Published in final edited form as:
IBM J Res Dev. 2018 ; 62(1): . doi:10.1147/JRD.2017.2769320.

An interpretable health behavioral intervention policy for mobile device users

X. Hu, **P.-Y. S. Hsueh**, **C.-H. Chen**, **K. M. Diaz**, **F. E. Parsons**, **I. Ensari**, **M. Qian**, and **Y.-K. K. Cheung**

Abstract

An increasing number of people use mobile devices to monitor their behavior, such as exercise, and record their health status, such as psychological stress. However, these devices rarely provide ongoing support to help users understand how their behavior contributes to changes in their health status. To address this challenge, we aim to develop an interpretable policy for physical activity recommendations that reduce a user's perceived psychological stress, over a given time horizon. We formulate this problem as a sequential decision-making problem and solve it using a new method that we refer to as threshold Q-learning (TQL). The advantage of the TQL method over traditional Q-learning is that it is "doubly robust" and interpretable. This interpretability is achieved by making model assumptions and incorporating threshold selection into the learning process. Our simulation results indicate that the TQL method performs better than the Q-learning method given model misspecification. Our analyses are performed on data collected from 79 healthy adults over a 7 week period, where the data comprise physical activity patterns collected from mobile devices and self-assessed stress levels of the users. This work serves as a first step toward a computational health coaching solution for mobile device users.

Introduction

There is a growing interest in using mobile technologies, such as smartphones and wearable devices, for collecting health-relevant data and delivering health interventions [1]. In particular, mobile technologies that are able to continuously collect data over an extended period of time could deliver interventions in an adaptive manner. As a result, behavioral scientists and healthcare researchers are interested in using mobile technologies as an interactive platform for supporting health behavior change [2–5]. For example, by combining a fitness tracking wearable device (such as those made by Fitbit) that records a user's exercise patterns with a fitness tracking smartphone app that records the user's caloric intake and health status, the user can be provided with personalized coaching plans that suggest exercise or food based on the person's ongoing performance. As another example, data collected from mobile devices can be used to personalize messages to patients [6] and support their self-care management of such concerns as obesity [7], chronic obstructive pulmonary disease [8], and post-traumatic stress disorder [9].

Researchers have used mobile devices to deliver behavioral interventions in the areas of substance use disorders [10], physical activities [11], mental health [12], cardiovascular disease management [13], and stress management [14]. The ability to use algorithms to prescribe interventions based on the behavioral data collected on mobile devices is still an active area of research. The early evidence has started to emerge and stimulate the field through best practices [15, 16]. More recently, precision health applications have been developed to fuel closed-loop feedback systems that can help patients accurately follow the estimated treatment rules. Despite that the developed applications have performed well on observational data sets, the actual adoption in real life is often slow, and patient compliance is low. As demonstrated in recent large-scale trials such as the Innovative Approaches to Diet, Exercise, and Activity (IDEA) trial and the Tailored Rapid Interactive Mobile Messaging (TRIMM) trial for weight control [17, 18] and the Scripps Health study for healthcare cost and utilization control [19], it is insufficient to simply rely on unmasked data and algorithmic output from devices to improve outcomes. The data-driven insights need to be delivered with their interpretations in personal contexts (e.g., through text messaging or coaching advice with live discussion) to help patients. It is therefore imperative to develop interpretable analytics approaches that can provide highly relevant, person-specific insights [20, 21]. As medical decisions often involve unclear choices, developing interpretable analytics that consider human psychology is necessary to implement effective patient engagement technologies beyond simply addressing clinical efficacy.

Two key challenges emerge. One concerns how to generate adaptive feedback for a target user’s choice of actions by learning from real-world experiences that have led to desired health outcomes. The other regards how to make the feedback more interpretable by representing the learned behavior intervention decision rules in a simpler form that can be digested by users with ease and providing them with additional actionable guidance such as performance sub-goals. Both challenges lead to many subsequent questions regarding the interpretability of patient-generated health data in real-world healthcare applications.

To address the first challenge, a long line of research studies in adaptive clinical trials has established the benefits of incorporating adaptive treatment regimes in sequential multiple adaptive randomized trials (SMARTs) to achieve the best mean outcomes [22, 23]. More recently, researchers have considered applying adaptive rules to the delivery of behavioral intervention [24]. To address the second challenge, recent studies have started paying attention to the development of interpretable machine learning models for healthcare applications [25, 26]. However, the definitions of interpretability are often targeted at measuring some model characteristics (e.g., complexity), but unrelated to actual user interpretation [20, 21]. In the healthcare domain, an interpretable decision rule has the benefit of being easily communicated with domain experts and allowing machine learning algorithms to be trusted and widely implemented. There is some recent research work trying to address this problem, among which tree-based methods are known for interpretability and are easily deployed and disseminated. This approach was first introduced by Laber and Zhao [27] for one-stage treatment assignment problems. Zhang et al. [28] proposed a decision list based method to estimate the optimal dynamic treatment regimens. The decision list is easy to interpret and can be expressed as a flow chart of if-else clauses. This method has interpretability, but it requires a pre-mined class of decision-list-based regimes. In addition, tree-based methods, although they have low bias, are often highly variable. To bridge the gaps, in this study we consider ways that can help learn the interpretable decision rules and to generate adaptive and interpretable feedback according to the observations of target users.

Our proposed method has an easier formulation compared to the above methods and is expected to have balanced bias-variance trade-off.

We propose a learning framework to construct a health behavioral intervention policy (i.e., a set of decision rules that specifies actions to be taken under different conditions) that is adaptive and interpretable. By “interpretable,” what we mean is that the rules can not only be expressed using terms that are generally understood by a layperson, but that the meaning of the terms can be made concrete, rather than abstract. For example, a rule that says, “do moderate exercise for 30 minutes” if “your stress level is greater than 5” is more interpretable than “do moderate exercise for 30 minutes” if “the combination of the numbers you see on your screen has a value greater than 7.” In the latter case, the number that was displayed on the screen might be the result of evaluating a complicated and abstract mathematical expression that cannot be easily described in terms that a lay person could relate to.

Our intervention policy comprises a series of dynamic, interpretable decision rules that can be used to recommend behavior to users that can lead them to achieve their desired health outcomes. Q-learning (a model-free reinforcement learning technique) is commonly used to solve such sequential decision-making problems [29]. Q-learning uses a sequence of state-action-reward triples to determine a sequence of decision rules that maximizes cumulative rewards. However, the resulting decision rules generated by the learning algorithm can be difficult to explain, rendering them unappealing in practical settings. In the domain of healthcare and health behavior, recommended actions need to be made understandable to patients and practitioners to promote adoption and engagement with them.

In this paper, we present a new method for estimating an interpretable policy for health behavior interventions. This method applies a linear approximation of the well-known Q-learning algorithm with statistical regression modeling approaches [30, 31]. By using a regression model to approximate the Q-learning function, we can produce a result that is interpretable. The novelty of our work is that the estimated policy is adaptive and interpretable, and provides users with information on performance sub-goals (e.g., intermediate stress reduction). In contrast to a static, one-size-fits-all policy, an adaptive policy uses the most current information from each individual to tailor the intervention based on the individual’s ongoing performance. In the proposed method, we incorporate threshold finding into the Q-learning model and consider the threshold to be the sub-goal option. Therefore, we refer to our method as the Threshold Q-learning (TQL) method.

The remainder of this paper is organized as follows. In the next section, we describe the structure of the data set. Then, we explain the Q-learning and TQL methodological framework. After this, we compare the proposed TQL method with the Q-learning method using Monte Carlo simulations and the stress study data set. In the discussion section, we describe the advantages, potential, and limitations of the TQL method. In the last section, we present summary remarks regarding the TQL method.

The observational data set

Psychological stress contributes to the development and progression of cardiovascular diseases, whereas good health behavior can improve physical health and reduce the negative consequences of stress. Research thus far has focused on the effect of regular exercise (i.e., a subset of physical activity that is planned, structured, and repetitive with the objective of improving or maintaining physical fitness) and physical activity on mental health and demonstrated the benefit of regular exercise and physical activity on emotional well-being and stress reduction [32, 33]. We use data collected from an observational cohort study [34], which aimed to understand the bidirectional relationship between stress exposure/perception and exercise behavior, to illustrate the potential effectiveness of our method. Access to the study data set and information about the study's execution and materials is publicly available at [35]. We refer to this data set as the stress study data set.

The stress study data set comprises daily objective measures of exercise and self-reported stress levels from 79 healthy adults over a 7-week period from January 2014 to July 2015. The healthy adults participating in the study were reported to have exercised at least 6 times per month but did not have regular exercise schedules (e.g., intermittent exercisers). In this study, users were continuously monitored using a wrist-based accelerometer (Fitbit Flex, http://www.fitbit.com/) to detect whether they had engaged in a total of at least 24 minutes of moderate or vigorous physical activity (MVPA) within a 30-minute time interval each day. Each such observed instance of MVPA is referred to as an "MVPA bout" (and is akin to exercise). The duration of each MVPA bout is also recorded. Different levels of physical activity represent different actions that may be taken. Users were asked to report daily on their perceived psychosocial stress via a smartphone-based electronic diary. To measure psychosocial stress, the users were asked how stressed they felt at three randomized times during the day and were asked again at the end of the day how stressful their day had been overall. The stress level is a score ranging from 0 to 10 and is based on self-evaluation. In our method, the reported stress levels represent the individual's health status. We use the end-of-the-day stress level as the health outcome. If this value is missing, then the averaged stress levels during the day are used. The Fitbit Flex wirelessly transmits activity data from users in real time. This prevents loss of data for users who may otherwise not return the device, provides a means to monitor wear compliance in real time, and allows for the continuous assessment of physical activity over numerous weeks. As such, the Fitbit Flex provides numerous benefits as compared to traditional research-grade accelerometers. Researchers have demonstrated that the Fitbit Flex is an accurate and reliable device for measuring physical activity [36–40].

Methodological framework

The data we observe are sequential. To initialize the sequence, information about each user's characteristics and baseline health status (e.g., baseline stress level) is collected. At each stage, an action (e.g., MVPA pattern) and an intermediate health outcome (e.g., stress level reduction from baseline) are observed. In the notation that follows, lower-case variables are the realizations of their corresponding upper-case random variables. For example, o is a particular realization of the random variable O. Additionally, variables in bold font are vectors or matrices (as opposed to scalars). In a T-stage study for each user i, we observe the data

$$\{O_{i1}, A_{i1}, O_{i2}, A_{i2}, \ldots, O_{iT}, A_{iT}, Y_i\}\,.$$

O_{i1} is a scalar of a baseline health outcome measure, and O_{it}, where $1 < t \leq T$, is a scalar of an intermediate health outcome measure at the end of stage t. A_{it} is a scalar of action for user i at stage t. To simplify the problem, we only consider binary actions, $A_{it} = 0$ or 1. Y_i is a scalar of a final health outcome measure for user i we aim to optimize. Let $\mathbf{H}_{it}$ denote the historical information for user i up to stage t. That is, $\mathbf{H}_{it} = \{O_{i1}, A_{i1}, \ldots, O_{it}\}$. In our data set, for each user i, O_{it} is the stress level reduction from baseline at stage t for $1 < t \leq T$, and the final health outcome Y_i is the final stress level reduction from baseline. The propensity score, $P(A_{it}|\mathbf{H}_{it})$, is defined as the probability of assigning some intervention given a user's historical information. In a randomized study, the propensity score is known. In contrast to a randomized study, an observational study is not intervened in but observed. All the related features and the outcome of interest are observed in order to assess the relationship between the exposure and the outcome. Therefore, in an observational study, the propensity score is unknown and to be estimated. In this context, a policy, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T)$, is defined as a set of decision rules that takes the historical information of a user $\mathbf{H}_{it}$ as input and outputs a recommended action A_{it} at stage t for $t = 1, \ldots, T$. Our goal is to estimate the policy, when implemented in the study population, that will optimize the expected outcome $E_{\pi}(Y)$. This policy is called the optimal policy, and denoted by $\pi^* = (\pi_1^*, \ldots, \pi_T^*)$. For notation simplicity, the subscript i in the variables is omitted for the rest of the section.

Q-learning notation and framework

Q-learning is a learning method to construct high-quality policies. The purpose of Q-learning is to model the interaction between an agent and environment. The Q-learning method learns an action-state value function of an agent, referred to as a Q function (quality of the action-state), and uses backward induction to estimate a policy that maximizes a cumulative reward obtained by interacting with the environment. In the health domain, an agent refers to a subject, and environment refers to the system of human body and external source of observations. Under our problem settings, the action is the health intervention, the state is implicitly represented by a function of historical health outcomes and actions, and the reward is the increased value of a beneficial health outcome or the decreased value of a hazardous health outcome. We use the Q-learning methodology framework from [30], and assume a parametric regression model for the Q function at each stage to learn the optimal policy that maximizes the reward for a finite time horizon (i.e., the final stage T is a finite number). In a T-stage study, the Q function is specified as follows.

In a final stage, T, we have

$$Q_T(\mathbf{h}_T, a_T) = E(\mathbf{Y} \mid \mathbf{H}_T = \mathbf{h}_T, A_T = a_T)\,.$$

In a previous stage, t, we have

$$Q_t(\mathbf{h}_t, a_t) = E(\max_{a_{t+1}} Q_{t+1}(\mathbf{h}_{t+1}, a_{t+1}) \mid \mathbf{H}_t = \mathbf{h}_t, A_t = a_t).$$

It can be shown that the optimal policy π^* will choose the action that maximizes the Q function at each stage. That is,

$$\pi_t^*(\mathbf{h}_t) = \arg\max_{a_t} Q_t(\mathbf{h}_t, a_t).$$

We assume a parametric regression model for the Q function and denote the Q function at stage t as $Q_t(\mathbf{h}_t, a_t; \theta_t)$, where θ_t is a vector of regression coefficients in the linear regression model and it consists of both the regression coefficients of the main effect and the action effect; i.e., $\theta_t = (\theta_{t0}, \theta_{t1})$. An example of this linear approximation for the case where $T = 2$ is provided below.

In the second stage,

$$\mathrm{Q}_2(\mathbf{h}_2, a_2; \boldsymbol{\theta}_2) = \mathbf{M}_2^{\mathrm{T}}\boldsymbol{\theta}_{20} + \mathbf{M}_2^{\mathrm{T}} a_2 \boldsymbol{\theta}_{21}.$$

In the first stage,

$$Q_1(\mathbf{h}_1, a_1; \boldsymbol{\theta}_1) = \mathbf{M}_1^{\mathrm{T}}\boldsymbol{\theta}_{10} + \mathbf{M}_1^{\mathrm{T}} a_1 \boldsymbol{\theta}_{11},$$

where $\mathbf{M}_1 = (1, o_1)^{\mathrm{T}}$ and $\mathbf{M}_2 = (1, a_1, o_1, o_2, o_1 o_2, a_1 o_2)^{\mathrm{T}}$. Linear regression can be used to estimate the parameters, denoted by $\hat{\boldsymbol{\theta}}_t^Q$, and the optimal policy at stage, t, can be estimated by

$$\hat{\pi}_t^Q\left(\mathbf{h}_t; \hat{\boldsymbol{\theta}}_t^Q\right) = \mathrm{argmax}_{a_t} Q_t\left(\mathbf{h}_t, a_t; \hat{\boldsymbol{\theta}}_t^Q\right) = I(\mathbf{M}_t^{\mathbf{T}} \hat{\boldsymbol{\theta}}_{t1}^Q > 0).$$

Threshold Q-learning framework

The TQL method is based on Q-learning with a linear approximation and has threshold selection added into the learning process. The benefit of the threshold selection step is to obtain interpretability. In addition, we select the threshold by maximizing a doubly robust estimate of the expected outcome under a policy. “Doubly robust” means that we can still obtain a consistent estimation of the expected outcome as long as the Q functions or the propensity scores are correctly specified. For ease of explaining our framework, we will continue our exposition under the assumption that $T = 2$. To formulate the problem in mathematical terms, we now define R_{it} to be the dichotomized health outcome of the user i at the stage t, where $R_{it} = I(O_{it} > c_t)$, c_t is the threshold to be estimated at each stage, and $I(.)$ is an indicator function. Therefore, $R_{it} = 1$ if the outcome O_{it} exceeds c_t; otherwise, $R_{it} = 0$. The threshold, c_t, can be considered the outcome goal option (e.g., stress level reduction) at stage t. Therefore, R_{it} indicates whether the user i has met the goal in the stage t. We use the notation p_t to denote the propensity score at stage t. The resulting policy takes the form of a linear combination of the dichotomized outcomes. We model the Q function at the second stage as

$$\mathrm{Q}_2(\mathbf{h}_2, a_2; \boldsymbol{\theta}_2, c_1, c_2) = \mathbf{M}_2^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{20}} + \mathbf{S}_2^{\mathrm{T}}(a_2 - p_2)\boldsymbol{\theta}_{\mathbf{21}}$$

and the Q function in the first stage as

$$Q_1(\mathbf{h}_1, a_1; \boldsymbol{\theta}_1, c_1) = \mathbf{M}_1^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{10}} + \mathbf{S}_1^{\mathrm{T}}(a_1 - p_1)\boldsymbol{\theta}_{\mathbf{11}},$$

where $\mathbf{S}_1 = (1, r_1)^{\mathrm{T}}$ and $\mathbf{S}_2 = (1, a_1, r_1, r_2, r_1r_2, a_1r_2)^{\mathrm{T}}$, $r_1 = I(o_1 > c_1)$ and $r_2 = I(o_2 > c_2)$.

The estimated optimal policy in the second stage takes the following form:

$$\pi_2(\mathbf{h}_2; \boldsymbol{\theta}_2, c_1, c_2) = \mathrm{argmax}_{a_2}\mathbf{S}_2^{\mathrm{T}}a_2\boldsymbol{\theta}_{\mathbf{21}} = I(\mathbf{S}_2^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{21}} > 0)$$

The estimated optimal policy at the first stage takes the following form:

$$\pi_1(\mathbf{h}_1; \boldsymbol{\theta}_1, c_1) = \mathrm{argmax}_{a_1}\mathbf{S}_1^{\mathrm{T}}a_1\boldsymbol{\theta}_{\mathbf{11}} = I(\mathbf{S}_1^{\mathrm{T}}\boldsymbol{\theta}_{\mathbf{11}} > 0)$$

The optimal policy is estimated using backward induction in order to maximize the reward over the time horizon. We are interested in the parameters that contain information about how different actions affect the value of the Q function. Therefore, the parameters of our main interest are the regression coefficients $\{\theta_{11}, \theta_{21}\}$ and the threshold parameters $\{c_1, c_2\}$.

The estimation of an optimal policy consists of the estimation of regression parameters $\{\theta_{11}, \theta_{21}\}$ and the estimation of threshold parameters $\{c_1, c_2\}$. The regression parameters are estimated as least squares estimators. The propensity scores are estimated using a logistic regression. The threshold parameters are estimated to maximize a doubly robust estimate of the beneficial expected outcome under a policy [41]. For the policy $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, the doubly robust estimate of the expected outcome is defined as

$$\hat{\mu}_{\pi} = \frac{1}{n}\left(\sum_{i=1}^{n} W_{i2}(Y_i - \hat{Q}_2(\mathbf{h}_{\mathrm{i}2}, a_{\mathrm{i}2})) - \sum_{i=1}^{n} W_{i1}(\hat{Q}_1(\mathbf{h}_{\mathrm{i}1}, a_{\mathrm{i}1}) - \max_{a_{\mathrm{i}2}}\hat{Q}_2(\mathbf{h}_{\mathrm{i}2}, a_{\mathrm{i}2})) + \sum_{i=1}^{n} \max_{a_{\mathrm{i}1}}\hat{Q}_1(\mathbf{h}_{\mathrm{i}1}, a_{\mathrm{i}1})\right),$$

where

$$W_{i2} = \frac{\mathrm{I}(a_{i1} = \pi_1(\mathbf{h}_{i1}))\mathrm{I}(a_{i2} = \pi_2(\mathbf{h}_{i2}))}{\mathrm{p}_1(a_{i1} \mid \mathbf{h}_{i1})\mathrm{p}_2(a_{i2} \mid \mathbf{h}_{i2})}$$

and

$$W_{i1} = \frac{\mathrm{I}(a_{i1} = \pi_1(\mathbf{h}_{i1}))}{\mathrm{p}_1(a_{i1} \mid \mathbf{h}_{i1})}.$$

Assuming the final outcome is a beneficial outcome, the threshold parameters are chosen to maximize $\hat{\mu}$ so that the optimal mean outcome is achieved. The expected outcome under a policy is the value of the policy. The estimate of the expected outcome is nonparametric and does not depend on the linear approximation for the Q function. It is doubly robust with respect to the misspecification of the regression model or the propensity score. Thus, the estimate is used to evaluate policies that are estimated under different model assumptions for the Q functions.

The detailed TQL algorithm is given below.

1. Estimate the propensity score $p_t(\mathbf{H}_t)$ by $\hat{p}_t(\mathbf{H}_t)$ using logistic regression with $\mathbf{A}_t$ as outcome and $\mathbf{H}_t$ as predictors, with $t = 1, 2$.
2. For each pair of $c = (c_1, c_2)$, do the following:
   - a. Regress Y on $\mathbf{M}_2^{\mathrm{T}}\boldsymbol{\theta}_{20} + \mathbf{S}_2^{\mathrm{T}}(a_2 - \hat{p}_2(\mathbf{H}_2))\boldsymbol{\theta}_{21}$ to obtain the estimate $\hat{\boldsymbol{\theta}}_{21}(\boldsymbol{c})$.
   - b. Construct the pseudo outcome $\hat{Y} = Y - \mathbf{S}_2 a_2 \hat{\boldsymbol{\theta}}_{21}(\boldsymbol{c}) + \mathbf{S}_2\hat{\boldsymbol{\theta}}_{21}(\boldsymbol{c}) I(\mathbf{S}_2\hat{\boldsymbol{\theta}}_{21}(\boldsymbol{c}) > 0)$.
   - c. Regress $\hat{Y}$ on $\mathbf{M}_1^{\mathrm{T}}\boldsymbol{\theta}_{10} + \mathbf{S}_1^{\mathrm{T}}(a_1 - \hat{p}_1(\mathbf{H}_1))\boldsymbol{\theta}_{11}$ to obtain the estimate $\hat{\boldsymbol{\theta}}_{11}(\boldsymbol{c})$.
   - d. Construct the policy $\boldsymbol{\pi}(\boldsymbol{c}) = (\boldsymbol{\pi}_1(\boldsymbol{c}), \boldsymbol{\pi}_2(\boldsymbol{c}))$ where $\boldsymbol{\pi}_t(\boldsymbol{c}) = I(\mathbf{S}_t\hat{\boldsymbol{\theta}}_{t1}(\boldsymbol{c}) > 0)$, $t = 1, 2$.
3. Choose the threshold $\hat{\boldsymbol{c}} = (\hat{c}_1, \hat{c}_2)$ that maximizes $\hat{\mu}_{\boldsymbol{\pi}(\boldsymbol{c})}$.
4. Output the estimated policy $\hat{\boldsymbol{\pi}}_1(\mathbf{h}_1; \hat{\boldsymbol{\theta}}_1, \hat{c}_1) = I(\mathbf{S}_1\hat{\boldsymbol{\theta}}_{11}(\hat{c}_1) > 0)$ and $\hat{\boldsymbol{\pi}}_2(\mathbf{h}_2; \hat{\boldsymbol{\theta}}_2, \hat{c}_1, \hat{c}_2) = I(\mathbf{S}_2\hat{\boldsymbol{\theta}}_{21}(\hat{c}_1, \hat{c}_2) > 0)$.

The optimization in Step 3 of the TQL algorithm is performed using a genetic algorithm (GA) [42]. The GA is a heuristic searching algorithm and commonly used to find solutions to optimization questions. In a genetic algorithm, potential solutions are evolved iteratively towards the better ones. The evolution starts from a group of randomly selected solutions and mutates each solution to achieve a better value of the objective function. GA is implemented in the "GA" package available in R, an open-source statistical computing environment [43].

Experiments

Method evaluation based on Monte Carlo simulation

To assess the quality of the estimated policy using the TQL method, we performed a series of Monte Carlo simulations and then compared the expected outcome following the estimated policy identified by the TQL method with that of the policy identified by the Q-learning method. We aim to maximize the final outcome; thus, the estimated policy that results in a high value of the expected outcome is desired. The simulation procedure is specified as the following: We first generated the training data sets and a test data set. Then we estimated the optimal policy in each of the training data sets using the TQL method and the Q-learning method. The estimated policies were evaluated on the test set. The mean and the standard deviation of the expected outcome associated with the estimated policy from the two methods were compared.

To generate a training data set, we generated a set of actions, A_{i1} and A_{i2}, and outcomes, O_{i1}, O_{i2}, and Y_i, independently and identically distributed (*i.i.d.*) for each user *i*. To simulate the actions taken by a user, we considered binary actions, denoted by {0, 1}. A_{i1} is an action generated from a Bernoulli distribution with probability

$$\frac{\exp((1, h_{i1})\boldsymbol{\alpha}_1^{\mathrm{T}})}{1 + \exp((1, h_{i1})\boldsymbol{\alpha}_1^{\mathrm{T}})}.$$

A_{i2} is an action generated from a Bernoulli distribution with probability

$$\frac{\exp((1, \mathbf{h}_{i2})\boldsymbol{\alpha}_2^{\mathrm{T}})}{1 + \exp((1, \mathbf{h}_{i2})\boldsymbol{\alpha}_2^{\mathrm{T}})}.$$

where $\boldsymbol{\alpha}_1 = (0.4, -0.1)$ and $\boldsymbol{\alpha}_2 = (-1.5, 0.2, 1.8, -0.1)$.

The baseline stress level, O_{i1}, was *i.i.d.* generated from a normal distribution N(3, 1). The intermediate stress level, O_{i2}, was *i.i.d.* generated from a normal distribution N(0.2O_{i1}, 1), so O_{i2} depends only on O_{i1} but not A_{i1}. We generated the final outcome, Y_i, from a model with a linear and a nonlinear action effect function separately. Specifically, in the model of a linear action effect function

$$Y_i = (1, A_{i1}, O_{i1}, O_{i2}, O_{i1}O_{i2}, A_{i1}O_{i2})^{\mathrm{T}}\boldsymbol{\theta}_{20} + (1, A_{i1}, O_{\mathrm{i}1}, O_{\mathrm{i}2}, O_{\mathrm{i}1}O_{\mathrm{i}2}, A_{i1}O_{\mathrm{i}2})^{\mathrm{T}}(A_{i2} - p_2(\mathbf{H}_{i2}))\boldsymbol{\theta}_{21} + \varepsilon_{\mathrm{i}},$$

where $\boldsymbol{\theta}_{20} = (2.6, 0.4, -2.2, -0.03, 0.3, 0.02)$, $\boldsymbol{\theta}_{21} = (-0.4, -0.3, 1.5, -0.3, 0.03, 0.06)$ and $\varepsilon_{\mathrm{i}} \sim N(0, 1)$.

In the model of a nonlinear action effect function

$$Y_i = (1, A_{i1}, O_{i1}, O_{i2}, O_{i1}O_{i2}, A_{i1}O_{i2})^{\mathrm{T}}\boldsymbol{\theta}_{20} + {O_{i2}}^2(A_{i2} - p_2(\mathbf{H}_{i2}))\theta_{21} + \varepsilon_{\mathrm{i}},$$

where $\boldsymbol{\theta}_{20} = (2.6, 0.4, -2.2, -0.03, 0.3, 0.02)$, $\theta_{21} = 1$ and $\varepsilon_{\mathrm{i}} \sim N(0, 1)$.

We used the training sample size of 50, 100, and 500 to account for possible effects caused by different sample sizes. We also simulated a large test data set with a sample size of 2000. A summary of the mean and standard deviation of the expected outcome following the estimated policy using different methods, sample sizes, and models is provided in Table 1.

Two models were used to generate the data. One used a linear action effect function (referred to as the linear model), and the other used a nonlinear action effect function (referred to as the nonlinear model). Table 1 indicates that the TQL method results in a higher expected outcome on average compared to the Q-learning method in the nonlinear model setting. In the linear model setting, the TQL method and the Q-learning method result in similar expected outcomes on average. Additionally, the standard deviation of the expected outcome using the TQL method is similar to the one using the Q-learning method in the linear model case. However, in the nonlinear case the standard deviation of the expected outcome using the TQL method is higher.

Figure 1 shows the boxplot of the expected outcome following the estimated optimal policy over 500 Monte Carlo replications [44] when the training sample size is 500. It is worth noting that in the nonlinear model case, the policy derived from the TQL method is less stable but often yields higher expected outcome compared to the one derived from the Q-learning method. This is because the threshold is selected by maximizing a doubly robust estimate of the expected outcome. As a result, the TQL method is expected to be more robust to model misspecification (i.e., less biased), by “paying a price” of larger variance as compared to the Q-learning method. The novelty of the TQL method is the added interpretability of the policy to provide interpretable guidance. This simulation study shows that in addition to interpretability, the TQL method can also yield policies that can lead to desirable expected outcomes.

Policy estimation from observational data

In this section, for the purpose of illustrating the proposed methodology, we applied the TQL method to the stress study data set. That is, we show how the TQL method can be used to generate a policy for recommending an action to a user in each decision stage, so as to maximize their final stress level reduction from baseline.

We considered a two-stage decision problem, where the time interval for each stage is 1 week. We used the first 3 weeks of study users’ data as the training data and the fifth to the seventh week of the data as the test data. At the start of the first stage, we observed the baseline stress level O_{i1}, which is an averaged stress level in the first week. For each stage, we observed the user’s actual physical activity level and stress level reduction from baseline. The physical activity level is converted into a binary action, A_{it}, by computing the dichotomized MVPA bout duration, using a cutoff of 150 minutes per week. $A_{it} = 1$ means “active.” It indicates that the weekly MVPA bout duration is greater than 150 minutes. $A_{it} = 0$ means “inactive.” It indicates that the weekly MVPA bout duration is not greater than 150 minutes. We choose the cutoff of 150 minutes since the American Heart Association recommends 150 minutes of moderate exercise per week for an adult to improve cardiovascular health [45]. We observed the averaged self-reported stress level reduction from baseline in the second week and in the third week. Using the difference of the recorded stress level can adjust for heterogeneity of self-reports across users.

After applying our TQL method and the Q-learning method separately on the stress study data set, we compared the two resulting policies for interpretability. The policy generated by the TQL method is shown in Figure 2. The interpretation of the estimated policy in Figure 2 is as follows: In the first stage, the user is suggested to be inactive regardless of the baseline stress level. In the second stage, if a user's baseline stress level is greater than 0.7, and she does not follow the suggestion of being inactive, then she is suggested to be active in the second stage, regardless of the intermediate outcome. For the case that the user follows the suggestion of being inactive, if his stress level increases from baseline greater than 1.1, then he is suggested to be active in the second stage, otherwise to stay inactive. If the user's baseline stress level is less than or equal to 0.7 and his stress level at the end of the first stage increases greater than 1.1, then he is encouraged to be active in the second stage, otherwise to be inactive. In this situation, whether the user follows the recommended action in the first stage does not matter.

The estimated policy using the Q-learning method is as follows: In the first stage, the user is suggested to be active if the baseline stress level is less than 3.9, otherwise to be inactive. In the second stage, if ($0.33 - 0.06o_1 - 1.01o_2 + 0.17a_1 + 1.08a_1o_2 + 0.16o_1o_2 > 0$), then be active, otherwise be inactive. While the decision rule used in the second stage is implementable, we consider it to be uninterpretable because the expression "$0.33 - 0.06o_1 - 1.01o_2 + 0.17a_1 + 1.08a_1o_2 + 0.16o_1o_2$" does not have a concrete meaning.

As we can see from the descriptions of the two policies, the policy estimated using the TQL method is more interpretable than the one generated using the Q-learning method. Therefore, the policy generated by the TQL method can be more easily translated into a meaningful set of guidelines for the user. To assess the effect of the estimated policies, we find the mean stress level reduction from baseline across the users at the end of the 2-week period following the policy estimated by the TQL method is 3.7, the one following the policy estimated by the Q-learning method is 0.5, while in the original data the mean stress level reduction across the users at the end of the 2-week period is 0.5.

Discussion

In this paper, we introduce a new learning method that not only helps identify an adaptive policy for behavior recommendation, but also generates feedback that is interpretable to a target user. The advantages of the new TQL method are three-fold. First, the policy estimated using the TQL method generates adaptive feedback that can account for individual observations directly. Taking the stress data as an example: At each stage wherein the target user's action and averaged stress level reduction are observed, the expected utilities of the different next-step actions can then be calculated accordingly to drive behavior recommendations. Second, another strength of the TQL method is that it incorporates threshold finding in the problem formulation, which leads to the estimation of performance sub-goals and in turn helps generate detailed guidance for the users. Whether the sub-goal at each stage is achieved or not may affect the next-stage action recommendation. Finally, the policy learned by the TQL method can be interpreted to the target user to help the user understand the recommendation. We believe that policies that are less interpretable are less likely to be adopted by a user because it makes it difficult for them to understand the relationship between their experiences and the resulting recommendation. Thus, the learned policy with concrete meaning can be explained to individual users as a motivator for better compliance to the recommendation.

Despite all the advantages and potentials, there still exist limitations of the TQL method that warrant future investigation. For one, the performance of the TQL method depends on the underlying nature of the data set since we make model assumptions on the Q function. If the linear combination of the indicator functions cannot characterize the true relationship between the stress outcome and the observed history, then the TQL method may not outperform the methods without model assumptions. The doubly robust estimate is used to account for potential confounders. In a future study, it would be important to assess to what extent the findings can be generalizable. In the perspective of the development of interpretable machine learning methods for healthcare, it is also imperative to start developing patient subgroup analysis and calibration methods that can incorporate the evaluation metrics of some interpretability measures to inform on the trade-off between policy performance and interpretability [46].

Going forward, there is certainly still a need to further develop finer-granular strategies to translate the policies into messages or interfaces that can help explain the relationship between recommendations and users' experiences. For example, based on the different motivational frames (e.g., emotional, social, and informational) [47], the adaptive and interpretable feedback would still need to be further tailored in different ways to engage and "nudge" the users more effectively. This is also aligned with the interest of the greater healthcare community in the pursuit of the vision in patient-centered care [48, 49]. Opportunities are especially rich in the application areas that are traditionally performed with survey-based ecological momentary assessment (EMA) [50]. By coupling EMA with mobile devices, the TQL method can help make sense of the incoming streams of repeatedly collected exposure data (e.g., psychosocial stressors) in ecologically valid settings such as home and work. Compared to the survey-based EMA, the TQL-enhanced mobile EMA applications with more contextual information from real-world experience are expected to further reduce recall bias and improve the efficacy of *N*-of-1 study [51, 52]. Users and field validation studies would be needed to demonstrate the value of adaptive and interpretable feedback for improving patient experience, healthcare quality and care coordination in the care management flow.

Conclusion

In an effort to bridge the gap between observed individual-level evidence and practical application, we designed and implemented an interpretable TQL method to estimate a recommendation policy and provide individualized recommendations for stress reduction. To meet the specific challenge of learning an interpretable policy, this work uses observations obtained from a stress study including self-reported stress levels and physical activity measurements from mobile devices. The method estimates a health intervention policy and computes optimal thresholds to set sub-goals. Personalized behavioral recommendations are made based on both the estimated policy and the sub-goals being set, thus avoiding making one-size-fits-all recommendations. Compared to the Q-learning method, the proposed method provides a more interpretable policy, which can in turn serve as the foundation for a more progressive version of behavioral coaching in the future.

Acknowledgments

This work was supported by the following Grants: R01 MH109496, R01 NS072127, R01 HL111195, R21 MH108999, and R01 HL115941.

References

1. Riley WT, Rivera DE, Atienza AA, et al. Health behavior models in the age of mobile interventions: Are our theories up to the task? Translational. 2011; 1(1):53–71.
2. Mohr DC, Cheung K, Schueller SM, et al. Continuous evaluation of evolving behavioral intervention technologies. Amer. J. Prev. Med. 2013; 45(4):517–523. [PubMed: 24050429]
3. Kennedy CM, Powell J, Payne TH, et al. Active assistance technology for health-related behavior change: An interdisciplinary review. J. Med. Internet Res. 2012; 14(3) Art. no. 80.
4. Skinner, C., Finkelstein, J. Review of mobile phone use in preventive medicine and disease management; Proc. IASTED Int. Conf. Telehealth/Assistive Technol; 2008. p. 180-189.
5. Depp CA, Mausbach B, Granholm E, et al. Mobile interventions for severe mental illness: Design and preliminary data from three approaches. J. Nervous Mental Dis. 2010; 198(10):715–721.
6. Piwek L, Ellis DA, Andrews S, et al. The rise of consumer health wearables: Promises and barriers. PLoS Med. 2016; 13(2) Art. no e1001953.
7. Teixeira PJ, Carraça EV, Marques MM, et al. Successful behavior change in obesity interventions in adults: A systematic review of self-regulation mediators. BMC Med. 2015; 13(1) Art. no. 84.
8. Chen, B., Patel, S., Toffola, LD., et al. Long-term monitoring of COPD using wearable sensors; Proc. 2nd Conf. Wireless Health; 2011. Paper 19
9. Lange A, van de Ven JP, Schrieken B. Interapy: Treatment of post-traumatic stress via the internet. Cognitive Behav. Therapy. 2003; 32(3):110–124.
10. Litvin EB, Abrantes AM, Brown RA. Computer and mobile technology-based interventions for substance use disorders: An organizing framework. Addictive Behav. 2013; 38(3):1747–1756.
11. King AC, Hekler EB, Grieco LA, et al. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. PLOS ONE. 2013; 8(4) Art. no. e62613.
12. Ben-Zeev D, Davis KE, Kaiser S, et al. Mobile technologies among people with serious mental illness: Opportunities for future services. Adm. Policy Mental Health Mental Health Serv. Res. 2013; 40(4):340–343.
13. Winter SJ, Sheats JL, King AC. The use of behavior change techniques and theory in technologies for cardiovascular disease prevention and treatment in adults: A comprehensive review. Prog. Cardiovascular Dis. 2016; 58(6):605–612.
14. Christmann CA, Hoffmann A, Bleser G. Stress management apps with regard to emotion-focused coping and behavior change techniques: A content analysis. JMIR mHealth uHealth. 2017; 5(2) Art. no. e22.
15. Hsueh PS, Cheung YK, Dey S, et al. Added value from secondary use of person generated health data in consumer health informatics. Yearb Med. Informat. 2017; 28(1):160–171.
16. Choi E, Schuetz A, Stewart WF, et al. Medical concept representation learning from electronic health records and its application on heart failure prediction. 2016 [Online]. Available: arXiv: 1602.03686.
17. Jakicic JM, Davis KK, Rogers RJ, et al. Effect of wearable technology combined with a lifestyle intervention on long-term weight loss. JAMA. 2016; 316(11):1161–71. [PubMed: 27654602]
18. Lin M, Mahmooth Z, Dedhia N, et al. Tailored, interactive text messages for enhancing weight loss among African American adults: The TRIMM randomized controlled trial. Amer. J. Med. 2015; 128(8):896–904. [PubMed: 25840035]
19. Bloss CS, Wineinger NE, Peters M, et al. A prospective randomized trial examining health care utilization in individuals using multiple smartphone-enabled biosensors. PeerJ. 2016; 4 Art. no. e1554.

20. Lipton ZC. The mythos of model interpretability. Proc. ICML Workshop Human Interpretability Mach. Learn. 2016:96–100.
21. Hsueh PS, Das S, Dey S, et al. Making sense of Patient Generated Health Data (PGHD) with better interpretability: The transition from more to better. Proc. MEDINFO. 2017
22. Cheung YK, Chakraborty B, Davidson KW. Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. Biometrics. 2015; 71(2):450–459. [PubMed: 25354029]
23. Schulte PJ, Tsiatis AA, Laber EB, et al. Q-and A-learning methods for estimating optimal dynamic treatment regimes. Statist. Sci., Rev. J. Inst. Math. Statist. 2014; 29(4) Art. no. 640.
24. Wagner LI, Duffecy J, Penedo F, et al. Coping strategies tailored to the management of fear of recurrence and adaptation for E-health delivery: The FoRtitude intervention. Cancer. 2017; 123(6): 906–910. [PubMed: 28207157]
25. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proc. KDD. 2015:1721–1730.
26. Varshney KR, Alemzadeh H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. 2016 [Online]. Available: arXiv:1610.01256.
27. Laber EB, Zhao YQ. Tree-based methods for individualized treatment regimes. Biometrika. 2015; 102:501–514. [PubMed: 26893526]
28. Zhang Y, Laber EB, Tsiatis A, et al. Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics. 2015; 71(4):895–904. [PubMed: 26193819]
29. Sutton, RS., Barto, AG. Reinforcement Learning: An Introduction. Vol. 1. Cambridge, MA, USA: MIT Press; 1998.
30. Murphy SA. A generalization error for Q-learning. J. Mach. Learn. Res. 2005; 6(Jul):1073–1097. [PubMed: 16763665]
31. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann. Statist. 2011; 39(2) Art. no. 1180.
32. Salmon P. Effects of physical exercise on anxiety, depression, and sensitivity to stress: A unifying theory. Clin. Psychol. Rev. 2001; 21(1):33–61. [PubMed: 11148895]
33. Scully D, Kremer J, Meade MM, et al. Physical exercise and psychological well being: A critical review. Brit. J. Sports Med. 1998; 32(2):111–120. [PubMed: 9631216]
34. Burg MM, Schwartz JE, Kronish IM, et al. Does stress result in you exercising less? Or does exercising result in you being less stressed? Or is it both? Testing the bi-directional stress-exercise association at the group and person (N of 1) level. Ann. Behav. Med. 2017; 51(6):799–809. [PubMed: 28290065]
35. Navarrete, S., Diaz, K. Ecological link of psychosocial stress to exercise. 2016. [Online]. Available: https://osf.io/kmszn
36. Diaz KM, Krupka DJ, Chang MJ, et al. Fitbit: An accurate and reliable device for wireless physical activity tracking. Int. J. Cardiol. 2015; 185:138–40. [PubMed: 25795203]
37. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int. J. Behav. Nutrition Phys. Activity. 2015; 12(1) Art. no. 159.
38. Sushames A, Edwards A, Thompson F, et al. Validity and reliability of Fitbit Flex for step count, moderate to vigorous physical activity and activity energy expenditure. PLOS ONE. 2016; 11(9) Art. no. e0161224.
39. Alharbi M, Bauman A, Neubeck L, et al. Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population. Eur. J. Prev. Cardiol. 2016; 23 Art. no. 2047487316634883.
40. Reid RE, Insogna JA, Carver TE, et al. Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J. Sci. Med. Sport. 2017; 20(6):578–582. [PubMed: 27887786]
41. Murphy SA, van der Laan MJ, Robins JM. Marginal mean models for dynamic regimes. J. Amer. Statist. Assoc. 2001; 96(456):1410–1423.
42. Goldberg DE, John HH. Genetic algorithms and machine learning. Mach. Learn. 1988; 3(2):95–99.

43. Scrucca L. GA: A package for genetic algorithms in R. J. Statist. Softw. 2013; 53(4):1–37.

44. Berg, BA. Markov Chain Monte Carlo Simulations and Their Statistical Analysis (With Web-Based Fortran Code). Hackensack, NJ, USA: World Scientific; 2004.

45. Haskell WL, Lee IM, Pate RR, et al. Physical activity and public health. Updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Circulation. 2007; 39:1423–1434.

46. Wagstaff, K. Machine learning that matters; Proc. 29th Int. Conf. Mach. Learn; 2012. arXiv preprint arXiv:1206.4656

47. King AC, Hekler EB, Grieco LA, et al. Effects of three motivationally targeted mobile device applications on initial physical activity and sedentary behavior change in midlife and older adults: A randomized trial. PLOS ONE. 2016; 11(6) Art. no. e0156370.

48. Berwick DM. What patient-centered should mean: Confessions of an extremist. Health Affairs. 2009; 28(4) Art. no. 560.

49. Coelho T. A patient advocate's perspective on patient-centered comparative effectiveness research. Health Affairs. 2010; 29(10):1885–1890. [PubMed: 20921490]

50. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu. Rev. Clin. Psychol. 2008; 4(1):1–32. [PubMed: 18509902]

51. Cheung YK, Hsueh PS, Qian M, et al. Are Nomothetic or ideographic approaches superior in predicting daily exercise behaviors? Analyzing N-of-1 health data. Methods Inf. Med. 2017; 56(5)

52. Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? Personalized Med. 2011; 8(2):161–173.

Biographies

Xinyu Hu *Department of Biostatistics, Columbia University, New York, NY 10032 USA* (xh2194@cumc.columbia.edu). Ms. Hu is a Ph.D. student in biostatistics at Columbia University. She received an M.S. degree in biostatistics from Columbia University in 2014. Her research interest is in machine learning in sequential decision-making problems. This paper is based on the work she performed while she was a research summer intern at the IBM Thomas J. Watson Research Center.

Pei-Yun S. Hsueh *Computational Health Behavior & Decision Science, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA* (phsueh@us.ibm.com). Dr. Hsueh is currently a Research Staff Member and the Health Informatics Professional Interest Community Chair of the IBM T. J. Watson Research Center, leading the development of evidence-based behavioral insights generation analytics in the Center for Computational Health. She is also an IBM Academy of Technology Member. Her current research focuses on innovative approaches for computing personalization and incorporating personalization analytics into service design. Her research interest ties closely to the marriage of artificial intelligence and human–computer interaction, with a focus on integrating machine learning and empirical analysis approaches for natural language understanding.

Ching-Hua Chen *Computational Health Behavior & Decision Science, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA* (chinghua@us.ibm.com). Dr. Chen leads an interdisciplinary team of psychologists, data scientists, and medical researchers, whose goal is to develop and test technology-enabled approaches for supporting health behavior change. She joined IBM Research in 2005 as a Research Staff Member in the Business Analytics and Mathematical Sciences Department.

Keith M. Diaz *Center for Behavioral Cardiovascular Health, Columbia University Medical Center, New York, NY 10032 USA* (kd2442@cumc.columbia.edu). Dr. Diaz is an Assistant Professor of Behavioral Medicine at Columbia University Medical Center. He oversees the Exercise Testing Laboratory at the Center for Behavioral Cardiovascular Health. His research focuses on the role of prolonged sedentary behavior in the pathogenesis of cardiovascular disease, with a specific focus of iteratively optimizing feasible, sustainable, and cost-effective guidelines for reducing prolonged sitting. His current work involves examining prolonged sedentary behavior as a prognostic risk factor for recurring events or early death after a cardiac event and elucidating the biological mechanisms through which cardiorespiratory fitness buffers and sedentary behavior augments the deleterious cardiovascular consequences of negative emotions.

Faith E. Parsons *Center for Behavioral Cardiovascular Health, Columbia University, New York, NY 10032 USA* (fep2110@cumc.columbia.edu). Ms. Parsons is a Data Manager at the Center for Behavioral Cardiovascular Health (CBCH), where she oversees the data management of multiple interventional and observational studies. Her professional interests include the utilization of both existing and emerging technologies to maximize the efficiency of research data collection and management. She is also pursuing a master's degree in biostatistics at the Mailman School of Public Health at Columbia University.

Ipek Ensari *Center for Behavioral Cardiovascular Health, Columbia University, New York, NY 10032 USA* (ie2145@cumc.columbia.edu). Dr. Ensari is a Postdoctoral Research Scientist at Columbia University. Her research focuses on the associations between cardiovascular function, emotional health, and physical activity. She investigates the effect of varying intensities, durations, and modalities of exercise on negative mood symptoms and their physiological correlates in various adult populations. Finally, she has been involved in exercise training and behavioral intervention studies for improving physical activity levels in adults with multiple sclerosis.

Min Qian *Department of Biostatistics, Columbia University, New York, NY 10032 USA* (mq2158@cumc.columbia.edu). Dr. Qian is an Assistant Professor of Biostatistics in the Mailman School of Public Health at Columbia University. Her primary research interest is in the area of medical decision-making, where the goal is to develop individualized treatment policies that specify which type and/or intensity of treatment should be offered over time. These treatment policies take patient information, such as demographics, preference, intermediate response, and adherence as input and output treatment decisions at each decision point. Her current research work includes design of clinical trials and development of novel statistical methodologies that can be used to construct optimal treatment policies.

Ying-Kuen K. Cheung *Department of Biostatistics, Columbia University, New York, NY 10032 USA* (yc632@cumc.columbia.edu). Dr. Cheung is a Professor of Biostatistics in the Mailman School of Public Health at Columbia University. He has general interests in the development and evaluation of evidence-based treatments, interventions, and policies at all phases of translational research. He is an expert in adaptive designs in clinical trials of treatments for cancer, stroke, and other neurological disorders; sequential multiple assignment randomized trial (SMART) designs for behavioral intervention technologies; *N*-of-1 study designs; implementation study designs; and the analysis of high-dimensional physical activity data. He is a recipient of an IBM Faculty Award on Big Data and Analytics. He is a member of the American Association for the Advancement of Science, the American Heart Association, American Statistical Association, the International Biometric Society, and the Society for Clinical Trials. He is an elected Fellow of the American Statistical Association. He serves as an Associate Editor for *Biometrics, Clinical Trials*, and the *Journal of the Korean Statistical Society*.

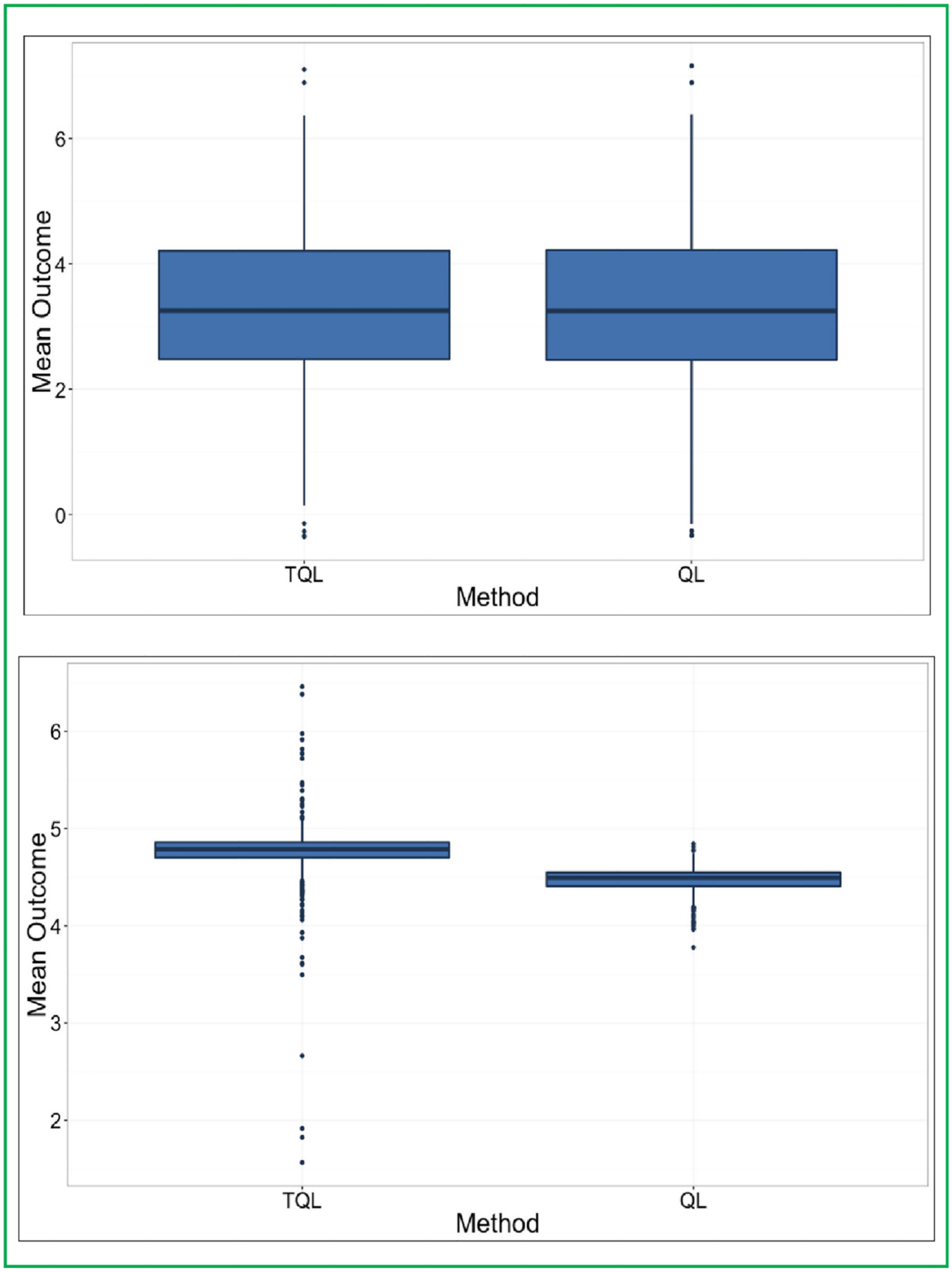

Figure 1.

Boxplot of 500 Monte Carlo replications of the expected outcome under the estimated policy using the TQL method and QL method with a training sample size equal to 500 (*top*: in a linear model setting; *bottom*: in a nonlinear model setting).

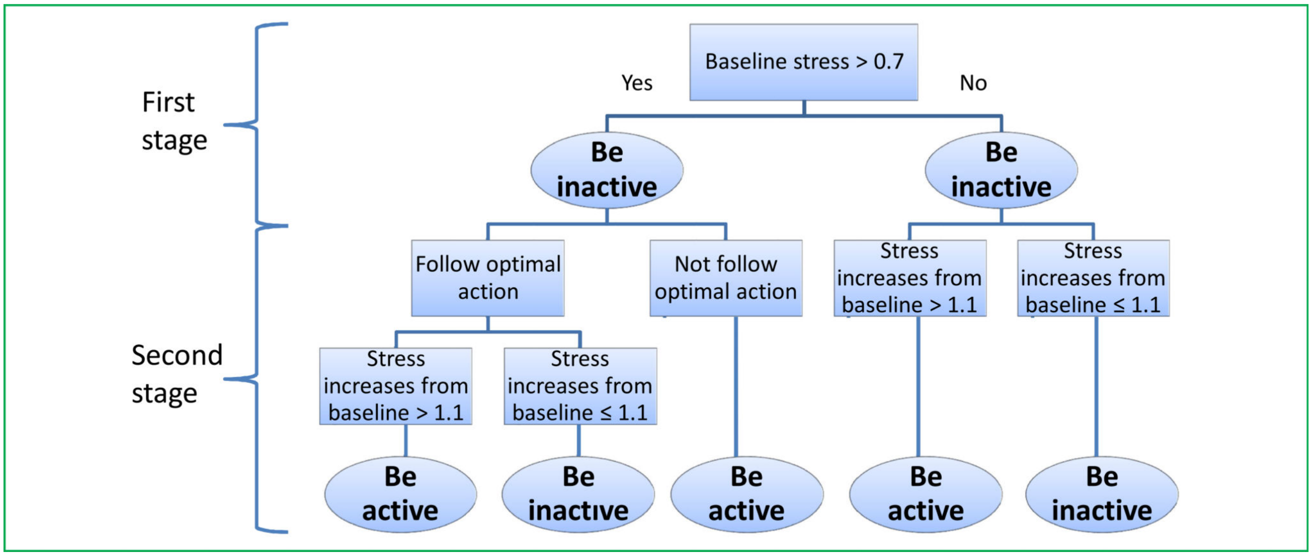
First stage
Second stage
Baseline stress > 0.7
Yes
No
Be inactive
Be inactive
Follow optimal action
Not follow optimal action
Stress increases from baseline > 1.1
Stress increases from baseline ≤ 1.1
Stress increases from baseline > 1.1
Stress increases from baseline ≤ 1.1
Be active
Be inactive
Be active
Be active
Be inactive

Figure 2.

The estimated interpretable policy by the TQL method using the two-stage stress data.

Table 1

Estimated mean (standard deviation) of the expected outcomes using the Threshold Q-learning (TQL) method and the Q-learning (QL) method.

	Linear model		Nonlinear model	
n	**TQL**	**QL**	**TQL**	**QL**
50	3.16(1.32)	3.16(1.32)	4.66(0.91)	4.38(0.69)
100	3.14(1.37)	3.15(1.37)	4.72(0.62)	4.42(0.28)
500	3.30(1.28)	3.30(1.29)	4.75(0.36)	4.47(0.14)