}) = A$ tend to have larger benefit than those who receive $D(\boldsymbol{X}) \neq A$. Denote an existing decision function by $f_e(\boldsymbol{X}_e)$ and the corresponding decision rule by $D_e(\boldsymbol{X}_e)$ based on the existing variables $\boldsymbol{X}_e$. Likewise, denote an updated decision function by $f_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$ and the corresponding updated decision rule by $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$ by involving new variables $\boldsymbol{X}_u$. Under decision rules $D_e(\boldsymbol{X}_e)$ and $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$, there exist two subgroups of patients who receive the same treatment allocation, that is, $D_e(\boldsymbol{X}_e) = A$ and $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u) = A$, respectively. For those patients who are assigned the same treatment by $D_e(\boldsymbol{X}_e)$ and $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$, the inclusion of a new biomarker does not lead to any benefit gain. Thus, they should be excluded from the assessment of the difference caused by the biomarker. In other words, only those patients who are assigned a different treatment by $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$ from that by $D_e(\boldsymbol{X}_e)$ should be used to define an effective amount of clinical benefit.

Figure 2 illustrates a simple example showing the different groups of subjects included in the calculation of $V(D)$ when new biomarker X_2 is used in the learning of an updated decision function $f(X_1, X_2)$ compared to $f(X_1)$. Subjects randomly assigned to $A = 1$ and $A = -1$ in the trial are denoted by circles and triangles. The sizes of circles and triangles reflect the magnitude of clinical benefit, with a bigger size corresponding to a larger benefit. Only those subjects in black are effectively included in the calculation of $V(D)$ since they are assigned the concordant treatment $D(\boldsymbol{X}) = A$. By comparing the black circles and triangles in Figures 2(a) and 2(b), we can find that a circle and a triangle (pointed by arrows) are newly included in the calculation of $V(D)$, indicating that there is a gain as the consequence of reallocation by $f(X_1, X_2)$. At the same time, another circle and another triangle (pointed by arrows) are excluded from the calculation of $V(D)$, indicating that there is a loss. Since the newly included subjects have larger benefit (larger size), the gain exceeds the

loss in the expected benefit value. Note that only subjects with $f(X_1) \times f(X_1, X_2) < 0$ will be included into the respective groups "gain" and "loss" since they have switching allocations from $D(\boldsymbol{X}_e)$ to $D(\boldsymbol{X}_e, \boldsymbol{X}_u)$. Technically, they are the ones responsible for a difference in the calculation of $V(D)$. It is conceptually appealing to quantify the contrast between "gain" and "loss" to understand the influence of a new biomarker for added value in personalized treatment allocation. We assume the following ethics conditional on reallocation treatment.

Assumption 1. *Let $D(\boldsymbol{X})$ be an allocation rule based on variable $\boldsymbol{X}$, and let $B(A|\boldsymbol{X})$ be the observed benefit when $D(\boldsymbol{X}) = A$. Suppose $\forall \epsilon > 0$, $\exists \delta(\epsilon) < \epsilon$, such that $P\{B(A^c|\boldsymbol{X}) \geq B(A|\boldsymbol{X}) | B(A|\boldsymbol{X}) < \epsilon\} \geq 1 - \delta(\epsilon)$, where A^c is the alternative treatment to A.*

Assumption 1 implies that when the clinical benefit of a patient receiving treatment A tends to zero, with probability approaching to 1 there is no loss of benefit for allocating the alternative treatment A^c to the patient.

3.3 | Net benefit index (NBI)

Let B_i be the observed benefit value for each patient, $i = 1, \ldots, n$. Denote $\mathcal{S}_{\text{gain}}$ as the sample of patients in group "gain," and $\mathcal{S}_{\text{loss}}$ as the sample of patients in group "loss." An NBI for a new biomarker X_u is defined as follows:

Definition (NBI).

$$\text{NBI}(X_u) = \frac{\sum_{i \in \mathcal{S}_{\text{gain}}} B_i / P(A_i|\boldsymbol{X}_i)}{\sum_{i \in \mathcal{S}_{\text{gain}}} 1/P(A_i|\boldsymbol{X}_i)} - \frac{\sum_{i \in \mathcal{S}_{\text{loss}}} B_i / P(A_i|\boldsymbol{X}_i)}{\sum_{i \in \mathcal{S}_{\text{loss}}} 1/P(A_i|\boldsymbol{X}_i)}. \tag{1}$$

Remark 1. In the case of a randomized clinical trial, propensity $P(A_i|\boldsymbol{X}_i) = P(A_i) \equiv 1/2$ for $i = 1, \ldots, n$. Let $n_{\text{gain}} = |\mathcal{S}_{\text{gain}}|$ and $n_{\text{loss}} = |\mathcal{S}_{\text{loss}}|$, NBI becomes:

$$\text{NBI} = \sum_{i \in \mathcal{S}_{\text{gain}}} B_i / n_{\text{gain}} - \sum_{i \in \mathcal{S}_{\text{loss}}} B_i / n_{\text{loss}} = \overline{B}_{\mathcal{S}_{\text{gain}}} - \overline{B}_{\mathcal{S}_{\text{loss}}}, \tag{2}$$

which is actually the difference of the average observed benefits of $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$. Clearly, NBI > 0 suggests that a new biomarker is potentially valuable to improve ITR.

Remark 2. To account for sampling variability, we propose a standardized NBI as: standardized-NBI$(X_u) = (\overline{B}_{\mathcal{S}_{\text{gain}}} - \overline{B}_{\mathcal{S}_{\text{loss}}}) / \sqrt{\frac{s^2_{\text{gain}}}{n_{\text{gain}}} + \frac{s^2_{\text{loss}}}{n_{\text{loss}}}}$, where s^2_{gain} and s^2_{loss} are sample variances of observed benefits for $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$, respectively.

Algorithm 1 Calculation of standardized-NBI(X_u)

1: Establish models $f_e(\mathbf{X}_e)$ and $f_u(\mathbf{X}_e, X_u)$ using OWL on a training dataset.
2: Get classifications $D_e(\mathbf{X}_e)$ and $D_u(\mathbf{X}_e, X_u)$ for subjects in a NBI evaluation dataset.
3: Characterize samples $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$ by comparing $D_e(\mathbf{X}_e)$ and $D_u(\mathbf{X}_e, X_u)$.
4: **if** $n_{\text{gain}} \geq 5$ and $n_{\text{loss}} \geq 5$ **then**
5: Calculate standardized-NBI(X_u) based on the values of B in $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$.
6: **else** Set standardized-NBI(X_u)=0.

The calculation of NBI and standardized NBI for X_u is given by Algorithm 1. Note that the minimal sample size, $n_{\text{gain}} \geq 5$ and $n_{\text{loss}} \geq 5$, is required in the calculation of standardized NBI to have numerical stability.

3.4 | Test for significant NBI

NBI > 0 is only suggestive subjective to sampling uncertainty, which may further be made rigorous by hypothesis testing. For a practical point of view, we hypothesized that $D_u(\boldsymbol{X}_e, \boldsymbol{X}_u)$ should not be inferior to $D_e(\boldsymbol{X}_e)$. Thus, we consider the following hypotheses: H_0: the new biomarker does not improve ITR; against H_a: the new biomarker improves ITR. Let μ_{gain} and μ_{loss} be the expected benefits in the "gain" and the "loss" population, respectively. The hypotheses can be stated as a two-sample mean comparison: $H_0 : \mu_{\text{gain}} = \mu_{\text{loss}}$; against $H_a : \mu_{\text{gain}} > \mu_{\text{loss}}$. We will apply the standardized NBI to perform the hypothesis of the two-sample mean comparison. However, it is difficult to derive the distribution of standardized NBI. Therefore, we invoke the empirical null distribution of the standardized NBI to generate the p-values.

Remark 3. Different from the standard two-sample test, here $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$ are random sets generated by a resulting optimal reallocation of treatments under a common overall optimal benefit objective function. Thus, there exists a certain shared action in group membership labeling, which leads to a dependence between these two sets. However, when conditional on the memberships of $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$, we have the conditional independence, which leads to a standard unequal variance two-sample t-statistic, $t|\mathcal{S}_{\text{gain}}, \mathcal{S}_{\text{loss}} = \{\delta - (\overline{B}_{\text{gain}} - \overline{B}_{\text{loss}})\} / \sqrt{s^2_{\text{gain}}/n_{\text{gain}} + s^2_{\text{loss}}/n_{\text{loss}}} \sim t(\nu)$, where $\delta = \mu_{\text{gain}} - \mu_{\text{loss}}$ and $\nu = \nu(n_{\text{gain}}, n_{\text{loss}}, s^2_{\text{gain}}, s^2_{\text{loss}})$ is degrees of freedom. Clearly, the sizes of both sets, n_{gain} and n_{loss}, are random and correlated in ν. Since the labels in $\mathcal{S}_{\text{gain}}$ and $\mathcal{S}_{\text{loss}}$ have

a rather complicated and unknown joint distribution, the marginal distribution of the t-statistic is not available to make inference. Thus, we invoke the empirical null distribution to obtain p-values.

To do so, we propose to create a null variable X_{null} via the means of permutation with projected residuals $r_i = X_{u,i} - \widehat{E}(X_{u,i}|\boldsymbol{X}_{e,i}), i = 1, \dots, n$ as detailed in Algorithm 2.

Algorithm 2 Generation of empirical null distribution for standardized-NBI(X_u)

1: **if** $\boldsymbol{X}_e \neq Null$ **then**
2: Model $X_u = g(\boldsymbol{X}_e) + \epsilon$, where $g(\cdot)$ is a suitable function independent of A and B.
3: Get residuals $r_i = X_{u,i} - \hat{E}(X_{u,i}|\boldsymbol{X}_{e,i}), i = 1, \dots, n$.
4: **else** Let $r_i = X_{u,i}$.
5: **for** $l = 1, \dots, L$ **do** (L is the number of permutation replicates).
6: Permute the residuals conditional on A; get the permuted residuals $r^p_{l,i}, i = 1, \dots, n$.
7: Values of $X_{\text{null},l}$ are generated as $X_{\text{null},l,i} = \widehat{E}(X_{u,i}|\mathbf{X}_{e,i}) + r^p_{l,i}, i = 1, \dots, n$.
8: Calculate standardized-NBI($X_{\text{null},l}$) using Algorithm 1.

Assumption 2. *The new variable X_u can be expressed by an additive model $X_u = g(\boldsymbol{X}_e) + \epsilon$ of the existing variables $\boldsymbol{X}_e$ and the error term ϵ. Discussion of violations of Assumption 2 is included in Section 6.*

Algorithm 2 outputs the empirical null distribution of the standardized NBI, and the p-value is given as $p = \#\{\text{standardized-NBI}(X_{\text{null}}) > \text{standardized-NBI}(X_u)\}/L$. The invocation of stratification by treatment arm in the permutation test is to retain the difference between the underlying distributions of the residuals across two treatment groups. Pooling the residual distributions together while performing permutation test would ruin the interaction effect between treatment and biomarkers. Since our major interest is to evaluate the added value of a biomarker when we have an existing ITR, we will focus our method on the situation when $\boldsymbol{X}_e \neq Null$ in the following simulation studies and real data analysis. Simulations results with $\boldsymbol{X}_e = Null$ are included in Table S.1 in the Supporting Information.

When there are several new variables under screening, say m, it is necessary to control FDR to ensure a balance of sensitivity and specificity. To proceed, we propose a forward selection method, Algorithm 3, that sequentially adds the currently most significant variable with the smallest p-value into a current model at each step until no more variables are to be added. The significant variables at each step are identified as those passing the FDR control through the Benjamini-Hochberg procedure.

4 | SIMULATION EXPERIMENT

In this section, we conducted extensive simulations to evaluate the finite sample performance of the proposed NBI method.

Algorithm 3 Sequential forward variable selection based on NBI test

1: Set $m = \dim(\boldsymbol{X}_u)$.
2: **while** $m > 0$ **do**
3: Get p_j for $X_{u,j}, j = 1, \dots, m$ with the existing model involving $\boldsymbol{X}_e$ by NBI test.
4: Order $p_{(1)} \leq \cdots \leq p_{(m)}$, each corresponding to $H_{(j)}$: $\boldsymbol{X}_{u,(j)}$ does not improve ITR.
5: Find $j_{\max} = \max_j\{j : p_{(j)} \leq \frac{j}{m}q\}$, where $q \in (0,1)$ is the chosen target FDR control.
6: **if** $j_{\max}$ exists **then**Update $\boldsymbol{X}_e = \{\boldsymbol{X}_e, X_{u,(1)}\}, \boldsymbol{X}_u = \boldsymbol{X}_u \setminus X_{u,(1)}$, and set $m = m - 1$.
7: **else** Stop.

4.1 | Single-variable–based decision rule evaluation

The first simulation concerns a setting in which an existing ITR consists of two variables X_1 and X_2, where $X_i \overset{i.i.d}{\sim} U(0,1), i = 1, 2$. Suppose that a new variable $X_u \sim U(0,1)$ becomes available, which is correlated with X_2, namely, $\text{Corr}(X_2, X_u) = \rho$ with $\rho \in \{0.0, 0.2, 0.5, 0.8\}$. The following types of X_u are considered: (i) $X_u = X_3$, an important feature related to benefit B; (ii) $X_u = X_4$, a noise variable unrelated to B. Our goal is to assess the sensitivity (ie, rate of detecting X_3) and specificity (ie, rate of not detecting X_4) by the proposed NBI test. Allocation of treatment $A \in \{-1, 1\}$ is independent of $\boldsymbol{X}$ with $P(A|\boldsymbol{X}) = 1/2$. B is generated from a normal distribution with mean $\mu = 0.5 + X_1 + 2.0Af(\boldsymbol{X})$ and standard deviation $\sigma = 1.0$. Interaction term $Af(\boldsymbol{X})$ specifies a bimodal expected benefit that generates bifurcated benefit outcomes. We consider the following three scenarios of true decision function f:

(1) (Linear) $f(\boldsymbol{X}) = 1 - X_1 + X_2 - 2X_3$;
(2) (Binary) $f(\boldsymbol{X}) = 4\{I(X_1 > 0.1 \cap X_2 < 0.75 \cap X_3 > 0.25) - 0.5\}$;
(3) (Nonlinear) $f(\boldsymbol{X}) = 2.5\{(X_1 - 0.5)^+ + (X_2 - 0.2)^+ - (X_3 - 0.1)^+\}$.

The sample size is set at $n = 800, 1000, 1200$. Fivefold cross-validation is used to determine the training data set to learn f and the NBI evaluation data set to assess X_u. We set type I error rate $\alpha = 0.05$. Simulation is repeated for 1000 times.

TABLE 1 Discovery rates for X_3 and X_4 in the single-variable–based decision rule evaluation (discovery rate for X_4 equals 1-specificity)

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.8$	
Scenario	***n***	X_3	X_4	X_3	X_4	X_3	X_4	X_3	X_4
Linear	800	0.987	0.047	0.991	0.056	0.984	0.057	0.931	0.053
	1000	0.998	0.054	0.997	0.043	0.992	0.051	0.958	0.055
	1200	0.999	0.059	1.000	0.051	0.998	0.052	0.974	0.053
Binary	800	0.932	0.051	0.946	0.049	0.956	0.050	0.815	0.050
	1000	0.956	0.049	0.967	0.034	0.971	0.046	0.866	0.051
	1200	0.968	0.054	0.975	0.048	0.972	0.045	0.927	0.050
Nonlinear	800	0.977	0.053	0.987	0.055	0.973	0.055	0.854	0.050
	1000	0.984	0.055	0.996	0.058	0.983	0.042	0.906	0.044
	1200	0.995	0.043	0.997	0.049	0.994	0.054	0.934	0.036

TABLE 2 NBI values for X_3 and X_4 in the single-variable–based decision rule evaluation

		$\rho = 0.0$		$\rho = 0.2$	
Scenario	***n***	X_3 **mean (*SD*)**	X_4 **mean (*SD*)**	X_3 **mean (*SD*)**	X_4 **mean (*SD*)**
Linear	800	1.708 (0.547)	−0.150 (0.636)	1.724 (0.556)	−0.136 (0.653)
	1000	1.700 (0.517)	−0.139 (0.622)	1.693 (0.509)	−0.194 (0.611)
	1200	1.657 (0.445)	−0.161 (0.614)	1.655 (0.442)	−0.140 (0.611)
Binary	800	3.139 (1.553)	−0.155 (1.922)	3.192 (1.489)	−0.217 (1.815)
	1000	3.090 (1.449)	−0.127 (1.827)	3.121 (1.382)	−0.160 (1.835)
	1200	3.075 (1.386)	−0.081 (1.779)	3.069 (1.287)	−0.046 (1.776)
Nonlinear	800	1.793 (0.883)	−0.224 (0.888)	1.862 (0.804)	−0.265 (0.894)
	1000	1.793 (0.791)	−0.227 (0.883)	1.856 (0.740)	−0.293 (0.889)
	1200	1.770 (0.723)	−0.228 (0.827)	1.855 (0.695)	−0.277 (0.791)
		$\rho = 0.5$		$\rho = 0.8$	
Scenario	***n***	X_3 **mean (*SD*)**	X_4 **mean (*SD*)**	X_3 **mean (*SD*)**	X_4 **mean (*SD*)**
Linear	800	1.594 (0.628)	−0.127 (0.651)	1.125 (0.677)	−0.158 (0.688)
	1000	1.551 (0.552)	−0.148 (0.626)	1.154 (0.563)	−0.145 (0.626)
	1200	1.503 (0.487)	−0.171 (0.616)	1.152 (0.511)	−0.132 (0.618)
Binary	800	3.329 (1.499)	−0.224 (1.855)	2.285 (1.888)	−0.208 (1.760)
	1000	3.172 (1.318)	−0.056 (1.751)	2.345 (1.685)	−0.154 (1.813)
	1200	3.252 (1.231)	−0.029 (1.756)	2.573 (1.622)	−0.139 (1.653)
Nonlinear	800	1.791 (0.796)	−0.284 (0.903)	1.188 (0.832)	−0.264 (0.932)
	1000	1.801 (0.689)	−0.329 (0.873)	1.293 (0.743)	−0.343 (0.912)
	1200	1.823 (0.672)	−0.277 (0.841)	1.336 (0.692)	−0.310 (0.811)

Table 1 summarizes the discovery rates of X_3 and X_4 across different f based on the proposed NBI method. It is shown that the sensitivity is high in detecting the useful variable X_3, and type I error has been well controlled for the noise variable X_4 at the nominal level 0.05. Table 2 reports the NBI values for X_3 and X_4. Aligned with the high sensitivity, the corresponding NBI(X_3) are all positive, implying that the inclusion of X_3 results in an improved ITR(X_1, X_2, X_3), in which more patients are assigned into their beneficial treatment arm in comparison to the previous ITR(X_1, X_2). In contrast, all the NBI(X_4) values are negative, indicating that the inclusion of X_4 results in a worse updated ITR(X_1, X_2, X_4) that assigns more patients into their nonbeneficial treatment arm. When FDR is controlled, the chance of X_4 entering the updated ITR is slim, and the resulting decline in outcome of benefit is indeed ignorable.

4.2 | Multiple-variable–based decision rule evaluation

The second simulation uses the same setup of the base ITR(X_1, X_2) specified in Section 4.1. Now we

TABLE 3 Size, TDR, MCC, and CCR for variable selection based on NBI test, SAS, and riskRFE in the multiple-variable–based decision rule evaluation

NBI					
Scenario	***n***	**Size (*SD*)**	**TDR (*SD*)**	**MCC (*SD*)**	**CCR (*SD*)**
Linear	800	1.745 (0.709)	0.906 (0.222)	0.801 (0.242)	0.835 (0.081)
	1000	1.813 (0.689)	0.922 (0.194)	0.828 (0.225)	0.852 (0.076)
	1200	1.904 (0.622)	0.934 (0.170)	0.868 (0.202)	0.870 (0.070)
Binary	800	1.844 (0.747)	0.894 (0.232)	0.813 (0.248)	0.765 (0.098)
	1000	1.917 (0.705)	0.906 (0.215)	0.847 (0.232)	0.786 (0.095)
	1200	1.924 (0.650)	0.919 (0.199)	0.863 (0.222)	0.805 (0.090)
Nonlinear	800	1.805 (0.744)	0.904 (0.220)	0.802 (0.245)	0.818 (0.081)
	1000	1.926 (0.738)	0.910 (0.199)	0.833 (0.225)	0.832 (0.075)
	1200	1.913 (0.616)	0.929 (0.186)	0.869 (0.217)	0.847 (0.073)
SAS					
Scenario	***n***	**Size (*SD*)**	**TDR (*SD*)**	**MCC (*SD*)**	**CCR (*SD*)**
Linear	800	2.817 (0.887)	0.774 (0.212)	0.831 (0.165)	0.971 (0.012)
	1000	2.525 (0.732)	0.846 (0.193)	0.887 (0.146)	0.976 (0.010)
	1200	2.339 (0.609)	0.898 (0.169)	0.926 (0.126)	0.979 (0.010)
Binary	800	3.662 (1.244)	0.613 (0.215)	0.693 (0.187)	0.743 (0.014)
	1000	3.286 (1.095)	0.677 (0.219)	0.751 (0.180)	0.744 (0.014)
	1200	3.052 (1.007)	0.723 (0.217)	0.790 (0.174)	0.744 (0.014)
Nonlinear	800	3.238 (1.102)	0.688 (0.219)	0.760 (0.180)	0.943 (0.013)
	1000	2.915 (0.942)	0.752 (0.215)	0.813 (0.169)	0.948 (0.012)
	1200	2.654 (0.817)	0.814 (0.205)	0.862 (0.157)	0.949 (0.011)
riskRFE					
Scenario	***n***	**Size (*SD*)**	**TDR (*SD*)**	**MCC (*SD*)**	**CCR (*SD*)**
Linear	800	3.091 (1.003)	0.660 (0.222)	0.704 (0.222)	0.852 (0.056)
	1000	2.815 (0.895)	0.732 (0.221)	0.773 (0.209)	0.869 (0.057)
	1200	2.586 (0.781)	0.797 (0.218)	0.828 (0.204)	0.883 (0.052)
Binary	800	3.498 (1.135)	0.636 (0.211)	0.716 (0.178)	0.737 (0.123)
	1000	3.200 (0.990)	0.686 (0.209)	0.761 (0.169)	0.755 (0.124)
	1200	2.909 (0.847)	0.745 (0.206)	0.810 (0.159)	0.779 (0.117)
Nonlinear	800	3.142 (1.083)	0.651 (0.235)	0.689 (0.243)	0.834 (0.061)
	1000	2.837 (1.009)	0.723 (0.230)	0.755 (0.219)	0.844 (0.059)
	1200	2.576 (0.823)	0.791 (0.224)	0.812 (0.215)	0.857 (0.055)

consider multiple signal and noise candidate biomarkers $X_j \sim U(0,1), j = 3, \ldots, 12$, in which only X_3 and X_4 are signal biomarkers involved in the optimal ITR. The correlation structure of the variables is that $\text{Corr}(X_3, X_5) = \text{Corr}(X_4, X_6) = 0.5$, and $\text{Corr}(X_s, X_t) = 0.2, s, t \in \{7, \ldots, 12\}, s \neq t$. The mean parameter of benefit B is set as $\mu = 0.5 + X_1 + 2.0Af(X)$, where $f(\boldsymbol{X})$ is given as follows:

(4) (Linear) $f(\boldsymbol{X}) = 0.5(1 + X_1 + X_2 - 1.8X_3 - 2.2X_4)$;
(5) (Binary) $f(\boldsymbol{X}) = 6\{I(X_1 > 0.12 \cap X_2 < 0.88 \cap X_3 > 0.2 \cap X_4 < 0.8) - 0.5\}$;
(6) (Nonlinear) $f(\boldsymbol{X}) = (X_1 - 0.9)^+ - (X_2 - 0.78)^+ + (X_3 - 0.1)^+ - (X_4 - 0.22)^+$.

We draw summary statistics under the FDR control set at 0.10. In addition to those basic performance measures considered in Section 4.1, we add a comparison of our NBI method on biomarker selection with SAS mentioned in Section 1 and riskRFE (Dasgupta *et al.*, 2019), a backward elimination method for variable selection developed for SVM.

Table 3 reports some summary statistics, including (a) size: the total number of selected biomarkers; (b) true discovery rate (TDR): the number of correctly selected biomarkers over size; (c) Matthers correlation coefficient (MCC): $\text{MCC} = \frac{(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$, where TP is true positive, TN is true negative, FP is false positive, and FN is false negative; (d) correct classification rate (CCR).

The gold numbers are size = 2, TDR = 1, MCC = 1, and CCR = 1. Our NBI test tends to give slightly conservative results with smaller size, a known consequence of FDR control (Benjamini and Hochberg, 1995). Clearly, SAS and riskRFE pay a price of overfitting with a large number of noise features selected, resulting in larger size and smaller TDR. One lesson we learn from the simulation is that we may first run SAS or riskRFE to select a relatively large pool of potential biomarkers, and then apply NBI to control FDR. In this way, we could reach a desirable balance of sensitivity and specificity. In regard to MCC, the proposed NBI test outperforms SAS and riskRFE, except for the linear scenario, where SAS gives the highest MCC. In addition, the estimated ITR derived from the NBI method gives the highest CCR for the binary scenario, but not for the linear and nonlinear scenarios, which is a limitation of OWL. Some additional simulation studies, including the small sample cases where $n = 200$, the scenarios where $\boldsymbol{X}_e = Null$ are included in the Supporting Information.

5 | ANALYSIS OF DIABETES TRIAL DATA

We apply the proposed NBI methodology to analyze the motivating diabetes trial described in Section 2. The outcome of benefit is the average reduction rate of FPG over the 52 weeks of treatment. The base ITR is driven by three variables $\boldsymbol{X}_e = \{\text{b.FPG, age, BMI}\}$. Among those candidate biomarkers listed in Section 2, we want to select some important ones and evaluate their added values to improve ITR.

We first performed a prescreening of these candidate biomarkers using SAS. Under the cut-off point 0.01 for the proportion of the incremental sequential advantage, SAS selects five variables potentially useful to update ITR, including baseline HbA1c, fasting insulin, AST, triglycerides, and creatinine, denoted by $\boldsymbol{X}_u^{\text{SAS}}$. The resulting decision rule is

$$\begin{aligned}\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}}) = 0.13 &- 0.14\text{b.FPG} - 0.02\text{age} - 0.17\text{BMI} \\ &- 0.05\text{HbA1c} - 0.12\text{b.fasting insulin} \\ &- 0.13\text{AST} + 0.04\text{triglycerides} \\ &+ 0.18\text{creatinine},\end{aligned}$$

We would allocate a patient with Type 2 diabetes to take pioglitazone if $\widehat{f} > 0$ and to take gliclazide if $\widehat{f} < 0$. $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$ assigns 586 patients to pioglitazone and 244 patients to gliclazide. Following Murphy *et al.* (2001), we further calculate the estimated value function by $\mathbb{E}_n^*[I(A = D(\boldsymbol{X}))B/P(A|\boldsymbol{X})]/\mathbb{E}_n^*[I(A = D(\boldsymbol{X}))/P(A|\boldsymbol{X})]$, where $\mathbb{E}_n^*$ is the empirical average value. In order to make the comparison from the same baseline, the same method (eg, SVM, which is the learning algorithm for both NBI and riskRFE) is used to calculate the estimated value function. $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$ gives an estimated value function of 0.049, meaning that the expected average FPG reduction rate would be 0.049 mmol/L/week over 52 weeks if $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$ were implemented for the whole population. The estimated value functions given by complete random allocation and $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e)$ are 0.048 and 0.052, respectively, indicating that $\boldsymbol{X}_u^{\text{SAS}}$ does not improve the existing decision rule as far as the estimated value function concerns. We then performed a biomarker screening using riskRFE, which selected three variables potentially useful to update the existing ITR, including baseline fasting insulin, creatinine, and GGT. The estimated decision rule is

$$\begin{aligned}\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}}) = -1.48 &- 0.08\text{b.FPG} + 0.37\text{age} \\ &+ 1.09\text{BMI} + 0.45\text{b.fasting insulin} \\ &+ 0.53\text{creatinine} + 0.97\text{GGT},\end{aligned}$$

$\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$ assigns 527 patients to take pioglitazone and 303 patients to take gliclazide. The estimated value function given by $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$ is 0.051, even slightly smaller than that given by $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e)$, which is 0.052, indicating that $\boldsymbol{X}_u^{\text{riskRFE}}$ is not actually improving the existing ITR.

A clinically relevant question to the above estimated decision rules is whether the selected candidate variables in $\boldsymbol{X}_u^{\text{SAS}}$ and $\boldsymbol{X}_u^{\text{riskRFE}}$ are really influential to ITR, or whether we may further reduce $\boldsymbol{X}_u^{\text{SAS}}$ and $\boldsymbol{X}_u^{\text{riskRFE}}$ to achieve a more cost-effective ITR. We then applied the proposed NBI test. It turns out that baseline fasting insulin is the only candidate variable selected by our NBI method. This gives a simpler decision rule,

$$\begin{aligned}\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}}) = -1.30 &- 0.07\text{b.FPG} + 0.39\text{age} \\ &+ 2.04\text{BMI} + 0.35\text{b.fasting insulin},\end{aligned}$$

which assigns 481 patients to pioglitazone and 349 patients to gliclazide. The estimated value function of this OWL-updated treatment regime is 0.053, which is higher than that given by $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$ and $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$, although $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}})$ only uses a single biomarker, baseline fasting insulin, to improve ITR. The estimated value functions given by $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e)$ is 0.052, indicating that the inclusion of $\boldsymbol{X}_u^{\text{NBI}}$ in the decision rule also improves

the ITR with respect to the expected average FPG reduction rate.

Comparing the coefficients in the estimated decision rules, we notice that the signs of the coefficients for some common variables are different in $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$, $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$, and $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}})$ (eg, age has a negative coefficient in $\widehat{f}_{\text{SAS}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{SAS}})$ but positive coefficients in $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$ and $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}})$). It may due to the following reasons. First, loading coefficients are estimated by conditioning on other variables in the decision rule, and thus signs of these coefficients are possibly different with different sets of predictors (which are correlated) used in the construction of the decision functions. What matters the most is indeed the maximum treatment benefit, which is the primary objective of this learning procedure. Although the signs of the coefficients are not directly interpretable in this type of methodology, we would still see a great deal of concordance, in particular between $\widehat{f}_{\text{riskRFE}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{riskRFE}})$ and $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}})$. This is because that NBI and riskRFE are both based on support vector machine (SVM), while SAS is based on regularized linear regression. Thus, the training procedures and estimating criteria are different between SAS and NBI/riskRFE. With the inclusion of the selected variables $\boldsymbol{X}_u^{\text{NBI}}$, $\widehat{f}_{\text{NBI}}(\boldsymbol{X}_e, \boldsymbol{X}_u^{\text{NBI}})$ improves the estimated value function by about 10% compared to complete randomness. In regard to the clinical impact of our results, we would think that the demonstrated improvement is clinically meaningful, especially for people on the border line of diabetes, that is, the so-called prediabetes. It is known that approximately 88 million adults—more than one in three—have prediabetes in the United States. Of those with prediabetes, more than 80% do not know they have it. The 10% change may help prediabetic people whose diagnostic values just cross the border line to be controlled at the normal level.

In addition to the expected reduction rate of FPG, we also compare the expected reduction rate of HbA1c, another outcome of interest in the trial. Preliminary analysis identifies no significant difference between the HbA1c reduction rates for patients receiving pioglitazone or gliclazide. The inclusion of $\boldsymbol{X}_u^{\text{NBI}}$, $\boldsymbol{X}_u^{\text{SAS}}$, and $\boldsymbol{X}_u^{\text{riskRFE}}$ does not improve the existing ITRs with all the existing and updated decisions rules giving the same estimated reduction of HbA1c, which is 0.031 mmol/L/week over the 52-week treatment. But it is still slightly higher than the value of 0.029 given by complete random allocation A.

6 | CONCLUDING REMARKS

In this paper, we proposed a new biomarker assessment tool, termed as NBI, that enables to evaluate added values of biomarkers for improving existing ITRs. This new method can be used in both single and multiple-variable–based decision rule evaluations. Extensive simulation studies demonstrate that our method can correctly identify signal biomarkers under various scenarios with desirable performances in comparison to existing methods. Application of the proposed method to a real diabetes clinical trial reveals that baseline fasting insulin is an important biomarker that can significantly improve an existing ITR involving age, BMI, and baseline fasting FPG, for the allocation of pioglitazone or gliclazide to patients with Type 2 diabetes. It results significant clinical benefit of average reduction rate of FPG during the 52 weeks of treatment.

NBI is an analog to net reclassification improvement (NRI), a seminal index that has been extensively used to evaluate the usefulness of new markers for predicting risk of developing diseases (Pencina *et al.*, 2010). The proposed NBI is fundamentally different from NRI in the sense that NBI pertains to reclassification with respect to treatment group when class labels are not directly observed, rather based on outcome of treatment. Pepe *et al.* (2014) demonstrated that false-positive conclusions based on the NRI statistic were unacceptably high. However, our simulation studies have illustrated that the FDR is well controlled using the NBI method. Vickers and Pepe (2014) pointed out that NRI weights reclassification (ie, false positive and false negative) inappropriately, which may also be an underlying problem of NBI. However, with no information of true label available in the setting of NBI, appropriate weighting of false positive and false negative may be infeasible in practice. Decision curve analysis (DCA) (Vickers and Elkin, 2006) is another commonly used method for comparing multiple treatment decision rules to select the optimal one that maximizes the outcome of interest. The formulation of DCA relies on the calculation of a net benefit, which is the relative harm of a false positive and a false negative. Similarly, it is infeasible to apply DCA in our setting since we never know the underlying true labels for patients in a clinical trial in practice.

NBI is naturally applicable for nonlinear decision rules due to the invocation of SVM, in which different kernels (eg, Gaussian kernel) can be easily applied. We would like to clarify two points in the usage of kernels: (a) kernel is used exclusively to model the conditional distribution of R given $\boldsymbol{X}$ and (b) the assumption of additive errors in the generation of X_{null} is imposed on the conditional distribution of $\boldsymbol{X}_u$ given $\boldsymbol{X}_e$. Thus, the additivity assumption does not influence the relationship (or the decision rule) between R and $\boldsymbol{X}$ characterized by kernel in the generation of the empirical null distribution. In order to generate the null distribution for NBI, a certain assumption on the influence of random errors on signals is inevitable. In this paper, we adopted the classical additive error

assumption, which can be violated in practice. In order to check the validity of the additive error assumption, we suggest first to run a residual diagnosis. If it indicates that the error is not additive, that is, $X_u = f(\boldsymbol{X}_e, \epsilon)$, we can apply Taylor Expansion to the function $f(\boldsymbol{X}_e, \epsilon)$ and use a generalized additive model (GAM) to model X_u on $\boldsymbol{X}_e$. Then, permutation test can be performed based on the new error term $\epsilon' = X_u - \widehat{\text{GAM}}(\boldsymbol{X}_e)$, or $\epsilon'' = \{X_u - \widehat{\text{GAM}}(\boldsymbol{X}_e)\}/\sqrt{\widehat{\theta}(\boldsymbol{X}_e)}$ if we assume the variance of ϵ' can be modeled by a function $\theta(\boldsymbol{X}_e)$.

One future direction of this study is to extend this methodology to multiple treatment settings since clinical trials sometimes have more than two treatments in practice. In addition, it may be desirable to extend the method to situations where clinical benefit outcomes are categorical or time-to-event. With increasing interest and research in dynamic treatment regimes, we may also extend the NBI test to settings with multiple decision time points. Due to the number of replicates required by the proposed permutation test as well as the computation time needed for SVM, the proposed method may run into high computational demand. The algorithm can become faster if a theoretically justified null distribution is available for the NBI test statistic, which is worth an exploration in our future research.

ACKNOWLEDGMENTS

This research is supported by NIH grant (R01ES024732) and NSF grant (DMS 1811734). The authors would like to thank the coeditor, the associate editor, and two anonymous referees for their valuable comments that helped improve this paper significantly.

DATA AVAILABILITY STATEMENT

The data used in this paper cannot be freely shared with the public since the study protocol restricts the distribution of these data with no approved data use license. Readers who want to analyze the data should submit a formal research proposal to the sponsor Eli Lilly and Company via the corresponding author.

ORCID

Peter X.K. Song https://orcid.org/0000-0001-7881-7182

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Berkane, M. (Ed.). (2012) *Latent Variable Modeling and Applications to Causality* (Vol. 120). New York, NY: Springer Science & Business Media.

Charbonnel, B.H., Matthews, D.R., Schernthaner, G., Hanefeld, M., Brunetti, P. and QUARTET Study Group., (2005) A long-term comparison of pioglitazone and gliclazide in patients with Type 2 diabetes mellitus: a randomized, double-blind, parallel-group comparison trial. *Diabetic Medicine*, 22(4), 399–405.

Dasgupta, S., Goldberg, Y. and Kosorok, M.R. (2019) Feature elimination in kernel machines in moderately high dimensions. *The Annals of Statistics*, 47(1), 497–526.

Fan, A., Lu, W. and Song, R. (2016) Sequential advantage selection for optimal treatment regime. *The Annals of Applied Statistics*, 10(1), 32.

Gunter, L., Zhu, J. and Murphy, S.A. (2011) Variable selection for qualitative interactions. *Statistical Methodology*, 8(1), 42–55.

Lu, W., Zhang, H.H. and Zeng, D. (2013) Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5), 493–504.

Murphy, S.A. (2003) Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 331–355.

Murphy, S.A., van der Laan, M.J., Robins, J.M. and Conduct Problems Prevention Research Group., (2001) Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456), 1410–1423.

Pencina, M.J., D'agostino, R.B. and Vasan, R.S. (2010) Statistical methods for assessment of added usefulness of new biomarkers. *Clinical Chemistry and Laboratory Medicine*, 48(12), 1703–1711.

Pepe, M.S., Janes, H. and Li, C.I. (2014) Net risk reclassification p values: valid or misleading? *Journal of the National Cancer Institute*, 106(4), dju041.

Qian, M. and Murphy, S.A. (2011) Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2), 1180.

Vickers, A.J. and Elkin, E.B. (2006) Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.

Vickers, A.J. and Pepe, M. (2014) Does the net reclassification improvement help us evaluate models and markers? *Annals of Internal Medicine*, 160(2), 136–137.

Watkins, C.J.C.H. (1989) *Learning from Delayed Rewards*. Cambridge: King's College.

Zhao, Y., Zeng, D., Rush, A.J. and Kosorok, M.R. (2012) Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499), 1106–1118.

SUPPORTING INFORMATION

Web Appendices, Tables and Python code referenced in Sections 4.2 are available with this paper at the Biometrics website on Wiley Online Library. One Python coding example of NBI used for Table 3 is available at https://github.com/yiwangz/NBI for a free download.

How to cite this article: Zhou Y, Song PXK, Fu H. Net benefit index: Assessing the influence of a biomarker for individualized treatment rules. *Biometrics*. 2021;77:1254–1264. https://doi.org/10.1111/biom.13373