The Annals of Applied Statistics
2019, Vol. 13, No. 2, 824–847
https://doi.org/10.1214/18-AOAS1216

FUSED COMPARATIVE INTERVENTION SCORING FOR HETEROGENEITY OF LONGITUDINAL INTERVENTION EFFECTS[1]

BY JARED D. HULING*, MENGGANG YU† AND MAUREEN SMITH†

*Ohio State University** *and University of Wisconsin-Madison*†

With the growing cost of health care in the United States, the need to improve efficiency and efficacy has become increasingly urgent. There has been a keen interest in developing interventions to effectively coordinate the typically fragmented care of patients with many comorbidities. Evaluation of such interventions is often challenging given their long-term nature and their differential effectiveness among different patients. Furthermore, care coordination interventions are often highly resource-intensive. Hence there is pressing need to identify which patients would benefit the most from a care coordination program. In this work we introduce a subgroup identification procedure for long-term interventions whose effects are expected to change smoothly over time. We allow differential effects of an intervention to vary over time and encourage these effects to be more similar for closer time points by utilizing a fused lasso penalty. Our approach allows for flexible modeling of temporally changing intervention effects while also borrowing strength in estimation over time. We utilize our approach to construct a personalized enrollment decision rule for a complex case management intervention in a large health system and demonstrate that the enrollment decision rule results in improvement in health outcomes and care costs. The proposed methodology could have broad usage for the analysis of different types of long-term interventions or treatments including other interventions commonly implemented in health systems.

1. Introduction. Health care costs in the United States have continued to rise and at the same time health outcomes remain relatively poor. This trend has led to an increasingly urgent need to improve both the quality and efficiency of care. Much work towards this aim has focused on the care of complex patients with many comorbidities, chronic conditions, or serious illnesses. These patients, often

Received December 2017; revised September 2018.

[1]Research reported in this article was partially funded through two Patient-Centered Outcomes Research Institute (PCORI) Awards (ME-1409-21219 and HSD-1603-35039). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

This project was also supported by the UW Health Office of Population Health, the Health Innovation Program, the UW School of Medicine and Public Health from The Wisconsin Partnership Program.

Key words and phrases. Fused lasso, precision medicine, comparative effectiveness research, electronic health records, interaction modeling.

referred to as "high-need, high-cost" (HNHC) patients account for approximately half of all US health care spending (Cohen and Yu (2012)) yet their needs are often unmet. To better meet these needs it is imperative to focus on care delivery strategies that are commensurate with the particular needs of HNHC patients.

Due to their often wide-ranging comorbidities, HNHC patients require complex and multifaceted care. Often, patients with many chronic conditions and diseases visit a multitude of different doctors for their different conditions. Each doctor works to address the individual conditions and diseases of the patient, typically using the latest advances in care for each specific condition. However, a care delivery system that works on solving the individual problems of a patient without considering treating the patient as a whole tends to result in high costs and poor health outcomes (Stange (2009)). The tendency of patients to have their individual conditions treated separately often serves HNHC patients poorly, resulting in a model of care which can be significantly fragmented and uncoordinated (Cheung et al. (2009), U. S. Department of Health and Services (2012), Stange (2012)). To help mitigate this problem, much research has focused on developing interventions that reduce care fragmentation and encourage treatment of the *patient* and not the individual diseases and conditions of the patient.

One common approach to accomplishing this is through a complex case management (CCM) intervention, a class of interventions which seek to coordinate the care of complex patients and provide them with a system of increased support, advocacy, and education (Hickham et al. (2013)). In CCM interventions, a nurse or social worker takes responsibility for coordinating and implementing the care plan of individual patients, thus reducing care fragmentation. CCM interventions have been widely adopted across health systems in the US, however, a recent systematic review of such interventions suggests that these programs have a wide range of efficacy and cost (Hickham et al. (2013)).

While many CCM interventions show little benefit in improving care, the programs which are successful in reducing costs and improving patient outcomes tend to target the enrollment of patients who are believed to be likely to benefit (Blumenthal et al. (2016)). This suggests that CCM interventions have different effects across different subsets of the population of HNHC patients. Indeed it is well-established that whether or not patients benefit from an intervention often depends on their individual characteristics, yet there is little guidance towards effectively identifying patients who might benefit from CCM interventions (Hickham et al. (2013)). In this paper we seek to fill this gap by introducing a statistical framework for identifying which patients benefit from long-term interventions.

However, the long-term nature of many health system interventions, such as the CCM, presents many challenges for such identification. First, patient outcomes are typically observed at regular time intervals and thus the repeated measures of patient health and utilization outcomes must be accounted for. Second, the effect of an intervention is *expected* to change over time due to institutional learning,

that is, nurses involved in care coordination may improve their coordination efforts with time. Furthermore, there may be delayed intervention effects which may differ based on what comorbidities specific patients have and the nurses and doctors involved in the intervention may need time to build trust with certain patients. Another problem in reliably identifying a subgroup of patients who would benefit from an intervention is the fact that randomized controlled trials are often impossible to conduct due to time, resource, or administrative constraints. Thus researchers are often faced with the difficult task of using existing observational studies to establish a causal relationship between patient characteristics and benefit from enrollment in an intervention.

In this paper we develop a methodology to overcome these challenges in identifying patients likely to benefit from long-term health system interventions like CCM. We extend the statistical methodology for subgroup identification developed by Chen et al. (2017) to handle longitudinal intervention effects which may change smoothly over time. Our proposed method results in the estimation of a score which can be used to assess the expected benefit of an intervention using patient baseline characteristics. We propose two versions of our methodology, one which relies on propensity score weighting and another which relies on matching, such as propensity score matching (Rosenbaum and Rubin (1983)).

The remainder of this paper is organized as follows. In Section 2 we introduce the underlying outcome model that guides our methodology and also introduce background information on a framework for subgroup identification for single time-point outcomes. In Section 3 we introduce our working model and our estimator for subgroup identification in addition to a technique for evaluating the validity of our estimated subgroups. In Section 4 we investigate the finite-sample performance of our estimator using a numerical study. Finally, in Section 5 we use our estimator to identify a subgroup of patients who benefit from a CCM intervention in UW Health, the health system affiliated with the University of Wisconsin-Madison, and demonstrate that it identifies a subgroup of patients which benefit from CCM in terms of multiple patient outcomes.

2. Preliminaries.

2.1. *Notation and model for longitudinal outcomes.* In many long-term interventions, it is conceivable and often expected that the effect of the intervention may change gradually over time. Some interventions may have a pronounced effect in the beginning with less effect over time. Some interventions may involve an intensive process, whose effects take time to establish and may build over time. In such settings, we are interested in modeling the longitudinal patient outcomes as a time-varying function of patient characteristics and the intervention status. Let Y_t be an outcome of interest measured at time t, $\boldsymbol{X} \in \mathcal{X}$ be a length p vector of baseline covariates, and $A \in \{1, -1\}$ be a treatment status where $A = 1$ indicates enrollment in an intervention and $A = -1$ indicates a patient is in the control

group. Then the outcome over time can be modeled as

$$E(Y_t|\boldsymbol{X}, A) = \phi_t(\boldsymbol{X}) + \boldsymbol{\gamma}_0(t)'\boldsymbol{X} \cdot A/2, \tag{1}$$

where t indexes the temporal domain and $\phi_t(\boldsymbol{X})$ is an unspecified function representing covariate main effects. Model (1) allows for covariates and the intervention-covariate interactions to have temporally changing effects on the longitudinal outcomes where the functional form of the covariate main effects is unspecified and the form of the treatment-covariate interactions is a varying coefficient model. This is similar to a varying coefficient model (Hastie and Tibshirani (1993), Fan and Zhang (1999, 2008)) which is often used to model longitudinal data. The key difference is Model (1) specifies a varying coefficient model for only the treatment-covariate interactions and the main covariate effects ϕ_t are left unspecified.

In the context of subgroup identification, the key interest is in estimating a contrast function which elicits the covariate-specific effect of an intervention at time t: $\Delta(\boldsymbol{X}, t) = E(Y_t|A = 1, \boldsymbol{X}) - E(Y_t|A = -1, \boldsymbol{X})$. Under Model (1), this contrast is simply $\boldsymbol{\gamma}_0(t)'\boldsymbol{X}$. Here, although CCM interventions are often long-term, the intervention status A does not change over time and hence we assume that the intervention status has no time dynamics. In the formulation of Model (1), the effects which determine the contrast function are specified as a function of time. In settings where data is limited in sample size and in time, it may be challenging to estimate $\boldsymbol{\gamma}(\cdot)$ nonparametrically, as hundreds or even thousands of covariates may be available from healthcare claims or electronic health records. Furthermore, in the evaluation of health system-wide interventions, such data are often measured at regular intervals (monthly in the case of utilization outcomes such as hospitalization events) so we can instead focus on simply estimating this function over time on a regular grid. Thus it is natural to model the response as

$$E(Y_t|\boldsymbol{X}, A) = \phi_t(\boldsymbol{X}) + \boldsymbol{\gamma}'_{0t}\boldsymbol{X} \cdot A/2 \qquad \text{where } t = 1, \ldots, K. \tag{2}$$

We now use the notation $\boldsymbol{\gamma}_{0t}$ instead of $\boldsymbol{\gamma}_0(t)$ for the interaction effects due to data being collected on a discrete grid of time points. Using the model above is equivalent to fitting a model for each point in time where data is observed. Here the contrast function can be written as $\Delta_t(\boldsymbol{X}) = \boldsymbol{\gamma}'_{0t}\boldsymbol{X}$. While this parameterization is flexible, due to a large number of parameters when the number of covariates and time points is large, the parameters in this model also may be challenging to estimate efficiently when the sample size is limited even when using variable selection techniques such as the lasso (Tibshirani (1996)). In essence, estimating the parameters in this model to change freely over time may be too flexible. Yet there is additional information about CCM interventions that we can incorporate into our modeling strategy to help reduce this complexity. In health system intervention settings we would not expect rapid temporal changes in the intervention effects and the subgroup of patients who benefit. In particular, case managers involved in the implementation of CCM believe that for most patients the benefits of

CCM may slowly accumulate over time. This belief is substantiated by evidence more broadly that care management interventions may take time to realize benefit (Bodenheimer and Berry-Millett (2009)). In Section 3 we will introduce a flexible smoothing approach to mitigate this problem by encouraging coefficients closer in time to be more similar, thus borrowing strength over time.

2.2. *Subgroup identification via treatment scoring*. The overarching goal of subgroup identification is to relate patient characteristics to an optimal treatment or intervention decision. This is often achieved by modeling the interactions between patient characteristics and a treatment or intervention. With a suitable model for these interactions one can then identify for any new patient whether that patient is expected to realize an improvement in his or her outcomes if given the treatment or intervention. We now provide some background information on a robust estimation procedure to accomplish this goal for outcomes measured at a single time point $K = 1$. Assume the continuous response Y, where without loss of generality positive values are associated with better outcomes, is generated from the following model:

$$E(Y|\boldsymbol{X}, A) = \phi(\boldsymbol{X}) + \boldsymbol{\gamma}_0'\boldsymbol{X} \cdot A/2, \tag{3}$$

where $\phi(\boldsymbol{X})$ represents main covariate effects, $\boldsymbol{\gamma}_0$ reflects linear intervention-covariate interaction effects, and $\boldsymbol{X}$ contains an intercept. We make a few further assumptions to clarify how our estimands relate to causal quantities using the familiar potential outcome framework of Rubin (2005). $Y^{(1)}$ and $Y^{(-1)}$ denote the potential outcomes for a patient under intervention $A = 1$ and control $A = -1$ respectively. A comparison of $Y^{(-1)}$ and $Y^{(1)}$ reveals the causal effect of A. However for any individual, only one of the potential outcomes is observed. To relate the potential outcomes to the observed outcome Y, we make the following assumptions:

(i) $Y = I(A = 1)Y^{(1)} + I(A = -1)Y^{(-1)}$, which is essentially a claim that the intervention status of one patient does not affect the potential outcomes of other patients. This assumption is also known as the Stable Unit Treatment Value Assumption (SUTVA).

(ii) $A \perp\!\!\!\perp (Y^{(1)}, Y^{(-1)})|\boldsymbol{X}$; this assumption is typically called the "strongly ignorable assumption" and essentially means that there are no unmeasured confounders.

(iii) The treatment assignment is guided by $\pi(\boldsymbol{x}) = P(A = 1|\boldsymbol{X} = \boldsymbol{x})$, where $0 < \pi(\boldsymbol{x}) < 1$ for all $\boldsymbol{x} \in \mathcal{X}$.

Under Model (3) and Assumptions (i)–(iii), the contrast function $\Delta(\boldsymbol{x}) = E(Y|A = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y|A = -1, \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{\gamma}_0'\boldsymbol{x}$ for those with covariates $\boldsymbol{X} = \boldsymbol{x}$ has a causal interpretation and equals the conditional average causal effect $E(Y^{(1)} - Y^{(-1)}|\boldsymbol{X} = \boldsymbol{x})$.

The term $\Delta(\boldsymbol{x})$ acts as a "benefit score", in the sense that positive values of $\Delta(\boldsymbol{x})$ correspond to a positive benefit for a patient with characteristics $\boldsymbol{x}$ and negative

values indicate a negative impact on the average causal effect of an intervention. Hence, $\Delta(\boldsymbol{x})$ can be used to identify which patients are expected to benefit from an intervention. Furthermore, since the benefit score has a direct interpretation regarding the magnitude of effect, it can also be used to order patients by *how much* they are expected to benefit from an intervention. Chen et al. (2017) proposed to estimate $\Delta(\boldsymbol{x})$ by estimating $\hat{\boldsymbol{\gamma}}$ as the minimizer of

$$L(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N} \frac{M[Y_i, \boldsymbol{\gamma}'\boldsymbol{x}_i \cdot A_i/2]}{A_i\pi(\boldsymbol{x}_i) + (1 - A_i)/2} \tag{4}$$

with respect to $\boldsymbol{\gamma}$, where $M(y, v)$ is a loss function satisfying a) $M_v(y, v) = \partial M(y, v)/\partial v$ is increasing in v for any given y and b) $U(y) \equiv M_v(y, 0)$ is monotone in y. When the outcome Y is continuous, a reasonable choice for $M(y, v)$ is the squared error loss $(y - v)^2$. Many other choices are available and correspond to existing literature (Chen et al. (2017)). Different outcomes can be accommodated by the use of corresponding losses, such as the logistic loss if binary outcomes are of interest. While the main effects $\phi(\cdot)$ are not being estimated in (4), minimization of (4) still results in valid estimates of $\Delta(\boldsymbol{x})$ and can often be better than when incorporating the main effects of the covariates if the model for the main effects is misspecified, as modeling the full relationship between covariates and outcome emphasizes accuracy in predicting the clinical outcome over optimizing treatment decisions (Zhao et al. (2012)). The population version of (4) is $\ell(\boldsymbol{\gamma}) = E[\ell(\boldsymbol{\gamma}, \boldsymbol{x})]$, where

$$\ell(\boldsymbol{\gamma}, \boldsymbol{x}) = E\left[\frac{M(Y, \boldsymbol{\gamma}'\boldsymbol{x} \cdot A/2)}{A\pi(\boldsymbol{x}) + (1 - A)/2}\Big|\boldsymbol{X} = \boldsymbol{x}\right]. \tag{5}$$

Under Assumptions (i)–(iii), the minimizer $\hat{\boldsymbol{\gamma}}_0$ of $\ell(\boldsymbol{\gamma})$ has the property that if $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} < 0$ then $E[U(Y^{(1)})|\boldsymbol{X} = \boldsymbol{x}] > E[U(Y^{(-1)})|\boldsymbol{X} = \boldsymbol{x}]$ and if $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} > 0$ then $E[U(Y^{(1)})|\boldsymbol{X} = \boldsymbol{x}] < E[U(Y^{(-1)})|\boldsymbol{X} = \boldsymbol{x}]$. Thus $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x}$ reflects for each level of covariate values which of the treatment options yields larger expected *potential outcomes* and hence it can be used to identify which patients are expected to benefit from a treatment. Thus, the average potential outcome is larger among patients with $A = 1$ and $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} > 0$ than those with $A = -1$ and $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} > 0$. Similarly, the average potential outcome is larger among patients with $A = -1$ and $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} \leq 0$ than those with $A = 1$ and $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} \leq 0$. Furthermore, $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x}$ can be used to estimate the magnitude of the treatment effect. For example, if $M(y, v) = (y - v)^2$, then $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x} = \Delta(\boldsymbol{x})$. For other loss functions, $\hat{\boldsymbol{\gamma}}_0'\boldsymbol{x}$ will not necessarily be the treatment effect itself, but a monotone transformation of it.

In observational studies, $\pi(\boldsymbol{x})$ is generally unknown and thus must be estimated. This can be accomplished in the same way as is typically done in analyses focusing on estimating average treatment effects from observational studies. As noted in Rosenbaum and Rubin (1983), this involves modeling choices to relate the treatment status A to the covariates. A typical choice is a logistic regression model, however, in some applications a more flexible model may be required. High di-

mensionality can be handled using variable selection techniques. When using (4) in an analysis, one can straightforwardly deal with data scenarios with a large number of potential treatment-effect modifiers. In the analysis of health system interventions, it is reasonable to assume that the major patient characteristics which modify the effect of an intervention are a subset of the available information at hand. In such high-dimensional scenarios we can add a penalty term, $\lambda \sum_{j=1}^{p} |\boldsymbol{\gamma}_j|$, to (4) in order to perform variable selection. This causes the estimated benefit score to be a function of a smaller number of the available covariates.

The loss function (4) is clearly designed for data where the effect of a treatment or intervention is expected to be constant in time. This may not be appropriate for long-term interventions with longitudinally measured outcomes. We aim to adapt this general approach of benefit score estimation to appropriately handle data where the effects of an intervention may change over time.

3. Methodology.

3.1. *Fused comparative intervention scoring.* Denote $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_K')'$ and denote $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{01}', \ldots, \boldsymbol{\gamma}_{0K}')'$ as the true coefficients over time. We assume that the data are generated from model (2). Under this model we allow both the effect of the intervention and its interactions with patient characteristics to change freely over time.

We propose an extension of $\ell(\boldsymbol{\gamma})$ and thus (5) that accounts for longitudinally changing intervention effects as $\ell(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) = E[\ell(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{x})]$, where

$$\ell(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{x}) = E\left[\frac{1}{K}\sum_{t=1}^{K} \frac{M(Y_t, \boldsymbol{\gamma}_t'\boldsymbol{x} \cdot A/2)}{A\pi(\boldsymbol{x}) + (1-A)/2} \middle| \boldsymbol{X} = \boldsymbol{x}\right]. \tag{6}$$

Here each set of coefficients $\boldsymbol{\gamma}_k$ represents the treatment-covariate interaction for the kth time point. Thus the minimizer $(\hat{\boldsymbol{\gamma}}_{01}, \ldots, \hat{\boldsymbol{\gamma}}_{0K})$ of $\ell(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$ provides a benefit score for each time point. Although different patients may benefit differentially from an intervention over time, for example, some patients may reap benefits from a program early on while others may require months in the intervention to begin to see benefit, decisions about whether to enroll patients only occur once at baseline. We focus on baseline decisions in part because the effects of CCM are expected to last even after disenrollment. Thus, an attempt to optimize when patients should be disenrolled may be a secondary concern. We can then make personalized enrollment decisions by summarizing the time-specific benefit scores into a single benefit score $d(\boldsymbol{X})$, which reflects the overall benefit. The single score can be used to create an individualized intervention rule (IIR), or a map from baseline patient characteristics to an intervention decision, such as $\operatorname{sgn}(d(\boldsymbol{X}))$, where $\operatorname{sgn}(\cdot)$ is the sign function that takes value 1 for positive arguments and -1 otherwise. For example, a simple average of the benefit scores $d_{\text{avg}}(\boldsymbol{X}) = K^{-1}\sum_{t=1}^{K} \hat{\boldsymbol{\gamma}}_{0t}'\boldsymbol{X}$ would reflect the average benefit over time.

Furthermore, if Assumptions (i)–(iii) hold for $Y_t, Y_t^{(1)}$, and $Y_t^{(-1)}$ for all $t = 1, \ldots, K$ the IIR $\operatorname{sgn}(d_{\text{avg}}(\boldsymbol{X}))$ (i.e., patients with an average benefit score greater

than zero are assigned to the intervention) maximizes the average value function $\mathbb{V}_U(d) = E[K^{-1}\sum_{t=1}^{K}\{U(Y_t^{(1)}) - U(Y_t^{(-1)})\}d(\boldsymbol{X})]$. This can be seen from the fact that the minimizer of $\ell(\boldsymbol{\gamma}_t, \ldots, \boldsymbol{\gamma}_K)$ has the property that if $\sum_{t=1}^{K}\hat{\boldsymbol{\gamma}}_{0t}'\boldsymbol{x} < 0$ then $E[\sum_{t=1}^{K}U(Y_t^{(1)})|\boldsymbol{X}=\boldsymbol{x}] > E[\sum_{t=1}^{K}U(Y_t^{(-1)})|\boldsymbol{X}=\boldsymbol{x}]$ and if $\sum_{t=1}^{K}\hat{\boldsymbol{\gamma}}_{0t}'\boldsymbol{x} > 0$ then $E[\sum_{t=1}^{K}U(Y_t^{(1)})|\boldsymbol{X}=\boldsymbol{x}] < E[\sum_{t=1}^{K}U(Y_t^{(-1)})|\boldsymbol{X}=\boldsymbol{x}]$. Thus, the estimated IIR reflects a mapping from baseline patient characteristics to intervention decisions that, if implemented, would cause the largest expected average potential outcomes over time. If certain times, such as early on in the intervention, are deemed more important, then one could consider a weighted average such as $d_{\text{wavg}}(\boldsymbol{X}) = \{\sum_{t=1}^{K}w_t\}^{-1}\sum_{t=1}^{K}w_t\hat{\boldsymbol{\gamma}}_{0t}'\boldsymbol{X}$ for $w_t > 0$.

The observed data consists of N_t samples at time t: $(Y_{it}, \boldsymbol{x}_i, A_i)$, where Y_{it} is the response for unit i at time t, A_i is the intervention status, and $\boldsymbol{x}_i$ are the patient covariate values. The sample size is a function of t only allowing for dropout due to noninformative issues such as patients being enrolled at different calendar times, which results in missing outcome information for recently enrolled patients who have not been followed for the entire follow-up period. We do not allow for temporally changing post-baseline covariate values.

As mentioned in Section 2.1, if there is a limited sample size, it may be challenging to estimate all parameters in this model efficiently. Therefore we propose to use an estimator which encourages smoothness of the estimated coefficients over time, and thus in turn smoothness of the benefit scores over time. To do so we utilize the fused lasso (Tibshirani et al. (2005)) to encourage the interaction effects of each variable to be more similar for closer points in time. Additionally, high numbers of variables are present in many health care services applications. As such we utilize a lasso penalty to induce some coefficients to be zero. We propose the following estimator: $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1', \ldots, \hat{\boldsymbol{\gamma}}_K')'$ of $\boldsymbol{\gamma}_0$ as the minimizer of

$$\begin{aligned}(7)\qquad &\sum_{t=1}^{K}\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{M(Y_{it}, \boldsymbol{\gamma}_t'\boldsymbol{x}_i \cdot A_i/2)}{A_i\pi(\boldsymbol{x}_i) + (1-A_i)/2} + \lambda_1\sum_{j=1}^{p}\sum_{t=2}^{K}|\boldsymbol{\gamma}_{jt} - \boldsymbol{\gamma}_{j(t-1)}| \\ &\quad + \lambda_2\sum_{j=1}^{p}\sum_{t=1}^{K}|\boldsymbol{\gamma}_{jt}|\end{aligned}$$

with respect to $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$, where N_t is the sample size at time t. If there is differential loss to follow-up, immortal time bias may be a concern. In such a case further measures such as inverse probability of censoring weighting may be required (Robins and Finkelstein (2000)); however, this issue is beyond the scope of this paper. The third term in (7) is a standard lasso penalty which induces variable selection. The second term in (7) encourages the effects of individual variables to be similar over time, thus borrowing strength across observations in time. This stabilizes the estimated coefficients $\hat{\boldsymbol{\gamma}}_t$ over time which thus also stabilizes the estimated benefit scores and also helps to reduce variance in estimation. The

propensity function $\pi(\boldsymbol{x}_i)$ is unknown and must be estimated. Since we restrict our focus to single time-point enrollments, the propensity function can be estimated in the same way as described in Section 2.2.

In some modeling scenarios, it may be necessary to consider an estimator based on propensity score matching instead of weighting. To handle matching-based analyses, consider $\ell_m(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) = E[\ell_m(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{x})]$, where

$$\ell_m(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{x}) = E\left[\frac{1}{K}\sum_{t=1}^{K} wM\big(Y_t, \boldsymbol{\gamma}_t'\boldsymbol{x} \cdot A/2\big)\Big|\boldsymbol{X} = \boldsymbol{x}\right], \tag{8}$$

where w is a weight which either adjusts for the size of the matched groups or can be set as $w = (A\pi(\boldsymbol{x}) + (1 - A)/2)^{-1}$. However, if the weights depend on A_i, both $E[w(A)I(A = 1)|\boldsymbol{X}]$ and $E[w(A)I(A = -1)|\boldsymbol{X}]$ must be constants conditional on $\pi(\boldsymbol{X}) = a$. Commensurate with this population-based objective, we propose the estimator $\hat{\boldsymbol{\gamma}}_m = (\hat{\boldsymbol{\gamma}}_{1m}', \ldots, \hat{\boldsymbol{\gamma}}_{Km}')'$ as the minimizer of

$$\begin{aligned}\sum_{t=1}^{K}\frac{1}{N_t}\sum_{i=1}^{N_t} w_i M\big(Y_{it}, \boldsymbol{\gamma}_t'\boldsymbol{x}_i \cdot A_i/2\big) + \lambda_1\sum_{j=1}^{p}\sum_{t=2}^{K}|\boldsymbol{\gamma}_{jt} - \boldsymbol{\gamma}_{j(t-1)}| \\ + \lambda_2\sum_{j=1}^{p}\sum_{t=1}^{K}|\boldsymbol{\gamma}_{jt}|,\end{aligned} \tag{9}$$

where again w_i is an individual weight which is either $(A_i\pi(\boldsymbol{x}_i) + (1 - A_i)/2)^{-1}$ or is calculated to adjust for the size of each matched group, for example, $w_i = w(A_i) = I(A_i = 1)(C_i + 1) + I(A_i = -1)(C_i + 1)/C_i$ where C_i is the number of controls in the matched cluster to which unit i belongs. The above conditions required for the weights $w(A)$ hold for the above examples. The minimizer of (9) targets a slightly different quantity. In particular, the population version $\ell_m(\boldsymbol{\gamma}, \ldots, \boldsymbol{\gamma}_K)$ of the first term in (9) is such that its minimizer $(\hat{\boldsymbol{\gamma}}_{01m}, \ldots, \hat{\boldsymbol{\gamma}}_{0Km})$ has the property that if $\sum_{t=1}^{K}\hat{\boldsymbol{\gamma}}_{0tm}'\boldsymbol{x} < 0$ then $E[\sum_{t=1}^{K}U(Y_t^{(1)})|\boldsymbol{X} = \boldsymbol{x}, A = 1] > E[\sum_{t=1}^{K}U(Y_t^{(-1)})|\boldsymbol{X} = \boldsymbol{x}, A = 1]$ and if $\sum_{t=1}^{K}\hat{\boldsymbol{\gamma}}_{0tm}'\boldsymbol{x} > 0$ then $E[\sum_{t=1}^{K}U(Y_t^{(1)})|\boldsymbol{X} = \boldsymbol{x}, A = 1] < E[\sum_{t=1}^{K}U(Y_t^{(-1)})|\boldsymbol{X} = \boldsymbol{x}, A = 1]$. This is related to the treatment effect on the treated, but conditional on patient covariates. For more detailed derivations regarding the matching estimator, please see the Supplementary Material (Huling, Yu and Smith (2019)).

The expected improvement in outcome for a new patient with baseline covariates $\boldsymbol{x}^*$ if given the intervention A is then estimated as $\boldsymbol{x}^{*\prime}\hat{\boldsymbol{\gamma}}_t$ at time t. This improvement may change over the course of an intervention. Hence, if we wish to recommend whether a patient should enroll in an intervention, we can average their expected improvement in outcome over time as $\widehat{B}(\boldsymbol{x}^*) = K^{-1}\sum_{j=1}^{K}\hat{\boldsymbol{\gamma}}_t'\boldsymbol{x}^*$. Then we may recommend patients to the intervention if we expect them to have a positive average expected improvement in their outcomes. The recommended

enrollment status, or IIR, is $\widehat{d}_{\text{avg}}(\boldsymbol{x}^*) = \text{sgn}(\widehat{B}(\boldsymbol{x}^*))$. The term $\widehat{B}(\boldsymbol{x}^*)$ is thus a benefit score for recommending the intervention. One may also consider other summaries of the expected longitudinal outcomes, such as the median or other quantiles. There of course is uncertainty in our estimate of which patients benefit from an intervention and which do not and thus in any finite sample there may be a subgroup of patients for whom the proper recommendation is ambiguous. Model selection plays a role in (7), and thus a standard bootstrap approach may fail to obtain correct prediction intervals for the benefit score. It may be possible to obtain prediction intervals for benefit scores for individuals with an approach similar to that developed in Efron (2014). Identification of such patients is an interesting area of research but is beyond the scope of this paper.

Similar to Chen et al. (2017), we can augment the loss function $M(y, v)$ to allow for efficiency improvements. That is, we can work with $\widetilde{M}(y, v, \boldsymbol{x}) = M(y, v) + g(\boldsymbol{x}, v)$, where the augmentation function g satisfies that $\partial g(\boldsymbol{x}, v)/\partial v$ is nondecreasing in v. For the squared loss, this includes a common practice of shifting the outcome by a function of the covariates: $\widetilde{M}(y, v, \boldsymbol{x}) = \{y - a(\boldsymbol{x}) - v\}^2$. In our setting, one can let $a(\boldsymbol{x})$ vary for different t and estimate it using both $\boldsymbol{x}$ and outcome information up to time t. The loss function (7) can be straightforwardly minimized with minor modifications to an alternating direction method of multipliers (ADMM) algorithm as described in Section 6.4.1 of Boyd et al. (2011) for the generalized lasso (Tibshirani and Taylor (2011)).

3.2. *Assessment of IIR impact with bootstrap bias correction.* Evaluating the validity of the estimated subgroups using the same data used to identify the subgroups will result in overly optimistic estimates of the benefit of the treatment assignments by our models. As such, we use the bootstrap bias correction method of Harrell, Lee and Mark (1996), and further explained in Foster, Taylor and Ruberg (2011), to adjust for any overfitting bias which might occur. The bootstrap procedure works by first estimating the amount of bias present in a particular statistic and then subtracting that amount from the statistic. In general, suppose we want to estimate the amount of bias for any statistic S. In subgroup identification, S may represent the expected improvement in the patient outcomes if patients are enrolled in an intervention based on the estimated IIR. Denote $S_{\text{train}}(\boldsymbol{X})$ to be the estimate of the statistic S based on our training data of n subjects and evaluated on the data $\boldsymbol{X}$. For $b = 1, \ldots, B$, let $\boldsymbol{X}_b$ be bootstrap samples (with replacement) of size n and based on which, we can similarly construct S_b. Now let $S_b(\boldsymbol{X}_b)$ and $S_b(\boldsymbol{X})$ be evaluations of S_b on the bth bootstrap sample $\boldsymbol{X}_b$ and our original training data $\boldsymbol{X}$ respectively. Then the bootstrap estimate of the amount of bias with regards to the statistic S is

$$\text{bias}(\boldsymbol{X}) = \frac{1}{B}\sum_{b=1}^{B}\big[S_b(\boldsymbol{X}_b) - S_b(\boldsymbol{X})\big]. \tag{10}$$

A bias-corrected estimate of the statistic S is $S_{\text{train}}(\boldsymbol{X}) - \text{bias}(\boldsymbol{X})$.

4. Simulation. Data were simulated according to the following model:

$$Y_t = \boldsymbol{\beta}_{0t}'\boldsymbol{X} + \boldsymbol{\gamma}_{0t}'\boldsymbol{X} \cdot A/2 + \varepsilon_t, \tag{11}$$

where ε_t are i.i.d. $N(0, 3^2)$ random variables. Model (11) is a special case of Model (2) with main effects $\phi_t(\boldsymbol{X}) = \boldsymbol{\beta}_{0t}'\boldsymbol{X}$ and $K = 6$ in accordance with the motivating study. The initial sample size is 100 for all scenarios. As patients in the CCM study are enrolled at different calendar times, resulting in some patients with missing outcomes in later time points due to incomplete follow-up, we include a scenario where 10 observations are lost to dropout after each time point. In addition, we investigate a setting where the sample size is fixed for each time point. The number of variables at each time point is set to 50 where 10 of the variables have nonzero coefficients at each time point for their corresponding interaction effects and 12 variables have nonzero coefficients for their main effects. The nonzero coefficients for the first five intervention-covariate interactions will be as described below and the second five intervention-covariate interactions coefficients are the negation of the first five.

In our data generation, we chose the nonzero main effects $\boldsymbol{\beta}_{0t}'$ as $c \cdot (1, -1, 1, 1, 1, -1, 1, -1, 1, 1, 1, -1)$ for $t \in \{1, 2, 3\}$ and $c \cdot (2, -2, -1, 1, 1, -1, 2, -2, -1, 1, 1, -1)$ for $t \in \{4, 5, 6\}$. The value c is set to 1 for simulations with large main effects and $1/3$ for simulations with small main effects. Then we use the following three separate scenarios of intervention-covariate interactions with varying degrees of smoothness. The first scenario exhibits the most smoothness in intervention effects over time and we chose

$$(\boldsymbol{\gamma}_{01}, \ldots, \boldsymbol{\gamma}_{06}) = \begin{bmatrix} 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ -1.00 & -1.00 & -1.00 & -0.75 & 0.75 & 0.75 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.75 & 0.75 & 0.75 & 0.75 \\ -0.50 & -0.50 & -0.50 & -0.75 & -0.75 & -0.75 \end{bmatrix}.$$

The second scenario sets $\boldsymbol{\gamma}_{0t}'$ as $(1, -1, 1, 1, -0.5, -1, 1, -1, -1, 0.5)$ and for $t \in \{1, 2, 3\}$ and $(1.5, -0.5, -1, 1, -0.5, -1.5, 0.5, 1, -1, 0.5)$ for $t \in \{4, 5, 6\}$. The third scenario sets

$$(\boldsymbol{\gamma}_{01}, \ldots, \boldsymbol{\gamma}_{06}) = \begin{bmatrix} 0.70 & 0.80 & 0.90 & 1.25 & 1.50 & 1.50 \\ -1.00 & -1.00 & -0.50 & -0.50 & 0.50 & 0.50 \\ 1.00 & 1.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ -0.75 & -0.75 & -0.75 & -0.50 & -0.25 & -0.25 \end{bmatrix}.$$

The third scenario represents the least smooth situation in terms of the interactions.

The proportion of nonzero interaction effects is chosen to be similar to the CCM study and that both binary and continuous covariates affect the intervention heterogeneity. The first two and last eight covariates in $\boldsymbol{X}$ are binary with success probability 0.25 and the remaining covariates in $\boldsymbol{X}$ are multivariate normal with variance covariance matrix 1 along the diagonal and $\rho^{|i-j|}$ for the element in the

row i column j position for $i \neq j$. The covariates are generated in this way as there are both binary covariates and continuous covariates with high multicollinearity in the CCM study. The intervention statuses $A \in \{1, -1\}$ are generated from the propensity model $\text{logit}(\Pr(A = 1|\boldsymbol{X})) = \nu_0 + \boldsymbol{\nu}'\boldsymbol{X}$ with the first 15 elements of $\boldsymbol{\nu}$ as $(-1, 0, 1, 0, 0, 0, 0, 0, 0, 1, -1, 1, -1, 1, -1) \cdot 0.75$ and the remaining 0. Therefore the particular nonzero structure of $\boldsymbol{\nu}$ indicates that there are overlapping covariates that impact both the treatment assignment and outcome. In the simulation, the outcome Y has a range of approximately $[-23.7, 23.5]$, where we assume that larger is better.

We investigate the performance of the proposed fused lasso estimator in comparison with an ad hoc approach in which a separate model is fit for each time point. In the ad hoc approach, we minimize the loss

$$\sum_{t=1}^{K} \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{(Y_{it} - \boldsymbol{\gamma}_t' \boldsymbol{x}_i A_i)^2}{A_i \pi(\boldsymbol{x}_i) + (1 - A_i)/2} + \sum_{t=1}^{K} \lambda_t \sum_{j=1}^{p} |\boldsymbol{\gamma}_{jt}| \tag{12}$$

with respect to $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$, which is equivalent to fitting a separate model for each time point with the lasso for variable selection. All tuning parameters λ_t for (12) and λ_1 and λ_2 for (7) are chosen by 10-fold cross validation with mean-squared error as the criterion. For both our method and the ad hoc approach, we use a simple outcome-shifted augmentation function for each time point: $a_t(\boldsymbol{x}) = N_t^{-1} \sum_{i=1}^{N_t} Y_{it}$. For both methods we estimate $\pi(\boldsymbol{x}_i)$ using a penalized logistic regression model with a lasso penalty and tuning parameter selected by 10-fold cross validation with out-of-fold deviance as the criterion. The simulation is run 500 times for each method and data-generating scenario.

The performance of each approach is compared based on two criteria. The first is with regards to the accuracy in identifying the subgroup of patients who will benefit from the treatment under the true simulation model. In particular, we define those who truly benefit from the intervention as those covariate values such that the "oracle" average benefit score $K^{-1} \sum_{t=1}^{K} \Delta(\boldsymbol{X}, t) = K^{-1} \sum_{t=1}^{K} \boldsymbol{\gamma}_{0t}' \boldsymbol{X} > 0$. Hence the accuracy over the simulations is $acc = 500^{-1} \sum_{s=1}^{500} I\{\text{sgn}(\sum_{t=1}^{K} \boldsymbol{\gamma}_{0t}' \boldsymbol{X}_s^* > 0) = \text{sgn}(\sum_{t=1}^{K} \hat{\boldsymbol{\gamma}}_{ts}' \boldsymbol{X}_s^* > 0)\}$ where $\hat{\boldsymbol{\gamma}}_{ts}$ is the estimate of $\boldsymbol{\gamma}_{0t}$ for a particular method based on the sth simulated dataset and $\boldsymbol{X}_s^*$ is the sth independent test set of size 100,000. Thus an accuracy of 1 indicates the estimated IIR identifies all patients whose average potential outcomes under the intervention $K^{-1} \sum_{t=1}^{K} Y_t^{(1)}$ are higher than their potential outcomes under no intervention $K^{-1} \sum_{t=1}^{K} Y_t^{(-1)}$. The second criteria is the overall benefit of the assigned treatments defined as

$$\frac{1}{N_t^*} \sum_{i=1}^{N_t^*} \frac{Y_{it}}{A_i \pi(\boldsymbol{x}_i) + (1 - A_i)/2} \{\text{sgn}(\widehat{d}(\boldsymbol{x}_i)) = A_i - \text{sgn}(\widehat{d}(\boldsymbol{x}_i)) \neq A_i\}, \tag{13}$$

where $N_t^* = 100{,}000$ is the sample size in the test set for time t. The quantity (13) is a measure of the improvement in terms of average potential outcomes for patients whose treatment statuses are concordant to the estimated IIR versus pa-

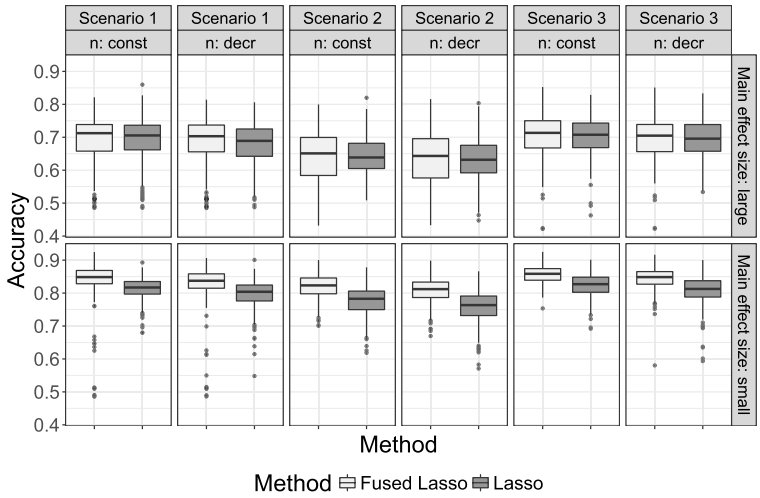

FIG. 1. *Accuracy results from the simulation for each method and scenario. The accuracy of the estimated subgroups is evaluated on an independent test set of size* 100,000 *for each simulation and each scenario. The columns of plots labeled "n: const" have sample sizes which are fixed for all time points, that is, have no dropout. The columns of plots labeled "n: decr" have sample sizes which decrease over time, that is, a number of samples drop out after each time point.*

tients whose treatment statuses are not. Thus higher values of (13) indicate more effective IIRs that yield better overall patient outcomes. Maximization of (13) is equivalent in an asymptotic sense to maximization of the value function $\mathbb{V}_u$. We compare the values of (13) under both our proposed and ad hoc approaches with that under the oracle IIR, $\operatorname{sgn}(K^{-1}\sum_{t=1}^{K}\boldsymbol{\gamma}_{0t}'\boldsymbol{X})$.

The accuracy results evaluated on the test set for all simulations are depicted as boxplots in Figure 1. In the simulation, the true subgroups have equal proportions in the treatment group and the control group and hence random guessing for each patient would result in an accuracy of 0.50 on average. The benefit of the proposed estimator is more pronounced when the size of the main covariate effects on the outcome is smaller. For scenarios with a large main effect size, the results can be improved by constructing a model for the main effects $\phi_t(\boldsymbol{X})$ to use as the augmentation function. The proposed estimator performs better when the sample size decreases over time, as the fused lasso penalty takes advantage of the underlying smoothness in the interactions over time, thus information from time points with more observations is leveraged in later time points with fewer observations. In all small main effect scenarios and all decreasing sample size scenarios the proposed estimator always results in a higher proportion of optimal IIRs and tends to be less variable over the simulations.

It is possible for two models to yield a similar accuracy and at the same time result in different average improvements in outcomes across a population. Thus in evaluating benefit score estimation approaches, evaluating the improvement in

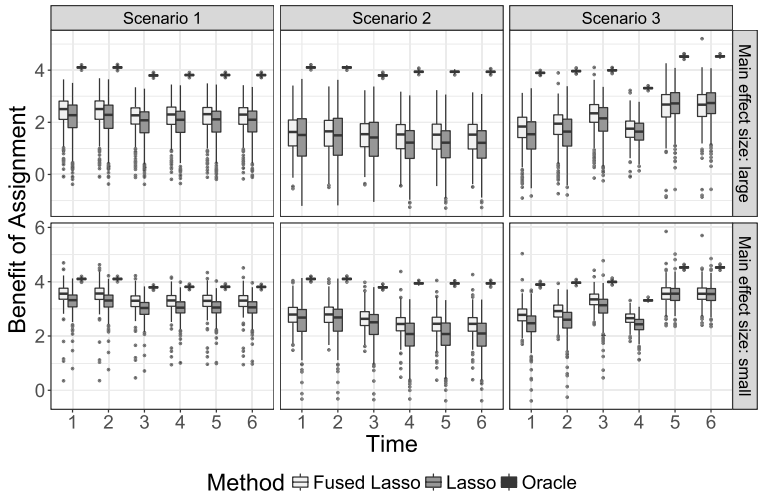

FIG. 2. *Benefits of treatment assignment results from the simulation for each time point and method with sample sizes that decrease by* 10 *after each time point. The benefit of assignments is evaluated on an independent test set of size* 100,000 *for each simulation and each time point.*

outcomes is highly informative about the quality of subgroups found. The results for the estimated benefit of treatment assignments for the scenarios with sample sizes decreasing over time are depicted as boxplots in Figure 2 and the corresponding results for scenarios with no decrease in sample size are in Figure 3. These

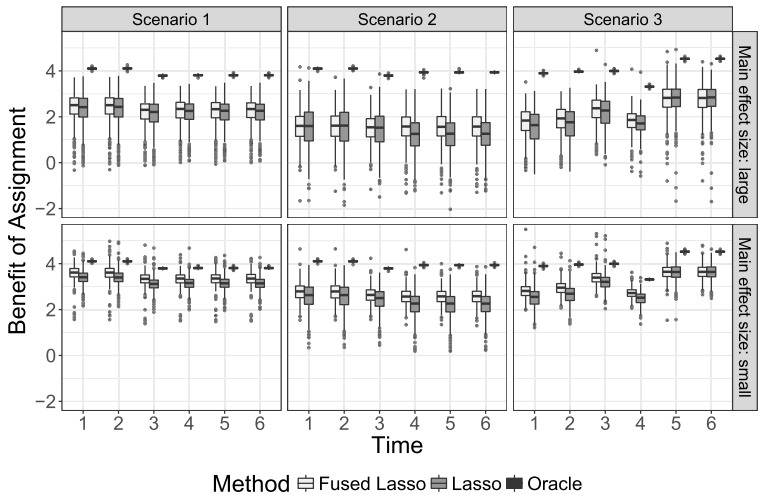

FIG. 3. *Benefits of treatment assignment results from the simulation for each time point and method with constant sample sizes in time. The benefit of assignments is evaluated on an independent test set of size* 100,000 *for each simulation and each time point.*

boxplots are based on the independent test set. The simulation results for the benefit of treatment assignments show a similar story as for accuracy. For simulation scenarios with large main covariate effects, there are a few time points for both scenarios 2 and 3, the scenarios with the least smoothness in interactions, when the proposed approach results in a lower proportion of optimal IIRs, especially in the last two time periods for scenario 3. The proposed fused lasso approach performs better for time periods where the true effects change slowly over time, and does worse only when the true interaction effects have both a rapid change in a particular time point and the magnitudes of the main effects are large, as can be seen in time points 5 and 6 in simulation scenario 3 with large main effects. There is more variability in the results for scenarios with large main effects; this is due to the fact that the signal-to-noise ratio of the interaction effects, defined as $\sqrt{\text{Var}(\boldsymbol{\gamma}_0(t)'\boldsymbol{X} \cdot A/2)}/\sqrt{\text{Var}(\varepsilon_t + \phi_t(\boldsymbol{X}))}$, is lower when the main effects are larger, all else being equal. As mentioned previously, the decrease in performance when main effects are large may be remedied by specifying a model for the main effects to be used as an augmentation function $a_t(\boldsymbol{x})$ for each time point. If specified correctly, or near-correctly, this type of augmentation function can mitigate the variability in estimating the benefit scores induced by the large main covariate effects (Chen et al. (2017), Tian et al. (2014)).

Note that both the proposed and naive approaches do not reach the performance of the oracle treatment. This is experienced across the literature of subgroup identification. Please see Qian and Murphy (2011), Zhou et al. (2017), Zhou and Kosorok (2017), among many others, for further empirical evidence of this phenomenon. Given the high dimensionality of our simulation and the low signal-to-noise ratio of the interaction effects (ranging between 0.417 and 0.854 across the simulation scenarios), our simulation setup is quite challenging and thus it is not expected for any method to achieve fully optimal results with small samples.

In the Supplementary Material (Huling, Yu and Smith (2019)) we investigate the performance of the proposed and naive approaches when the main effects $\phi_t(\boldsymbol{X})$ have a nonlinear form, in particular $\phi_t(\boldsymbol{X}) = (\beta_{0t}'\boldsymbol{X})^2/4$. The proposed approach is robust to such nonlinearities, enabling the use of the proposed approach for outcomes that have a complex relationship with covariates, provided that $\Delta_t(\boldsymbol{X})$ has a linear form.

5. Analysis of complex case management data. In this section we analyze the implementation of a CCM intervention at UW Health. The intervention is an ongoing effort and was designed as an intensive system of coordinated care for highly complex patients, such as those with many chronic conditions. There is substantial evidence that patients may be more or less likely to benefit from an intervention depending on their personal and clinical characteristics (Hickham et al. (2013)). However, there is a lack of effective tools for choosing candidates for case management, as no studies have rigorously examined patient subgroups to

learn which patients achieve the greatest benefit from enrollment in case management programs. In this section we seek to fill this gap by utilizing the proposed approach of Section 3. Our analysis focuses on 198 patients enrolled in CCM. In this analysis we seek to use information from these patients to identify what patients are likely to benefit from enrollment in CCM. Information regarding these outcomes was collected once per month for six months from Medicare claims. The timescale of interest is time since enrollment.

Given the size of our pool of control patients and the large number of covariates, we found it more effective to utilize propensity score matching, instead of weighting. Matching can increase the robustness of parametric propensity score modeling and decrease sensitivity to extreme weights as recommended by Rosenbaum and Rubin (1983) and Imbens and Rubin (2015), Chapter 15. Thus, we utilize the matching-based objective (9) for estimation.

Seven hundred fifty-three variables from Medicare claims and electronic health records were considered in the propensity score model. These 753 variables were screened from a much larger pool of variables based on knowledge and experience of the case managers involved in the CCM implementation and also based on a measure of their potential as confounders based on the procedure of Schneeweiss et al. (2009). Further, based on consultation with case managers, we know that certain patient characteristics such as healthcare utilization, specific diagnoses, and social issues are considered when making enrollment decisions. These characteristics are also included in the 753 variables. Variables were selected into the propensity score model using the lasso with the tuning parameter selected by 10-fold cross validation, resulting in a propensity score model with 41 variables. Summary information regarding standard predetermined demographic and medical covariates of the CCM group and comparison group are presented in Table 1; however, these covariates are not the 41 selected in the propensity score model. Each case is matched to C control patients with $C = 4$ in most cases and $C = 3$ when 4 close matches were not available.

The baseline information is collected over a period of 12 months and the outcomes are collected monthly for a total of six months starting after the baseline period. The endpoints of interest are the monthly average event rate, the monthly average number of event days, and the monthly average Medicare payment amount in thousands of dollars. An event is defined as a hospitalization or visit to the Emergency Department. Event days are defined as the number of days spent in an event (i.e., a hospitalization or an Emergency Department visit). The mean event rate is between 0.068 (0.005) and 0.056 (0.005) across the months, the mean number of monthly event days is between 0.175 (0.018) and 0.137 (0.015) over the months, and the mean monthly payment amount is between 655 (51) and 545 (42), where the numbers in parentheses are standard errors. In practice some patients enrolled in CCM rarely actively participate in the intervention. We are interested in understanding the effect of CCM as it is implemented in practice and furthermore the

TABLE 1
Average covariate values for the CCM and comparison groups. Binary covariates are summarized in terms of percents

	CCM $n = 198$	**Comparison** $n = 759$	**P Value**
Sociodemographics			
Mean age (SD)	70.035 (13.603)	72.401 (14.132)	0.031
Female	66.667	64.954	0.714
Non-hispanic white	89.394	93.676	0.055
Other race/ethnicity	10.606	6.324	0.055
Medicaid insurance ever	37.374	31.489	0.137
Disability entitlement	37.374	27.141	0.006
Mean percent of patient zip code with a high school degree or more	93.713	93.047	0.025
Mean HCC score (SD)	3.101 (1.937)	3.121 (2.094)	0.899
Utilization			
Average number of emergency dept visits (SD)	2.076 (2.927)	1.854 (3.073)	0.348
Average number of hospitalizations (SD)	1.394 (1.738)	1.387 (1.591)	0.962
Average number of days in hospital (SD)	7.030 (11.987)	6.827 (10.977)	0.829
Average number of OBS stays (SD)	0.308 (0.630)	0.271 (0.597)	0.462
Average number of days in outpatient observation stay (SD)	0.354 (0.771)	0.381 (1.109)	0.689
Average payment amount (SD)	32.937 (40.285)	36.949 (42.529)	0.218
Chronic conditions			
Congestive heart failure	41.919	32.279	0.014
COPD/Asthma	47.475	46.377	0.845
Chronic kidney disease	49.495	51.252	0.718
Anxiety	52.020	47.694	0.315
ESRD	2.020	7.773	0.006
Anemia	44.444	44.005	0.976
Rheumatoid arthritis/vasculitis	13.131	13.439	1.000
Chronic blood loss anemia	4.040	4.084	1.000
Coagulopathy	9.596	8.432	0.707
Depression	44.949	37.418	0.064
Diabetes with chronic complication	17.677	13.439	0.161
Diabetes without chronic complication	27.273	25.955	0.776
Hypertension	72.222	81.950	0.003
Hypothyroidism	37.374	31.884	0.168
Liver Disease	7.071	6.719	0.987
Lymphoma	1.515	2.635	0.512
Fluid or electrolyte disorders	52.020	49.144	0.521
Metastatic cancer	5.556	5.007	0.896
Other neurological disorders	34.848	32.543	0.596
Obesity	36.869	37.022	1.000
Paralysis	7.576	6.456	0.688
Pulmonary circulation disease	18.182	15.415	0.402

TABLE 1
(*Continued*)

	CCM $n = 198$	Comparison $n = 759$	P Value
Psychosis	26.263	20.158	0.077
Peripheral vascular disease	27.778	30.435	0.522
Renal failure	32.828	35.441	0.546
Solid tumor without metastasis	8.586	7.510	0.722
Valvular disease	18.687	21.344	0.470
Weight loss	17.677	12.516	0.077

effects of CCM are expected by case managers to last after disenrollment. Hence we conduct an intention-to-treat analysis of its clinical effectiveness.

All cases and controls were required to have Medicare claims, to be alive in the baseline period, to have primary medical insurance coverage Medicare Part A and Part B throughout the study period or until death, and for this to be their first enrollment in the CCM intervention. Each patient is required to have at least 12 months of baseline data available prior to enrollment, or prior to a potential enrollment time for control patients. Patients are enrolled at different calendar times, with a few patients being enrolled within six months of the data collection time, resulting in recently enrolled patients having missing outcome information (20 dropouts). A smaller number of patients were lost to follow-up due to death (4 dropouts). The dropout rate of the case and control groups is remarkably similar. As covariate information contains missing values, we utilize a simple imputation approach. Covariate values are imputed within each decile of a Hierarchical Condition Categories risk variable (Pope et al. (2000)) and a missingness indicator is included. However, not all missingness indicators were screened into the smaller list of 753 variables.

A model estimated by minimizing (7) for each of the three outcomes separately, resulting in three sets of estimated benefit scores. The squared error loss was used for each outcome. Although the full regression model for each of these outcomes may be complex, our estimation approach is robust to nonlinearities in the main effects, furnishing confidence in the appropriateness of our modeling approach. See the Supplementary Material (Huling, Yu and Smith (2019)) for simulation evidence of such robustness. The two tuning parameters controlling the level of sparsity in the estimates and the level of similarity of the estimates over time are selected by five-fold cross validation. The folds were chosen by randomly selecting cases with all of their matched controls. We investigated use of the naive approach (12), however, this resulted in no selected variables.

The estimated subgroups are relatively consistent across all three outcomes (for each of the six months between 53.9% and 55.5% of all patients were assigned the

same intervention status for all of the three models), lending credence to the validity of the results. For the payment amount model, between 366 and 370 variables were selected for each of the six months, between 248 and 250 for the event rate model, and between 299 and 305 were selected for the event day count model. The number of variables selected in common across all three models was 101, further suggesting consistency of estimated subgroups. For each of the three models, the vast majority of selected variables had coefficients estimated to be nearly constant for all six months, indicating a consistency of the subgroups over time. Furthermore, the models for different outcomes tended to result in estimated decision rules that were consistent for individuals. The pairwise correlations of estimated benefit scores (averaged over the six months) for the different models ranged from 0.504 to 0.79, indicating a reasonably high concordance with each other. The overall IIRs for each outcome are constructed as $\widehat{d}_{\text{avg}}(\boldsymbol{x}) = \text{sgn}(\widehat{B}(\boldsymbol{x}))$, in other words a patient with a positive average benefit score, that is, $\widehat{B}(\boldsymbol{x}) > 0$, will be "recommended" CCM and any patient with a negative benefit score, that is, $\widehat{B}(\boldsymbol{x}) \leq 0$, will be recommended not to enroll in CCM.

As mentioned above, many covariates were selected for each model. Here we describe some of the effects of covariates which had the biggest impact on the estimated IIRs. Characteristics of patients who are estimated to benefit from CCM include those: with lupus, who had Bilirubin tests ordered, with a high number of congestive heart failure admissions, who infrequently visited the electronic health record online tool for communicating with their doctors, a high distinct number of analgesic or anesthetic prescriptions, and those with a high number of unique therapeutic or pharmaceutical drug classes administered in the baseline period. Characteristics of patients who are estimated *not* to benefit from CCM include those: who had Emergency Department visits in the baseline period that did not lead to hospitalization, who have a low resting pain level, those with a low recent Body Mass Index, with a low number of unique prescribing providers, with a large number of blood calcium level tests ordered, and those with a low fall risk. These characteristics are sensible and indicate that patients with a high level of complexity in their care with many different potentially interacting prescriptions, for example, are likely to benefit from CCM, which is designed to mitigate these types of problems. Similarly, those who have indication of a low level of complexity are less likely to benefit from CCM.

To determine the impact of the estimated IIRs, we evaluate the difference of the average outcomes in the subgroup of patients whose assigned intervention statuses are concordant to the model recommendations, compared with the subgroup of patients whose assigned intervention statuses are not. If our model finds a meaningful subgroup, we would expect a difference. Furthermore, we also look at refined subgroups of patients. In particular we want to evaluate whether patients who we recommend CCM and actually received CCM fare better than patients we recommend CCM and did not receive it. If this is the case, we can have confidence that our recommendations to CCM are helpful. Similarly we want to evaluate whether

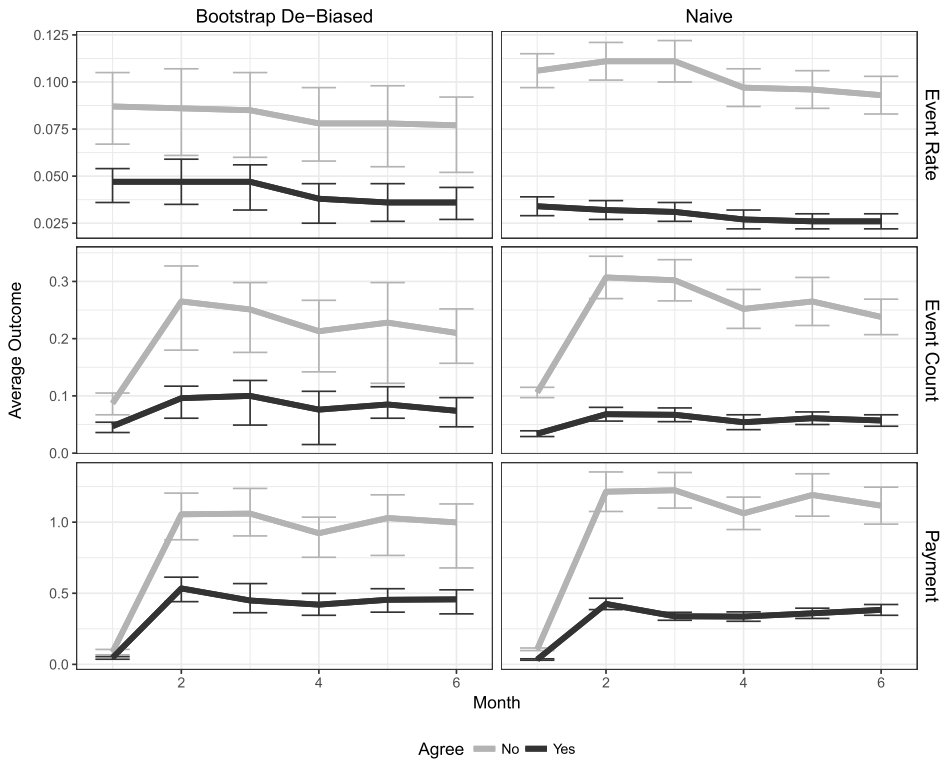

FIG. 4. *Dark gray lines represent the empirical averages over time among the subgroup of patients whose recommended intervention status (by the model constructed with the corresponding outcome) agrees with their actual intervention status and the light gray lines represent the empirical averages over time among the subgroup of patients whose recommended intervention status disagrees with their actual intervention status. The error bars for the bootstrap debiased means represent a* 95% *confidence interval estimated using bootstrap resampling. The error bars for the naive means represent the mean plus and minus one standard deviation. The payment outcome is in thousands of dollars.*

patients we recommend not enroll in CCM and actually did not receive CCM fare better or worse than patients we recommend not enroll in CCM and did, in fact, enroll. It would be reassuring if these two groups have similar outcomes, although it is possible that CCM could result in increased healthcare utilization. Figures 4 and 5 display these comparisons. In addition to the empirical averages, the bootstrap debiased estimates of these outcome means are displayed. It is important to note that the purpose of the debiasing step is to estimate a population-level quantity and hence while it results in estimates of the benefit of the estimated IIRs which are less pronounced, it does not change the subgroups themselves.

In Figure 5 we can see that after adjusting for bias due to overfitting, those who were recommended to be enrolled in CCM and were actually enrolled in CCM have fewer events, fewer event days, and smaller payment amounts than those

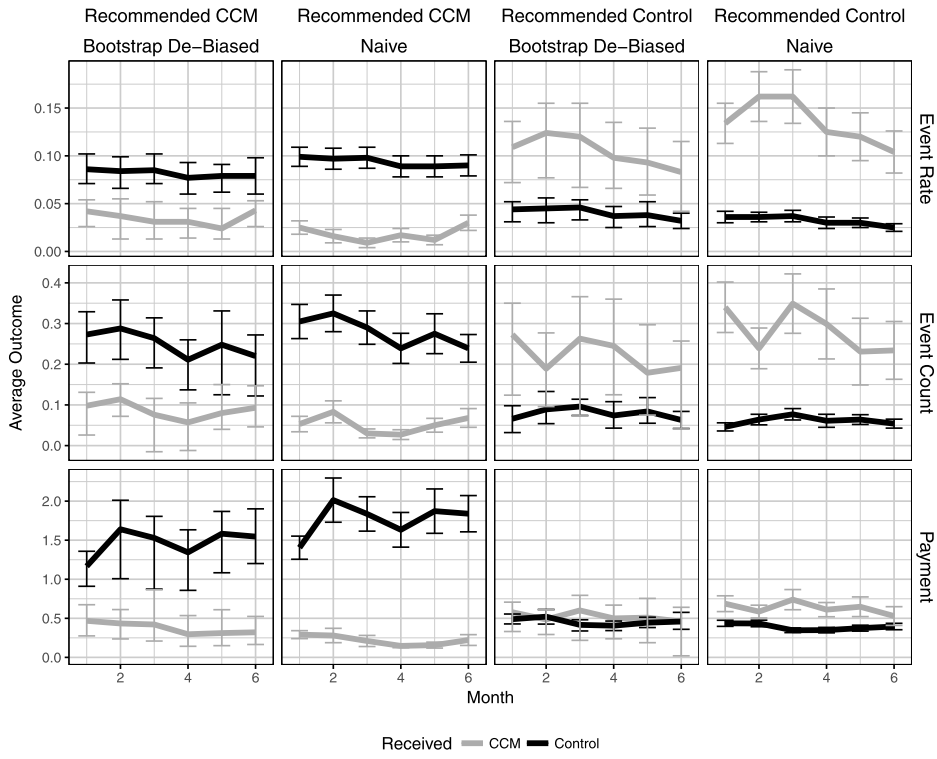

FIG. 5. *Dark gray lines represent the empirical averages over time among various subgroups of patients who were in the control group and the light gray lines represent the empirical averages over time among various subgroups of patients who were enrolled in CCM. These averages are stratified based on the recommendations of the estimated IIRs. All outcomes are on a monthly average basis and payments are in thousands of dollars per month. The error bars for the bootstrap debiased means represent a* 95% *confidence interval estimated using bootstrap resampling. The error bars for the naive means represent the mean plus and minus one standard deviation. The payment outcome is in thousands of dollars.*

who were recommended CCM and were not actually enrolled in CCM. However, patients who were not recommended to be enrolled in CCM and did enroll in CCM did not see any significant worsening of their overall payments after adjusting for overfitting, although some patients tended to have more events in earlier months when enrolled in CCM and not recommended CCM. In Figure 4, we see that, for months two through six, the concordant patients tended to have better outcomes than discordant patients. This indicates that implementing our estimated IIR should result in better outcomes overall. It is also interesting to note that there is less overall benefit to CCM in the first month, substantiating the notion that CCM interventions require time for patients to accumulate health improvements. Additional information on a health system's application of an estimated IIR (i.e., benefit score) will be available at https://www.hipxchange.org/BenefitScore.

6. Conclusion. In this paper we have introduced an approach for modeling the heterogeneity of longitudinal intervention effects. The proposed approach can be effective in scenarios with a small sample size, many longitudinal observations, and many covariates and can be used for subgroup analyses of data from either randomized controlled trials or observational studies. Health systems across the US are actively assessing the effectiveness of their interventions and seeking new ways of improving the implementation of them. Our proposed method for identifying subgroups of patients who benefit the most from such interventions can provide a new avenue for the improvement of health system interventions more broadly. Currently our approach only considers scenarios where the effects governing the subgroups are linear; however, it may be appropriate in some circumstances to consider a more flexible model for the benefit score. This would require more involved considerations to ensure smoothness of the estimated regression function over time.

The intervention studied in this paper was not plagued by informative censoring issues, such as death that may be attributable to the intervention of interest. This may not be an issue in many studies of health system interventions, however, such issues may arise for studies with extraordinarily long follow-up or studies involving interventions with greater health risks. Handling such scenarios would require extensions of our framework to help mitigate from biases resulting from informative censoring. Depending on the type of intervention and the nature of the censoring involved, an inverse-probability-of-censoring weighting approach or a semi-competing risks approach may be warranted.

Acknowledgments. We are deeply appreciative of the constructive comments of the anonymous referees.

SUPPLEMENTARY MATERIAL

Supplement A: "Fused comparative intervention scoring for heterogeneity of longitudinal intervention effects" (DOI: 10.1214/18-AOAS1216SUPPA; .pdf). We provide derivation of the validity of the matching version of our estimator and additional simulation results under nonlinear main effects.

Supplement B: personalizedLong_0.0.1 (DOI: 10.1214/18-AOAS1216SUPPB; .zip). We provide an `R` implementation of the proposed methodology.

REFERENCES

BLUMENTHAL, D., ANDERSON, G., BURKE, S., et al. (2016). Tailoring complex-care management, coordination, and integration for high need, high cost patients. Technical report, Discussion paper, National Academy of Medicine, Washington, DC.

BODENHEIMER, T. and BERRY-MILLETT, R. (2009). Care management of patients with complex health care needs. In *The Synthesis Project*. Research synthesis report 19. Robert Wood Johnson Foundation, Princeton, NJ.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B., ECKSTEIN, J.. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534

CHEUNG, W. Y., NEVILLE, B. A., CAMERON, D. B., COOK, E. F. and EARLE, C. C. (2009). Comparisons of patient and physician expectations for cancer survivorship care. *J. Clin. Oncol.* **27** 2489–2495.

COHEN, S. B. and YU, W. (2012). Statistical brief # 354 2008–2009. Agency for Healthcare Research and Quality, Washington, DC.

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

FAN, J. and ZHANG, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface* **1** 179–195. MR2425354

FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689

HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HICKHAM, D., WEISS, J. W., GUISE, J., BUCKLEY, D., MOTU'APUAKA, M., GRAHAM, E., WASSON, N. and SAHA, S. (2013). Case management for adults with medical illness and complex care needs. Comparative Effectiveness Reviews No. 99, AHRQ Publication No. 13-EHC031-EF. Accessed on November 13, 2015. Available at http://www.ncbi.nlm.nih.gov/pubmed/23346604.

HULING, J. D., YU, M. and SMITH, M. (2019). Supplement to "Fused comparative intervention scoring for heterogeneity of longitudinal intervention effects." DOI:10.1214/18-AOAS1216SUPPA, DOI:10.1214/18-AOAS1216SUPPB.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

POPE, G. C., ELLIS, R. P., ASH, A. S., AYANIAN, J. Z., BATES, D. W., BURSTIN, H., IEZZONI, L. I., MARCANTONIO, E. and WU, B. (2000). Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Health Economics Research, Inc., Waltham, MA.

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351

ROBINS, J. M. and FINKELSTEIN, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56** 779–788.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

SCHNEEWEISS, S., RASSEN, J. A., GLYNN, R. J., AVORN, J., MOGUN, H. and BROOKHART, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20** 512.

STANGE, K. C. (2009). The problem of fragmentation and the need for integrative solutions. *Ann. Fam. Med.* **7** 100–103.

STANGE, K. C. (2012). In this issue: Challenges of managing multimorbidity. *Ann. Fam. Med.* **10** 2–3.

TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. MR3293607

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2012). *Multiple Chronic Conditions Initiative. Private Sector Activities Focused on Improving the Health of Individuals with Multiple Chronic Conditions: Innovative Profiles*. U.S. Dept. Health and Human Services, Washington, DC. Available at http://www.hhs.gov/ash/initiatives/mcc/.

ZHOU, X., KOSOROK, M. R. (2017). Augmented outcome-weighted learning for optimal treatment regimes. Preprint. Available at arXiv:1711.10654.

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. MR3646564

J. D. HULING
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
USA
E-MAIL: huling.7@osu.edu

M. YU
DEPARTMENT OF BIOSTATISTICS &
MEDICAL INFORMATICS
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: meyu@biostat.wisc.edu

M. SMITH
DEPARTMENT OF POPULATION HEALTH SCIENCES
DEPARTMENT OF FAMILY MEDICINE &
COMMUNITY HEALTH
HEALTH INNOVATION PROGRAM
UNIVERSITY OF WISCONSIN-MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: maureensmith@wisc.edu