Received: 15 October 2020 | Revised: 1 October 2021 | Accepted: 14 October 2021

DOI: 10.1111/biom.13586

BIOMETRIC METHODOLOGY

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

Functional additive models for optimizing individualized treatment rules

Hyung Park[1] | Eva Petkova[1] | Thaddeus Tarpey[1] | R. Todd Ogden[2]

[1] Division of Biostatistics, Department of Population Health, New York University, New York, USA

[2] Department of Biostatistics, Columbia University, New York, USA

Correspondence
Hyung Park, Division of Biostatistics, Department of Population Health, New York University, New York, NY 10016, USA.
Email: parkh15@nyu.edu

Funding information
National Institute of Mental Health, Grant/Award Number: R01 MH099003

Abstract

A novel functional additive model is proposed, which is uniquely modified and constrained to model nonlinear interactions between a treatment indicator and a potentially large number of functional and/or scalar pretreatment covariates. The primary motivation for this approach is to optimize individualized treatment rules based on data from a randomized clinical trial. We generalize functional additive regression models by incorporating treatment-specific components into additive effect components. A structural constraint is imposed on the treatment-specific components in order to provide a class of additive models with main effects and interaction effects that are orthogonal to each other. If primary interest is in the interaction between treatment and the covariates, as is generally the case when optimizing individualized treatment rules, we can thereby circumvent the need to estimate the main effects of the covariates, obviating the need to specify their form and thus avoiding the issue of model misspecification. The methods are illustrated with data from a depression clinical trial with electroencephalogram functional data as patients' pretreatment covariates.

KEYWORDS

functional additive regression, individualized treatment rules, sparse additive models, treatment effect-modifiers

1 | INTRODUCTION

We propose a flexible functional regression approach to optimizing individualized treatment decision rules (ITRs) where the treatment has to be chosen to optimize the expected treatment outcome. We focus on the situation in which a potentially large number of patient characteristics are available as pretreatment functional and/or scalar covariates. Recent advances in biomedical imaging and high-throughput gene expression technology produce massive amounts of data on individual patients, opening up the possibility of tailoring treatments to the biosignatures of individual patients from individual-specific data (McKeague and Qian, 2014). Notably, some randomized clinical trials (RCTs) (e.g., Trivedi *et al.*, 2016) are designed to discover biosignatures that characterize patient heterogeneity in treatment responses from vast amounts of patient pretreatment characteristics. In this paper, we focus on some specific types of high-dimensional pretreatment patient characteristics observed in the form of curves or images, for instance, electroencephalogram (EEG) measurements. Such data can be viewed as functional (e.g., Ramsay and Silverman, 1997) and are becoming increasingly prevalent in modern RCTs as pretreatment covariates.

Much work has been carried out to develop methods for optimizing ITRs using data from RCTs. Regression-based methodologies are intended to optimize ITRs by estimating treatment-specific response (e.g., Lu *et al.*, 2011; Qian and Murphy, 2011; Tian *et al.*, 2014; Shi *et al.*, 2016; Jeng *et al.*, 2018) while attempting to maintain robustness with respect to model misspecification. Machine learning approaches for optimizing ITRs are often framed as a classification problem (e.g., Zhang *et al.*, 2012; Zhao *et al.*, 2019), including outcome weighted learning (OWL) (e.g., Zhao *et al.*, 2012, 2015; Song *et al.*, 2015) based on support vector machines, tree-based classification (e.g., Laber and Zhao, 2015) and adaptive boosting (Kang *et al.*, 2014), among others. However, to date there has been relatively little research on ITRs that directly utilize pretreatment functional covariates. McKeague and Qian (2014) proposed methods for optimizing ITRs that depend on a single pretreatment functional covariate. Ciarleglio *et al.* (2016) considered a flexible regression for a single functional covariate. Focusing on a single covariate, Laber and Staicu (2018) considered sparse, noisy, and irregularly spaced functional data, treating patient longitudinal information as a sparse functional covariate. Ciarleglio *et al.* (2015) proposed a method that allows for multiple functional/scalar covariates, which was then extended to incorporate a simultaneous covariate selection for ITRs in Ciarleglio *et al.* (2018). However, both of these approaches are limited to a stringent linear model assumption on the treatment-by-covariates interaction effects that limits its flexibility in optimizing ITRs and to two treatment conditions.

In this paper, we allow for nonlinear interactions between the treatment and multiple pretreatment functional covariates on the outcome and also for more than two treatment conditions. We incorporate a simultaneous covariate selection for ITRs through an L^1 regularization to deal with a large number of functional and/or scalar covariates. In a review by Morris (2015) on functional regression, the functional additive regression of Fan *et al.* (2015) and the functional generalized additive model of McLean *et al.* (2014) are two popular approaches to functional additive regression. In this paper, we base our method on the functional additive regression model of Fan *et al.* (2015) that utilizes one-dimensional (1D) data-driven functional indices and the associated additive link functions. In Ciarleglio *et al.* (2016), nonlinear effects are presented with the functional additive regression of McLean *et al.* (2014), for a single covariate. However, the approach of McLean *et al.* (2014) requires more parameters for estimation and is based on an L^2 penalty rather than on L^1 penalties, which is less suitable in the context of many functional covariates and when sparsity is desired. In this paper, we develop a flexible approach to optimizing ITRs that can easily impose structural constraints in modeling nonlinear heterogenous treatment effects with functional and/or scalar pretreatment covariates.

2 | CONSTRAINED FUNCTIONAL ADDITIVE MODELS

Let $Y^{(a)} \in \mathbb{R}$ be the potential outcome under treatment $A = a$ $(a = 1, \dots, L)$. We consider a set of p functional-valued pretreatment covariates $\boldsymbol{X} = (X_1, \dots, X_p)$, and q scalar-valued pretreatment covariates $\boldsymbol{Z} = (Z_1, \dots, Z_q) \in \mathbb{R}^q$. These pretreatment covariates $(\boldsymbol{X}, \boldsymbol{Z})$ are considered as potential biomarkers for optimizing ITRs. We will assume that each functional covariate X_j is a square integrable random function, defined on a compact interval, say, [0,1], without loss of generality. The L available treatment options are assigned with associated randomization probabilities $(\pi_1, \dots, \pi_L)$, such that $\sum_{a=1}^{L} \pi_a = 1$, $\pi_a > 0$, independent of $(\boldsymbol{X}, \boldsymbol{Z})$ (see Supporting Information Section A.17 for a dependent case).

In this context we focus on optimizing ITRs based on $(\boldsymbol{X}, \boldsymbol{Z}) \in \mathcal{X}$. Without loss of generality, we assume that a larger value of the outcome $Y^{(a)}$ is better. The goal is then (for a single decision point) to find an optimal ITR $\mathcal{D} : \mathcal{X} \mapsto \{1, \dots, L\}$, such that the treatment assignment $A = \mathcal{D}(\boldsymbol{X}, \boldsymbol{Z})$ maximizes the expected treatment outcome, the so-called value (V) function (Murphy, 2005), $V(\mathcal{D}) := E[Y^{(\mathcal{D})}]$. Under the standard causal inference assumptions in the Supporting Information Section A.16, $V(\mathcal{D}) = E[E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = \mathcal{D}(\boldsymbol{X}, \boldsymbol{Z})]]$, and the optimal ITR $\mathcal{D}^{opt}$, that maximizes $V(\mathcal{D})$, satisfies: $\mathcal{D}^{opt}(\boldsymbol{X}, \boldsymbol{Z}) = \arg\max_{a \in \{1,\dots,L\}} E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = a]$ (Qian and Murphy, 2011). In particular, $\mathcal{D}^{opt}$ does not depend on the "main" effect of the covariates $(\boldsymbol{X}, \boldsymbol{Z})$ and depends only on the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect (Qian and Murphy, 2011) in the mean response function $E[Y|\boldsymbol{X}, \boldsymbol{Z}, A]$. However, if this mean response model inadequately represents the interaction effect, the associated ITR may perform poorly.

Thus, we will focus on modeling possibly nonlinear $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effects, while allowing for an unspecified main effect of $(\boldsymbol{X}, \boldsymbol{Z})$. We base the model on the functional additive model (FAM) of Fan *et al.* (2015) allowing for nonlinear $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interactions:

$$E[Y|\boldsymbol{X},\boldsymbol{Z},A] = \underbrace{\mu(\boldsymbol{X},\boldsymbol{Z})}_{(\boldsymbol{X},\boldsymbol{Z})\text{ "main" effect}} + \underbrace{\sum_{j=1}^{p} g_j(\langle X_j,\beta_j\rangle, A) + \sum_{k=1}^{q} h_k(Z_k, A).}_{(\boldsymbol{X},\boldsymbol{Z})\text{-by-}A\text{ interaction effect}} \tag{1}$$

In model (1), the treatment a-specific (with $a \in \{1, \dots, L\}$) component functions $\{g_j(\cdot, a), j = 1, \dots, p\} \cup \{h_k(\cdot, a), k = 1, \dots, q\}$ are unspecified smooth 1D functions. Specifically, each function X_j appears as a 1D projection $\langle X_j, \beta_j \rangle := \int_0^1 X_j(s)\beta_j(s)ds$, via the standard L^2 inner product with a coefficient function $\beta_j \in \Theta$, where Θ is the space of square integrable functions over [0,1], restricted, without loss of generality, to a unit L^2 norm. (This is to ensure model identifiability; due to the unspecified nature of the functions β_j and g_j, β_j is only identifiable up to multiplications by nonzero constants.) The form of the function μ in (1) is left unspecified. For model (1), we assume an additive noise, $Y = E[Y|\boldsymbol{X}, \boldsymbol{Z}, A] + \epsilon$, where $\epsilon \in \mathbb{R}$ is a zero-mean noise with finite variance.

In model (1), to separate the nonparametric $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect from the additive $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect components, and to obtain an identifiable representation, we will constrain the $p+q$ component functions $\{g_j, j = 1, \dots, p\} \cup \{h_k, k = 1, \dots, q\}$ associated with the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect to satisfy the following identifiability conditions:

$$E\left[g_j(\langle X_j, \beta_j\rangle, A) \mid X_j\right] = 0\, (\forall \beta_j \in \Theta)(j = 1, \dots, p) \quad \text{and}$$
$$E[h_k(Z_k, A) \mid Z_k] = 0\, (k = 1, \dots, q) \tag{2}$$

(almost surely), where the expectation is taken with respect to the distribution of A given X_j (or Z_k). Condition (2) implies $E[\sum_{j=1}^{p} g_j(\langle X_j, \beta_j\rangle, A) + \sum_{k=1}^{q} h_k(Z_k, A) \mid \boldsymbol{X}, \boldsymbol{Z}] = 0$ (almost surely), which makes not only representation (1) identifiable but also the two effect components in model (1) orthogonal to each other. We call model (1) subject to the constraint (2), a *constrained functional additive model* (CFAM), which is the main model of the paper.

Notation 1. For a fixed β, let us denote the L^2 space of component functions, $g(\cdot,\cdot)$, over the random variables $(\langle X, \beta\rangle, A)$ as $\mathcal{H}^{(\beta)} = \{g \mid E[g(\langle X, \beta\rangle, A)] = 0, \|g\| < \infty\}$, with the norm $\|g\| = \sqrt{E[g^2(\langle X, \beta\rangle, A)]}$, where the expectation is taken with respect to the joint distribution of $(\langle X, \beta\rangle, A)$ and the inner product of the space defined as $\langle g, g'\rangle = E[g(\langle X, \beta\rangle, A) g'(\langle X, \beta\rangle, A)]$. Similarly, let us denote the L^2 space of component functions, $h(\cdot,\cdot)$, over (Z, A) as $\mathcal{H} = \{h \mid E[h(Z, A)] = 0, \|h\| < \infty\}$ with the norm $\|h\| = \sqrt{E[h^2(Z, A)]}$, where the expectation is with respect to the distribution of (Z, A), and similarly defined inner product. Without loss of generality, we suppress the treatment-specific intercepts in model (1), by removing the treatment a-specific means from Y, and assume $E[Y|A = a] = 0$ $(a = 1, \dots, L)$, that is, the main effect of A is 0 (see Supporting Information Section A.15 for the model with the treatment-specific intercepts).

Under the formulation (1) subject to the constraint (2), the "true" (i.e., optimal) functions, denoted as $\{g_j^*, j = 1, \dots, p\} \cup \{\beta_j^*, j = 1, \dots, p\} \cup \{h_k^*, k = 1, \dots, q\}$ that constitute the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect, can be viewed as the solution to the constrained optimization:

$$\{g_j^*, \beta_j^*, h_k^*\} = \underset{g_j \in \mathcal{H}_j^{(\beta_j)}, \beta_j \in \Theta, h_k \in \mathcal{H}_k}{\operatorname{argmin}}$$
$$E\left\{Y - \sum_{j=1}^{p} g_j(\langle X_j, \beta_j\rangle, A) - \sum_{k=1}^{q} h_k(Z_k, A)\right\}^2,$$
subject to
$$E\left[g_j(\langle X_j, \beta_j\rangle, A)|X_j\right] = 0\ \forall \beta_j \in \Theta (j = 1, \dots, p) \text{ and}$$
$$E[h_k(Z_k, A)|Z_k] = 0\ (k = 1, \dots, q). \tag{3}$$

Specifically, representation (3) does not involve the "main" effect functional μ, due to the orthogonal representation (1) implied by (2) (see Supporting Information Section A.1 for additional detail). If μ in (1) is a complicated functional subject to model misspecification, exploiting the representation on the right-hand side of (3) for $\{g_j^*, j = 1, \dots, p\} \cup \{\beta_j^*, j = 1, \dots, p\} \cup \{h_k^*, k = 1, \dots, q\}$ on the left-hand side is particularly appealing, as it provides a means of estimating the interaction terms without having to specify μ, thereby avoiding any issue of possible model misspecification for μ. The function μ can also be specified similar to (3) and estimated separately (see Supporting Information Section A.11), due to orthogonality in model (1). In particular, estimators of $\{g_j^*, \beta_j^*, h_k^*\}$ based on optimization (3) can be improved in terms of efficiency if Y in (3) is replaced by a "residualized" response $Y - \widehat{\mu}(\boldsymbol{X}, \boldsymbol{Z})$, where $\widehat{\mu}$ is some estimate of μ (see also Supporting Information Section A.11). However, for simplicity, we will focus on the representation (3) with the "unresidualized" Y.

Under model (1), the potential treatment effect-modifiers among $\{X_j, j = 1, \dots, p\} \cup \{Z_k, k = 1, \dots, q\}$ appear in the model only through the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect terms in (1). Ravikumar *et al.* (2009) proposed a sparse additive model (SAM) for relevant covariate selection in a high-dimensional additive regression. As in SAM, to deal with a large $p+q$ and to achieve treatment effect-modifying variable selection, under the often reasonable assumption that most covariates are inconsequential as treatment effect-modifiers, we impose sparsity on the set of component functions $\{g_j, j = 1, \dots, p\} \cup \{h_k, k = 1, \dots, q\}$ of CFAM (1). This sparsity structure on the set of component functions can be usefully incorporated into the optimization-based representation (3) of $\{g_j^*, \beta_j^*, h_k^*\}$:

$$\{g_j^*, \beta_j^*, h_k^*\} = \underset{g_j \in \mathcal{H}_j^{(\beta_j)}, \beta_j \in \Theta, h_k \in \mathcal{H}_k}{\operatorname{argmin}}$$
$$E\left\{Y - \sum_{j=1}^{p} g_j(\langle X_j, \beta_j\rangle, A) - \sum_{k=1}^{q} h_k(Z_k, A)\right\}^2$$

$$+\lambda\left\{\sum_{j=1}^{p}\|g_j\|+\sum_{k=1}^{q}\|h_k\|\right\},$$

$$\text{subject to } E[g_j(\langle X_j,\beta_j\rangle,A)|X_j]=0\ \forall\beta_j\in\Theta(j=1,\dots,p)$$

$$\text{and}\quad E[h_k(Z_k,A)|Z_k]=0\ (k=1,\dots,q), \tag{4}$$

for some sparsity-inducing parameter $\lambda\geq 0$. In (4), the component $\sum_{j=1}^{p}\|g_j\|+\sum_{k=1}^{q}\|h_k\|$ behaves like an L^1 *ball* across different functional components $\{g_j, j=1,\dots,p; h_k, k=1,\dots,q\}$ to encourage functional sparsity. For example, a relatively large value of λ in (4) will result in many components to be exactly zero, thereby enforcing sparsity on the set of functions $\{g_j^*, h_k^*\}$ on the left-hand side of (4). Specifically, Equation (4) can help model selection when dealing with potentially many functional/scalar pretreatment covariates. Potentially, separate sparsity tuning parameters λ_j and $\check{\lambda}_k$ (for X_j and Z_k) can be employed in (4). However, we restrict our attention to the case of a single sparsity tuning parameter that treats all X_j and Z_k on the equal footing for treatment effect-modifier selection.

3 | ESTIMATION

We first consider a population characterization of the algorithm for solving (4) in Section 3.1 and then a sample counterpart of the population algorithm in Section 3.2.

3.1 | Population algorithm

For a set of fixed coefficient functions $\{\beta_j, j=1,\dots,p\}$, the minimizing component function $g_j\in\mathcal{H}_j^{(\beta_j)}$ (and $h_k\in\mathcal{H}_k$) for each j (and each k) of the constrained objective function of (4) has a component-wise closed-form expression, as indicated next.

Theorem 1. *Given $\lambda\geq 0$ and a set of fixed single-index coefficient functions $\{\beta_j, j=1,\dots,p\}$, the minimizing component function $g_j\in\mathcal{H}_j^{(\beta_j)}$ of the constrained objective function of (4) satisfies:*

$$g_j(\langle X_j,\beta_j\rangle,A)$$
$$=\left[1-\frac{\lambda}{\|P_j\|}\right]_+ P_j(\langle X_j,\beta_j\rangle,A)\ \text{(almost surely)}, \tag{5}$$

where the function $P_j\in\mathcal{H}_j^{(\beta_j)}$:

$$P_j(\langle X_j,\beta_j\rangle,A):=E[R_j|\langle X_j,\beta_j\rangle,A]-E[R_j|\langle X_j,\beta_j\rangle], \tag{6}$$

in which

$$R_j=Y-\sum_{j'\neq j}g_{j'}(\langle X_{j'},\beta_{j'}\rangle,A)-\sum_{k=1}^{q}h_k(Z_k,A) \tag{7}$$

represents the jth (functional covariate's) partial residual; similarly, the minimizing component function $h_k\in\mathcal{H}_k$ of the constrained objective function of (4) satisfies:

$$h_k(Z_k,A)=\left[1-\frac{\lambda}{\|\check{P}_k\|}\right]_+\check{P}_k(Z_k,A)\ \textit{(almost surely)}, \tag{8}$$

where the function $\check{P}_k\in\mathcal{H}_k$:

$$\check{P}_k(Z_k,A):=E[\check{R}_k|Z_k,A]-E[\check{R}_k|Z_k], \tag{9}$$

and

$$\check{R}_k=Y-\sum_{j=1}^{p}g_j(\langle X_j,\beta_j\rangle,A)-\sum_{k'\neq k}h_{k'}(Z_{k'},A) \tag{10}$$

represents the kth (scalar covariate's) partial residual. (In (5) and (8), $[u]_+=\max(0,u)$ represents the positive part of u.)

The proof of Theorem 1 is in Supporting Information Section A.2. Given a sparsity tuning parameter $\lambda\geq 0$, optimization (4) can be split into two iterative steps (Fan *et al.*, 2014, 2015). First (*Step 1*), for a set of fixed single-indices $\langle X_j,\beta_j\rangle$ $(j=1,\dots,p)$, the component functions $\{g_j, j=1,\dots,p\}\cup\{h_k, k=1,\dots,q\}$ of the model can be found by a coordinate descent procedure that fixes $\{g_{j'}; j'\neq j\}\cup\{h_k, k=1,\dots,q\}$ and obtains g_j by Equation (5) (and that fixes $\{g_j, j=1,\dots,p\}\cup\{h_{k'}; k'\neq k\}$ and obtains h_k by Equation (8)), and then iterates through all j and k until convergence. This step (*Step 1*) amounts to fitting a SAM (Ravikumar *et al.*, 2009) subject to the constraint (2). Second (*Step 2*), for a set of fixed component functions $\{g_j, j=1,\dots,p\}\cup\{h_k, k=1,\dots,q\}$, the jth single-index coefficient function $\beta_j\in\Theta$ can be optimized by solving, for each $j\in\{1,\dots,p\}$ separately:

$$\underset{\beta_j\in\Theta}{\text{minimize}}\ E\left\{R_j-g_j(\langle X_j,\beta_j\rangle,A)\right\}^2\ (j=1,\dots,p), \tag{11}$$

where the jth partial residual R_j is defined in (7). These two steps can be iterated until convergence to obtain a population solution $\{g_j^*,\beta_j^*,h_k^*\}$ on the left-hand side of (4).

To obtain a sample version of the population solution, we can insert sample estimates into the population algorithm, as in standard backfitting in estimating generalized additive models (Hastie and Tibshirani, 1999), which we describe in the next subsection.

3.2 | Sample version of the population algorithm

To simplify the exposition, we describe only the optimization of $g_j(\langle X_j, \beta_j\rangle, A)$ $(j = 1, \dots, p)$ associated with the functional covariates X_j $(j = 1, \dots, p)$. The components $h_k(Z_k, A)$ $(k = 1, \dots, q)$ associated with the scalar covariates Z_k $(k = 1, \dots, q)$ in (4) are optimized in the same way, except that we do not need to perform *Step 2* of the alternating optimization procedure; i.e., when optimizing $h_k(Z_k, A)$ $(k = 1, \dots, q)$, we only perform *Step 1*.

3.2.1 | Step 1

First, we consider a sample version of *Step 1* of the population algorithm. Suppose we are given a set of estimates $\{\widehat{\beta}_j, j = 1, \dots, p\}$ and the data version of the jth partial residual R_j in (7): $\widehat{R}_{ij} = Y_i - \sum_{j' \neq j} \widehat{g}_{j'}(\langle X_{ij'}, \widehat{\beta}_{j'}\rangle, A_i) - \sum_{k=1}^{q} \widehat{h}_k(Z_{ik}, A_i) (i = 1, \dots, n)$, where $\widehat{g}_{j'}$ represents a current estimate for $g_{j'}$ and $\widehat{h}_k$ that for h_k. For each j, we update the component function g_j in (5) in two steps: first, estimate the function P_j in (6); second, plug the estimate of P_j into $[1 - \frac{\lambda}{\|P_j\|}]_+$ in (5), to obtain the soft-thresholded estimate $\widehat{g}_j$.

Although any linear smoothers can be utilized to obtain estimators $\{\widehat{g}_j, j = 1, \dots, p\}$ (see Supporting Information Section A.3), we will focus on regression spline-type estimators, which are simple and computationally efficient to implement. For each j and $\beta_j = \widehat{\beta}_j$, we will represent the component function $g_j \in \mathcal{H}_j^{(\widehat{\beta}_j)}$ on the right-hand side of (4) as

$$g_j(\langle X_j, \widehat{\beta}_j\rangle, a) = \boldsymbol{\Psi}_j(\langle X_j, \widehat{\beta}_j\rangle)^{\top} \boldsymbol{\theta}_{j,a} \; (a = 1, \dots, L) \tag{12}$$

for some prespecified d_j-dimensional basis $\boldsymbol{\Psi}_j(\cdot)$ (e.g., cubic B-spline basis with $d_j - 4$ interior knots, evenly placed over the range (scaled to, say, [0,1]) of the observed values of $\langle X_j, \widehat{\beta}_j\rangle$) and a set of unknown treatment a-specific basis coefficients $\{\boldsymbol{\theta}_{j,a} \in \mathbb{R}^{d_j}\}_{a\in\{1,\dots,L\}}$. Based on representation (12) of $g_j \in \mathcal{H}_j^{(\widehat{\beta}_j)}$ for fixed $\widehat{\beta}_j$, the constraint $E[g_j(\langle X_j, \beta_j\rangle, A) | X_j] = 0$ in (4) on g_j, for fixed $\beta_j = \widehat{\beta}_j$, can be simplified to $E[\boldsymbol{\theta}_{j,A}] = \sum_{a=1}^{L} \pi_a \boldsymbol{\theta}_{j,a} = \mathbf{0}$. If we fix $\beta_j = \widehat{\beta}_j$, the constraint in (4) on the function g_j can then be succinctly written in matrix form:

$$\boldsymbol{\pi}^{(j)} \boldsymbol{\theta}_j = \mathbf{0}, \tag{13}$$

where $\boldsymbol{\theta}_j := (\boldsymbol{\theta}_{j,1}^{\top}, \boldsymbol{\theta}_{j,2}^{\top}, \dots, \boldsymbol{\theta}_{j,L}^{\top})^{\top} \in \mathbb{R}^{d_j L}$ is the vectorized version of the basis coefficients $\{\boldsymbol{\theta}_{j,a}\}_{a\in\{1,\dots,L\}}$, and the $d_j \times d_j L$ matrix $\boldsymbol{\pi}^{(j)} := (\pi_1 \mathbf{I}_{d_j}; \pi_2 \mathbf{I}_{d_j}; \dots; \pi_L \mathbf{I}_{d_j})$ where $\mathbf{I}_{d_j}$ is the $d_j \times d_j$ identity matrix.

The details provided in Supporting Information Section A.4 (where constraint (13) is incorporated in the estimation) yield an estimate of the treatment a-specific function $g_j(\cdot, a) (a = 1, \dots, L)$ that appears in model (1):

$$\widehat{g}_j(\cdot, a) = \boldsymbol{\Psi}_j(\cdot)^{\top} \widehat{\boldsymbol{\theta}}_{j,a} \; (a = 1, \dots, L) \; (j = 1, \dots, p) \tag{14}$$

estimated within the class of functions (12), for a given tuning parameter $\lambda \geq 0$, resulting in the estimates of the component functions $\{\widehat{g}_j, j = 1, \dots, p\} \cup \{\widehat{h}_k, k = 1, \dots, q\}$; this completes Step 1 of the alternating optimization procedure.

3.2.2 | Step 2

We now consider a sample version of *Step 2* of the population algorithm that optimizes the coefficient functions $\{\beta_j, j = 1, \dots, p\}$ on the right-hand side of (4), for a fixed set of the component function estimates $\{\widehat{g}_j, j = 1, \dots, p\} \cup \{\widehat{h}_k, k = 1, \dots, q\}$ provided by *Step 1*. As an empirical approximation to (11), we consider

$$\underset{\beta_j \in \Theta}{\text{minimize}} \sum_{i=1}^{n} \left\{\widehat{R}_{ij} - \widehat{g}_j(\langle X_{ij}, \beta_j\rangle, A_i)\right\}^2 \; (j = 1, \dots, p), \tag{15}$$

where $\widehat{R}_{ij}$ is given from the previous *Step 1* at convergence. For this alternating step, solving (15) for β_j can be approximately achieved based on the first-order Taylor series approximation of the term $\widehat{g}_j(\langle X_{ij}, \beta_j\rangle, A_i)$ at the current estimate of β_j, which we denote as $\widehat{\beta}_j^{(c)} \in \Theta$:

$$\begin{aligned} & \sum_{i=1}^{n} \left\{\widehat{R}_{ij} - \widehat{g}_j\big(\langle X_{ij}, \beta_j\rangle, A_i\big)\right\}^2 \\ & \approx \sum_{i=1}^{n} \Big\{\widehat{R}_{ij} - \widehat{g}_j\big(\big\langle X_{ij}, \widehat{\beta}_j^{(c)}\big\rangle, A_i\big) \\ & \quad - \dot{\widehat{g}}_j\big(\big\langle X_{ij}, \widehat{\beta}_j^{(c)}\big\rangle, A_i\big)\big\langle X_{ij}, \beta_j - \widehat{\beta}_j^{(c)}\big\rangle\Big\}^2 \\ & = \sum_{i=1}^{n} \left\{\widehat{R}_{ij}^{*} - \big\langle X_{ij}^{*}, \beta_j\big\rangle\right\}^2, \end{aligned} \tag{16}$$

where the "modified" residuals $\widehat{R}^*_{ij}$ and the "modified" covariates X^*_{ij} are defined as

$$\widehat{R}^*_{ij} = \widehat{R}_{ij} - \widehat{g}_j(\langle X_{ij}, \widehat{\beta}^{(c)}_j \rangle, A_i) + \dot{\widehat{g}}_j(\langle X_{ij}, \widehat{\beta}^{(c)}_j \rangle, A_i)\langle X_{ij}, \widehat{\beta}^{(c)}_j \rangle \quad (i = 1, \dots, n),$$

$$X^*_{ij} = \dot{\widehat{g}}_j(\langle X_{ij}, \widehat{\beta}^{(c)}_j \rangle, A_i)\, X_{ij} \quad (i = 1, \dots, n), \tag{17}$$

in which each $\dot{\widehat{g}}_j(\cdot, a)$ denotes the first derivative of $\widehat{g}_j(\cdot, a)$ in (14) given from *Step 1*. We can perform a functional linear regression (e.g., Cardot *et al.*, 2003) with scalar response $\widehat{R}^*_{ij}$ and (functional) covariate X^*_{ij} to minimize the right-hand side of (16) over $\beta_j \in \Theta$. Specifically, we represent the smooth coefficient function β_j in (16) by a prespecified and normalized m_j-dimensional B-spline basis $B_j(s) = (b_{j1}(s), \dots, b_{jm_j}(s))^\top \in \mathbb{R}^{m_j}$, where m_j depends only on the sample size n (Fan *et al.*, 2015):

$$\beta_j(s) = \sum_{r=1}^{m_j} b_{jr}(s)\gamma_{jr} \; s \in [0, 1], \tag{18}$$

with an unknown basis coefficient vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jm_j})^\top \in \mathbb{R}^{m_j}$. Suppose the functional covariate $X_{ij}(s)$ $(i = 1, \dots, n)$ is discretized at points $\{s_l : 0 = s_0 < s_1 < s_2 < \cdots < s_{r_j} = 1\}$, with the distance between two adjacent discretization points denoted as Δ_l. Based on the approximation $\langle X_{ij}, \widehat{\beta}^{(c)}_j \rangle \approx \sum_{l=1}^{r_j} \Delta_l X_{ij}(s_l)\widehat{\beta}^{(c)}_j(s_l)$, we approximate $\widehat{R}^*_{ij}$ and X^*_{ij} in (17). Let $\boldsymbol{\Delta X}^*_j$ be the $n \times r_j$ matrix whose ith $(i = 1, \dots, n)$ row is the length-r_j vector $(\Delta_1 X^*_{ij}(s_1), \dots, \Delta_{r_j} X^*_{ij}(s_{r_j}))^\top$, corresponding to the ith subject's $X^*_{ij}(s)$ evaluated at the discretization points $\{s_1, \dots, s_{r_j}\}$ where each evaluation is multiplied by the corresponding Δ_l. Let $\boldsymbol{B}_j$ be the $r_j \times m_j$ matrix whose lth $(l = 1, \dots, r_j)$ row is the length-m_j vector $(b_{j1}(s_l), b_{j2}(s_l), \dots, b_{jm_j}(s_l))^\top$, corresponding to the vector of basis in (18) evaluated at s_l. Given $\beta_j(s)$ in (18) discretized at $\{s_1, \dots, s_{r_j}\}$, we can represent the right-hand side of (16) as

$$\|\widehat{\boldsymbol{R}}^*_j - \boldsymbol{U}^*_j \boldsymbol{\gamma}_j\|^2, \tag{19}$$

where $\widehat{\boldsymbol{R}}^*_j := (\widehat{R}^*_{1j}, \dots, \widehat{R}^*_{nj})^\top \in \mathbb{R}^n$ and $\boldsymbol{U}^*_j := \boldsymbol{\Delta X}^*_j \boldsymbol{B}_j \in \mathbb{R}^n \times \mathbb{R}^{m_j}$. Minimizing (19) over $\boldsymbol{\gamma}_j \in \mathbb{R}^{m_j}$ for each j separately $(j = 1, \dots, p)$ provides estimates $\{\widehat{\beta}_j, j = 1, \dots, p\}$ of the coefficient functions under (18). Here, the minimizer $\widehat{\boldsymbol{\gamma}}_j$ for (19) is scaled to $\|\widehat{\boldsymbol{\gamma}}_j\| = 1$, so that the resulting $\widehat{\beta}_j(s) = \sum_{r=1}^{m_j} b_{jr}(s)\widehat{\gamma}_{jr}$ $(s \in [0, 1])$ satisfies the identifiability constraint $\widehat{\beta}_j \in \Theta$. This completes *Step 2* of the alternating optimization procedure.

3.2.3 | Initialization and convergence criterion

At the initial iteration, we need some estimates $\{\widehat{\beta}_j, j = 1, \dots, p\}$ of the single-index coefficient functions to initialize the single indices $\{u_j = \langle \widehat{\beta}_j, X_j \rangle, j = 1, \dots, p\}$, in order to perform *Step 1* (i.e., the coordinate-descent procedure) of the estimation procedure described in Section 3.2.1. At the initial iteration, we take $\widehat{\beta}_j(s) = 1$ $(s \in [0, 1])$, that is, we take $u_j = \int_0^1 X_j(s)ds$ $(j = 1, \dots, p)$, which corresponds to the common practice of taking a naïve scalar summary of each functional covariate. The proposed algorithm alternating between *Steps 1* and *2* terminates when the estimates $\{\widehat{\beta}_j, j = 1, \dots, p\}$ converge. To be specific, the algorithm terminates when $\max_{j=1,\dots,p, r=1,\dots,m_j} \|(\widehat{\gamma}_{jr} - \widehat{\gamma}^{(c)}_{jr})/\widehat{\gamma}_{jr}\|$ is less than a prespecified convergence tolerance; here, $\widehat{\gamma}^{(c)}_{jr}$ represents the current estimate for γ_{jr} in (18) at the beginning of *Step 1*, and $\widehat{\gamma}_{jr}$ is the updated estimate at the end of *Step 2*. The proposed computational procedure is summarized as Algorithm 1 in (with discussion on computational time and convergence provided in Sections A.7 and A.9) Supporting Information Section A.6. The sparsity tuning parameter $\lambda \geq 0$ can be chosen to minimize an estimate of the expected squared error of the models over a dense grid of λ's, estimated, for example, by a 10-fold cross-validation.

4 | SIMULATION STUDY

4.1 | ITR estimation performance

In this section, we assess the optimal ITR estimation performance of the proposed method based on simulations. We generate n independent copies of p functional-valued covariates $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ $(i = 1, \dots, n)$, where we use a 4D Fourier basis, $\boldsymbol{\Phi}(s) = (\sqrt{2}\sin(2\pi s), \sqrt{2}\cos(2\pi s), \sqrt{2}\sin(4\pi s), \sqrt{2}\cos(4\pi s))^\top \in \mathbb{R}^4$ $(s \in [0, 1])$, and random coefficients $\widetilde{\boldsymbol{x}}_{ij} \in \mathbb{R}^4$, each independently following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_4)$, to form the functions $X_{ij}(s) = \boldsymbol{\Phi}(s)^\top \widetilde{\boldsymbol{x}}_{ij}(s \in [0, 1]) \qquad (i = 1, \dots, n; j = 1, \dots, p)$. Then these covariates are evaluated at 50 equally spaced points $\{s_l\}_{l=1}^{50}$ between 0 and 1. We also generate n independent copies of q scalar covariates $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iq})^\top \in \mathbb{R}^q$ $(i = 1, \dots, n)$, based on the multivariate normal distribution with each component having mean 0 and variance 1, with correlations between the components

$\mathrm{corr}(Z_{ij}, Z_{ik}) = 0.5^{|j-k|}$. We generate the outcomes Y_i $(i = 1, \dots, n)$ from

$$Y_i = \epsilon_i + \delta\left\{\sum_{j=1}^{8} \sin(\langle \eta_j, X_{ij}\rangle) + \sum_{k=1}^{8} \sin(Z_{ik})\right\}$$

$$+ 4(A_i - 1.5)\,[\sin(\langle \beta_1, X_{i1}\rangle) - \sin(\langle \beta_2, X_{i2}\rangle) + \cos(Z_{i1})$$

$$- \cos(Z_{i2}) + \xi\{\cos(\langle X_{i1}, X_{i2}\rangle) + \sin(Z_{i1}Z_{i2})\}], \quad (20)$$

where the treatments $A_i \in \{1, 2\}$ are generated with equal probability, independently of $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\epsilon_i \sim \mathcal{N}(0, 0.5^2)$. In (20), there are only four "signal" covariates (X_{i1}, X_{i2}, Z_{i1}, and Z_{i2}) influencing the effect of A_i on Y_i (i.e., four treatment effect-modifiers). The other $p + q - 4$ covariates are "noise" covariates not critical in optimizing ITRs. We set $p = q = 20$, therefore we consider a total of 40 pretreatment covariates in this example. In (20), we set the single-index coefficient functions, β_1 and β_2, to be $\beta_1(s) = \boldsymbol{\Phi}(s)^\top(0.5, 0.5, 0.5, 0.5)$ and $\beta_2(s) = \boldsymbol{\Phi}(s)^\top(0.5, -0.5, 0.5, -0.5)$, respectively (see Figure 2). We set the coefficient functions η_j $(j = 1, \dots, 8)$ associated with the X_j "main" effect to be: $\eta_j(s) = \boldsymbol{\Phi}(s)^\top\tilde{\boldsymbol{\eta}}_j$, with each $\tilde{\boldsymbol{\eta}}_j \in \mathbb{R}^4$ $(j = 1, \dots, 8)$ following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_4)$ and then rescaled to a unit L^2 norm $\|\tilde{\boldsymbol{\eta}}_j\| = 1$. The data model (20) is indexed by a pair (δ, ξ). The parameter $\delta \in \{1, 2\}$ controls the contribution of the $(\boldsymbol{X}, \boldsymbol{Z})$ main effect component, $\delta\{\sum_{j=1}^{8} \sin(\langle \eta_j, X_{ij}\rangle) + \sum_{k=1}^{8} \sin(Z_{ik})\}$, to the variance of Y, in which $\delta = 1$ corresponds to a relatively *moderate* $(\boldsymbol{X}, \boldsymbol{Z})$ main effect (about four times greater than the interaction effect when $\xi = 0$) and $\delta = 2$ corresponds to a relatively *large* $(\boldsymbol{X}, \boldsymbol{Z})$ main effect (about 16 times greater than the interaction effect when $\xi = 0$). In (20), the parameter $\xi \in \{0, 1\}$ determines whether the A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect component has an additive structure ($\xi = 0$) of the specified form (1) or whether it deviates from an additive structure ($\xi = 1$). In the case of $\xi = 0$, the proposed CFAM (1) is correctly specified, whereas, for the case of $\xi = 1$, it is misspecified. For each simulation replication, we consider the following four approaches to estimating $\mathcal{D}^{opt}$:

1. The proposed approach (4) estimated via Algorithm 1 in Supporting Information Section A.6, where the dimensions of the cubic B-spline basis for $\{g_j, h_k, \beta_j\}$ are set at $d_j = d_k = m_j = 4 + (2n)^{1/5}$ (rounded to the closest integer) following the conditions of Corollary 3 of Fan *et al.* (2015). The sparsity tuning parameter $\lambda > 0$ is chosen to minimize 10-fold cross-validated prediction error of the fitted models.
2. The functional linear regression approach of Ciarleglio *et al.* (2018),

$$\underset{\beta_j \in L^2[0,1], \alpha_k \in \mathbb{R}}{\text{minimize}}\; E\left\{Y - \sum_{j=1}^{p}\langle \beta_j, X_j\rangle(A - 1.5)\right.$$

$$\left. - \sum_{k=1}^{q} \alpha_k Z_k (A - 1.5)\right\}^2$$

$$+ \lambda\left\{\sum_{j=1}^{p}(\|\beta_j\| + \rho_j \boldsymbol{\gamma}_j^\top \boldsymbol{S}_j \boldsymbol{\gamma}_j) + \sum_{k=1}^{q} |\alpha_k|\right\},$$

which tends to yield a sparse set $\{\beta_j\} \cup \{\alpha_k\}$, estimated based on representation (18) for β_j with $m_j = 10$ and an associated $m_j \times m_j$ P-spline penalty matrix $(\boldsymbol{S}_j)$ that ensures appropriate smoothness. The tuning parameters $\lambda > 0$ and $\rho = \rho_j > 0$ $(j = 1, \dots, p)$ are chosen to minimize a 10-fold cross-validated prediction error (Ciarleglio *et al.*, 2018). As the component functions $\{g_j, h_k\}$ associated with Ciarleglio *et al.* (2018) are restricted to be linear (i.e., we restrict them to $g_j(\langle \beta_j, X_j\rangle, A) = \langle \beta_j, X_j\rangle(A - 1.5)$ and $h_k(Z_k, A) = \alpha_k Z_k (A - 1.5)$) corresponding to a special case of CFAM, we call the model of Ciarleglio *et al.* (2018), a CFAM with *linear* component functions (CFAM-lin) for the notational simplicity.
3. The OWL (Zhao *et al.*, 2012) method based on a linear kernel (OWL-lin), implemented in the R-package `DTRlearn`. As there is no currently available OWL method that deals with functional covariates, we compute a scalar summary of each functional covariate, that is, $\bar{X}_j = \int_0^1 X_j(s)ds \in \mathbb{R}$, and use $\bar{X}_j$ along with the other scalar covariates Z_k as inputs to the augmented (residualized) OWL procedure. To improve its efficiency, we employ the augmented OWL approach of Liu *et al.* (2018), which amounts to prefitting a linear model for μ in (1) via Lasso (Tibshirani, 1996) and residualizing the response Y. The tuning parameter κ in Zhao *et al.* (2012) is chosen from the grid of $(0.25, 0.5, 1, 2, 4)$ (the default setting of `DTRlearn`) based on a 10-fold cross-validation.
4. The same approach as in approach 3 but based on a Gaussian radial basis function kernel (OWL-Gauss) in place of a linear kernel . The inverse bandwidth parameter σ_n^2 in Zhao *et al.* (2012) is chosen from the grid of $(0.01, 0.02, 0.04, \dots, 0.64, 1.28)$ and κ is chosen from the grid of $(0.25, 0.5, 1, 2, 4)$, based on a 10-fold cross-validation.

Throughout the paper, for CFAM and CFAM-lin, we fit the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect on Y based on the

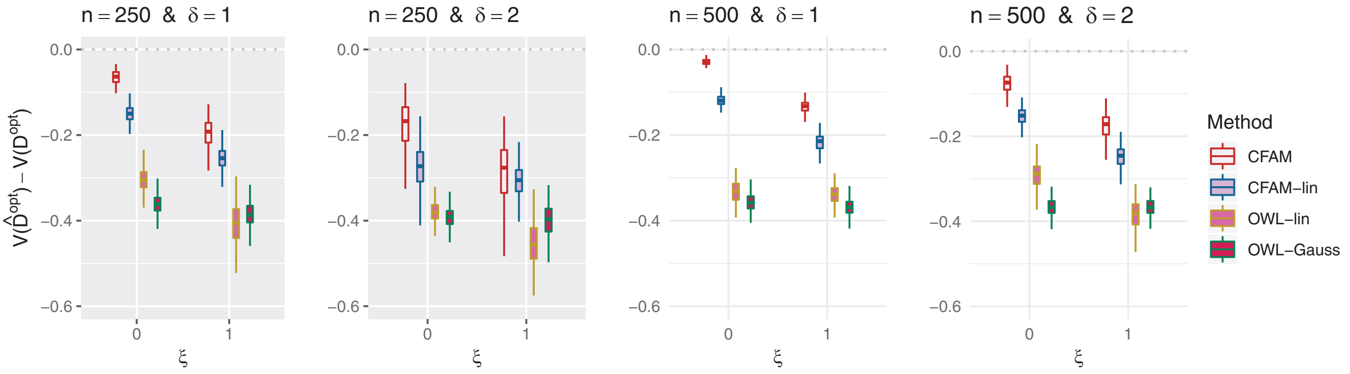

FIGURE 1 Boxplots obtained from 200 Monte Carlo simulations comparing 4 approaches to estimating $\mathcal{D}^{opt}$, given each scenario indexed by $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and $n \in \{250, 500\}$. The dotted horizontal line represents the optimal value corresponding to $\mathcal{D}^{opt}$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

(misspecified) linear model with the naïve scalar averages of X_j, that is, $\bar{X}_j$, along with Z_k, fitted via Lasso with 10-fold cross-validation for the sparsity parameter and utilize the "residualized" response $Y - \widehat{\mu}(\boldsymbol{X}, \boldsymbol{Z})$. For each simulation run, we estimate $\mathcal{D}^{opt}$ from each of the above four methods based on a training set (of size $n \in \{250, 500\}$), and to evaluate these methods, we compute the value $V(\widehat{\mathcal{D}}^{opt}) = E[E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = \widehat{\mathcal{D}}^{opt}(\boldsymbol{X}, \boldsymbol{Z})]]$ of each estimate $\widehat{\mathcal{D}}^{opt}$, based on a Monte Carlo approximation using a separate random sample of size 10^3. As we know the true data-generating model in simulation studies, the optimal $\mathcal{D}^{opt}$ can be determined for each simulation run. Given each estimate $\widehat{\mathcal{D}}^{opt}$ of $\mathcal{D}^{opt}$, we report $V(\widehat{\mathcal{D}}^{opt}) - V(\mathcal{D}^{opt})$, as the performance measure of $\widehat{\mathcal{D}}^{opt}$. A larger (i.e., less negative) value of the measure indicates better performance.

In Figure 1, we present boxplots, obtained from 200 simulation runs, of the normalized values $V(\widehat{\mathcal{D}}^{opt})$ (normalized by the optimal values $V(\mathcal{D}^{opt})$) of the decision rules $\widehat{\mathcal{D}}^{opt}$ based on the four approaches, for each combination of $n \in \{250, 500\}$, $\xi \in \{0, 1\}$ (corresponding to *correctly specified* or *misspecified* CFAM interaction models, respectively) and $\delta \in \{1, 2\}$ (corresponding to *moderate* or *large* main effects, respectively). The results in Figure 1 indicate that the proposed method (CFAM) outperforms all other approaches. In particular, if the sample size is relatively large ($n = 500$), for a correctly specified CFAM ($\xi = 0$), the method gives a close-to-optimal performance with respect to $\mathcal{D}^{opt}$. With nonlinearities present in the underlying model (20), CFAM-lin is outperformed by CFAM that utilizes the flexible component functions $g_j(\cdot, a)$ and $h_k(\cdot, a)$, although it substantially outperforms the OWL-based approaches.

In Supporting Information Section A.12, we have also considered a set of similar experiments under a *linear* A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect, in which CFAM-lin outperforms CFAM, but by a relatively small amount, whereas if the underlying model deviates from the exact linear structure and $n = 500$, CFAM tends to outperform CFAM-lin. This suggests that, in the absence of prior knowledge about the form of the interaction effect, the more flexible CFAM that accommodates nonlinear treatment effect-modifications can be set as a default approach over CFAM-lin for optimizing ITRs. The estimated values of the OWL methods using linear and Gaussian kernels, respectively, are similar to each other; however, because the current OWL methods do not directly deal with the functional pretreatment covariates, both are outperformed by CFAM, even when CFAM is incorrectly specified (i.e., when $\xi = 1$). When the $(\boldsymbol{X}, \boldsymbol{Z})$ "main" effect dominates the A-by-$(\boldsymbol{X}, \boldsymbol{Z})$ interaction effect (i.e., when $\delta = 2$), although the increased magnitude of this nuisance effect dampens the performance of all approaches to estimating $\mathcal{D}^{opt}$, the proposed approach outperforms all other methods.

In Table S.1 of Supporting Information Section A.10, we additionally illustrate the estimation performance for model parameters β_1 and β_2 (and g_1, g_2, h_1, and h_2) when $\xi = 0$ (i.e., when CFAM is correctly specified) with varying $\delta \in \{1, 2\}$ and $n \in \{250, 500, 1000\}$, with respect to the root squared error $\text{RSE}(\beta_j) = \sqrt{\int (\widehat{\beta}_j(s) - \beta_j(s))^2 ds}$ $(j = 1, 2)$ (similarly for $\text{RSE}(g_j)$ and $\text{RSE}(h_k)$). In Figure 2, we display typical CFAM estimates $\widehat{\beta}_j$ of β_j from 10 random samples, for each sample size n (for the case of $\delta = 1$). With sample size increasing, the estimators $\widehat{\beta}_j$ get close to the true coefficient functions β_j. (Similar results are provided for g_j and h_k in Supporting Information Table S.1.)

4.2 | Treatment effect-modifier variable selection performance

In this subsection, we will report simulation results for the treatment effect-modifier selection among $\{X_j, j = 1, \ldots, p\} \cup \{Z_k, k = 1, \ldots, q\}$. The complexity of the

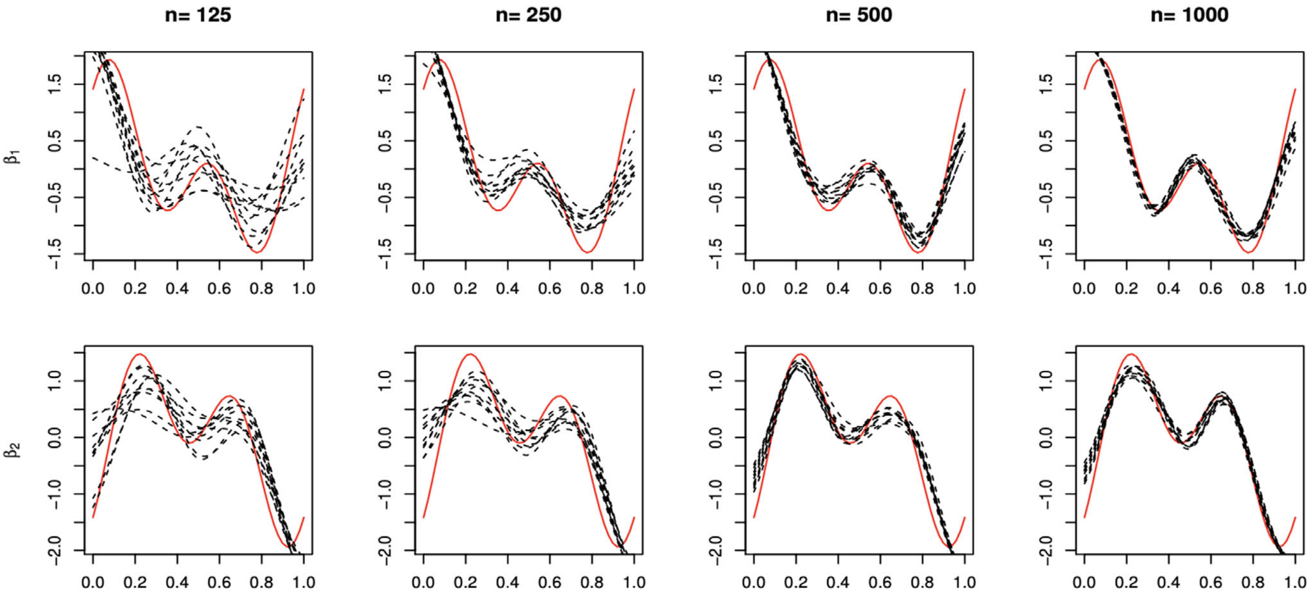

FIGURE 2 An illustration of typical 10 CFAM sample estimates $\widehat{\beta}_j(s)$ (black dashed curves) for the parameters $\beta_j(s)$ (the red solid curves), for $j = 1$ and 2 in the top and bottom panels, respectively, with a varying training sample size $n \in \{125, 250, 500, 1000\}$ for the case of $\delta = 1$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

$(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction terms of CFAM (1) can be summarized in terms of the size (cardinality) of the index set of $\{g_j, j = 1, \ldots, p\} \cup \{h_k, k = 1, \ldots, q\}$ that are not identically zero, each of which can be either correctly or incorrectly estimated to be equal to zero. As in Section 4.1, we generate 200 datasets based on (20), with varying $\xi \in \{0, 1\}$, $\delta \in \{1, 2\}$ and sample size $n \in \{50, 100, 200, \ldots, 700, 800\}$ and $p = q = 20$, that is, we consider a total of $p + q = 40$ potential treatment effect-modifiers, among which there are only four "true" treatment effect-modifiers.

Figure 3 summarizes the results of the treatment effect-modifier covariate selection performance with respect to the true/false positive rates (the top/bottom panels, respectively), comparing the proposed CFAM and the CFAM-lin of Ciarleglio *et al.* (2018). The results are reported as the averages (and ±1 standard deviations) across the 200 simulated datasets, for each simulation scenario. Figure 3 illustrates that the proportion of *correct* selection out of the four *true* treatment effect-modifiers (i.e., the "true positive" rate; the top gray panels) of CFAM (the red solid curves) tends to 1, as n increases from $n = 50$ to $n = 800$, whereas the proportion of *incorrect* selection (i.e., the "false positive" rate; the bottom white panels) of the 36 *irrelevant* "noise" covariates tends to 0; these proportions tend to either 1 or 0 quickly for *moderate* main effect ($\delta = 1$) scenarios compared to *large* main effect ($\delta = 2$) scenarios. On the other hand, the proportion of correct selections for CFAM-lin (the blue dotted curves), even with a large n, tends to be only around 0.55, due to the stringent linear model assumption on the form of the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interaction effect. Figure 3 appears in color in the electronic version of this article, and any mention of color refers to that version.

5 | APPLICATION

In this section, we illustrate the utility of CFAM for optimizing ITRs, using data from an RCT (Trivedi *et al.*, 2016) comparing an antidepressant and placebo for treating major depressive disorder. The study collected various scalar and functional patient characteristics at baseline, including EEG data. Study participants were randomized to either placebo ($A = 1$) or an antidepressant (sertraline) ($A = 2$). Subjects were monitored for 8 weeks after initiation of treatment. The primary endpoint of interest was the Hamilton Rating Scale for Depression (HRSD) score at week 8. The outcome Y was taken to be the improvement in symptoms severity from baseline to week 8 taken as the difference: week 0 HRSD score −week 8 HRSD score (larger values of the outcome Y are considered desirable).

There were $n = 180$ subjects. We considered $p = 19$ pretreatment functional covariates consisting of the current source density (CSD) amplitude spectrum curves over the alpha frequency range (observed while the participants' eyes were open), measured from a subset of EEG channels from a total of 72 EEG electrodes, which gives a fairly good spatial coverage of the scalp. The locations for these 19 electrodes are indicated in the top panel of Figure 4. The alpha frequency band (8–12 Hz) considered as a potential biomarker of antidepressant response (e.g., Wade and Iosifescu, 2016) was scaled to $[0, 1]$, hence each of the

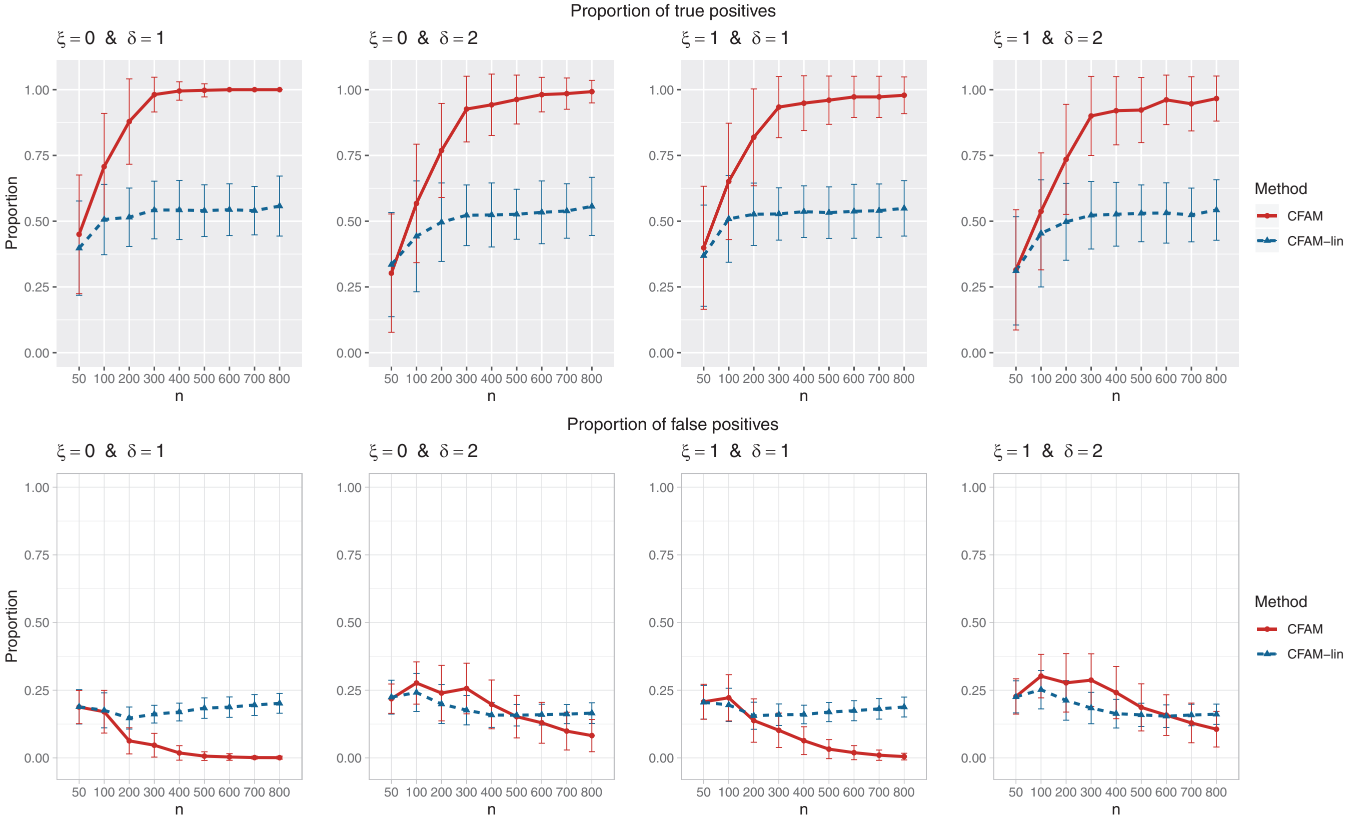

FIGURE 3 The proportion of the relevant covariates (i.e., the treatment effect-modifiers) correctly selected (the "true positives"; the top gray panels), and the "noise" covariates incorrectly selected (the "false positives"; the bottom white panels), respectively (and ±1 standard deviation), with a varying sample size $n \in \{50, 100, 200, \dots, 800\}$, for each combination of $\xi \in \{0, 1\}$ and $\delta \in \{1, 2\}$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

functional covariates $X = (X_1(s), \dots, X_{19}(s))$ was defined on the interval $[0, 1]$. We also considered $q = 5$ baseline scalar covariates consisting of the week 0 HRSD score (Z_1), sex (Z_2), age at evaluation (Z_3), word fluency (Z_4), and Flanker accuracy (Z_5) cognitive test scores, which were identified as predictors of differential treatment response in a previous study (Park *et al.*, 2020). In this dataset, 49% of the subjects were randomized to the sertraline ($A = 2$). The average outcomes Y for the sertraline and placebo groups were 7.41 and 6.29, respectively. The means (and standard deviations) of Z_1, Z_3, Z_4, and Z_5 were 18.59 (4.44), 37.7 (13.57), 38 (11.42), and 0.19 (0.11), respectively, and 67% of the subjects were female.

The proposed CFAM approach (4) selected two functional covariates: "C3" (X_4) and "P3" (X_5) (the selected electrodes are indicated by the dashed circles in the top panel of Figure 4), and a scalar covariate: "Flanker accuracy test" (Z_5). In the left two columns of Figure 4, we display the treatment arm-specific CSD curves for the selected two functional covariates, $X_4(s)$ and $X_5(s)$ (measured before treatment), from the 180 subjects. In the third column of Figure 4, we display the associated coefficient function estimates, $\hat{\beta}_4(s)$ and $\hat{\beta}_5(s)$. The coefficient function $\beta_j(s)$, discretized at $\{s_1, s_2, \dots, s_{r_j}\}$, is represented by $\boldsymbol{B}_j \hat{\boldsymbol{\gamma}}_j \in \mathbb{R}^{r_j}$ in (19), whose variance estimate, $\boldsymbol{B}_j \hat{\boldsymbol{V}} \boldsymbol{B}_j^\top$, is used to construct a 95% point-wise normal approximated confidence band in Figure 4, where $\hat{\boldsymbol{V}}$ represents the covariance of the length-m_j vector $\hat{\boldsymbol{\gamma}}_j$, i.e., the minimizer of (19) scaled to unit norm (see Supporting Information Section A.8 for discussion on this confidence band).

In this example, the coefficient functions $\hat{\beta}_j(s)$ summarizing the $X_j(s)$ lead to data-driven *indices* $u_j = \langle \hat{\beta}_j, X_j \rangle \in \mathbb{R}$ that are linked to differential treatment response via two estimated nonzero component functions, $\hat{g}_j(u_j, A)$ ($j = 4, 5$; i.e., for X_4 and X_5), displayed in the left two panels on the top row of Figure 5. Roughly put, the placebo ($A = 1$) effect tends to slightly increase with the index u_j ($j = 4, 5$), whereas the sertraline ($A = 2$) effect slightly decreases with the index. In the third column of Figure 4, β_4 puts a bulk of its negative weight on lower frequencies (8–9 Hz), meaning that patients whose CSD values are small in those frequency regions would have large values of $\langle \beta_4, X_4 \rangle$, over the values which the placebo effects are predicted to be relatively strong, in comparison to the sertraline effects. In the third panel on the top row of Figure 5, the

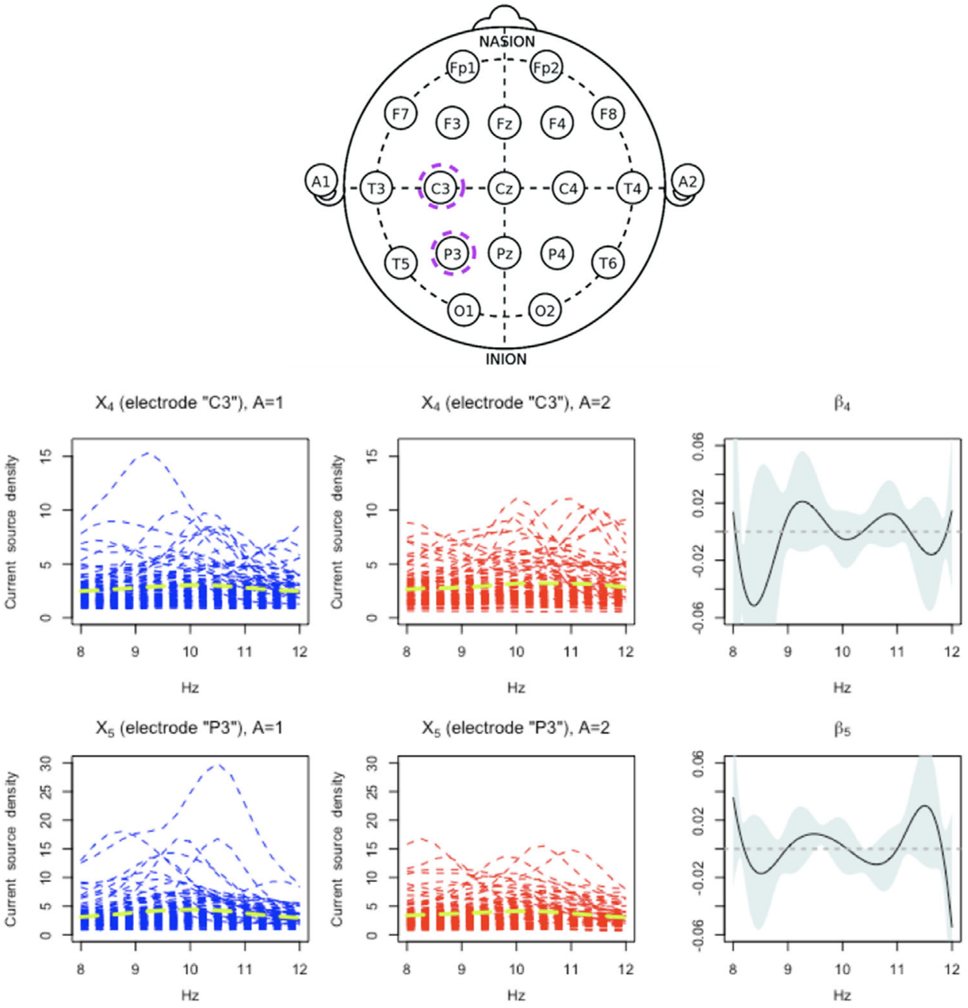

FIGURE 4 *Top row*: The locations for the 19 electrode channels ("A1" and "A2" were not used). Those two electrodes ("C3" and "P3") highlighted in dashed violet circles are the selected electrodes from the proposed approach. *Bottom rows*: first two columns—observed current source density (CSD) curves from the selected electrodes X_4 (C3) and X_5 (P3) (each electrode corresponds to each row), over the Alpha band (8–12 Hz), for the placebo $A = 1$ arm (in the first column) and the active drug $A = 2$ arm (in the second column), measured before treatment. The arm-specific mean functions are overlaid as dashed green curves. Third column: the estimated single-index coefficient functions (β_4 and β_5) for the selected channels X_4 and X_5 (with the associated 95% confidence bands, conditioning on the jth partial residual and $\hat{g}_j$). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

estimated component function $\widehat{h}_5(Z_5, A)$ associated with the selected scalar covariate Z_5 is displayed, where the placebo ($A = 1$) effect tends to increase with Z_5.

In model (1), without loss of generality, the treatment-specific intercept was suppressed. Let τ_a ($a = 1, 2$) represent the treatment a-specific intercept, so that $\tau_2 - \tau_1$ represents the marginal treatment effect (comparing $a = 2$ with $a = 1$). For the most common situation of binary treatment conditions (i.e., $L = 2$), let us define a 1D index $f(\boldsymbol{X}, \boldsymbol{Z}) := \sum_{j=1}^{p} g_j(\langle X_j, \beta_j\rangle, a = 2) + \sum_{k=1}^{q} h_k(Z_k, a = 2)$ that parameterizes the treatment effect "contrast" according to (1), $E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = 2] - E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = 1] = \tau_2 - \tau_1 + f(\boldsymbol{X}, \boldsymbol{Z})\frac{1}{\pi_1}$, as a linear function (see Supporting Information Section A.15 for this parametrization). The index $f(\boldsymbol{X}, \boldsymbol{Z})$ provides a continuous gradient of the benefit from one treatment to another. The bottom row of Figure 5 displays the observed treatment-specific outcome $Y^{(a)}$ versus this combined index $f(\boldsymbol{X}, \boldsymbol{Z})$. In those panels, the treatment benefit (comparing $a = 2$ vs. $a = 1$) corresponds to the contrast between the solid and dotted lines, and the benefit increases monotonically with this combined index: an index greater than the crossing point (group "Benefit") indicates that the patient is expected to benefit from Sertraline, and an index smaller than the crossing point (group "No Benefit") indicates that the patient is expected to benefit from placebo. Given this monotone relationship between the treatment benefit and the continuous index $f(\boldsymbol{X}, \boldsymbol{Z})$, a more refined decision using three or more groups (e.g., benefit level groups "B1," "B2," and "B3,"

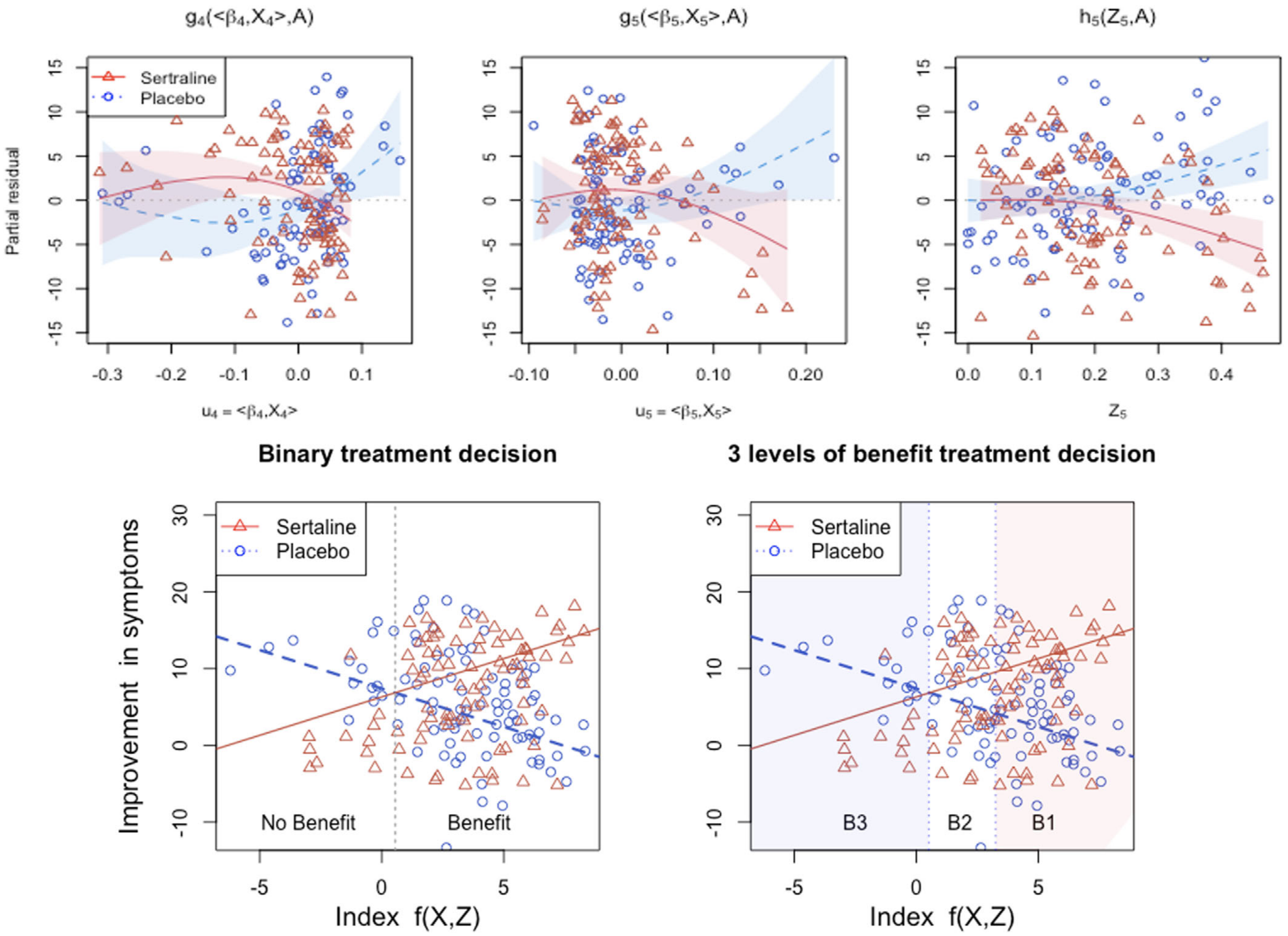

FIGURE 5 *Top row*: The scatter plots of the (jth; $j = 4, 5$) (and kth; $k = 5$) partial residual versus the estimated functional indices $u_4 = \langle X_4, \beta_4 \rangle$ and $u_5 = \langle X_5, \beta_5 \rangle$ (and Z_5), respectively, for the placebo $A = 1$ (blue circles) and sertraline $A = 2$ (red triangles) treated individuals. The estimated nonzero treatment-specific component functions $g_4(u_4, A)$, $g_5(u_5, A)$ and $h_5(Z_5, A)$ are overlaid separately for the $A = 1$ (placebo; blue dotted curve) condition and the $A = 2$ (sertraline; red solid curve) condition (with the associated 95% confidence bands, given the partial residuals and $\widehat{\beta}_j$). *Bottom row*: The scatter plots of the observed treatment-specific response Y versus the "index" $f(\boldsymbol{X}, \boldsymbol{Z}) = g_4(\langle X_4, \beta_4 \rangle, 2) + g_5(\langle X_5, \beta_5 \rangle, 2) + h_5(Z_5, 2)$, with possible two-group recommendation grouping (in the left panel; the cut point was the crossing point = 0.56 between the two treatment-specific expected responses) and possible three-group recommendation grouping (in the right panel; the cut point for B2 and B1 was 3.15, which gives the difference (= 7.42) in the two treatment-specific expected responses larger than the expected marginal response under sertraline (= 7.41); note that this cut-point choice was just for an illustration of the idea of the benefit stratification). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

specified in the right panel, where the associated cut points were determined based on the treatment-specific expected responses) than a simple binary recommendation can be also considered when recommending treatments to patients, which can help triage of patients according to the expected benefit by the treatment.

To evaluate the ITR performance of the four different approaches described in Section 4, we randomly split the data into a training set and a testing set (of size $\widetilde{n}$) with a ratio of 5:1, replicated 500 times, each time estimating an ITR $\widehat{\mathcal{D}}^{opt}$ based on the training set, and its "value" $V(\widehat{\mathcal{D}}^{opt}) = E[E[Y|\boldsymbol{X}, \boldsymbol{Z}, A = \widehat{\mathcal{D}}^{opt}(\boldsymbol{X}, \boldsymbol{Z})]]$, by an inverse probability weighted estimator (Murphy, 2005) $\widehat{V}(\widehat{\mathcal{D}}^{opt}) = \sum_{i=1}^{\widetilde{n}} Y_i I_{(A_i = \widehat{\mathcal{D}}^{opt}(\boldsymbol{X}_i, \boldsymbol{Z}_i))} / \sum_{i=1}^{\widetilde{n}} I_{(A_i = \widehat{\mathcal{D}}^{opt}(\boldsymbol{X}_i, \boldsymbol{Z}_i))}$, computed based on with the testing set (of size $\widetilde{n}$). For comparison, we also include two naïve rules: treating all patients with placebo (All PBO) and treating all patients with the active drug (All DRUG), each regardless of the individual patient's characteristics $(\boldsymbol{X}, \boldsymbol{Z})$. The resulting boxplots obtained from the 500 random splits are illustrated in Figure 6.

The results in Figure 6 demonstrate that CFAM and CFAM-lin perform at a similar level, showing a clear advantage over the both OWL-lin and OWL-Gauss, suggesting that regression utilizing the functional nature of

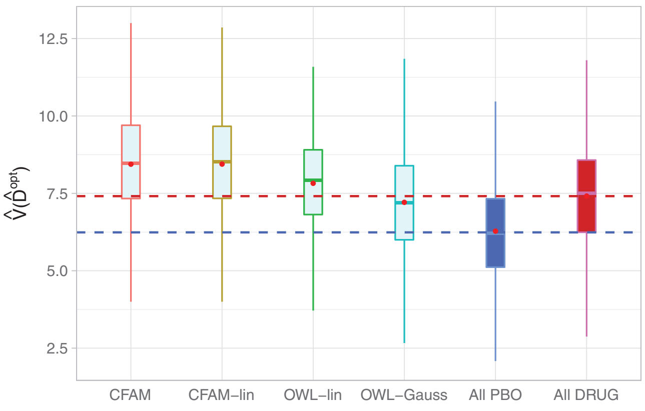

FIGURE 6 Boxplots of the estimated values of the treatment rules $\widehat{D}^{opt}$ estimated from six approaches, obtained from 500 randomly split testing sets. Higher values are preferred. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

the EEG measurements, that targets the treatment-by-functional covariates interaction effects is well suited in this example. Specifically, in Figure 6, the superiority of CFAM (or CFAM-lin) over the policy of treating everyone with the drug (All DRUG) was of similar magnitude of the superiority of All DRUG over All PBOs. This suggests that accounting for patient characteristics can help treatment decisions. The estimated model parameters (β_j, g_j, h_k) for CFAM-lin are provided in Supporting Information Section A.13 (see also for the proportion of agreement of the recommended treatments from the four ITR approaches considered). In this example, the estimated nonlinear treatment effect-modification is rather modest, as can be observed from the first row of Figure 5. As a result, the performances of CFAM and CFAM-lin are comparable to each other. However, as demonstrated in Section 4, the more flexible CFAM can be employed as a default approach over CFAM-lin, allowing for potentially important nonlinearities when modeling treatment effect-modification.

6 | DISCUSSION

We have developed a functional additive regression approach specifically focused on extracting possibly nonlinear pertinent interaction effects between treatment and multiple functional/scalar covariates, which is of paramount importance in developing effective ITRs for precision medicine. This is accomplished by imposing appropriate structural constraints, performing treatment effect-modifier selection and extracting 1D functional indices. The estimation approach utilizes an efficient coordinate-descent for the component functions and a functional linear model estimation procedure for the coefficient functions. The proposed functional regression for ITRs extends existing methods by incorporating possibly nonlinear treatment-by-functional covariates interactions. Encouraged by our simulation results and the application, future work will investigate the asymptotic properties of the method related to variable selection and estimation consistency. The main theoretical challenge that the working model associated with the proposed estimation criterion is misspecified (see Supporting Information Section A.18 for discussion). Another important direction is the development of a Bayesian framework for the model accounting for the posterior uncertainty in β_j g_j, h_k and the unmodeled noise variance in predicting the individualized treatment benefit using the index $f(\boldsymbol{X}, \boldsymbol{Z})$, and making inference on the $(\boldsymbol{X}, \boldsymbol{Z})$-by-$A$ interactions.

The proposed method is not directly applicable to the functional covariates irregularly or sparsely sampled or observed with nonnegligible error, and an initial step to denoise and re-construct the underlying curves is required, as is done in Goldsmith *et al.* (2011) using the principal component decomposition of the observed functions (see Supporting Information Section A.14 for discussion) in such cases.

The proposed approach to optimizing ITRs can also accommodate data from observational studies, under condition $Y^{(a)} \perp A$ given additive measurable functions of $\langle X_j, \beta_j \rangle (j = 1, \dots, p)$ and $\boldsymbol{Z}$ (see Supporting Information Section A.16 for discussion). For more general cases, with treatment propensity information available, we can reparametrize model (1) and accommodate the treatment propensities in the estimation (see Supporting Information Section A.17).

ACKNOWLEDGMENT

This work was supported by National Institute of Health (NIH) grant 5 R01 MH099003.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are available from the corresponding author upon reasonable request.

ORCID

Hyung Park https://orcid.org/0000-0002-8994-9583

REFERENCES

Cardot, H., Ferraty, F. and Sarda, P. (2003) Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571–592.

Ciarleglio, A., Petkova, E., Ogden, R.T. and Tarpey, T. (2015) Treatment decisions based on scalar and functional baseline covariatesecisions based on scalar and functional baseline covariates. *Biometrics*, 71, 884–894.

Ciarleglio, A., Petkova, E., Ogden, R.T. and Tarpey, T. (2018) Constructing treatment decision rules based on scalar and functional predictors when moderators of treatment effect are unknown. *Journal of Royal Statistical Society: Series C*, 67, 1331–1356.

Ciarleglio, A., Petkova, E., Tarpey, T. and Ogden, R.T. (2016) Flexible functional regression methods for estimating individualized treatment rules. *Stat*, 5, 185–199.

Fan, Y., Foutz, N., James, G. and Jank, W. (2014) Functional response additive model with online virtual stock markets. *The Annals of Applied Statistics*, 8, 2435–2460.

Fan, Y., James, G.M. and Radchanko, P. (2015) Functional additive regression. *The Annals of Statistics*, 43, 2296–2325.

Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B. and Reich, D. (2011) Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20, 830–851.

Hastie, T. and Tibshirani, R. (1999) *Generalized Additive Models*. London: Chapman & Hall.

Jeng, X., Lu, W. and Peng, H. (2018) High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12, 2074–2089.

Kang, C., Janes, H. and Huang, Y. (2014) Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70, 696–707.

Laber, E.B. and Staicu, A. (2018) Functional feature construction for individualized treatment regimes. *Journal of the American Statistical Association*, 113, 1219–1227.

Laber, E.B. and Zhao, Y. (2015) Tree-based methods for individualized treatment regimes. *Biometrika*, 102, 501–514.

Liu, Y., Wang, Y., Kosorok, M.R., Zhao, Y. and Zeng, D. (2018) Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, 37, 3776–3788.

Lu, W., Zhang, H. and Zeng, D. (2011) Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22, 493–504.

McKeague, I. and Qian, M. (2014) Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24, 1461–1485.

McLean, M., Hooker, G., Staicu, A., Scheipel, F. and Ruppert, D. (2014) Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23, 249–269.

Morris, J.S. (2015) Functional regression. *Annual Review of Statistics and Its application*, 2, 321–359.

Murphy, S.A. (2005) A generalization error for q-learning. *Journal of Machine Learning*, 6, 1073–1097.

Park, H., Petkova, E., Tarpey, T. and Ogden, R.T. (2020) A sparse additive model for treatment effect-modifier selection. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxaa032

Qian, M. and Murphy, S.A. (2011) Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39, 1180–1210.

Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis*. New York: Springer.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *Journal of Royal Statistical Society: Series B*, 71, 1009–1030.

Shi, C., Song, R. and Lu, W. (2016) Robust learning for optimal treatment decision with np-dimensionality. *Electronic Journal of Statistics*, 10, 2894–2921.

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E.B. and Yuan, M. (2015) On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4, 59–68.

Tian, L., Alizadeh, A., Gentles, A. and Tibshrani, R. (2014) A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109, 1517–1532.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

Trivedi, M., McGrath, P., Fava, M., Parsey, R., Kurian, B., Phillips, M. et al. (2016) Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *Journal of Psychiatric Research*, 78, 11–23.

Wade, E. and Iosifescu, D. (2016) Using electroencephalography for treatment guidance in major depressive disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1, 411–422.

Zhang, B., Tsiatis, A.A., Davidian, M., Zhang, M. and Laber, E. (2012) Estimating optimal treatment regimes from classification perspective. *Stat*, 1, 103–114.

Zhao, Y., Laber, E., Ning, Y., Saha, S. and Sands, B. (2019) Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, 20, 1–23.

Zhao, Y., Zheng, D., Laber, E.B. and Kosorok, M.R. (2015) New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110, 583–598.

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, 1106–1118.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2, 3, 4, 5 and 6, and a zip file containing R-codes for running the examples presented in this paper are available with this paper at the Biometrics website on Wiley Online Library. The R-package `famTEMsel` (Functional Additive Models for Treatment Effect-Modifier Selection) for the methods proposed in this paper is publicly available on `GitHub` (`syhyunpark/famTEMsel`).

How to cite this article: Park, H., Petkova, E., Tarpey, T., Ogden, R.T. (2023) Functional additive models for optimizing individualized treatment rules. *Biometrics*, 79, 113–126. https://doi.org/10.1111/biom.13586