*Biostatistics* (2017) **18**, 1, *pp.* 105–118
doi:10.1093/biostatistics/kxw035
Advance Access publication on July 27, 2016

# Generated effect modifiers (GEM's) in randomized clinical trials

EVA PETKOVA*

*Department of Child and Adolescent Psychiatry, New York University, 1 Park Ave., New York, NY 10016, USA and Nathan Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA*
eva.petkova@nyumc.org

THADDEUS TARPEY

*Department of Mathematics and Statistics, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA and Department of Child and Adolescent Psychiatry, New York University, 1 Park Ave., New York, NY 10016, USA*

ZHE SU

*Department of Child and Adolescent Psychiatry, New York University, 1 Park Ave., New York, NY 10016, USA*

R. TODD OGDEN

*Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th St., New York, NY 10032, USA*

## Summary

In a randomized clinical trial (RCT), it is often of interest not only to estimate the effect of various treatments on the outcome, but also to determine whether any patient characteristic has a different relationship with the outcome, depending on treatment. In regression models for the outcome, if there is a non-zero interaction between treatment and a predictor, that predictor is called an "effect modifier". Identification of such effect modifiers is crucial as we move towards precision medicine, that is, optimizing individual treatment assignment based on patient measurements assessed when presenting for treatment. In most settings, there will be several baseline predictor variables that could potentially modify the treatment effects. This article proposes optimal methods of constructing a composite variable (defined as a linear combination of pre-treatment patient characteristics) in order to generate an effect modifier in an RCT setting. Several criteria are considered for generating effect modifiers and their performance is studied via simulations. An example from a RCT is provided for illustration.

*Keywords*: Biosignature; Moderator; Precision medicine; Treatment decision; Value.

## 1. Introduction

Precision medicine focuses on making treatment decisions for an individual patient based on the patient's measures (e.g., clinical and biological features). The idea underlies a long history of attempts to identify

*To whom correspondence should be addressed.

© The Author 2016. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

characteristics that exhibit interaction with treatment assignment in a regression model for the outcome of interest. Such baseline characteristics, called "treatment effect modifiers", indicate that the outcome under one treatment compared to another treatment depends on these characteristics. Measures with such interactions can aid decisions about which treatment to prescribe (Gail and Simon, 1985; Wellek, 1997; Song and Pepe, 2004; Wang and Ware, 2013).

Interest in precision medicine is growing rapidly, both in clinical research and in statistical methodology. An important component of precision medicine is the notion of an "optimal treatment regime", first formalized by Murphy (2003) and Robins (2004). Given a vector $\boldsymbol{x}$ of baseline covariates, a treatment decision can be based a decision function $d(\boldsymbol{x})$ that maps $\boldsymbol{x}$ to a treatment indicator, say $A$. Treatment decisions can be compared using the "value" of a decision $d$, denoted $V(d)$. The value of a decision is the expected value of an outcome variable $y$ (with respect to the joint distribution of $(y, \boldsymbol{x})$) when all patients are treated according to a decision function $d$ and Qian and Murphy (2011) show that the value can be expressed as

$$V(d) = E[E(y|\boldsymbol{x}, A = d(\boldsymbol{x}))], \tag{1.1}$$

where $(y|\boldsymbol{x}, a)$ is the outcome of a patient given treatment $A = a$ with covariates $\boldsymbol{x}$. Here we consider outcome variables $y$ that are continuous, where higher values of $y$ are preferred, as per convention. Determining optimal individual treatment decisions using data from RCTs is a topic that is the subject of active research (see Robins *and others*, 2008; Zhao *and others*, 2012; Zhang *and others*, 2012b; Kang *and others*, 2014; Zhao *and others*, 2015, among others). The "optimal treatment decision" is the one that, when applied to the target population, has the largest value.

It has long been recognized that features that are important for predicting outcome might not be necessarily be useful for making treatment decisions (e.g., Wellek, 1997; Song and Pepe, 2004). Much recent research has focused on identification of individual baseline covariates related to the treatment effect (i.e., variables that exhibit interactions with the treatment indicator in predicting treatment outcome) in contrast to being important in the baseline model. A major challenge in precision medicine is that most baseline measures typically have small moderating effects and individually contribute little to informed treatment decisions. Unconstrained regression models with $p$ predictors (plus treatment and predictor-by-treatment interactions) become unwieldy, unstable and difficult to interpret when $p$ is moderate to large. Various strategies have been proposed to deal with the problem (see Qian and Murphy, 2011; Gunter *and others*, 2011; Lu *and others*, 2011, among others). Extensions of the methodology that allow functional data objects to be incorporated as baseline features have also been developed (e.g., McKeague and Qian, 2014; Ciarleglio *and others*, 2015).

A parsimonious alternative to these previous methods that has received little attention is to use a simple model with only a single "composite" predictor. Herein, a methodology is developed for combining several baseline predictors into a single treatment effect modifier in the context of the classic linear model, which we call a *generated effect modifier* (GEM). Given a vector of $p$ predictors $\boldsymbol{x} = (x_1, \ldots, x_p)'$, we consider linear combinations of the predictors $z = \boldsymbol{\alpha}'\boldsymbol{x}$ for $\boldsymbol{\alpha} \in \Re^p$ as potential GEMs. The idea of combining covariates was proposed by Tukey (1991, 1993) for balancing and increasing the precision of the estimates of treatment effect in RCTs. A closely related approach was proposed by Tian and Tibshirani (2011) who developed a method of constructing binary "markers" from continuous variables (via cut-off values) and forming an index to detect treatment–marker interactions. Emura *and others* (2012) introduced a compound covariate approach for predicting survival time in the case when there are too many covariates, for example, gene expression data. In contrast to this work, we propose to combine covariates with the goal of obtaining a single moderating variable, a GEM, that would aid in deciding which treatment is appropriate for any particular patient. Although the GEM model is more restrictive than an unconstrained model, it provides a parsimonious single index approach for making individualized treatment decisions.

Alternative approaches to optimal treatment decision estimation have been proposed that fall in the realm of machine learning and can often be framed in the context of classification problems (Zhang *and others*, 2012a). Examples are the outcome weighted learning (OWL) (e.g., Zhao *and others*, 2015; Song *and others*, 2015) based on support vector machines, tree-based classification (e.g. Laber and Zhao, 2015), and the Kang *and others* (2014) method based on adaptive boosting. Although these approaches can be appealing options in many settings, we base our general approach on the linear model as it is most frequently utilized in practice and lends itself very well to interpretability. This paper fulfills the practical need of providing a simple treatment effect modifier methodology in the classic linear model setting for making precision medicine decisions. Also, the GEM approach provides the benefit of a visual presentation that is familiar to clinicians.

In efficacy studies, after the primary analysis of treatment efficacy has been performed, the usual practice is to seek individual effect modifiers (single patient baseline characteristics) with the ultimate goal of informing treatment decisions. When no single variable has a strong modifying effect, the GEM is an appealing and novel approach for secondary exploratory analysis to find a strong treatment effect modifier. The GEM can be particularly useful for analysis of studies designed to discover biosignatures for treatment response.

## 2. Criteria for choosing a GEM

Here we introduce several optimality criteria for defining a GEM $z = \boldsymbol{\alpha}'\boldsymbol{x}$. For notational simplicity, we present the model in terms of the centered (at zero within treatment group) outcomes $\boldsymbol{y}_k$ and predictors matrix $\boldsymbol{X}_k$. The unrestricted linear model for the $K$ treatment groups is

$$E(\boldsymbol{y}_k|\boldsymbol{X}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k, \text{ with } \boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})', \quad \text{for } k = 1, \ldots, K, \tag{2.2}$$

while the GEM model under consideration can be parameterized as

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_K \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{X}_K \end{pmatrix} \boldsymbol{\gamma} \otimes \boldsymbol{\alpha} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_K \end{pmatrix}, \tag{2.3}$$

where $\otimes$ denotes the Kronecker product. The vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_K)'$ is the vector of the scaling coefficients for the GEM model (2.3). Because the predictors might be measured on different scales, a natural constraint that ensures identifiability is that the GEM $\boldsymbol{\alpha}'\boldsymbol{x}$ has a unit variance constraint

$$\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\alpha} = 1, \tag{2.4}$$

where $\boldsymbol{\Psi}_x$ denotes the predictor covariance matrix (assumed equal across treatment groups as in a RCT). An unrestricted multiple regression model for $K$ treatment groups (e.g. model (2.2)) with $p$ predictors and all interactions between treatment indicators and predictors, has $pK$ regression coefficients (not counting intercepts), whereas the restricted GEM model (2.3) is more parsimonious with only $K + p - 1$ parameters (constraint (2.4) reduces the number of free parameters in $\boldsymbol{\alpha}$ by one). Model (2.3) was considered by Follmann and Proschan (1999), but from a different perspective, where the vectors of regression coefficients $\boldsymbol{\beta}_k$ from (2.2) are all equal under the null hypothesis and (2.3) is the alternative hypothesis model. In addition to being more parsimonious and providing an intuitive interpretation with easy visualization, GEMs can also be used for making straightforward treatment decisions. When $K = 2$, for a new subject with covariates $\boldsymbol{x}^{new}$, the estimated treatment decision based on an unrestricted regression model is $d^{UnR}(\boldsymbol{x}^{new}) = I\{\hat{\beta}_{02} + \hat{\boldsymbol{\beta}}_2'\boldsymbol{x}^{new} > \hat{\beta}_{01} + \hat{\boldsymbol{\beta}}_1'\boldsymbol{x}^{new}\} + 1$, where $I$ is an indicator function and $\hat{\beta}_{0k}$ and $\hat{\boldsymbol{\beta}}_k, k = 1, 2$ are the least squares (LS) estimates of the regression coefficients of model (2.2) written for the uncentered outcomes and predictors. Under a GEM model, the treatment decision is $d^{GEM}(\boldsymbol{x}^{new}) = I\{\hat{\gamma}_{02} + \hat{\gamma}_2\hat{\boldsymbol{\alpha}}'\boldsymbol{x}^{new} > \hat{\gamma}_{01} + \hat{\gamma}_1\hat{\boldsymbol{\alpha}}'\boldsymbol{x}^{new}\} + 1$, where $\hat{\gamma}_{0k}$ and $\hat{\gamma}_k, k = 1, 2$ are the LS estimates of the scaling coefficients in model (2.3) for non-centered outcomes and predictors.

Since the GEM is defined as a linear combination of predictors, the GEM model lends itself most naturally to continuous predictors. In the results that follow, there is nothing that precludes the use of discrete predictors; only care must be taken in how discrete predictors are coded and how the corresponding GEM is to be interpreted. It is very common in clinical practice that categorical variables are actually discretized versions of continuous variables. If this is the case, we recommend that the original variable is used in the GEM instead of its discretized version.

There are several principled criteria one can use for choosing $\boldsymbol{\alpha}$ for optimizing the GEM. A natural choice obviously, in terms of moderator analysis, is to maximize the magnitude of interaction in the GEM model. Alternatively, $\boldsymbol{\alpha}$ can be choosen to provide the best fit to the data using a GEM model which is consistent with the classic goal in linear models of minimizing the error sum of squares. A third approach, also consistent with the linear model framework, is to determine $\boldsymbol{\alpha}$ that maximizes the statistical significance of the interaction effects via an $F$-test. Summarizing, we consider the following three criteria, which we refer to as the "numerator" (N), "denominator" (D) and "$F$-ratio" (F) criteria, respectively:

(N) **Maximizing the interaction effect:** Maximize the variability in the GEM scaling coefficients $\gamma_k$'s in (2.3), corresponding to maximizing the Numerator of an $F$-test for significance of interaction effects. When there are $K = 2$ treatment groups, this is the same as maximizing the squared difference between the scaling coefficients $\gamma_1$ and $\gamma_2$ in the GEM model.

(D) **Fidelity to the data:** Minimize the sum of squared residuals in the GEM model (2.3). This corresponds to the Denominator of an $F$-test for significance of interaction effects.

($F$) ***F*-ratio:** Combine the first two criteria and maximize the ratio of the variability of the GEM scaling coefficients relative to the sum of squared residuals for the GEM model. This criterion corresponds to choosing $\boldsymbol{\alpha}$ to maximize the $F$-test statistic when testing significance of interactions in the GEM model.

The method of LS is used to estimate the parameters of models (2.2) and (2.3). The common covariance matrix $\boldsymbol{\Psi}_x$ can be estimated by the pooled estimate of the predictor covariance matrix:

$$\hat{\boldsymbol{\Psi}}_x = \sum_{k=1}^{K} \boldsymbol{X}_k'\boldsymbol{X}_k/(N-K), \tag{2.5}$$

where $N = \sum_{k=1}^{K} n_k$, where $n_k$ is the sample size in group $k$. The following notation will be used: let $\boldsymbol{\Psi}_{xy_k}$ denote the vector of covariances between $\boldsymbol{x}$ and the $y_k$ and $\sigma_{y_k}^2$ denote the variance of $y_k$ in the $k$th group. Then the usual unconstrained vector of slope coefficients in the $k$th treatment group in terms of population parameters and the weighted average coefficient vector are respectively

$$\boldsymbol{\beta}_k = \boldsymbol{\Psi}_x^{-1}\boldsymbol{\Psi}_{xy_k} \quad \text{and} \quad \bar{\boldsymbol{\beta}} = \sum_{k=1}^{K} \pi_k \boldsymbol{\beta}_k. \tag{2.6}$$

With a randomized experiment, equal weights ($\pi_k = 1/K$) are used for $\bar{\boldsymbol{\beta}}$ and that is the convention used in this article (although more flexible choices for weights are also possible). The GEM scaling coefficients $\gamma_k$ in (2.3) can be expressed equivalently, using (2.4), as

$$\gamma_k = \frac{\mathrm{cov}(\boldsymbol{X}_k\boldsymbol{\alpha}, y_k)}{\mathrm{var}(\boldsymbol{X}_k\boldsymbol{\alpha})} = \frac{\boldsymbol{\alpha}'\boldsymbol{\Psi}_{xy_k}}{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\Psi}_x^{-1}\boldsymbol{\Psi}_{xy_k}}{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\alpha}} = \boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\beta}_k.$$

### 2.1. *The "numerator" criterion: maximizing the interaction effect*

This section derives the expression for $\boldsymbol{\alpha}$ in the GEM model that maximizes the variance of a discrete random variable taking values $\gamma_1, \ldots, \gamma_K$ with respective probabilities $\pi_1, \ldots, \pi_K$ (i.e., the variance of the GEM slopes) which is given by

$$\sum_{k=1}^{K} \pi_k \left( \frac{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})}{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\alpha}} \right)^2 = \frac{\boldsymbol{\alpha}'\boldsymbol{\Psi}_x \left[ \sum_{k=1}^{K} \pi_k(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})' \right] \boldsymbol{\Psi}_x\boldsymbol{\alpha}}{(\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{\alpha})^2}. \tag{2.7}$$

Denote the "between" group covariance matrix for the unconstrained slope coefficients as

$$\boldsymbol{B} = \sum_{k=1}^{K} \pi_k(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})'. \tag{2.8}$$

Using (2.4), we seek $\boldsymbol{\alpha}$ that maximizes $\boldsymbol{\alpha}'\boldsymbol{\Psi}_x\boldsymbol{B}\boldsymbol{\Psi}_x\boldsymbol{\alpha} = \boldsymbol{\alpha}'\boldsymbol{\Psi}_x^{1/2}\left[\boldsymbol{\Psi}_x^{1/2}\boldsymbol{B}\boldsymbol{\Psi}_x^{1/2}\right]\boldsymbol{\Psi}_x^{1/2}\boldsymbol{\alpha}$, where $\boldsymbol{\Psi}_x^{1/2}$ is the symmetric square-root of $\boldsymbol{\Psi}_x$. The solution is $\boldsymbol{\alpha}^N = \boldsymbol{\Psi}_x^{-1/2}\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the eigenvector of $\boldsymbol{\Psi}_x^{1/2}\boldsymbol{B}\boldsymbol{\Psi}_x^{1/2}$ that is associated with the largest eigenvalue. To obtain an estimator $\hat{\boldsymbol{\alpha}}^N$, we can apply the plug-in principal, use the pooled estimator $\boldsymbol{\Psi}_x$ from (2.5) and the usual unrestricted LS estimators $\hat{\boldsymbol{\beta}}_k$ in place of the $\boldsymbol{\beta}_k$'s. The GEM $\gamma_k$'s and intercepts can be estimated via LS.

In the case of $K = 2$ groups, $\boldsymbol{B} = \sum_{k=1}^{K} \pi_k(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}})' = \pi_1\pi_2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'$, is a rank one matrix with eigenvector proportional to $(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$, in which case

$$\boldsymbol{\alpha}^N = \frac{\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2}{\sqrt{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'\boldsymbol{\Psi}_x(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}}. \tag{2.9}$$

Section 1.1 of the supplementary material shows that for $K = 2$, in terms of population parameters, the treatment decision based on the unrestricted regression is equivalent to the treatment decision based on the numerator GEM model. Minor differences in the empirical decision rules from these two methods are due to differences in the LS estimates using the GEM predictor versus using the original predictors in the unrestricted model.

### 2.2. *The "denominator" criterion: minimizing the residual error*

This subsection gives the LS expression for $\boldsymbol{\alpha}$ that minimizes the sum of squared residuals in a GEM model, that is, that provides the best fitting GEM model. Under the assumption of normality, the LS estimator coincides with the maximum likelihood estimator in the GEM linear model.

The sum of squared residuals from a standard linear model using LS can be written as $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, where $\boldsymbol{H}$ is the hat matrix and $\boldsymbol{I}$ is an identity matrix. This sum of squared residuals (when divided by its associated degrees of freedom) is an estimate of the quantity $\sigma_y^2 - \boldsymbol{\Psi}_{xy}'\boldsymbol{\Psi}_x^{-1}\boldsymbol{\Psi}_{xy}$. In the GEM model (2.3) with $K$ treatment arms, the hat matrix in the $k$th group is $\boldsymbol{H}_k(\boldsymbol{\alpha}) = (\boldsymbol{X}_k\boldsymbol{\alpha})(\boldsymbol{\alpha}'\boldsymbol{X}_k'\boldsymbol{X}_k\boldsymbol{\alpha})^{-1}(\boldsymbol{X}_k\boldsymbol{\alpha})'$. Letting

$\boldsymbol{D} = \sum_{k=1}^{K} \pi_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k' = \boldsymbol{B} + \bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'$, Section 1.2 of the supplementary material available at *Biostatistics* online shows that the $\boldsymbol{\alpha}$ minimizing the "denominator" criterion is given by $\boldsymbol{\alpha}^D = \boldsymbol{\Psi}_x^{-1/2}\boldsymbol{e}_2$, where $\boldsymbol{e}_2$ is the leading eigenvector of $\boldsymbol{\Psi}_x^{1/2}\boldsymbol{D}\boldsymbol{\Psi}_x^{1/2}$. As before, $\boldsymbol{\alpha}^D$ can be estimated by plugging in the LS estimators for $\boldsymbol{\beta}_k$ in the expression for $\boldsymbol{D}$ and using the sample covariance matrix of the pooled predictors (2.5) to estimate $\boldsymbol{\Psi}_x$.

### 2.3. *The "F-criterion": maximizing the F-statistic*

This section determines $\boldsymbol{\alpha}$ that maximizes the strength of the statistical evidence for the interaction effect in the GEM model (2.3) via an $F$-test. With $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)'$, we can consider the general linear hypothesis of $H_0 : \boldsymbol{L}\boldsymbol{\gamma} = \boldsymbol{0}$. If $K = 2$ and $\boldsymbol{L} = (1/2, -1/2)$, the null hypothesis above states that the two groups have the same coefficients with respect to the GEM $Z = \boldsymbol{X}\boldsymbol{\alpha}$ (i.e., no interaction). Thus, the goal is to determine $\boldsymbol{\alpha}$ that maximizes the $F$-ratio for testing $H_0$. From the two previous subsections, the $F$-ratio is proportional to the ratio with (2.7) in the numerator and a denominator corresponding to the residual sum-of-squares. The value of $\boldsymbol{\alpha}$ satisfying the "$F$-ratio" criterion is $\boldsymbol{\alpha}^F = \boldsymbol{\Psi}_x^{-1/2}\boldsymbol{e}_3$, where $\boldsymbol{e}_3$ is the leading eigenvector of

$$\left[\left(\sum_{k=1}^{K} \pi_k \sigma_{y_k}^2 \boldsymbol{I}_p\right) - \boldsymbol{\Psi}_x^{1/2}\boldsymbol{D}\boldsymbol{\Psi}_x^{1/2}\right]^{-1} \boldsymbol{\Psi}_x^{1/2}\boldsymbol{B}\boldsymbol{\Psi}_x^{1/2}. \tag{2.10}$$

The derivation is in Section 1.3 of the supplementary material. Once again, $\boldsymbol{\alpha}^F$ can be estimated by plugging parameter estimates into (2.10) and extracting the leading eigenvector.

## 3. FITTING A GEM WHEN THE GEM MODEL IS MISSPECIFIED

The GEM model allows us to combine several predictors into a single linear combination that has good treatment effect moderator properties. Generally, we do not expect the GEM model to be the true data generating model and (based on the above expressions), the "true" $\boldsymbol{\alpha}$ for the three criteria would differ. Consider two cases with $K = 2$ groups and $p = 2$ predictors $(x_1, x_2)$ from a Gaussian distribution with means 0, variances 1 and 2, respectively, and a covariance 0.2:

$$\text{Case 1: } \boldsymbol{\beta}_1 = \begin{pmatrix} -0.4 \\ 2.0 \end{pmatrix}, \ \boldsymbol{\beta}_2 = \begin{pmatrix} -0.6 \\ 2.5 \end{pmatrix}; \qquad \text{Case 2: } \boldsymbol{\beta}_1 = \begin{pmatrix} 1.5 \\ 2.5 \end{pmatrix}, \ \boldsymbol{\beta}_2 = \begin{pmatrix} -2.5 \\ 2.5 \end{pmatrix}.$$

The deviation from a GEM model is measured by the angle $\theta$ between the coefficient vectors $\boldsymbol{\beta}_k$, $k = 1, 2,$: $\theta = \arccos\left(\frac{\boldsymbol{\beta}_1'\boldsymbol{\beta}_2}{\|\boldsymbol{\beta}_1\|\|\boldsymbol{\beta}_2\|}\right)$. In Case 1, $\theta = 0.012\pi$, and in Case 2, $\theta = 0.422\pi$, so Case 1 is very "close" to a GEM model ($\theta = 0$ or $\pi$), while Case 2 is almost as far away from GEM as possible ($\theta = 0.5\pi$). The "true" $\boldsymbol{\alpha}$'s are:

$$\text{Case 1: } \begin{array}{rcl} \boldsymbol{\alpha}^N & = & (0.283, -0.707)' \\ \boldsymbol{\alpha}^D & = & (0.160, -0.714)' \\ \boldsymbol{\alpha}^F & = & (0.160, -0.714)' \end{array} \qquad \text{Case 2: } \begin{array}{rcl} \boldsymbol{\alpha}^N & = & (1.000, 0.000)' \\ \boldsymbol{\alpha}^D & = & (0.143, -0.714)' \\ \boldsymbol{\alpha}^F & = & (1.000, 0.000)' \end{array}.$$

From (2.10), $\boldsymbol{\alpha}^F$ depends on the error variance; the results above are for a coefficient of determination $R^2 = 0.8$. As expected, the $\boldsymbol{\alpha}^F$ is closer to $\boldsymbol{\alpha}^D$ when the data is from a GEM model since the GEM regression fits the data well in this case, while when the model is far from a GEM model, $\boldsymbol{\alpha}^F$ is closer to $\boldsymbol{\alpha}^N$. This observation together with results from simulations suggest the use of the $F$-criterion in practice.

## 4. Permutation testing for the interaction in a GEM model

The GEM model estimation seeks to determine a linear combination of predictors that maximizes the evidence of an interaction effect using one of the three criteria described above. If there are no interaction effects between predictors and treatment indicators, then the GEM procedure would tend to generate anti-conservative $p$-values. A straightforward remedy to this problem is to fit the GEM model on many data sets with permuted treatment labels. A permutation $p$-value for testing an interaction effect can then be calculated as

$$\text{Permutation } p \text{ value } = \{\text{Proportion of ``permuted'' } p \text{ values} < \text{original } p \text{ value}\}.$$

Theoretical details for using permutation tests for interaction effects in the presence of possible main effects have been investigated previously in the literature (e.g., Wang *and others*, 2015, p. 2046).

## 5. Simulation studies

An appealing feature of the GEM model is its utility for making individual treatment decisions, especially when $p$ is large. In this subsection we investigate the value (1.1) of treatment decisions based on the three GEM criteria for both GEM and non-GEM generating models. Data sets were simulated under a variety of parameter settings. We varied the coefficient of determination $R^2$ to be small (0.2), medium (0.5), and large (0.8). Another useful measure in the "effect size" (ES) of a moderator. For a regression model $y = \gamma_0 + \gamma_1 A + \gamma_2 z + \gamma_3(Az) + \epsilon$, with $\text{var}(z) = 1$ and a treatment indicator $A$ $(= \pm\frac{1}{2})$, the ES (Kraemer, 2013) of $Z$ as an effect modifier is the proportion of the outcome variance (after removing the variance due to treatment) that is explained by the different relationships between $y$ and $z$ in the two treatment groups, that is,

$$ES = \sqrt{\frac{(\gamma_3/2)^2}{(\gamma_2 + \gamma_3/2)^2 + (\gamma_3/2)^2 + \sigma^2}}, \tag{5.11}$$

where $\sigma^2$ is the error variance (assuming equal error variances for all values of $A$). The simulations are similar for the GEM and non-GEM settings, except that the GEM models are characterized with respect to the effect size of $\boldsymbol{\alpha}'\boldsymbol{x}$ (using ES = 0.1 and 0.3), while the non-GEM cases are characterized with respect to the angle between the vectors of regression coefficients as described in Section 3; we use a small $(\theta = 0.15\pi)$ and a large $(\theta = 0.48\pi)$ deviation from GEM.

The sample sizes per treatment group considered are $n : n_1 = n_2 = 100,\ 300,$ and 1000, mimicking typical situations in medical research. For each sample size, the number of predictors used were $p = 10$ and $p = 200$ (except when $p > n$, namely $n = 100$ and $p = 200$). The predictors are generated from $p$-variate normal distributions with mean zero and variances equal to 1, and small pairwise correlations (from $-0.2$ to 0.2) randomly selected, while ensuring a positive definite correlation matrix. For each $p$, $\boldsymbol{\beta}_1 = (1, \frac{1}{2}, \ldots, \frac{1}{p})$. Under GEM, $\boldsymbol{\beta}_2$ is computed to satisfy the respective $R^2$ and $ES$. Under non-GEM, $\boldsymbol{\beta}_2$ is obtained by adding random noise to the $p$ coefficients in $\boldsymbol{\beta}_1$ and computing the angle $\theta$ between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. More details about the values of $\boldsymbol{\beta}_2$ are given in Section 3.2 of the supplementary material. For each combination of $p$ and the $\boldsymbol{\beta}_k$'s $(k = 1, 2)$, a large sample $(N = 10^6)$ is generated with known outcome values under both treatments and it is used to evaluate the "true" optimal population average outcome $V^+$, which is the highest achievable value of any decision.

For each simulation configuration $(n, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2,\ R^2$ and ES), $B = 1000$ data sets are generated and estimates of $\boldsymbol{\alpha}^N, \boldsymbol{\alpha}^D,$ and $\boldsymbol{\alpha}^F$ are computed, as well as $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ coefficients of the unrestricted regression model (2.2). These estimates are used to define treatment decisions as described in Section 2. These

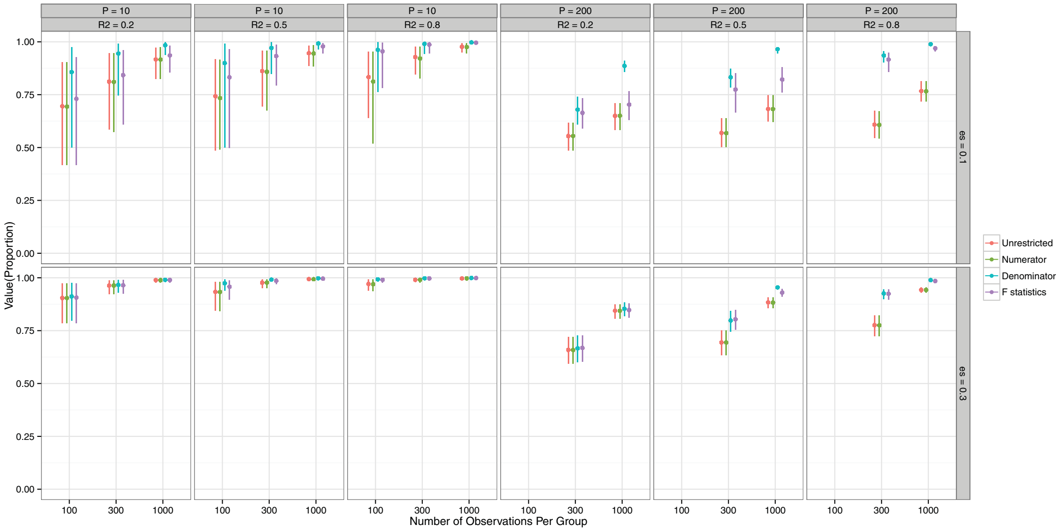

Fig. 1. GEM data generation model. Mean and 95% Monte Carlo (MC) confidence intervals (based on the $B = 1000$ MC runs) of the values $V$ of the decisions, as a proportion $(V - V^-)/(V^+ - V^-)$, for $p = 10$ (left half of panels) and 200 (right half of panels), and for ES = 0.1 (top half of panels) and ES=0.3 (bottom half of panels). The three panels per ($p$, ES) combination correspond to $R^2 = 0.2$ on the left, $R^2 = 0.5$ in the middle and $R^2 = 0.8$ on the right. The method based on the unrestricted regression and the three GEM approaches are denoted as: (i) unrestricted—red color, most left; (ii) numerator criteria—green, second from left; (iii) denominator criterion—blue, third from left; (iv) $F$ criterion—purple, most right. The "Number of observations" on the bottom horizontal axis is the sample size per group.

decisions are applied to the $N = 10^6$ cases in the large data set to obtain the estimated values $V$ of the respective decisions $V(d(\boldsymbol{x}; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2))$, $V(d^N(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}^N, \hat{\boldsymbol{\gamma}}^N))$, $V(d^D(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}^D, \hat{\boldsymbol{\gamma}}^D))$, and $V(d^F(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}^F, \hat{\boldsymbol{\gamma}}^F))$. For the sake of comparison, these values are expressed as a proportion of the "true" optimal average outcome $V^+$, and also taking into account the the worst average outcome $V^-$, which is obtained by choosing the worst (lower) outcome for each subject in the large data set. For example, the values of the treatment decision based on the "numerator" GEM approach are reported as $\left[V\left(d^N(\boldsymbol{x}; \hat{\boldsymbol{\alpha}}^N, \hat{\boldsymbol{\gamma}}^N)\right) - V^-\right]/(V^+ - V^-)$.

Figure 1 shows the means and the 95% Monte Carlo (MC) confidence intervals for the value of the decisions in the case of data generation from GEM models. A general observation is that for small ES of the GEM, the estimated decisions produce values that are about 10-20% lower than the "true" optimal value $V^+$ for $p = 10$ and still lower for $p = 200$. How much worse the estimated decisions are compared with the "true" optimal average population outcome depends on the sample size and $R^2$ (performance improves with increasing sample size and $R^2$). The "denominator" method is superior to the other two approaches, especially for larger $p$'s and smaller ES's, which is not be surprising since the denominator criterion is equivalent to the MLE objective when the error is normal and the true model is a GEM, as is the case here.

Figure 2 presents information similar to that on Figure 1, but here the data are generated under a general linear non-GEM model (2.2). It shows that even when the data is not generated from a GEM model, the criteria perform quite well for relatively small number of covariates $p$. For larger $p$, larger sample sizes and larger $R^2$ are needed to achieve good performance. The values of the decisions based on the denominator criterion are meaningfully inferior to the values of the decisions from the other methods as the deviation from GEM becomes large. The denominator's inferiority becomes more pronounced as

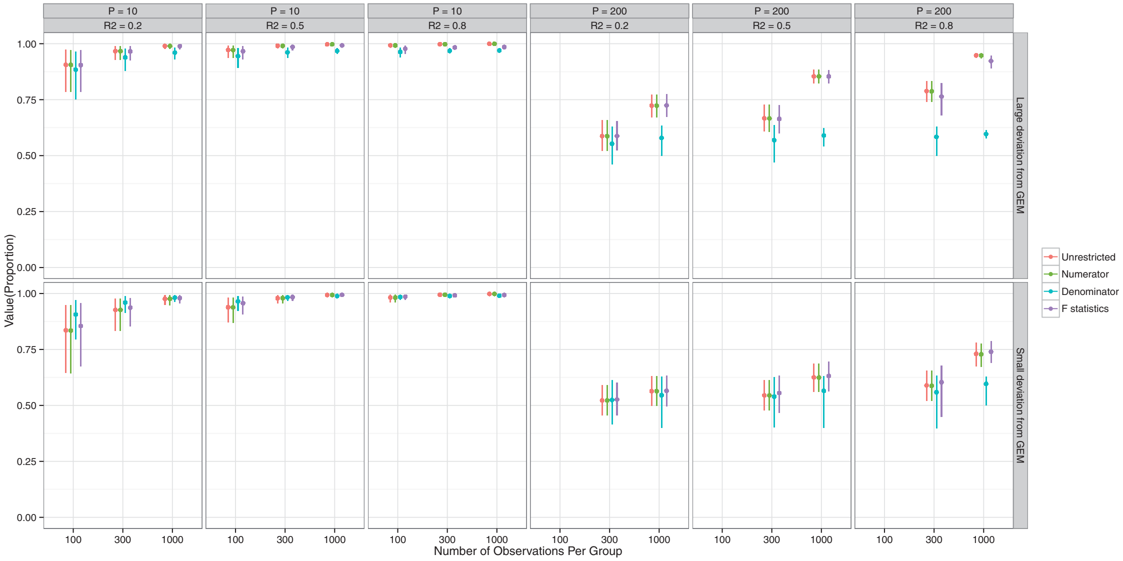

Fig. 2. Non-GEM data generation model. Mean and 95% Monte Carlo (MC) confidence intervals (based on the $B = 1000$ MC runs) of the values $V$ of the decisions, as a proportion $(V - V^-)/(V^+ - V^-)$, for $p = 10$ (left half of panels) and 200 (right half of panels), and for small deviation from GEM (top half of panels) and large deviation from GEM (bottom half of panels). The three panels per ($p$, deviation from GEM) combination correspond to $R^2 = 0.2$ on the left, $R^2 = 0.5$ in the middle and $R^2 = 0.8$ on the right. The method based on the unrestricted regression and the three GEM approaches are denoted as: (i) unrestricted—red color, most left; (ii) numerator criteria—green, second from left; (iii) denominator criterion—blue, third from left; (iv) $F$ criterion—purple, most right. The "Number of observations" on the bottom horizontal axis is the sample size per group.

$R^2$, $n$, and $p$ increase. Regardless of the data generating model, the values produced by the $F$ method are either the best or very close to the best values produced by either of the other methods compared here. Additionally, simulations were run using the non-GEM generating model except that a subset of predictors were discretized to be binary (5 out of 10 for $p = 10$ and 20 out of 200 when $p = 200$); the results are very similar to those when all predictors are continuous—details are provided in the supplementary material.

Section 4 of the supplementary material available at *Biostatistics* online presents results on the performance of the GEM methods in the case when the data generation is not from the linear model (2.2). There we show simulation results based on a doubly-robust estimation procedure using an augmented inverse probability weighted estimator (AIPWE) of the value $V(d)$ (Robins *and others*, 1994; Zhang *and others*, 2012b). Although the GEM approach based on the AIPWE does marginally worse than the unrestricted approach described in Zhang *and others* (2012b) using an example with $p = 3$ predictors, their approach becomes computationally infeasible for larger values of $p$. In cases with large $p$, the GEM reduces the dimensionality of the predictor space to 1 making the AIPWE approach fast and feasible.

## 6. Application to data from a RCT

We illustrate the three GEM procedures using data from a RCT for the treatment of depression comparing antidepressants of the class of selective serotonin reuptake inhibitors (SSRI) against placebo. In addition to establishing the overall efficacy of the SSRI, the investigators were interested in finding biosignatures for SSRI treatment response. The investigators defined "biosignature" as a baseline patient characteristic

Table 1. *SSRI Clinical biosignature: potential moderators of the efficacy of treatment with SSRI vs. placebo with respect to change in HRSD from baseline to week 8. The 3rd column gives the p-values for the interaction predictor-by-treatment term and the 4th column gives the effect size of the predictor as a moderator of treatment effect from a regression model with only that variable as a predictor in addition to treatment. The last two columns give the regression coefficients from models with all five baseline measures as predictors for treatment $A = 0$ (placebo) and $A = 1$ (SSRI) respectively*

| | Mean | St. dev. | Interaction | Effect | Reg. coefs | |
|---|---|---|---|---|---|---|
| | | | $p$ value | size | $A = 0$ | $A = 1$ |
| Anxiety | 5.36 | 1.80 | 0.797 | 0.020 | 1.06 | 1.44 |
| Anger attack | 3.05 | 2.12 | 0.671 | 0.034 | −0.59 | −0.09 |
| Suicide risk | 5.42 | 2.37 | 0.155 | 0.113 | 1.00 | −0.38 |
| Medical comorbidity score | 2.01 | 2.78 | 0.092 | 0.140 | 0.11 | −0.58 |
| Life pleasure score | 33.17 | 5.51 | 0.065 | 0.148 | −0.20 | 0.04 |

or a combination of such characteristics, that constitutes a moderator of the treatment effect of SSRI vs. placebo.

Data from 76 and 72 subjects randomized to placebo and SSRI, respectively, were available. The outcome was the change from baseline (week 0) to 8 week of treatment on the Hamilton Rating Scale for Depression (HRSD). High values of HRSD indicate higher depression severity and thus positive change (week 0–week 8) indicate reduction of depression. The following baseline clinical measures were proposed as potential moderators: (i) level of anxiety (ii) severity of anger attack; (iii) suicidal risk; (iv) medical comorbidity score; and (v) experience of pleasure score.

Outcome was modeled as a linear function of a baseline measure, treatment indicator (SSRI $A = 1$ vs. placebo $A = 0$) and the interaction between them for each measure individually. None of the interaction terms were statistically significant, see Table 1. A comparison of a full unrestricted model with all five predictors and their interactions with treatment against a reduced model without the interactions, yielded a non-significant $F$-test for the interactions ($F_{(5,136)} = 1.41$, $p$ value $= 0.14$). Thus, the usual approaches of treating each predictor separately or a full unrestricted model for all predictors fail to find evidence for heterogeneous effect of SSRI and consequently fails to identify patients who stand to benefit from or be harmed by it.

Next, the linear combinations $\boldsymbol{\alpha}$ for the 3 GEM criteria were estimated, see Table 2. The numerator and $F$-criteria give similar results, but only the $F$-criterion has a statistically significant permutation $p$ value ($p < 0.05$). Note, that the effect sizes for the GEMs based on the numerator and the $F$-criterion (which are very similar, both $ES = 0.27$), are double that of any individual predictor. The denominator GEM, on the other hand, does not produce a significant interaction $p$ value (and also has a very small estimated ES), which is consistent with the observation that, since the angle between the unrestricted regression coefficient vectors is relatively large ($0.35\pi$), the model deviates quite a bit from a true GEM model.

For the sake of comparison, estimates of the value for the three GEM criteria were obtained using an *Inverse Probability Weighted Estimator* (IPWE) IPWE $= \frac{1}{n}\sum_{i=1}^{n} \frac{C(\hat{d}(\boldsymbol{x}_i))y_i}{\pi^{A_i}(1-\pi)^{1-A_i}}$, where $C(\hat{d}(\boldsymbol{x}_i)) = 1$, if the treatment assignment $A$ and treatment decision $d$ coincide for subject $i$ with covariates $\boldsymbol{x}_i$. Here, $\pi^{A_i}$ is the probability of treatment assignment, which will be a constant for a RCT and is 0.5 in this example. Row 8 of Table 2 gives a 95% cross-validation bootstrap confidence interval (using 1000 bootstrap samples) for the value of each GEM criterion. The CIs were computed using a 10-fold cross-validation on each bootstrap sample, where treatment decisions were estimated by applying the respective GEM approach to 9 of 10 non-overlapping subsamples of equal size, and then applied to the remaining 10th subsample to

Table 2. *GEM Model for SSRI clinical biosignature. The estimated GEMs of the SSRI treatment effect on change in HRSD. The bottom rows give the GEM effect sizes (row 6), permutation-adjusted p-values (row 7); the estimated value (1.1) of the decision based on GEM criteria along with a 95% cross-validated bootstrap confidence interval (CI) (row 8); the difference in value and 95% cross-validated bootstrap CI for the difference between the decision based on the respective GEM and the decision (i) give everyone placebo (row 9), (ii) give everyone SSRI (row 10), and (iii) give everyone SSRI or placebo at random (row 11).*

| | Estimated $\boldsymbol{\alpha}$ | | |
|---|---|---|---|
| | $\hat{\boldsymbol{\alpha}}^N$ | $\hat{\boldsymbol{\alpha}}^D$ | $\hat{\boldsymbol{\alpha}}^F$ |
| Anxiety | 0.12 | 0.55 | 0.12 |
| Anger attack | 0.15 | −0.15 | 0.15 |
| Suicide risk | −0.42 | 0.14 | −0.42 |
| Medical comorbidity score | −0.21 | −0.10 | −0.21 |
| Life pleasure score | 0.07 | -0.04 | 0.07 |
| Effect size | 0.27 | 0.01 | 0.27 |
| Permutation $p$-value | 0.061 | 0.895 | 0.048 |
| Value of GEM | 8.03 | 7.60 | 8.03 |
| (95% CI) | (6.28, 9.78) | (5.62, 9.43) | (6.21, 9.68) |
| Value of GEM − Value of placebo | 2.02 | 1.57 | 2.00 |
| (95% CI) | (1.97, 2.06) | (1.52, 1.62) | (1.96, 2.05) |
| Value of GEM − Value of SSRI | 0.52 | 0.07 | 0.50 |
| (95% CI) | (0.48, 0.55) | (0.04, 0.10) | (0.46, 0.54) |
| Value of GEM − Value of random | 1.29 | 0.84 | 1.27 |
| (95% CI) | (1.25, 1.32) | (0.80, 0.87) | (1.24, 1.31) |

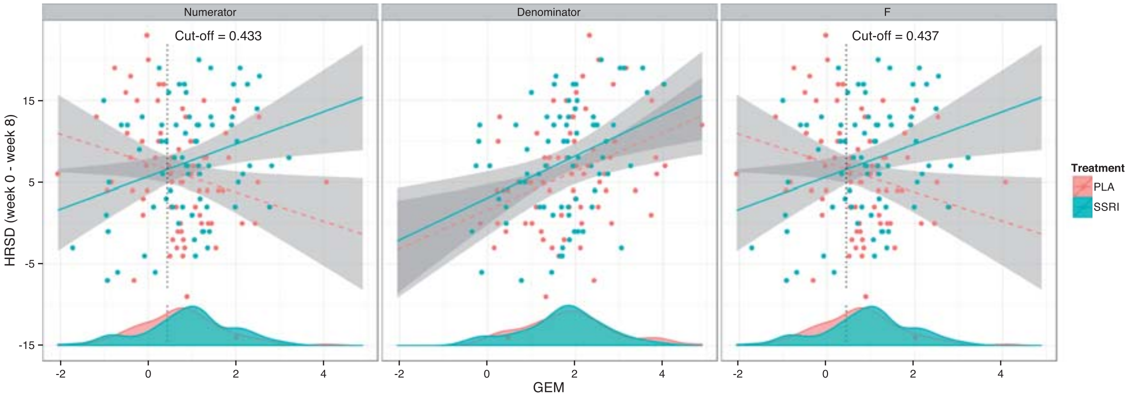

Fig. 3. The relationship between the GEMs obtained from the three criteria and the change in depression (HRSD) from baseline to week 8 for the SSRI (blue) and placebo (red) interventions. The GEMs corresponding to each of the criteria are plotted on the horizontal axis. The lines are the LS lines and the shaded areas indicate the 95% pointwise CIs. The densities of the respective GEMs for the two treatment groups are indicted at the lower part of each panel. The vertical lines indicate the cut-off point on the linear combinations of predictors above which a depressed patient would benefit from treatment with SSRI.

obtain an estimate of the value of the treatment decision and finally averaging those estimates across the 10 folds of the cross-validation. As Table 2 shows, the $F$ and numerator approaches produce very similar bootstrap confidence intervals for the value of the decision, while the denominator criterion results in a lower decision value that has a wider 95% CI. The last three rows of Table 2 show the differences between the values of the treatment decisions derived from each the three GEM approaches and the value of three commonly used comparison decisions (i) give everyone placebo; (ii) give everyone SSRI; and (iii) give placebo and SSRI at random estimated by the same cross-validation approach based on 1000 bootstrap samples.

The results from the GEM approaches are visually presented in Figure 3. The GEM analysis using the $F$-ratio criterion (similar to the numerator criterion) results in the conclusion that 30.4% of the target population (to the left of the vertical lines at GEM $= 0.44$) does not benefit from SSRI treatment. The decision based on the $F$ GEM could be not to prescribe SSRI to those subjects with $GEM_F < 0.44$; alternatively, one might choose to give SSRI only to patients with a $GEM_F$ scores in the range where the 95% CIs for placebo and SSRI GEM regressions do not overlap, that, GEM$\geq$ 1.4. These results are consistent with the fact that many antidepressant trials fail to show efficacy, or show only small benefits, for example, about 25–30% difference in response rates of the antidepressants vs. placebo (60–65% vs. 30–35% respectively).

## 7. DISCUSSION

This article has shown how to combine several baseline characteristics into a single generated effect moderator in the context of the classic linear model. Closed-form expressions have been derived for these GEMs that do not require complex iterative computations. The GEM offers a straightforward approach to determine beneficial treatments for patients. From this perspective, GEMs can be viewed as indices for treatment decisions. Of the three criteria, we generally recommend the $F$-criterion, because it simultaneously maximizes the interaction effect (the numerator) and also minimizes the prediction error (denominator) in the class of GEM models. Additionally, from our results, the $F$-criterion's performance is either optimal or very close to optimal with respect to making rules for treatment decisions with highest values.

In practice, after conducting the main hypotheses testing in efficacy studies, investigators attempt to discover baseline patient features that moderate the effect of treatment. Given that (if present) variables with large moderating effects of treatments for most illnesses have already been discovered, it is not surprising that researchers regularly fail to discover other moderators in studies where the primary goal is to establish efficacy. The proposed methods show that combining patient characteristics with little to no moderating effects of a treatment can result in a strong treatment effect modifier that can help with making treatment decisions. Of course, any treatment decision has to be validated in properly designed studies; for example, a 3-arm RCT where the experimental treatment, the control treatment and treatment according to the investigated treatment decision are compared. The proposed methodology is expected to be of particular utility in studies specifically designed to discover biosignatures for response to treatment, as discussed in the Introduction.

Several generalizations of the GEM procedure are currently under development, such as extending the GEM to generalized linear models and longitudinal outcomes. Work is also underway to allow the outcome to depend on nonparametric functions of GEMs, similar to generalized additive models. It will be useful to compare the linear GEM model developed here and a more flexible nonparametric GEM model to other methods for precision medicine for providing guidance in making treatment decisions.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

## Acknowledgments

The authors are thankful to the editors and three reviewers whose feedback has greatly improved this article. *Conflict of interest*: None declared.

## Funding

National Institutes of Health grant R01 MH099003.

## References

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

Emura, T., Chen, Y.-H. and Chen, H.-Y. (2012). Survival prediction based on compound covariate under cox proportional hazards models. *PLoS One* **7**, 247627.

Follmann, D. A. and Proschan, M. A. (1999). A multivariate test of interaction for use in clinical trials. *Biometrics* **55**, 1151–1155.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361–372.

Gunter, L., Zhu, J. and Murphy, S. A. (2011). Variable selection for qualitative interactions in presonalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics* **21**, 1063–1078.

Kang, C., Janes, H. and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics* **70**, 695–707.

Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. *Statistics in Medicine* **32**, 1964–1973.

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

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

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

Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 331–366.

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

Robins, J. M. (2004). Optimal structured nested models for optimal sequential decisions. In: P. J. Heagerty and D. Y. Lina (editors), *Proceedings of the Second Seattle Symposium on Biostatistics*. New York: Springer, pp. 189–326.

Robins, J., Orellana, L. and Rotnizky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* **27**, 4678–4721.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

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.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**, 874–883.

TIAN, L. AND TIBSHIRANI, R. J. (2011). Adaptive index models for market-based risk stratification. *Biostatistics* **12**, 68–86.

TUKEY, J. W. (1991). Use of many covariates in clinical trials. *International Statistical Review* **59**, 123–137.

TUKEY, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials* **14**, 266–285.

WANG, R., SCHOENFELD, D. A., HOEPPNER, B. AND EVINS, A. E. (2015). Detecting treatment-covariate interactions using permutation methods. *Statistics in Medicine* **34**, 2035–2047.

WANG, R. AND WARE, J. H. (2013). Detecting moderator effects using subgroup analyses. *Prevention Science* **14**, 111–120.

WELLEK, S. (1997). Testing for absence of qualitative interactions between risk factors and treatment effect. *Biometrical Journal* **39**, 809–821.

ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. AND LABER, E. (2012a). Estimating optimal treatment regimes from classification perspective. *Stat* **1**, 103–114.

ZHANG, B., TSIATIS, A. A., LABER, E. B. AND DAVIDIAN, M. (2012b). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.

ZHAO, Y., ZENG, D., RUSH, A. J. AND KOSOROK, M. P. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

ZHAO, Y., ZHENG, D., LABER, E. B. AND KOSORROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110**, 583–598.

[*Received December 4, 2015; revised June 9, 2016; accepted for publication June 12, 2016*]