HHS Public Access

Author manuscript

Eur J Oper Res. Author manuscript; available in PMC 2020 February 01.

Published in final edited form as:

Eur J Oper Res. 2019 February 1; 272(3): 1058–1072. doi:10.1016/j.ejor.2018.07.011.

Behavioral Modeling in Weight Loss Interventions

Anil Aswani[*], **Philip Kaminsky**, and **Yonantan Mintz**

Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720

Elena Flowers

Department of Physiological Nursing, School of Nursing, University of California, San Francisco, CA 94143

Yoshimi Fukuoka

Department of Physiological Nursing/Institute for Health and Aging, School of Nursing, University of fornia, San Francisco, CA 94143

Abstract

Designing systems with human agents is difficult because it often requires models that characterize agents' responses to changes in the system's states and inputs. An example of this scenario occurs when designing treatments for obesity. While weight loss interventions through increasing physical activity and modifying diet have found success in reducing individuals' weight, such programs are difficult to maintain over long periods of time due to lack of patient adherence. A promising approach to increase adherence is through the personalization of treatments to each patient. In this paper, we make a contribution towards treatment personalization by developing a framework for predictive modeling using utility functions that depend upon both time-varying system states and motivational states evolving according to some modeled process corresponding to qualitative social science models of behavior change. Computing the predictive model requires solving a bilevel program, which we reformulate as a mixed-integer linear program (MILP). This reformulation provides the first (to our knowledge) formulation for Bayesian inference that uses empirical histograms as prior distributions. We study the predictive ability of our framework using a data set from a weight loss intervention, and our predictive model is validated by comparison to standard machine learning approaches. We conclude by describing how our predictive model could be used for optimization, unlike standard machine learning approaches which cannot.

Keywords

OR in health services; predictive modeling; weight loss; inverse optimization; machine learning

[*]Corresponding author.

[1]aaswani@berkeley.edu (A. Aswani), kaminsky@ieor.berkeley.edu (P. Kaminsky), ymintz@berkeley.edu (Y. Mintz), elena.flowers@ucsf.edu (E. Flowers), Yoshimi.Fukuoka@ucsf.edu (Y. Fukuoka)

1. Introduction

Effective design of systems involving human agents often requires models that characterize the agents' varied responses to changes in the system's states and inputs. Most operations research (OR) models quantify agent behavior as decisions generated by optimizing static utility functions that depend upon time-varying system states and inputs. In contrast, researchers in the social sciences have found that the motivational psychology of agents changes in response to past states, decisions, and inputs from external agents (Kanfer, 1975; Ajzen & Fishbein, 1980; Gonzalez et al., 1990; Janz & Becker, 1984; Joos & Hickam, 1990; Bandura, 2001); however, these social science models are primarily qualitative in nature, making them challenging to incorporate into OR design and optimization approaches. In this paper, we focus on developing a predictive modeling framework that incorporates time-varying motivational states (which describe the changing efficiency or preferences of the agent) – thereby quantifying agent behavior as decisions generated by optimizing utility functions that depend upon time-varying system states, system inputs, and motivational states, all evolving according to some modeled process based on qualitative social science models of behavior change.

Our ultimate goal is to solve optimization problems to more effectively allocate resources in systems with human agents; to do this we need to develop behavioral models that can be integrated as constraints in standard optimization approaches. In this paper, we develop a modeling framework that inputs noisy and partially-missing data and uses this to estimate the parameters of a predictive model consisting of (a) a utility-function describing the decision-making process that depends upon time-varying system states, system inputs, and motivational states, and (b) temporal dynamics on agent's system state and motivational state (i.e., often referred to as the type of the agent). We consider two distinct but related kinds of estimates: estimation of the set of parameters for the utility function and dynamics, and separately, estimation of the distribution of future states.

The framework we develop in this paper is described within the context of modeling the behavior of individuals in a weight loss program; specifically, we are interested in using a short time-span (e.g., 15–30 days) of physical activity and weight data from an individual participating in a weight loss program in order to effectively characterize the likelihood of whether or not that individual will achieve clinically significant weight loss (i.e., 5% reduction in body weight) after a long period of time (e.g., 5 months). While machine learning approaches such as support vector machines (SVMs) (Hastie et al., 2009; Wang et al., 2017; Oztekin et al., 2018) and artificial neural networks can be used to make binary predictions of significant weight loss based on a short time span of data (Hastie et al., 2009) they have two significant limitations: first there is no obvious way to integrate them into an optimization model, and second these approaches are generally limited in their interpretability (Breiman et al., 2001). Here, we show that in contrast to these machine learning methods, our approach is interpretable since the equations are based on models from the social sciences, and can be incorporated into optimization models since it is posed as a mixed integer linear program (MILP), while maintaining comparable prediction accuracy.

1.1. Personalized Treatments and Obesity

Obesity is a significant problem in the United States. About 70% of American adults are overweight or obese (Flegal et al., 2012), and its annual cost to the health care system is estimated to be $350 billion (Valero-Elizondo et al., 2016). Currently, the most effective treatments for obesity are weight loss interventions composed of counseling sessions by clinicians and daily goals for physical activity and caloric consumption. The Diabetes Prevention Program Research Group (2002, 2009) showed that participating in these types of treatments results in significant weight loss of 5–7% and can prevent the onset of type-2 diabetes with few side effects. However, adherence to these clinician-set goals decreases over time (Acharya et al., 2009), and these programs are labor-intensive and expensive to sustain (McDonald et al., 2002; Diabetes Prevention Program Research Group, 2003). Making these interventions more *effective and efficient* will require designing treatments personalized to each individual's preferences.

While individualized goal-setting and personalized interventions are crucial to the success of these programs, these features are expensive to provide. Cost efficient programs will need automation of goal-setting and scheduling of counseling resources for individuals to succeed in reducing their weight. Such approaches will likely involve digital/mobile/wireless technologies, which already have high adoption rates (Lopez et al., 2013; Bender et al., 2014) and have shown promise for improving the quality of and adherence to weight loss programs (Fukuoka et al., 2011). These technologies allow clinicians and researchers to remotely collect real-time health data and communicate with individuals participating in the program. However, healthcare data sets generated by mobile devices have been underutilized to date, and little research has focused on effective ways to utilize individuals' health-related data patterns to improve and personalize weight loss interventions (Fukuoka et al., 2011; O'Reilly & Spruijt-Metz, 2013; Pagoto et al., 2013; Azar et al., 2013).

1.2. Overview

Ultimately, effective automated approaches will depend upon nuanced models to predict the effects different interventions (i.e., changes in activity and caloric goals, or specific types of counseling) will have on the weight loss trajectories of different individuals. In this paper, we present an initial step – specifically, we develop an approach for using a short time-span (e.g., 15–30 days) of physical activity and weight data from an individual participating in a weight loss program to effectively characterize the likelihood of whether or not that individual will achieve clinically significant (i.e., 5% reduction in body weight) weight loss after a long period of time (e.g., 5 months) as a function of the physical activity goals and amount of counseling given to the individual. (The Diabetes Prevention Program Research Group (2002, 2009) showed 5% weight loss provides substantial health benefits.) As discussed above, this type of predictive tool will ultimately enable the adaptive design of more effective and cost efficient interventions. Towards this end, we also show how our predictive model is able to predict the impact of changes in the intervention treatment on the weight loss trajectory of a specific individual.

A key feature of predicting future behavior is the inherent uncertainty due to having limited data. As a result, it is natural to consider predictive modeling approaches that generate

ranges or intervals of predictions. Though frequentist approaches can be used to construct confidence intervals, we instead propose a Bayesian approach that constructs a range of predictions characterized by a *posterior* distribution. An important benefit of our Bayesian (as compared to a frequentist) approach is that it can incorporate data from individuals that have been in the program for a longer period of time or have even completed a fixed duration (e.g., 5 months) of the program. We quantitatively show in Section 6 that incorporating the information of other individuals using a nonparametric Bayesian prior distribution improves the accuracy of predictions versus not using a Bayesian framework.

Our resulting predictive modeling approach is presented in Section 5. In the preceding sections, we develop essential elements for constructing the model. We first describe the structure of mobile phone-based weight loss interventions in Section 2. Section 3 describes our utility-maximizing model of the decisions of an individual participating in a weight loss intervention. Mathematically, we represent prior information in the Bayesian framework as histograms of parameter values for the utility functions of individuals that have completed the fixed duration of the program. To compute these parameters, we solve a maximum likelihood estimation (MLE) problem, which is the focus of Section 4. Our predictive modeling approach in Section 5 uses the utility-maximizing framework and corresponding histograms of parameter values to predict the weight loss trajectory of a single individual. Both the MLE in Section 4 and predictive model in Section 5 are computed by solving a mixed integer linear program (MILP).

To validate our predictive modeling approach, we use a longitudinal data set collected from a 5-month randomized controlled trial (RCT) of a mobile phone-based weight loss program. Section 6 begins with an overview of this RCT, and additional details are available in Fukuoka et al. (2015). Next, we evaluate the effectiveness of our approach for predicting whether or not an individual will achieve clinically significant (i.e., 5% or more) weight loss at the end of the intervention. We validate our approach by showing its binary predication accuracy is comparable to standard machine learning methods (i.e., linear SVM, decision tree, and logistic regression) in terms of prediction quality. In contrast to these machine learning methods, our predictive model is also able to determine the impact of changing intervention parameters for a specific individual on that individual's weight loss trajectory, and we conclude with a discussion of this aspect of our model and how it can be used to perform optimization.

1.3. Literature Review

Statistical classification methods (which include logistic regression, support vector machines, neural networks, and random forests) predict a binary $\{-1, +1\}$ output label based on an input vector (Hastie et al., 2009; Denoyel et al., 2017). In the context of weight loss interventions, these approaches could predict whether (+1) or not (−1) an individual will achieve 5% weight loss after 5 months, based on 30 days of an individual's data. However, these approaches lack interpretability (Breiman et al., 2001) and cannot be incorporated as constraints into standard optimization approaches. Our predictive modeling approach is similar in that it can be used as a classifier (i.e., it can predict whether or not an individual achieves 5% weight loss), but it differs in that its equations are based on models from the

social sciences, and can be incorporated into optimization models since it can be posed as a mixed integer linear program (MILP), making it more applicable for addressing the problem of intervention design.

A number of predictive models have been developed to determine the impact of changing a medical intervention on the health outcome for an individual, including: Markov chain models (Ayer et al., 2012; Mason et al., 2013; Deo et al., 2013; Andersen et al., 2017), dynamical systems models (Helm et al., 2015), decision tree models (Wu et al., 2013), graph-theoretic models (Fetta et al., 2018), bandit models (Negoescu et al., 2014), and dynamic programming models (Engineer et al., 2009). (This literature also studies the problem of designing optimal treatment plans, which we do not consider in our present paper.) Our work is similar in that we develop an approach to predict future body weight of an individual as physical activity goals and counseling scheduling are changed. One key difference is in the data available in weight loss programs. Existing approaches are designed for situations where data is collected infrequently (e.g., only during clinical visits), whereas in weight loss programs the data is collected daily using mobile devices. Our work seeks to develop a predictive modeling approach that can leverage this increased data availability in order to make improved predictions. Moreover, existing approaches focus either on motivational states (Mason et al., 2013) or health states (Ayer et al., 2012; Deo et al., 2013; Helm et al., 2015; Wu et al., 2013; Negoescu et al., 2014; Engineer et al., 2009). We seek to combine the notions of motivational and health states into a single predictive model, which is a modeling approach that has not been previously considered.

Previous approaches for automated exercise and diet management significantly differ in the goal of the predictive modeling. Bertsimas & O'Hair (2013) develop a system that learns a predictive model of an individual's dietary preferences and then designs a plan of what food to eat and how much time to exercise to maintain low blood glucose levels. The output of this predictive model is blood glucose levels and satisfaction of a given dietary plan, whereas we are interested in making predictions regarding future body weight. Additionally, this predictive model does not consider adherence to the prescribed plans (e.g., the individual may overeat or may not exercise the amount indicated by the plan), whereas our approach quantifies the level of adherence to prescribed physical activity goals and guidance on caloric intake. The Steptacular program (Gomes et al., 2012) used monetary incentives to encourage individuals to walk more, but a predictive model was not developed to design the incentives; our approach differs in that we seek to build a predictive model so that in the future we may be able to optimize the weight loss intervention for each individual.

1.4. Contributions

We develop a number of novel optimization modeling and analysis techniques that we believe will be useful for expanding the scope of predictive models of human decision-making in complex systems. For instance, much mobile phone data contains non-negligible noise and suffers from missing data points (Chen et al., 2012). Aswani et al. (2018) showed that statistically consistent estimation of model parameters in a utility-maximization framework requires joint estimation of the missing data and model parameters. It is known (see for instance Bickel & Doksum (2006)) that such joint estimation does not represent

statistical over-fitting, and in fact all regression approaches (even basic linear regression) jointly provide estimates of denoised data and model parameters; however, only the model parameter estimates are statistically consistent (Bickel & Doksum, 2006; Aswani et al., 2018). Existing approaches for dealing with missing data (e.g., the EM algorithm (Hastie et al., 2009)) generate an estimate by computing the local optimum of a suitably defined optimization problem that computes the parameters of the predictive model. Instead, we construct optimization models formulated as mixed integer linear programs (MILP's) that are able to simultaneously estimate missing/noisy data and parameters of the utility-maximizing framework; this yields global optima of the parameter computation optimization problem.

As mentioned above, we can likely improve trajectory predictions for a specific individual in a weight loss intervention by leveraging mobile phone data from other individuals who have already completed the intervention. This challenge can be posed in a Bayesian framework, but existing nonparametric approaches require computing numerically challenging integrals. In this paper, we provide what is to the best of our knowledge the first Bayesian estimation approach in which the prior distribution is purely data-driven and described by a histogram. For this Bayesian estimation, we use integer programming, and we show that a data-driven distribution can be represented as a piecewise constant function, which can then be formulated within a MILP (Vielma, 2015).

In many cases, patients favor behavior that does not improve (or is not optimal with respect to) their health outcomes. Non-adherence to a medical plan falls in to this category. Social scientists sometimes label such behavior "irrational" (Brock & Wartman, 1990); however, an argument has been made that many instances of "irrational" behavior are in fact rational decisions when considering a patient's actual utility function (Gafni, 1990; Cawley, 2004). In our case, we explicitly use a utility function in which the individual is assumed to heavily discount future health states, a behavior that is often characterized as "irrational" (Brock & Wartman, 1990). We note, however, that while these modeling choices may be controversial, the particular utility function framework that we develop has an alternative interpretation that does not make reference to utility maximization. In particular, our approach can alternatively be interpreted as leading to a model that has the best theoretical predictive accuracy given the set of underlying equations that characterize this framework. While this alternative interpretation is beyond the scope of this paper, we describe it in a companion paper focusing on inverse optimization problems with noisy data (Aswani et al., 2018). Thus, even if the behavioral argument we advance in subsequent sections of this paper does not accurately capture individuals' behavior, the framework we describe still enables us to make the most accurate set of predictions possible using the set of equations underlying the predictive model.

2. Structure of Mobile Phone-Based Weight Loss Interventions

Currently the healthcare community is refining a new class of weight loss interventions that rely on mobile phones and digital accelerometers (Gomes et al., 2012; Fukuoka et al., 2015; Flores Mateo et al., 2015). Though the particular features of these programs often differ, there is a growing consensus on the broad structure of these programs. In general, each

individual is provided with (i) a mobile phone app and a digital accelerometer, and (ii) in-person counseling sessions. The digital accelerometer is used to measure daily physical activity, and the digital aspect of the device simplifies data sharing and data uploading. The mobile phone app delivers physical activity goals, educational messages (such as those from (Diabetes Prevention Program Research Group, 2002, 2009)), and provide an interface for individuals to enter dietary and body weight information.

The accelerometer measures the number of seps taken each day since the majority of exercise for individuals in such weight loss interventions consists of walking. Individuals are also typically asked to input weight measurements multiple times a week into the mobile app. In principle, the data available for each individual consists of daily weight and step amounts; however, data for some dates is missing because individuals forget to enter weight data into the mobile app, wear the accelerometer, or because of a technical problem with the app. The age, gender, and height of each individual is also known data in these programs.

Individuals participating in such mobile phone-based weight loss interventions receive additional interaction. After an initial baseline period, exercise goals in terms of a minimum daily step count are provided to each individual. The goals change at regular intervals (e.g., every week). Individuals also have office visits (or phone calls) at regular intervals, during which they received behavioral counseling about their nutritional choices and physical activity. The exercise goals and timing of the office visits (or phone calls) are set in advance, and thus are also known data in these programs.

3. Formulating the Utility-Maximizing Framework

The utility-maximizing framework we propose has two components. The first describes how an individual makes decisions regarding the amount of steps and caloric intake, and this is formulated in terms of a utility-maximizing individual. The utility function contains heavy discounting of future health states, a behavior that is often characterized as “irrational” (Brock & Wartman, 1990). The second describes how the individual’s weight and *type* (a set of parameters describing each individual) evolve over time as a function of current states and decisions. This second part is formulated in terms of a linear dynamical system.

3.1. Summary of Framework

A subscript t denotes the value of a variable on the t-th day. Let $f_t \in \mathbb{R}_+$ denote the amount of calories consumed, $u_t \in \mathbb{R}_+$ be the number of steps, $w_t \in \mathbb{R}_+$ be the weight of the individual, $g_t \in \mathbb{R}_+$ be the given exercise goal in terms of number of steps, and $d_t \in \{0,1\}$ indicate whether or not an office visit occurred. We refer to $\theta_t = (k, q, s_0, s_t, p_t, \mu)$ as the *type* of the individual. The parameters $a,b,c,k \in \mathrm{R}$ describe the weight dynamics, are based on the physiology of the individual, and can be precomputed based on the age, gender, and height of the individual (Mifflin et al., 1990). Another set of the parameters are used in the utility function. These include $r_f, r_u \in \mathbb{R}$ which represent the marginal utility of quadratic terms, q, s_0 which represent baseline preferences in terms of physical activity and caloric consumption respectively, $p_t \in \mathbb{R}$ which represent the marginal dissutility of failing exercise goals, and $s_t \in \mathbb{R}$ which represents the current preference of caloric consumption. The last set of parameters describe the type dynamics, including $\mu \in \mathbb{R}_+$ that captures the impact of

achieving an exercise goal, and $0 < \gamma < 1$ which is a discount factor representing the diminishing effect of the intervention over time. The $\beta_t, \delta_t \in \mathbb{R}_+$ are random variables with finite variance that represent the impact of an office visit, and $z_t \in \mathbb{R}$ is a zero-mean random variable with finite variance that denotes weight fluctuations from unmodeled effects. These random variables β_t, δ_t, z_t are individual-specific, but we do not consider them to characterize the *type* of the individual. The expected behavior of any particular individual will not depend in a unique way upon these random variables. Using these quantities, we define the following utility functions and dynamics.

1. Individual decision-making when no exercise goals are given is

$$\mathbf{U_{no\ goals}} \qquad \begin{aligned}(u_t, f_t) = \arg\max_{u,f}\ & -w_{t+1}^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t \\ \text{s.t. } & w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k.\end{aligned}$$

Individual decision-making when exercise goals are given is

$$\mathbf{U_{goals}} \qquad \begin{aligned}(u_t, f_t) = \arg\max_{u,f}\ & -w_{t+1}^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t + p_t \cdot (u_t - g_t)^- \\ \text{s.t. } & w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k.\end{aligned}$$

Note that $\mathbf{U_{no\ goals}}$ and $\mathbf{U_{goals}}$ refer to the (u_t, f_t) that are computed by solving the corresponding optimization problems.

2. Weight and type are assumed to evolve according to the following:

$$w_{t+1} = a \cdot w_t + b \cdot u_t + c \cdot f_t + k + z_t \quad (1)$$

$$s_{t+1} = \gamma \cdot (s_t - s_0) + s_0 - \beta_{t+1} \cdot d_{t+1} \quad (2)$$

$$p_{t+1} = \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu \cdot \mathbb{1}(u_t \geq g_t). \quad (3)$$

Observe that the time index in (2), (3) for β, δ, d is $t+1$ because we assume that the impact of a clinical visit occurs on the day of the visit.

Note that in $\mathbf{U_{no\ goals}}$ is the caloric consumption preference s_t is time-varying, whereas the physical activity preference q is constant. The reason is that in clinically-supervised weight loss programs, individuals are encouraged to reduce their caloric consumption at the beginning of the program – in contrast, the individuals are asked to not increase their physical activity level until they begin to receive goals (Fukuoka et al., 2011). Thus, our predictive model assumes that the physical activity preference remains constant during the period in which no goals are given.

3.2. Structure of Utility Function

We assume an individual's utility function is separable with respect to weight, exercise amount, and caloric intake. An individual with *perfect knowledge* of his or her type θ_t may choose their exercise amount u and caloric intake f to maximize a utility of the form $\sum_{k=0}^{\infty} \alpha^{-k} \cdot \mathbb{E}(U_1(w_{t+k+1}, d_{t+k}, g_{t+k}; \theta_{t+k}) + U_2(u_{t+k}, d_{t+k}, g_{t+k}; \theta_{t+k}) + U_3(f_{t+k}, d_{t+k}, g_{t+k}; \theta_{t+k}))$

subject to weight $w_{t+k+1} = \eta(w_{t+k}, u_{t+k}, f_{t+k}, \xi_{t+k})$ and type dynamics $\theta_{t+k+1} = \zeta(\theta_{t+k}, w_{t+k}, u_{t+k}, f_{t+k}, \xi_{t+k}, d_{t+k}, g_{t+k})$, where $\xi_{t+k} = (z_{t+k}, \beta_{t+k}, \delta_{t+k})$ are random variables, $\alpha \in [0,1)$ is a discount factor, U_1,U_2,U_3 are utility functions, and η, ζ are functions that define the dynamics. Note that utility depends on weight one day ahead of the corresponding decision because future weight and present decisions affect utility.

However, it is not true that individuals make health care decisions with the goal of maximizing long term health benefits. Indeed, it is common for individuals to very heavily discount the impact of present decisions on future health outcomes (Chapman & Elstein, 1995). To capture this behavior that is sometimes characterized as "irrational" (Brock & Wartman, 1990), we explicitly use a utility function in which the individual is assumed to heavily discount future health states.

Proposition 1. *If the discount factor is* $\alpha = 0$, *then this is equivalent to an equation where the individual makes a decision considering only the one-day impact:*
$\max_{u,f} \left\{ \mathbb{E}(U_1(w_{t+1}, d_t, g_t; \theta_t) + U_2(u_t, d_t, g_t; \theta_t) + U_3(f_t, d_t, g_t; \theta_t)) \middle| w_{t+1} = \eta(x_t, u_t, f_t, \xi_t) \right\}$.

Proof. This follows by direct calculation.

3.3. Choice of Utility Function

Corresponding terms in the utility function are chosen to match to particular behaviors expected by social cognitive theory (Bandura, 2001): In this context, social cognitive theory asserts that caloric consumption and physical activity depend upon (1) *self-efficacy,* which is an individual's belief in their ability to achieve positive behavioral changes and is characterized by the coefficients p_t, q, s_t; and depend upon (2) receiving a positive reward from a small amount of weight loss for engaging in positive behavioral changes. We choose $U_1 = -w_{t+1}^2$, $U_3 = -r_f f_t^2 + s_t f_t$, $U_2 = -r_u u_t^2 + q u_t$ if no goal is given, and $U_2 = -r_u u_t^2 + q u_t + p_t \cdot (u_t - g_t)^-$ if a goal is given. Dislike for large amounts of steps and caloric intake is captured by the $-r_u u_t^2$ and $-r_f f_t^2$ terms. Positive satisfaction for increasing steps and caloric intake is represented by the q_{ut} and $s_t f_t$ terms. An individual's preference for lower weight is reflected by the $-w_{t+1}^2$ term. And an increase in satisfaction for getting closer to the exercise goal is captured by the $p_t \cdot (u_t - g_t)^-$ term.

Remark 1. *Observe that as* p_t *increases, the utility of meeting a step goal increases, and as* s_t *increases, the utility of higher caloric intake increases. Thus, we can interpret the values* p_t, s_t *as a quantification of the adherence of an individual to step goals and dietary goals, respectively.*

Remark 2. *An alternative choice is* $U_1 = -w_t^2 + 2w_b w_t$*, which has an additional linear term with coefficient* W_b*. After completing the square, this is equivalent to choosing* $U_1 = -(w_t - w_b)^2$*, which makes its interpretation clear: The* W_b *coefficient should be interpreted as the preferred weight of that individual. From a computational standpoint,* W_b *can be estimated using the same approach that we describe in later sections for estimation of* r_f, r_u*. However, we chose to not include the linear term for two reasons. The first is that including this linear term does make estimation more slow computationally. The second is that choosing* $W_b = 0$ *for all individuals (which makes* $U_1 = -w_t^2$*) and then scaling for each individual the other coefficients in* U_2*,* U_3 *can reasonably approximate within a finite range of weights a U1 with a linear term. Our second reason also explains why a purely linear* $U_1 = -w_t$ *is not an appropriate choice, because a purely linear* U_1 *cannot capture the diminishing returns to weight loss as weight decreases towards the desired weight. As we will show later, setting* $W_b = 0$ *leads to accurate predictions, which ultimately validates our choice.*

While other functional forms can represent the behaviors expected by social cognitive theory, these choices have several advantages. The choice that positive utility (qu_t and $s_t f_t$) increases at a slower rate than disutility decreases ($-r_u u_t^2$ and $-r_f f_t^2$) ensures that an individual takes a finite number of steps and consumes a finite amount of calories. (Other choices can lead to a situation where the individual is predicted to take an infinite number or steps or consume an infinite number of calories, which is clearly unreasonable.) Moreover, these choices ensure the objective is strictly concave, which ensures that an individual is predicted to make only one decision; if the utility function was merely concave, then there may be multiple maximizers that correspond to a set of different possible decisions on the number of steps and calories.

Additionally, this functional form has a relatively low parameter count, which facilitates estimation. For instance, there is no linear term for weight w_t. The utility term qu_t is kept constant because explicitly incorporating an increase in exercise utility (with an office visit) would be an over-parametrization due to the $p_t(u_t - g_t)^-$ term. Furthermore, we do not need to include a parameter for the $-w_t^2$ term because this would simply scale the function, and would not change the decision. Lastly, our choice implies that goal setting has no impact beyond the goal amount.

Remark 3. *Restated, the utility term* $p_t \cdot (u_t - g_t)^-$ *is at its maximum value for all* $u_t \geq g_t$*. This is a simplification to reduce the number of terms. A more detailed framework would also incorporate positive utility for exceeding the goal, such as by including the term* $\rho_t \cdot (u_t - g_t)^+$*. The reason we do not include a linear term* $\rho_t \cdot (u_t - g_t) = \rho_t \cdot u_t - \rho_t \cdot g_t$ *is that such a term inherently cannot capture the satisfaction of meeting a goal, because it has the same effect (due to* $p_t \cdot g_t$ *being a constant) as including the term* $\rho_t \cdot u_t$*.*

3.4. Dynamics of Weight

We also need to specify weight dynamics. Standard physiological arguments (i.e., weight change is proportional to “calories-in minus calories-out”) imply that the weight dynamics are given by $w_{t+1} = a\, w_t + b\, u_t + c \cdot f_t + k + z_t$, where $a, b, c, k \in \mathbb{R}$ are coefficients that can be

computed using existing physiological models, and z_t is a zero-mean random variable that captures unmodeled changes in weight (e.g., water fluctuation, physical activity in addition to steps, etc.). Suppose w_t, k, z_t are specified in units of kilograms, f_t is specified in units of kilocalories (also known as dietary calories), and u_t is specific in units of steps. Then a derivation given in the Supplementary Materials and based on the Mifflin St Jeor Equation (Mifflin et al., 1990) for the basal metabolic rate (BMR) gives $a = 0.9987$ and $k = —8.0357 \times 10^{-4} \cdot h + 6.4286 \times 10^{-4} \cdot a + s$, where h is height in centimeters, a is age in years, $s = —6.4286 \times 10^{-4}$ for males, and $s = 2.0700$ for females. To compute b, we note that 2000 steps is roughly equal to walking one mile and consumes about 100 calories, largely independent of the height, weight, age, and gender of an individual (Hill et al., 2003). This gives a value of $b = —6.4287 \times 10^{-6}$. Last, the value of $c = 1.2857 \times 10^{-4}$ is computed by performing the unit conversion that 3500 calories is 0.45 kilograms.

One consequence of linear weight dynamics is simplification of the utility-maximizing framework:

Proposition 2. *When the weight dynamics are linear, as in (1), we can rewrite the objective of the utility-maximizing framework as*

$$-(a \cdot w_t + b \cdot u_t + c \cdot f_t + k)^2 + U_2(u_t, d_t, g_t; \theta_t) + U_3(f_t, d_t, g_t; \theta_t) - \mathbb{E}(z_t^2).$$

Proof. This follows by first substituting the linear weight dynamics (1) and then noting that (i) the only stochasticity is in z_t, (ii) z_t is zero mean, (iii) z_t is unobservable at time t and cannot be used to make a decision at time t, and (iv) the terms involving z_t have an expectation of zero since the decisions are independent of z_t.

Remark 4. *The main insight from this substitution is that decisions made by an individual following the utility-maximizing framework do not depend on the stochasticity because* $-\mathbb{E}(z_t^2)$ *is a constant that does not depend on the decisions.*

Before describing the type dynamics, we discuss a more detailed model for the weight dynamics. Specifically, a phenomenon known as *adaptive thermogenesis* (Doucet et al., 2001; Rosenbaum et al., 2008) causes the metabolism of an individual who has lost weight to decrease. Our weight dynamics (1) can be modified to incorporate this phenomenon by allowing the z_t to have a non-zero mean. Though we do not use this more detailed model in this paper, we briefly outline how our MLE and Bayesian prediction formulations (that will be described in upcoming sections) would change: The first change is that the k term in the constraints would be replaced with k + m_t, where m_t is a new variable that represents the mean of z_t. This change allows the z_t in our formulations to have a non-zero mean. The interpretation of m_t is that it represents the amount of reduction in metabolism (in units of kilograms) as weight loss occurs, which is the above described phenomenon of adaptive thermogenesis. More precisely, a negative value of m_t indicates an increase in metabolism whereas a positive value of m_t indicates a decrease in metabolism. The reason for this interpretation is that z_t enters additively into the weight dynamics (1), and so a negative m_t means weight is decreased while a positive m_t means weight is increased. The second change is that an additional constraint $\sum_{t=1}^{n-1} |m_{t+1} - m_t| \leq \sigma_m$ is added to our formulations,

where n is the time step at which we are solving the formulation and σ_m is a constant the bounds the amount of metabolism change, and this constraint is known as *fused lasso* (Tibshirani et al., 2005) in the statistics and machine learning literature. This additional constraint has been show to have properties (Tibshirani et al., 2005) that would lead to estimates that ensure the estimated change in metabolism becomes roughly constant after an individual's weight stops changing, which is an important property because it matches what is clinically observed with changes in metabolism after weight loss (Doucet et al., 2001; Rosenbaum et al., 2008). As a final note, ensuring a monotonic dependence between the reduction in weight w_t and the reduction in metabolism m_t is more complicated. Constraints to exactly ensure such a dependence are too computationally expensive, but reasonable approximate constraints can be incorporated into the optimization model for estimation. For instance, let the measured weight data be $(t_i, \tilde{w}_{t_i})$, for $i = 1,\ldots, n_w$, where n_w is the number of weight measurements, t_i is the day of the i-th weight measurement, and $\tilde{w}_{t_i}$ is the weight measurement on the t_i-th day. If we define T_i to be an re-ordering of the $\left\{t_1, \ldots, t_{n_w}\right\}$ such that $\tilde{w}_{T_1} \leq \tilde{w}_{T_2} \leq \cdots \leq \tilde{w}_{T_{n_w}}$, then the constraints $m_{T_1} \geq m_{T_2} \geq \cdots \geq m_{T_{n_w}}$ ensure that lower weights correspond to lower metabolism via a larger m_t term.

3.5. Dynamics of Type

The type dynamics are as specified in (2),(3), where γ, so, μ are scalars and β_t, δ_t are random variables. Specific terms in these dynamics correspond to principles of social cognitive theory, which says in this context that self-efficacy as quantified by s_t, p_t will increase in response to social contact during office visits and in response to successfully achieving past goals. The uncertain impact of office visits is modeled by the stochastic β_t and δ_t. The fact that office visits sometimes make external goal-setting more effective and decrease interest in eating is described by the $\delta_{t+1} \cdot d_{t+1}$ and $-\beta_{t+1} \cdot d_{t+1}$ terms, respectively. Because the impact of a single office visit decreases to zero over time, the dynamics include the terms $\gamma \cdot (s_t - s_0) + s_0$ and $\gamma \cdot p_t$. Observe that these discounting terms are different because s_0, q are the baseline preferences for caloric consumption and physical activity, respectively. So the first discounting term ensures s_t goes to s_0 without more office visits, and the second discounting term ensures p_t goes to zero without more office visits since q already encodes the baseline coefficient for physical activity. Moreover, goal-setting can become more effective whenever the goal is met; this is characterized by the $\mu \cdot \mathbb{1}(u_t \geq g_t)$ term.

Multiple equation choices would lead to the behaviors suggested by social cognitive theory, but this set of choices ensures the dynamics are linear in s_t, p_t and reduces the parameter count. The latter objective is achieved through (i) using the same parameter γ for both the $\gamma \cdot (s_t - s_0)$ and $\gamma \cdot p_t$ terms, and (ii) using a constant parameter μ instead of allowing this to be a time varying quantity. Linearity in s_t, p_t is important for favorable computational properties. Though the term $\mu \cdot \mathbb{1}(u_t = g_t)$ is nonlinear, it has special structure that allows efficient computation.

4. Maximum Likelihood Estimation (MLE) for Utilility-Maximization

Estimating parameters of the utility-maximizing framework for a specific individual requires solving an optimization problem. However, formulating this model is challenging because the measurements suffer from noise and missing weight and step data. This can be overcome by formulating the optimization model so that its minimizer simultaneously estimates the values of weight, caloric intake, steps, type, and the random variables in the model for each individual. The optimization model for simultaneous estimation is generally a nonconvex, nonlinear program; and it is typical to generate an estimate by computing a local optimum (e.g., the EM algorithm (Hastie et al., 2009)). However, we show that simultaneous estimation can be modeled using as a MILP, allowing us to compute the global optimum of the optimization model for estimation.

We pose the estimation problem in the framework of MLE. Suppose that the data for a single individual consists of $(t_i, \tilde{w}_{t_i})$, for $i = 1, \ldots, n_w$, $(\tau_i, \tilde{u}_{\tau_i})$, for $i = 1, \ldots, n_u$, and (d_t, g_t) for t = 1, $\ldots, n$, where n_u are the number of weight measurements, n_u are the number of step measurements, and the noise model is $\tilde{w}_{t_i} = w_{t_i} + v_{t_i}$ and $\tilde{u}_{\tau_i} = u_{\tau_i} + \omega_{\tau_i}$, where v_{t_i}, ω_{τ_i} are zero-mean random variables with finite variance. Note that the times t_i, r_i do not coincide in general. Let $\psi_v(\cdot), \psi_w(\cdot), \psi_z(\cdot)$ by the probability density function (pdf) for the random variables v_t, ψ_t, z_t. The MLE problem seeks to estimate the type θ_t of each individual and the parameters r_f, r_u, γ, using the above described data; however, the MLE problem will also involve estimation of $f_t, u_t, w_t, \beta_t, \delta_t$ to deal with noise and missing weight and step data as described above. It is important to further discuss the interpretation of the type θ_t and parameters rf $, r_u, \gamma$ that are estimated. Clearly there will be additional factors beyond the ones we have included in our predictive model that influence how an individual decides their daily caloric intake and number of steps, and so the measured data cannot be expected to exactly match our predictive model. In this context, the type θ_t and parameters r_f, r_u, γ that are estimated for each individual should be interpreted as those that maximize the prediction accuracy of the predictive model (Aswani et al., 2018) – a concept sometimes known as *risk consistency* in the statistics literature.

4.1. Initial Optimization Model for Computing MLE

Let $n = \max\left\{t_{n_w}, \tau_{n_u}\right\}$ be the number of days of data used for estimation, and let m by the number of initial days before an exercise goal was given to the individual. For the utility-maximizing framework, the MLE is the minimizer of an optimization problem defined as

$$\mathbf{P_{mle}} \qquad \begin{array}{l} \min \ \sum_{i=1}^{n_w} -\log\psi_v(\tilde{w}_{t_i} - w_{t_i}) + \sum_{i=1}^{n_u} -\log\psi_w(\tilde{u}_{\tau_i} - u_{\tau_i}) + \sum_{t=1}^{n} -\log\psi_z(z_t) \\ \text{s.t. } \mathbf{U_{no\ goals}}, (1) \text{ for } t = 1, \ldots, m-1; \quad \mathbf{U_{goals}}, (1), (2), (3) \text{ for } t = m, \ldots, n. \end{array}$$

Recall that $\mathbf{U_{no\ goals}}$ captures decision-making without goals, U_{goals} captures decision-making model with goals, equations (1) are dynamics on weight, (2) and (3) are the dynamics of parameters s_t, p_t respectively. Note that the first office visit is on the same day

the first exercise goal is given. Since s_t, p_t cannot change until the start of the intervention their dynamics begin at time m.

The problem $\mathbf{P_{mle}}$ is more challenging to solve than may initially appear. The variables u_t, f_t are defined as the minimizing arguments of $\mathbf{U_{no\ goals}}$ and U_{goals}. This makes the MLE the solution to a bilevel optimization problem (Dempe, 2002). Among the bilevel optimization problems that have been considered in the literature include inverse optimization with linear objectives (Ahuja & Orlin, 2001) and inverse optimization for combinatorial problems like assignment and spanning tree problems (Heuberger, 2004). In the context of bilevel optimization problems for estimating utility functions, approaches have been derived under the assumption of small noise (Keshavarz et al., 2011; Bertsimas et al., 2014); more recently, statistically consistent approaches for noisy measurements have also been proposed (Aswani et al., 2018). Here, we develop a new integer programming approach for solving our specific bilevel optimization problem in $\mathbf{P_{mle}}$.

4.2. Choosing the Distribution of Random Variables Representing Noise

We first must select the distribution of random variables representing noise v_t, f_t, z_t. Their variances σ_1, σ_2, are constants that can be chosen based on our prior knowledge regarding the measurement accuracy of weight scales, measurement accuracy of accelerometers for measuring steps, and physiological information about the modeling errors of the Mifflin St Jeor Equation (Mifflin et al., 1990) for BMR. In our modeling, we used $\sigma_1 = 2$, $\sigma_2 = 0.1$, and $\sigma_3 = 0.1$.

Choosing zero-mean Gaussian random variables yields a quadratic objective for $\mathbf{P_{mle}}$: $\kappa_1 + \frac{1}{\sigma_1}\sum_{i=1}^{n_w}(\tilde{w}_{t_i} - w_{t_i})^2 + \frac{1}{\sigma_2}\sum_{i=1}^{n_u}(\tilde{u}_{\tau_i} - u_{\tau_i})^2 + \frac{1}{\sigma_3}\sum_{t=1}^{n}(z_t)^2$, where κ_1 is a constant. Alternatively, one could select v_t, ψ_t, z_t to be zero-mean Laplace random variables, which have a pdf of $\psi(x) = \frac{1}{\sqrt{2}\sigma}\exp(-|x|/\sqrt{\sigma/2})$ with variance σ. The resulting objective of $\mathbf{P_{mle}}$ is proportional to $\sigma_1^{-1/2}\sum_{i=1}^{n_w}\left|\tilde{w}_{t_i} - w_{t_i}\right| + \sigma_2^{-1/2}\sum_{i=1}^{n_u}\left|\tilde{u}_{\tau_i} - u_{\tau_i}\right| + \sigma_3^{-1/2}\sum_{t=1}^{n}|z_t|$.

Remark 5. *The objective function resulting from the Laplace case becomes a linear objective function after a minor reformulation (see, for example, Section 6.1.1 of (Boyd & Vandenberghe, 2004)).*

We assume the noise is Laplacian because this results in MILP optimization problems for estimation and prediction. Note that if we had assumed Gaussian noise, then this would have resulted in MIQP optimization problems for estimation and prediction. We have found that these resulting MIQP's are solvable using standard software, but that the prediction accuracy was not better than that of the MILP formulations arising from the Laplacian assumption. Hence we chose to assume Laplace noise because of the faster computation time for the resulting MILP's. The similar predictive accuracy under both assumptions is not surprising given that the difference in the objective is simply an absolute value of deviation versus the square of deviation.

4.3. Reformulating the MLE Using KKT

One approach to solving bilevel programs is to replace the convex optimization problems that are constraints by their corresponding necessary and sufficient optimality conditions (Dempe, 2002).

Proposition 3. *Necessary and sufficient optimality conditions for* $\mathbf{U_{no\ goals}}$ *can be written as*

$$\begin{aligned} &2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q = 0 \\ &2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_0 = 0. \end{aligned} \tag{4}$$

Proof. Because the constraints in $\mathbf{U_{no\ goals}}$ can be eliminated by rewriting the problem as $(u_t, f_t) = \arg\max_{u,f} -(a \cdot w_t + b \cdot u_t + c \cdot f_t + k)^2 - r_u u_t^2 + q u_t - r_f f_t^2 + s_t f_t$, the KKT conditions consist of only the stationarity conditions and are given by (4). A minor note is that $s_t = s_0$ here, because there are no dynamics on s_t when goals are not provided as in $\mathbf{U_{no\ goals}}$.

Proposition 4. *Neccesary and sufficient optimality conditions for* $\mathbf{U_{goals}}$ *can be written as*

$$\begin{aligned} &2b(aw_t + bu_t + cf_t + k) + 2r_u u_t - q - \lambda_t^2 = 0 \\ &2(aw_t + bu_t + cf_t + k) + 2r_f f_t - s_t = 0 \\ &g_t - \epsilon - (g_t - \epsilon) \cdot x_t^1 \le u_t \le M + (g_t - \epsilon - M) \cdot x_t^1 \\ &(g_t - \epsilon) \cdot x_t^2 \le u_t \le M + (g_t + \epsilon - M) \cdot x_t^2 \\ &(g_t + \epsilon) \cdot x_t^3 \le u_t \le g_t + \epsilon + (M - g_t - \epsilon) \cdot x_t^3 \\ &0 \le \lambda_t^2 \le p_t; \quad p_t - M \cdot (1 - x_t^1) \le \lambda_t^2 \le M \cdot (1 - x_t^3) \\ &x_t^1 + x_t^2 + x_t^3 = 1; \quad x_t^1, x_t^2, x_t^3 \in \{0, 1\}. \end{aligned} \tag{5}$$

Proof. Computing optimality conditions for U_{goals} requires reformulation as a quadratic program (QP) by using $p_t \cdot (u_t - g_t)^- = -\max\{-p_t \cdot (u_t - g_t), 0\}$. This QP reformulation has a differentiable, strictly concave objective and satisfies the linear independence constraint qualification (LICQ), and so the KKT conditions are necessary and sufficient for optimality. The KKT conditions can be rewritten after some manipulation as the first two lines of (5) combined with the following logical conditions on the Lagrange multipliers: $\lambda_t^2 = p_t$ if $u_t < g_t$, $0 \le \lambda_t^2 \le p_t$ if $u_t = g_t$, and $\lambda_t^2 = 0$ if $u_t > g_t$. Finally, let M be a constant such that $M \ge p_t$. Using a big-M formulation (Vielma, 2015), we can express these logical conditions as in (5). □

Remark 6. *We include an* $0 < \epsilon \ll 1$ *term to ensure all three regions for the integer program have a non-zero width. The resulting regions are* $u_t \le g_t - \epsilon$, $g_t - \epsilon < u_t \le g_t + \epsilon$, *and* $u_t \ge g_t + \epsilon$, *and note that the binary variables* x_1^t, x_t^2, x_t^3 *indicate if* u_t *respectively belongs to one of these three regions.*

Remark 7. *If g_t is not fixed, as would be the case in an optimization problem for personalizing physical activity goals, then the constraints (5) can be further reformulated as MILP constraints using the approach discussed in Section 4.5.*

4.4. Exercise Goal Inequalities to Constrain Integer Variables

We define an additional set of inequalities that lead to order of magnitude faster computation times when computing the MLE. Social cognitive theory suggests that if an exercise goal g_t is not achieved at a particular time point t (i.e., $u_t < g_t$), then it will not be achieved at time $t + 1$ unless the goal decreases $g_{t+1} < g_t$ or an office visit occurs $d_{t+1} = 1$. This insight leads to additional inequalities on the integer variables.

Proposition 5. *For fixed g_t, the logical constraint* $(u_t < g_t, g_{t+1} \geq g_t, d_{t+1} = 0) \Rightarrow (u_{t+1} < g_{t+1})$ *can be formulated as linear inequalities:*

$$\begin{aligned} x^1_{t+1} &\geq x^1_t - d_{t+1} - \mathbb{1}(g_{t+1} - g_t < 0) \\ x^2_{t+1} &\leq x^2_t + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0) \\ x^3_{t+1} &\leq x^3_t + d_{t+1} + \mathbb{1}(g_{t+1} - g_t < 0)\,. \end{aligned} \qquad (6)$$

Proof. The first inequality states $x^1_{t+1} \in \{0, 1\}$ (which indicates if $u_{t+1} < g_{t+1}$) can only decrease from x^1_{t+1} if the goal decreases ($g_{t+1} < g_t$) or there is an office visit ($d_{t+1} = 1$). Similarly, the second and third inequalities state $x^2_{t+1}, x^3_{t+1} \in \{0, 1\}$ can only increase from x^2_t, x^3_t if the goal decrease: is an office visit ($g_{t+1} < g_t$) or there is an office visit (d_{t+1}=1).

Remark 8. *When g_t is not fixed, the above constraints (6) can be further reformulated as MILP constraints_ using big-M formulations (Vielma, 2015).*

These inequalities further constrain the estimates beyond the equations of the utility-maximizing framework. Restated, depending upon the parameters the utility-maximizing framework could potentially predict that goals are not attained at t but then attained at $t + 1$ because of an increase in weight $w_{t+1} > w_t$. We constrain the parameters using the inequalities (6) so as to prevent such behavior in the utility-maximizing framework.

4.5. Addressing Bilinear Terms

Because we are jointly estimating noisy/missing data and parameters, our optimization model contains nonconvex quadratic terms. For instance, the dynamics on p_t (3) have the nonconvex quadratic term $\mu \cdot \mathbb{1}(u_t \geq g_t)$. It is difficult to directly solve nonconvex mixed-integer quadratically constrained quadratic programs (MIQCQP) problems, and so we discuss reformulations that allow us to solve the resulting problem more efficiently. We begin by reformulating (3).

Proposition 6. *The dynamics on p_t (3) can be represented by the linear constraints:*

$$
\begin{aligned}
p_{t+1} &\geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} \\
p_{t+1} &\leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + M \cdot (1 - x_t^1) \\
p_{t+1} &\geq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu - M x_t^1 \\
p_{t+1} &\leq \gamma \cdot p_t + \delta_{t+1} \cdot d_{t+1} + \mu \,.
\end{aligned} \tag{7}
$$

Proof. Recall that (3) has the nonconvex quadratic term $\mu \cdot \mathbb{1}(u_t \geq g_t)$. Using the integer variables x_t^1, x_t^2, x_t^3 from the integer-reformulated KKT conditions (5), we can express this as the bilinear term $\mu \cdot (1 - x_t^1)$. This term has the special structure of a binary variable multiplied by a continuous scalar, and so a standard exact-linearization approach (Glover, 1975; Torres, 1991) can be used to reformulate the dynamics on p_t as in (7). □

Finally, to eliminate bilinear terms in our MLE formulation note that the exact-linearization dynamics of p_t (7) have the term $\gamma \cdot p_t$, the dynamics of s_t (2) have the term $\gamma \cdot (s_t - s_0)$, and the integer-reformulated KKT conditions for decision-making with goals (5) have the terms $2r_u u_t$, $2r_f f_t$. When we fix the value of γ, r_f, r_u, the resulting MLE formulation will be a MILP. We use an enumeration approach, as described in the next subsection, to address these final bilinear terms.

4.6. MILP Formulation of MLE

We reformulate the initial MLE problem $\mathbf{P_{mle}}$ as optimization problem described below, $\mathbf{P_{mle\text{-}milp}}$. This is a MILP for fixed values of γ, r_f, r_u and after rewriting the absolute values using linear constraints (as in Section 6.1.1 of (Boyd & Vandenberghe, 2004), for example), because a, b, r_f, r_u, γ are constants when solving $\mathbf{P_{mle\text{-}milp}}$. The full MILP formulation for MLE can be found in the Supplementary Materials.

$$
\mathbf{P_{mle-milp}} \quad \begin{array}{l} \min \sigma_1^{-1/2} \sum_{i=1}^{n_w} \left| \tilde{w}_{t_i} - w_{t_i} \right| + \sigma_2^{-1/2} \sum_{i=1}^{n_u} \left| \tilde{u}_{t_i} - u_{t_i} \right| + \sigma_3^{-1/2} \sum_{t=1}^{n} |z_t| \\ \text{s.t.} \quad (1),\ (4), \quad \text{for } t = 1, \ldots, m-1; \quad (1),(2),(5),(6),(7), \quad \text{for } t = m, \ldots, n \,. \end{array}
$$

Recall that (1) are weight dynamics, (2) are dynamics on the s_t parameter, (4) are KKT conditions for the decision-making model without goals, (5) are integer-reformulated KKT conditions for the decision-making model with goals, (6) are the exercise goal inequalities that constrain the integer variables, and (7) are exact-linearization dynamics of p_t.

If γ, r_f, r_u are not fixed, then $\mathbf{P_{mle\text{-}milp}}$ is a nonconvex MIQCQP. To solve $\mathbf{P_{mle\text{-}milp}}$, observe that we can enumerate over γ, r_f, r_u and solve a series of MILP's. This is computationally viable because we only need to enumerate over three variables. We can gain an additional computational speedup by using a simple and accurate approximation that allows us to compute the MLE by solving a single MILP. The approximation is due to an observation we made while using enumeration to compute the MLE. We noticed that the MLE was insensitive to the values of γ, r_f, r_u: There was less than a 5% difference in the objective value

over a large range of values for $\gamma \in [0.8,1]$ and $r_f, r_u \in [1 \times 10^{-7}, 1 \times 10^{-5}]$, and the estimates of the type parameters were relatively constant over this range as well. As a result, we approximate this problem by fixing $\gamma = 0.85$, $r_f = 8.1633 \times 10^{-6}$, and $r_u = 1 \times 10^{-6}$ for all individuals: This allows us to compute the MLE by solving $\mathbf{P}_{\mathbf{mle\text{-}milp}}$ for a single value of γ, r, which is a single MILP. This approximation also reduces the number of parameters we are trying to estimate.

5. Bayesian Predictions of Individual Trajectories

Problem $\mathbf{P}_{\mathbf{mle\text{-}milp}}$ provides joint estimation of noisy/missing data and model parameters. However, this is not by itself useful for predicting the future weight loss trajectory of an individual given a short period of initial data. We would ideally like a framework to provide such predictions under different intervention scenarios, since this would support the adaptive design of personalized interventions. Moreover, we would like the predictions to be able to leverage past/historical data in order to improve the accuracy of predictions. Given this last constraint, a natural choice for predictions is to use this past data for a prior distribution in a Bayesian framework.

In particular, suppose we have past/historical data from many individuals that have completed the entire weight loss intervention. We can perform MLE to estimate the parameters for the utility-maximizing framework for each of these individuals. Then we can form our priors by computing histograms of these estimates. Let tf be the total length of an intervention, and define $\Theta = (\theta_1, \ldots, \theta_{tf})$. We use the pdf notation $\hat{\psi}(\Theta)$ to collectively refer to a set of histograms for the parameters Θ, because these histograms are be assumed to be normalized such that they are a pdf.

Now suppose we have an additional individual that has completed only T days of the intervention and has a remaining $t_f - T$ days left in the intervention, where t_f is the total days in the intervention. The data available for this new individual is $(t_i, \tilde{w}_{t_i})$, for $i = 1, \ldots, n_w$, and $(\tau_i, \tilde{u}_{\tau_i})$, for $i = 1, \ldots, n_u$, where n_w are the number of weight measurements, n_u are the number of step measurements, and the noise model is as before. We would like to construct an optimization model whose solution provides a prediction of the distribution of the individual's weight at the end of the intervention at time t_f using the histograms of the past individuals and the first T days of data for this new individual. In this section, we demonstrate that we can incorporate data-driven histograms as priors in Bayesian estimation using integer programming.

5.1. Initial Formulation for Bayesian Estimation

Our goal is to compute $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U})$, which is the *posterior distribution* of weight at the end of the intervention w_{t_f} conditioned (i) on the intervention parameters $C = (d_1, g_1, \ldots, d_{t_f}, g_{t_f})$, and (ii) on the data available for the new individual

$\widetilde{W} = ((t_i, \widetilde{w}_{t_i}),\ \text{for } i = 1, \ldots, n_w)$ and $\widetilde{U} = ((\tau_i, \widetilde{w}_{\tau_i}),\ \text{for } i = 1, \ldots, n_u)$. To accomplish this, we apply Bayes's theorem and then eliminate nuisance parameters by averaging over them.

First we calculate $\psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U})$, which is the joint posterior distribution of weight $W = (w_1, \ldots, w_{t_f})$, steps $U = (u_1, \ldots, u_{t_f})$, caloric intake $F = (f_1, \ldots, f_{t_f})$, and type Θ. This requires specifying prior distributions for W, U, F, Θ. The typical approach is to choose priors that admit efficient computation or are uninformative/non-constraining (Gelman et al., 2013). Because we have data from past individuals, we can use the histogram $\hat{\psi}(\Theta)$ as a prior distribution for Θ. We choose a uniform prior distribution for W, U, F because this is relatively uninformative/non-constraining (Gelman et al., 2013). Consequently, applying Bayes's theorem yields $\psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U}) = \frac{1}{Z} \cdot \psi(\widetilde{W}, \widetilde{U} | W, U, F, \Theta, C) \cdot \hat{\psi}(\Theta)$, where $\psi(\widetilde{W}, \widetilde{U} | W, U, F, \Theta, C)$ is the likelihood of the observations conditioned on the parameters of the utility-maximizing framework, Z is a normalization constant that ensures the integral of the posterior is one, and we have used the fact that $\psi(W) = \psi(U) = \psi(F) = 1$ over their supports since they are uniform. Recall that the log-likelihood (i.e., $\log \psi(\widetilde{W}, \widetilde{U} | W, U, F, \Theta, C)$) is given by the objective and constraints of $\mathbf{P_{mle\text{-}milp}}$.

The next step is to eliminate nuisance parameters, which can be accomplished in principle by averaging (Gelman et al., 2013). Averaging gives $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U}) = \int \psi(W, U, F, \Theta | C,\ \widetilde{W}, \widetilde{U}) \cdot dW_{-t_f} \cdot dU \cdot dF \cdot d\Theta$, where $W_{-t_f} = (w_1, \ldots, w_{t_f - 1})$. However, this integral is difficult to compute both symbolically (because of integer constraints in the formulation of the model) and computationally (the posterior $\psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U})$ can be sharply peaked and so Monte Carlo-based approaches converge slowly). (In fact, our initial approach was to use a Monte Carlo algorithm to compute the posterior distribution, but we found through empirical testing that the resulting posterior was simply a uniform distribution with a very broad support, which indicates convergence to the actual posterior was too slow for making accurate predictions with the posterior; such slow convergence is not surprising given the high-dimensionality of the nuisance parameters.) Our approach is to use the profile likelihood (Severini, 1999; Murphy & Vaart, 2000) as an approximation: The profile likelihood is computed by an optimization problem $\mathbf{P_{pl}}$ that is given by $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U}) \approx \max_{W - t_f, U, F, \Theta} \psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U})$, and our approximation can be justified by arguments relating the asymptotic consistency of Bayesian and MLE estimation under general conditions (Severini & Wong, 1992; Severini, 1999; Gelman et al., 2013). The key computational question is how to solve $\mathbf{P_{pl}}$. The normalizing factor Z can be computed by numerically integrating a one-dimensional function.

5.2. Histogram Construction

Before constructing the histograms defining $\hat{\psi}(\Theta)$, we need to specify which parameters are statistically independent. Assuming every parameter is correlated will not be successful because it would require high-dimensional histograms, which will be a statistically poor estimate of the true parameter distribution. Hence, specifying that some parameters are independent will enable expressing $\hat{\psi}(\Theta)$ in terms of low-dimensional histograms. Therefore,

we assume that μ, q, s_0, β_0, δ_0 are jointly independent. Furthermore, we assume that β_{K+1} conditioned on β_k is jointly independent with the other parameters. Similarly, β_{k+1} conditioned on β_k is assumed to be jointly independent with the other parameters. Lastly, we assume that the conditional relationships between β_{K+i}, β_k and between δ_{k+1}, δ_k are not a function of k. (Note that $\mathbf{P_{mle}}$ does not make any such assumptions about the conditional dependence of the β_k or the δ_k, by which we mean that $\mathbf{P_{mle}}$ assumes that μ, q, s_0, β_0, δ_0, β_1, δ_1, β_2, δ_2,… are jointly independent.)

Remark 9. *Under our assumptions, we can factor the histogram as* $\widehat{\psi}(\Theta) = \widehat{\psi}(\mu) \cdot \widehat{\psi}(q) \cdot \widehat{\psi}(s_0) \cdot \widehat{\psi}(\beta_0) \cdot \prod_{k=0}^{n_d} \widehat{\psi}(\beta_{k+1} | \beta_k) \cdot \widehat{\psi}(\delta_0) \cdot \prod_{k=0}^{n_d} \widehat{\psi}(\delta_{k+1} | \delta_k)$, *where* n_d *be the number of office visits.*

It will be the case that the objective function we use will involve the logarithm of $\widehat{\psi}(\Theta)$, and so the above remark implies that we can construct a MILP formulation of the resulting optimization problem as long as we are able to define MILP representations of $\widehat{\psi}(X), \widehat{\psi}(X_{k+1} | X_k)$, where X is a random variable. Observe, that these constituent histograms are piecewise constant:

Remark 10. *We can represent the one-dimensional histogram for parameter X (where X could be any of μ, q,* s_0, β_0, δ*) as* $\widehat{\psi}(X) = \sum_{i=1}^{m_x} \pi_i^x \cdot \mathbb{1}(h_i^x \leq X \leq h_{i+1}^x)$, *where* m_x *is the number of bins,* h_i^x *are the edges of these bins, and* π_i^x *is the value of the histogram in the i-th bin.*

Remark 11. *We can represent the histograms for parameter* X_{k+1} *conditioned on* X_k *(where X could be any of* β_k, δ_k*;) as* $\widehat{\psi}(X_{k+1} | X_k) = \sum_{i=1}^{m_x} \sum_{j=1}^{\eta_x} \pi_{i,j}^x \cdot \mathbb{1}(h_i^x \leq X_{k+1} \leq h_{i+1}^x) \cdot \mathbb{1}(\phi_j^x \leq X_k \leq \phi_{j+1}^x)$, *where* m_x *is the number of bin divisions in the* X_{k+1} *dimension,* X_k *is the number of bin divisions in the* X_k *dimension,* h_i^x *are the edges of the bins in the* X_{k+i} *dimension,* ϕ_i^x *are the edges of the bins in the* X_{k+1} *dimension, and* $\pi_{i,j}^x$ *is the value of the histogram in the (i,j)-th bin. Note that the histogram values* $\pi_{i,j}^x$ *should be normalized such that the above representation is a conditional distribution – an incorrect normalization would cause the above representation to be a joint distribution instead.*

5.3. MILP Formulation for Computing Posterior Distribution of Final Weigh

One of our goals is to show that data-driven prior distributions can be used to perform Bayesian estimation by formulating the problem as a MILP. Here, we focus on approximating the posterior $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U})$ by solving $\mathbf{P_{pl}}$. It is worth noting that an almost identical formulation can be used to perform Bayesian *maximum a posteriori* (MAP) estimation with data-driven priors by solving problem $\mathbf{P_{map}}$, which is given by $\max_{W,U,F,\Theta} \psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U})$; compare this problem to $\mathbf{P_{pl}}$. Observe that because a histogram is a piecewise constant function, it can be represented using inequality constraints

with integers (Vielma, 2015). This requires some minor reformulations, which we describe below, in order to ensure linearity of the optimization model.

Proposition 7. *The objective of* $\mathbf{P_{pl}}$ *(after computing its negative logarithm) is*

$$\sigma_1^{-1/2}\sum_{i=1}^{n_w}|\widetilde{w}_{t_i}-w_{t_i}|+\sigma_2^{-1/2}\sum_{i=1}^{n_u}|\tilde{u}_{\tau_i}-u_{\tau_i}|+\sigma_3^{-1/2}\sum_{t=1}^{n}|z_t|+ \quad (8)$$
$$2^{-1/2}\sum_{X\in\{\mu,q,s_0,\beta_0,\delta_0\}}\sum_{i=1}^{m_x}\log\pi_i^x\cdot y_i^x+2^{-1/2}\sum_{X\in\{\beta,\delta\}}\sum_{k=0}^{n_d-1}\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}\log\pi_{i,j}^x \cdot y_{i,j}^{x,k},$$

subject to constraints for one-dimensional histograms

$$\sum_{i=1}^{m_x}h_i^x\cdot y_i^x\le X\le\sum_{i=1}^{m_x}h_{i+1}^x\cdot y_i^x;\quad \sum_{i=1}^{m_x}y_i^x=1;\quad y_i^x\in\{0,1\},\ \forall i=1,\ldots,m_x, \quad (9)$$

for all $X\in\{\mu,q,s_0,\beta_0,\delta_0\}$*, and constraints for conditional histograms*

$$\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}h_{i,j}^x\cdot y_{i,j}^{x,k}\le X_{k+1}\le\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}h_{i+1}^{x,k}\cdot y_{i,j}^x \quad (10)$$
$$\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}\phi_{i,j}^x\cdot y_{i,j}^{x,k}\le X_k\le\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}\phi_{i+1}^x\cdot y_{i,j}^{x,k}$$
$$y_{i,j}^{x,k}\in\{0,1\},\ \forall i=1,\ldots,m_x,\ j=1,\ldots,\eta_x;\ \sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}y_{i,j}^{x,k}=1,$$

for all $X\in\{\beta\delta\}$ *and* $k=0,\ldots,n_d-1$.

Proof. First, note that the objective of $\mathbf{P_{pl}}$, after computing its negative logarithm, is proportional to:

$$\sigma_1^{-1/2}\sum_{i=1}^{n_w}|\widetilde{w}_{t_i}-w_{t_i}|+\sigma_2^{-1/2}\sum_{i=1}^{n_u}|\tilde{u}_{\tau_i}-u_{\tau_i}|+\sigma_3^{-1/2}\sum_{t=1}^{n}|z_t|,$$
$$+2^{-1/2}\sum_{X\in\{\mu,q,s_0,\beta_0,\delta_0\}}\sum_{i=1}^{m_x}\log\pi_i^x\cdot\mathbb{1}(h_i^x\le X\le h_{i+1}^x)$$
$$+2^{1/2}\sum_{X\in\{\beta,\delta\}}\sum_{k=0}^{n_d-1}\sum_{i=1}^{m_x}\sum_{j=1}^{\eta_x}\log\pi_{i,j}^x\cdot\mathbb{1}(h_i^x\le X_{k+1}\le h_{i+1}^x)\cdot\mathbb{1}(\phi_i^x\le X_k\le\phi_{i+1}^x)$$

where we have used the factorization of $\hat{\psi}(\Theta)$, the equation for one-dimensional histograms, the equation for the conditional histograms, and the equation for log $\log\psi(\widetilde{W},\widetilde{U}\,|\,W,U,F,\Theta,C)$ from $\mathbf{P_{mle\text{-}milp}}$. By defining $y_i^x\in\{0,1\}$ for parameters X ∈ {μ, q, so, βo, δ_0} and $y_{i,j}^{x,k}\in\{0,1\}$ for parameters $X\in\{\beta,\delta\}$ the objective function as in the hypothesis of the proposition. □

Remark 12. *An important benefit of the rewritten objective (8) and subsequent constraints (9), (10) is that they are linear in the decision variables. Note that the y decision variables in these equations are binary variables and indicate which bin of the histogram the corresponding variable belongs to.*

The posterior is computed by solving a series of MILPs and then using numerical integration to compute the normalization constant Z. In particular, define the following parametric (in ω) MILP:

$$\mathbf{P_{pl\text{–}milp}} \quad \begin{aligned} \ell(w_{t_f} = \omega) = \min\ & (8) \\ \text{s.t.}\ & (1),(4),\ \text{for } t = 1,\ldots,m-1; \quad (1),(2),(5),(6),(7),\ \text{for } t = m,\ldots,n \\ & (9),\ \text{for } X \in \left\{\mu, q, s_0, \beta_0, \delta_0\right\} \\ & (10),\ \text{for } X \in \left\{\beta, \delta\right\},\ k = 0,\ldots,n_d - 1; \quad w_{t_f} = \omega. \end{aligned}$$

The complete formulation can be found in the Supplementary Materials.

Let $\kappa_2 = \min_i \ell(w_{t_f} = \omega_i)$. If we solve $\mathbf{P_{pl\text{-}milp}}$ over a grid of values $\omega_1, \ldots, \omega_{n_g}$, then we can compute the normalization Z by numerically integrating the set of points $(\omega_i,\ \exp(-\ell(w_{t_f} = \omega_i) + \kappa_2))$, for $i = 1,\ldots, n_g$, where (i) we take the exponent of the negative of $\ell(\cdot)$ because we reformulated the objective for our MILP using a negative logarithm, and (ii) we scale this exponent using κ_2 because this improves the numerics of the computations. Finally, the posterior at ω_i is given by $\psi(w_{t_f} = \omega_i | C, \widetilde{W}, \widetilde{U}) = \exp(-\ell(w_{t_f} = \omega_i) + \kappa_2)/Z$. Consequently, we can approximate the posterior distribution of w_{t_f} by solving a series of problem $\mathbf{P_{pl\text{-}milp}}$. Observe that in this approximation process, we are in fact approximating the posterior likelihood of the final weight $w_{t_f} = \omega$ at different ω using different patient behaviors trajectories. Such an approximation approach has been previously proposed and is well-behaved asymptotically as more data is collected (Lindley, 1961; Tierney & Kadane, 1986; Evans & Swartz, 1995). The intuition from Evans & Swartz (1995) for why such an approximation is justified begins with the defining integral $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U}) = \int \psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U}) \cdot dW_{-t_f} \cdot dU \cdot dF \cdot d\Theta$. For a fixed w_{t_f}, by the law of large numbers most of the mass of $\psi(W, U, F, \Theta | C, \widetilde{W}, \widetilde{U})$ is concentrated about its maximizer, which corresponds to the minimizer of $\mathbf{P_{pl\text{-}milp}}$. Hence we can approximate this integral by considering its behavior at the optimizer. A complete proof of the theoretical validity of our approximation is found in our companion paper (Mintz et al., 2017).

6. Computational Results and Validation of Predictive Modeling

In this section, we first describe the data source used for the computational results and validation of our predictive model. Next, we provide computational results of solving $\mathbf{P_{pl\text{-}milp}}$ to compute MLE and of solving $\mathbf{P_{pl\text{-}milp}}$ to compute the Bayesian predictive model. Representative plots are shown in these first two subsections. Cross-validation (Hastie et al., 2009) is used to validate our approach through comparison to a benchmark approach from machine learning, and we specifically consider the prediction of 5% weight loss at 5 months based on the first 30 days of an individual's data. This validation compares all individuals in the data set. We conclude by demonstrating the ability of our approach to make predictions

on the weight loss trajectory of an individual as the number of counseling sessions is changed, and we discuss how this can be used for optimization.

6.1. Data Source of Mobile Phone Delivered Diabetes Prevention Program (mDPP) Trial

We used data from the mDPP trial (Fukuoka et al., 2015), which was a randomized controlled trial (RCT) to evaluate the efficacy of a 5-month mobile phone-based weight loss intervention among overweight English-speaking adults at risk for developing T2DM. The intervention was adapted from the Diabetes Prevention Program (DPP) 2002; 2009, but the frequency of in-person sessions was reduced from 16 to 6 sessions and group exercise sessions were replaced with a home based exercise program to reduce costs. Sixty-one overweight adults were randomized to an active control (accelerometer only) ($n = 31$) group or an mDPP mobile app plus accelerometer intervention (n = 30). Demographics are available in (Fukuoka et al., 2015), and changes in primary and secondary outcomes were promising: The intervention group lost an average of 6.2 ± 5.9 kg (−6.8% ± 5.7%) between baseline and 5-month follow-up compared to the control group's gain of 0.3 ± 3.0 kg (0.3% ± 5.7%) ($p < 0.001$). The intervention group's steps per day increased by 2551 ± 4712 compared to the control's group decrease of 734 ± 3308 steps per day ($p < 0.001$).

The data available from this RCT matches that described in Section 2. Specifically, we have step data from a digital accelerometer and body weight data recorded at least twice a week every week into the mobile app. We also have access to the age, gender, and height of each individual. After an initial two week period, exercise goals in desired number of steps per day were provided to each individual. The goals increased by 20% each week, starting at 1.2 times the average number of steps during the initial two weeks; the goals increased to a maximum of 12,000 steps a day (about 6 miles of walking). Individuals were also asked to make office visits (at 2, 4, 6, 10, 14, 18, and 20 weeks) during which they received behavioral counseling about their nutritional choices and physical activity.

6.2. Computational Results

We used the Gurobi solver (Gurobi Optimization, 2015) to solve $\mathbf{P}_{\mathbf{mle\text{-}milp}}$ and $\mathbf{P}_{\mathbf{pl\text{-}milp}}$. The CVX toolbox (Grant & Boyd, 2014) for MATLAB was used to generate each instance of the MILP. A 2.5GHz laptop computer with 4Gb of RAM was used to generate these results.

6.2.1. Results of MLE for Utility-Maximizing Model—The problem $\mathbf{P}_{\mathbf{mle\text{-}milp}}$ was solved foe rach individual in the mDPP. The fastest computation time was 3 sec, the slowest computation time was 550 sec, and the median computation time was 10 sec. The second and third quartiles of computation time were 6 sec and 70 sec, respectively. Overall, the computation was quick and can be easily parallelized because each MILP is solved independently.

Figure 1 shows a representative example of the weight, steps, and caloric intake trajectory estimated by solving $\mathbf{P}_{\mathbf{mle\text{-}milp}}$. The blue dots are measured data, and the red lines are estimated trajectories. The utility-maximizing framework captures increasing positive impacts from achieving exercise goals, as well as negative impacts from not meeting goals. The MLE reduces noise in measured data and estimates values for time points without data.

Observe that the large drops in caloric intake correspond to reductions in the preference of caloric consumption s_t that occurs after an office visit; however, the reductions are not constant for each office visit. This is because the impact of an office visit is characterized by β_t, δ_t, which are random variables. Moreover, when we computed the conditional histograms for β_{t+1} given β_t and for δ_{t+1} given δ_t, we empirically found that these histograms were such that they indicated subsequent office visits are generally less effective in encouraging increases in physical activity and reductions in caloric intake.

From a clinical standpoint, an additional benefit of our utility-maximizing framework is its ability to estimate caloric intake. Effective mobile technologies for directly measuring caloric intake are not commercially available, and self-reported caloric intake diaries are known to be highly inaccurate (Schoeller et al., 1990). Our approach indirectly estimates this by integrating physiology into the framework. This can be used to improve self-monitoring of an individual's food consumption.

6.2.2. Results of Bayesian Trajectory Prediction using MILP Formulation—

Problem $\mathbf{P_{pl\text{-}milp}}$ was solved using the first month of data for each individual in the intervention group of the mDPP in order to compute a posterior distribution of w_{t_f}. To generate the histograms for $\mathbf{P_{pl-milp}}$, we used the MLE parameters for the remaining individual computed using the entire data set for these individuals. We did *not* use an individual's data when computing the histogram used to make predictions for that particular individual; we constructed a different histogram for each individual by using the data excluding that individual.

For our computations, we chose $n_g = 100$ grid points at which we computed the posterior. The fastest, slowest, and median computation times were 190 sec, 1000 sec, and 360 sec, respectively. The second and third quartiles of computation time were 230 sec and 470 sec, respectively. Overall, the computation was relatively quick and can be easily parallelized because each MILP is solved independently.

A representative example of the posterior likelihood $\psi(w_{t_f} | C, \widetilde{W}, \widetilde{U})$ for the final weight of an individual (at 5 months) conditioned on 1 month of weight and step data is shown in Figure 2. The dashed line denotes the initial weight of the individual before starting the weight loss intervention, and the dotted line represents a final weight corresponding to 5% weight loss. We can also plot the entire weight, exercise, and caloric intake trajectories corresponding to the MAP estimate: This is shown in Figure 3. Data from the first month (dark blue and left of the dotted line) was used to compute the posterior and the MAP estimate of the past and future trajectories. The MAP prediction of the future trajectories is compared to the actual measurements (light blue and right of the dotted line); there is good agreement between the predicted and actual weight trajectories. An additional benefit of this approach is its ability to estimate past caloric intake.

6.3. Predicting Clinically Significant Weight Loss

This subsection evaluates the ability of our predictive model from Section 5 to predict whether an individual will achieve clinically significant weight loss at the end of the

intervention. We refer to a situation where an individual achieves 5% weight loss as a *positive*, and similarly if an individual does not achieve 5% weight loss then this is be a *negative.* We validate the predictive capabilities of our model by comparing it to three standard methods from machine learning. Specifically we consider a linear support vector machine (SVM) model, a decision tree model, and a logistic regression model for classification (Hastie et al., 2009). We additionally consider a version of our predictive model that does not incorporate a Bayesian prior in order to validate that our Bayesian approach improves prediction accuracy. For the purpose of comparison, we specifically consider a scenario in which the first 30 days of mobile phone data are used to predict whether an individual will achieve 5% weight loss after 5 months of participating in the weight loss intervention. Cross-validation (Hastie et al., 2009) is used to separate the data into a training set that is used to estimate the models and a hold-out set that is used to quantitatively validate the model.

6.3.1. Machine Learning Models—Let $x \in \mathbb{R}^2$ be a vector of percent weight loss to date and percent of step goals met, and let y be such that if y =1 then an individual has achieved at least 5% weight loss and y = —1 otherwise. Machine learning methods use data in this form to fit functions $f: \mathbb{R}^2 \rightarrow \{-1,1\}$ to best capture the relationship between x and y. We refer to the output of this function as $\hat{y}(x) = f(x)$, to signify that we are generating an estimate of the y values. The value $\hat{y}(x) = -1$ is a prediction that the individual *will not* achieve 5% weight loss after 5 months, and $\hat{y}(x) = +1$ is a prediction that the individual *will* achieve 5% weight loss after 5 months.

A *linear SVM* is the predictive model $\hat{y}(x) = \text{sign}(\beta_0 + x'\beta)$. The hyperplane $\beta_0 + x'\beta$ cuts the space $\mathbb{R}^p$ into two regions, and the two sides of the hyperplane are predicted to be positive or negative, respectively. The parameters β_0, β are computed by a quadratic program (Hastie et al., 2009), and we used the MATLAB Statistics and Machine Learning Toolbox to identify the SVM parameters using data from the mDPP trial. The identified parameters are 64 for percent weight loss to date and 1.715 for percent of step goals met; the parameter values normalized by sample standard deviation were similar. These magnitudes indicate that for predicting 5% weight loss: percent weight loss to date is the most important feature and percent of step goals met is the second most important. Because all parameters are positive, this means increased weight loss to date and percent of step goals met both lead to increased likelihood of achieving 5% weight loss.

A *decision tree* model (i.e., classification and regression trees or CART) is a sequential classifier. Each node of the tree partitions a different column of the data to ensure maximum separation between the two classes, and each leaf of the tree is assigned a label of 1 or −1. For prediction, data is compared along the nodes of the tree and then assigned a value that corresponds to the leaf of the final comparison. Computing the optimal decision tree model is NP-hard, and heuristics are used to construct these models (Hastie et al., 2009). For our implementation, we used the MATLAB Statistics and Machine Learning Toolbox to train the decision tree model from the mDPP data. Our trained decision tree model first branches on the percent of weight lost to date, with those who lost at least 2.6% being classified to the class which will achieve 5% weight loss. Next, the model branches on the average amount of

exercise goals met, with those who met at least 84% of their goals being classified as successful and the remainder as unsuccessful.

A *Logistic Regression* model specifies a classifier of the form $\hat{y}(x) = 2 \cdot 1\left\{\frac{1}{1+\exp(-\beta_0 - x'\beta}} \geq \frac{1}{2}\right\} - 1$. This probabilistic interpretation of this classifier is that the labels transformed to {0,1} follow a Bernoulli distribution with parameter p where $\log\left(\frac{p}{1-p}\right) = \beta_0 + x'\beta$. Hence if the probability that $y = 1$ is greater than 0.5 we predict $\hat{y} = 1$, and otherwise we predict $\hat{y} = -1$. The problem of training a logistic regression model can be posed as a convex optimization problem (Hastie et al., 2009) that can be solved by stochastic gradient descent. For our analysis, we used the MATLAB Statistics and Machine Learning Toolbox to train the logistic regression model. The coefficient for weight lost to date was 73 and the coefficient for percent of goals met was 0.889. These coefficients are similar to the SVM coefficients, which is unsurprising since logistic regression can be interpreted as a continuous relaxation of linear SVM (Hastie et al., 2009).

6.3.2. Adjusting True and False Positive Rate of Predictions—The quality of our models can be evaluated by estimating and comparing the true and false positive rates of different models. The *true positive rate* (TPR) specifies the probability of a model correctly predicting a positive, and the *false positive rate* (FPR) quantifies the probability of a model incorrectly predicting a positive. In making predictions, there is tradeoff between the TPR and FPR. It is customary for practitioners to choose the FPR, and this choice fixes the TPR (Bickel & Doksum, 2006; Lehmann & Romano, 2006). Choosing the FPR requires an understanding of how the model is used to make predictions and how parameters in the model impact the FPR. For instance, we can adjust the FPR of a linear SVM model by choosing the value of β_0. For example, if $\beta_0 = -\infty$, then the prediction will always be −1; similarly, if $\beta_0 = +\infty$, then the prediction will always be +1. By choosing intermediate values for β_0, we can adjust the FPR of the model. To specify the FPR of the Bayesian predictive model, we compute the posterior probability of 5% weight loss $\mathbb{P}(w_{t_f} \leq 0.95 w_0 | C, \widetilde{W}, \widetilde{U}) = \int_{-\infty}^{0.95 w_0} \psi(w_{t_f} | C, \widetilde{W}, \widetilde{U}) \cdot dw_{t_f}$ and then threshold this at successively lower levels. This is similar to the standard approach used to choose the FPR for logistic regression.

6.3.3. Estimating an ROC Curve—It is common to choose the FPR using a receiver operating characteristic (ROC) curve. An ROC curve explicitly displays the tradeoff between the TPR and FPR. We can estimate such a curve for the various machine learning models and our Bayesian predictive model both with and without the empirical prior distribution. In particular, we use leave-one-out cross-validation (Hastie et al., 2009) to estimate each ROC curve. The idea of this standard approach is that when making the prediction for each individual, we use a model that was computed using data from everyone excluding the present individual. The final result is a summation over the predictions for each individual. The benefit of this approach is we do not use data from a specific individual when making the prediction for that specific individual.

We estimated an ROC curve for each of the models using leave-one-out cross-validation, and these ROC curves are shown in Figure 4. These ROC curves compare the prediction accuracy for all individuals. The ROC curves have been smoothed using a binormal model (Metz et al., 1998), and the unsmoothed version of the ROC curves can be found in the Supplementary Materials. The results show that our predictive modeling framework is competitive in terms of prediction accuracy with the linear SVM, logistic regression, and decision tree models, which further justifies our choice of the utility-maximizing framework and its ability to capture "irrational" discounting in the decision-making of individuals participating in the intervention. Furthermore, our predictive model with the Bayesian empirical prior makes slightly better predictions than our predictive model without a Bayesian prior, though the difference in their ROC curves is not statistically significant ($P = 0.16$) when compared using a standard hypothesis testing approach developed by Hanley & McNeil (1983). In contrast, the difference in the ROC curves of our predictive model (with and without the prior) and the benchmark approaches of linear SVM, logistic regression, and decision tree models is statistically significant ($P = 0.001$). Our empirical results suggest that the Bayesian prior gives a slight improvement for this data set, but this is not expected to generally hold. Essentially, we expect that using a prior will give improvements in prediction accuracy when an individual is similar to those individuals used to construct the prior. On the other hand, if an individual is very different from those used to construct the prior, then we expect the prior to make predictions worse. However, the situation may be improved with a demographics-dependent prior: We could imagine constructing different priors for individuals with different demographics. Then when making predictions for an individual, we could either use a prior constructed by the data of those with matching demographics, or not use a prior if the individual has very different demographics than was used to construct any of the priors.

6.4. Personalizing Goal Setting Using the Predictive Model

One of our reasons for developing a predictive model is to enable the design of approaches for optimizing elements of large weight loss programs. In contrast to other predictive models, our behavioral framework can be used to formulate an optimization problem to determine the number of visits, timing of visits, and the physical activity goals for each individual in order to maximize the expected number of individuals that achieve clinically significant weight loss at the end of the program. It is in this way that our predictive model has the potential to be used to personalize the weight loss program for each individual.

Here, we present an example that demonstrates the ability of our model to make predictions about how future weight loss changes as the step goals for an individual are changed. Figure 5 shows the posterior likelihood of final weight of an individual conditioned on 50 days of data and on either having 12,000 steps/day goals after 50 days (dash dotted) or having 8,000 steps/day goals after 50 days (solid). When the goals are 8,000 steps/day, our model predicts a 51% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 86.6 kg. When goals are 12,000 steps/day, our model predicts a 3% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 86.8 kg. Our model predicts 8,000 steps/day

goals are superior to 12,000 steps/day goals for motivating this individual to increase their physical activity and consequently lose weight.

The significance of these results is they show how our predictive model forms a foundation for decision optimization problems that can personalize the weight loss program. Specifically, the weight loss program could be personalized by first solving our formulation for computing the MAP estimate of an individual's type. Next, we could solve the problem min w_{t_f} subject to constraints defined by our predictive model and the MAP estimate of the individual's type; this problem can be written as a MILP using reformulation techniques similar to the ones described in this paper. More details about such an approach can be found in (Mintz et al., 2017), and this approach was recently evaluated in a clinical trial where the goal was to increase the physical activity of individuals (Zhou et al., 2018).

6.5. Reducing the Number of Office Visits

Most of the costs and person hours spent on administering weight loss programs are associated with conducting office visits. Thus, it is essential to be able to optimize the total number of visits and when they are scheduled. Our model is able to capture differences in predicted weight loss trajectories that occur when changing the number of office visits. For instance, Figure 6a shows the posterior likelihood of final weight of an individual conditioned on 50 days of data and on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid). When scheduling 4 office visits on days 75, 105, 135, and 150, our model predicts a 96% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 59.3 kg. When scheduling 0 office visits, our model predicts a 94% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 59.3 kg. Our model predicts that for this individual the benefit of scheduling additional office visits is minor.

Another example is shown in Figure 6b, which displays the posterior likelihood of final weight of another individual conditioned on 50 days of data and on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid). When scheduling 4 office visits on days 75, 105, 135, and 150, our model predicts an 18% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 78.6 kg. When scheduling 0 office visits, our model predicts a 3% chance of achieving 5% weight loss and that the expected final weight conditioned on not achieving 5% weight loss is 78.7 kg. Our model predicts a clinically significant benefit of scheduling additional office visits for this particular individual.

These two examples demonstrate the ability of our predictive model to identify which individuals are responsive to office visits, and thus our predictive model can be combined with an optimization model to reduce the average number of office visits when considering a large number of individuals participating in a weight loss program. We briefly describe how an optimization model can be constructed (based on our predictive model) to reduce the average number of office visits; full details of the corresponding optimization models are available in our companion paper (Mintz et al., 2017). Specifically, we can use a

decomposition scheme: In the first step of the scheme, we vary the total number of office visits for each individual over a range of values, and we solve the problem min w_{tf} subject to constraints defined by our predictive model, the MAP estimate of the individual's type, and the total number of office visits. This problem can be written as a MILP using reformulation techniques similar to the ones described in this paper. In the second step of the scheme, we solve a knapsack-like problem that allocates the number of office visits to each individual based on the predicted effectiveness of different numbers of office visits.

7. Conclusion

We constructed a predictive model of individual behavior in a weight loss intervention, employing a utility-maximizing framework based on qualitative concepts from social cognitive theory. MILP formulations were developed to compute (i) parameters of the framework using MLE, and (ii) a Bayesian predictive model using an empirical histogram (constructed using parameters estimated by MLE) as a prior. Model prediction quality was assessed using leave-one-out cross-validation to compute an ROC curve, and the results show that the utility-maximizing framework leads to predictions on par with predictions of a linear SVM model.

We concluded by showing how our predictive model is able to capture differences in weight outcomes as the number of office visits is varied, and we briefly discussed how these models may be used in designing algorithms to optimize and personalize office visit schedules and exercise goals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors gratefully acknowledge the valuable feedback of the editorial and review team. The authors also acknowledge the support of NSF Award CMMI-1450963, UCSF Diabetes Family Fund for Innovative Patient Care-Education and Scientific Discovery Award, K23 Award (NR011454), and the UCSF Clinical and Translational Science Institute (CTSI) as part of the Clinical and Translational Science Award program funded by NIH UL1 TR000004.

References

Acharya S, Elci O, Sereika S, Music E, Styn M, Turk M, & Burke L (2009). Adherence to a behavioral weight loss treatment program enhances weight loss and improvements in biomarkers. Patient Preference and Adherence, 3, 151–160. [PubMed: 19936157]

Ahuja R, & Orlin J (2001). Inverse optimization. Operations Research, 49, 771–783.

Ajzen I, & Fishbein M (1980). Understanding Attitudes and Predicting Social Behavior. Prentice-Hall.

Andersen AR, Nielsen BF, & Reinhardt LB (2017). Optimization of hospital ward resources with patient relocation using markov chain modeling. European Journal of Operational Research, 260, 1152–1163.

Aswani A, Shen Z-J, & Siddiq A (2018). Inverse optimization with noisy data. Operations Research,. URL: http://arxiv.org/abs/1507.03266. To appear.

Ayer T, Alagoz O, & Stout N (2012). A POMDP approach to personalize mammography screening decisions. Operations Research, 60, 1019–1034.

Azar K, Lesser L, Laing B, Stephens J, Aurora M, Burke L, & Palaniappan L (2013). Mobile applications for weight management: theory-based content analysis./American Journal of Preventive Medicine, 45, 583–589. [PubMed: 24139771]

Bandura A (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1–26.

Bender M, Choi J, Arai S, Paul S, Gonzalez P, & Fukuoka Y (2014). Digital technology ownership, usage, and factors predicting downloading health apps among caucasian, filipino, korean, and latino americans: the digital link to health survey. JMIR Mhealth and Uhealth, 2, e43. [PubMed: 25339246]

Bertsimas D, Gupta V, & Paschalidis I (2014). Data-driven estimation in equilibrium using inverse optimization. Mathe-matical Programming, (pp. 1–39).

Bertsimas D, & O'Hair A (2013). Personalized diabetes management: A robust optimization approach. Submitted.

Bickel P, & Doksum K (2006). Mathematical Statistics: Basic Ideas And Selected Topics volume 1 (2nd ed.). Pearson Prentice Hall.

Boyd S, & Vandenberghe L (2004). Convex Optimization. Cambridge University Press.

Breiman L et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16, 199–231.

Brock D, & Wartman S (1990). When competent patients make irrational choices. New England Journal of Medicine, 322, 1595–1599. [PubMed: 2336090]

Cawley J (2004). An economic framework for understanding physical activity and eating behaviors. American Journal of Preventive Medicine, 27, 117–125. [PubMed: 15450622]

Chapman G, & Elstein A (1995). Valuing the future: temporal discounting of health and money. Medical Decision Making, 15, 373–386. [PubMed: 8544681]

Chen C, Haddad D, Selsky J, Hoffman J, Kravitz R, Estrin D, & Sim I (2012). Making sense of mobile health data: An open architecture to improve individual- and population-level health. Journal of Medical Internet Research, 14, e112. [PubMed: 22875563]

Dempe S (2002). Foundations of Bilevel Programming. Springer.

Denoyel V, Alfandari L, & Thiele A (2017). Optimizing healthcare network design under reference pricing and parameter uncertainty. European Journal of Operational Research, 263, 996–1006.

Deo S, Jiang T, Iravani S, Smilowitz K, & Samuelson S (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. Operations Research, 61, 1277–1294.

Diabetes Prevention Program Research Group (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine, 346, 393–403. [PubMed: 11832527]

Diabetes Prevention Program Research Group (2003). Costs associated with the primary prevention of type 2 diabetes mellitus in the diabetes prevention program. Diabetes Care, 26, 36–47. [PubMed: 12502656]

Diabetes Prevention Program Research Group (2009). 10-year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. Lancet, 374, 1677–1686. [PubMed: 19878986]

Doucet E, St-Pierre S, Almeras N, Despres J-P, Bouchard C, & Tremblay A (2001). Evidence for the existence of adaptive thermogenesis during weight loss. British Journal of Nutrition, 85, 715–723. [PubMed: 11430776]

Engineer F, Keskinocak P, & Pickering L (2009). Or practice – catch-up scheduling for childhood vaccination. Operations Research, 57, 1307–1319.

Evans M, & Swartz T (1995). Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. Statistical science, (pp. 254–272).

Fetta A, Harper P, Knight V, & Williams J (2018). Predicting adolescent social networks to stop smoking in secondary schools. European Journal of Operational Research, 265, 263–276.

Flegal K, Carroll M, Kit B, & Ogden C (2012). Prevalence of obesity and trends in the distribution of body mass index among U.S. adults, 1999–2010. Journal of the American Medical Association, 307, 491–497. [PubMed: 22253363]

Flores Mateo G, Granado-Font E, Ferre-Grau C, & na Carreras XM (2015). Mobile phone apps to promote weight loss and increase physical activity: A systematic review and meta-analysis. J Med Internet Res, 17, e253. [PubMed: 26554314]

Fukuoka Y, Gay C, Joiner K, & Vittinghoff E (2015). A novel diabetes prevention intervention using a mobile app: A randomized controlled trial with overweight adults at risk. American Journal of Preventive Medicine, 49, 223–237. [PubMed: 26033349]

Fukuoka Y, Komatsu J, Suarez L, Vittinghoff E, Haskell W, Noorishad T, & Pham K (2011). The mPED randomized controlled clinical trial: applying mobile persuasive technologies to increase physical activity in sedentary women protocol. BMC Public Health, 11.

Gafni A (1990). Correspondence to “competent patients and irrational choices”. New England Journal of Medicine, 323, 1353–1355.

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, & Rubin D (2013). Bayesian Data Analysis, Third Edition Chapman & Hall/CRC Texts in Statistical Science Taylor & Francis.

Glover F (1975). Improved linear integer programming formulations of nonlinear integer problems. Management Science, 22, 455–460.

Gomes N, Merugu D, O’Brien G, Mandayam C, Yue JS, Atikoglu B, Albert A, Fukumoto N, Liu H, Prabhakar B, & Wischik D (2012). Steptacular: An incentive mechanism for promoting wellness. In Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on (pp. 1–06).

Gonzalez V, Goeppinger J, & Lorig K (1990). Four psychosocial theories and their application to patient education and clinical practice. Arthritis Care and Research, 3, 132–143. [PubMed: 2285752]

Grant M, & Boyd S (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

Gurobi Optimization, I. (2015). Gurobi optimizer reference manual. URL: http://www.gurobi.com.

Hanley JA, & McNeil BJ (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148, 839–843. [PubMed: 6878708]

Hastie T, Tibshirani R, & Friedman J (2009). The Elements of Statistical Learning. (2nd ed.). Springer-Verlag.

Helm JE, Lavieri MS, Oyen MPV, Stein JD, & Musch DC (2015). Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. Operations Research,.

Heuberger C (2004). Inverse combinatorial optimization: A survey on problems, methods, and results. Journal of Combinatorial Optimization, 8, 329–361.

Hill J, Wyatt H, Reed G, & Peters J (2003). Obesity and the environment: Where do we go from here? Science, 299, 853–855. [PubMed: 12574618]

Janz N, & Becker M (1984). The health belief model: A decade later. Health Education & Behavior, 11, 1–47.

Joos S, & Hickam D (1990). How health professionals influence health behavior: patient provider interaction and health care outcomes In Health behavior and health education: theory, research and practice (pp. 216–241). Jossey-Bass.

Kanfer F (1975). Self-management methods In Helping People Change (pp. 309–316). Pergamon.

Keshavarz A, Wang Y, & Boyd S (2011). Imputing a convex objective function. In IEEE Multi-Conference on Systems and Control (pp. 613–619).

Lehmann E, & Romano J (2006). Testing Statistical Hypotheses. Springer Texts in Statistics. Springer.

Lindley DV (1961). The use of prior probability distributions in statistical inference and decisions. In Proc. 4th Berkeley Symp. on Math. Stat. and Prob (pp. 453–468).

Lopez M, Gonzalez-Barrera A, & Patten E (2013). Closing the digital divide: Latinos and technology adoption. Technical Report Pew Research Center.

Mason J, Denton B, Smith S, & Shah N (2013). Using electronic health records to monitor and improve adherence to medication. Working paper.

McDonald H, Garg A, & Haynes R (2002). Interventions to enhance patient adherence to medication prescriptions: scientific review. Journal of the American Medical Association, 288, 2868–2879. [PubMed: 12472329]

Metz C, Herman B, & Shen J-H (1998). Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. Statistics in Medicine, 17, 1033–1053. doi:10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z. [PubMed: 9612889]

Mifflin M, St Jeor S, Hill L, Scott B, Daugherty S, & Koh Y (1990). A new predictive equation for resting energy expenditure in healthy individuals. The American Journal of Clinical Nutrition, 51, 241–247. [PubMed: 2305711]

Mintz Y, Aswani A, Kaminsky P, Flowers E, & Fukuoka Y (2017). Behavioral analytics for myopic agents. *arXiv:1702.05496*,. URL: https://arxiv.org/abs/1702.05496.

Murphy SA, & Vaart AWVD (2000). On profile likelihood. Journal of the American Statistical Association, 95, 449–465.

Negoescu D, Bimpikis K, Brandeau M, & Iancu D (2014). Dynamic learning of patient response types: An application to treating chronic diseases.

O'Reilly G, & Spruijt-Metz D (2013). Current mHealth technologies for physical activity assessment and promotion. American Journal of Preventive Medicine, 45, 501–507. [PubMed: 24050427]

Oztekin A, Al-Ebbini L, Sevkli Z, & Delen D (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. European Journal of Operational Research, 266, 639–651.

Pagoto S, Schneider K, Jojic M, DeBiasse M, & Mann D (2013). Evidence-based strategies in weight-loss mobile apps. American Journal of Preventive Medicine, 45, 576–582. [PubMed: 24139770]

Rosenbaum M, Hirsch J, Gallagher DA, & Leibel RL (2008). Long-term persistence of adaptive thermogenesis in subjects who have maintained a reduced body weight—. The American journal of clinical nutrition, 88, 906–912. [PubMed: 18842775]

Schoeller D, Bandini L, & Dietz W (1990). Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. Canadian Journal of Physiology and Pharmacology, 68, 941–949. [PubMed: 2200586]

Severini T (1999). On the relationship between bayesian and non-bayesian elimination of nuisance parameters. Statistica Sinica, 9, 713–724.

Severini T, & Wong W (1992). Profile likelihood and conditionally parametric models. Annals of Statistics, 20, 1768–1802. URL: 10.1214/aos/1176348889. doi:10.1214/aos/1176348889.

Tibshirani R, Saunders M, Rosset S, Zhu J, & Knight K (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 91–108.

Tierney L, & Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. Journal of the american statistical association, 81, 82–86.

Torres F (1991). Linearization of mixed-integer products. Mathematical Programming, 49, 427–428.

Valero-Elizondo J, Salami J, Osondu C, Latif M, A A, Spatz E, Rana J, Virani S, Blankstein R, Blaha M, Veledar E, & Nasir K (2016). Abstract 146: Drivers of healthcare costs among adults with obesity in united states: 2012 medicales, expenditure panel survey. Circulation: Cardiovascular Quality and Out comes 9, A146.

Vielma J (2015). Mixed integer linear programming formulation techniques. SIAM Review, 57, 3–57.

Wang H, Zheng B, Yoon SW, & Ko HS (2017). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research,.

Wu S, Ell K, Gross-Schulman S, Sklaroff L, Katon W, Nezu A, Lee P-J, Vidyanti I, Chou C-P, & Guterman J (2013). Technology-facilitated depression care management among predominantly Latino diabetes patients within a public safety net care system: Comparative effectiveness trial design. Contemporary Clinical Trials,.

Zhou M, Fukuoka Y, Mintz Y, Goldberg K, Kaminsky P, Flowers E, & Aswani A (2018). Evaluating machine learning— based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. JMIR Mhealth Uhealth, 6, e28 URL: http://mhealth.jmir.org/2018/1/e28/. doi:10.2196/mhealth.9117. [PubMed: 29371177]

Highlights

- We develop data-driven models to predict behavior in weight loss programs
- These models can be used to optimize a weight loss program for each individual
- Estimation and Bayesian prediction with our models is computed using optimization
- We validate the model by comparison to common machine learning approaches

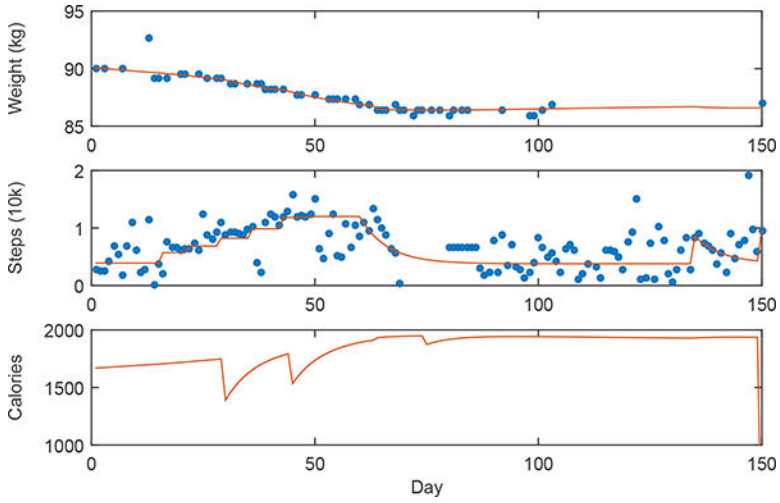

Figure 1:
Comparison of data (blue dots) with MLE estimates of weight, exercise, and caloric intake (red line).

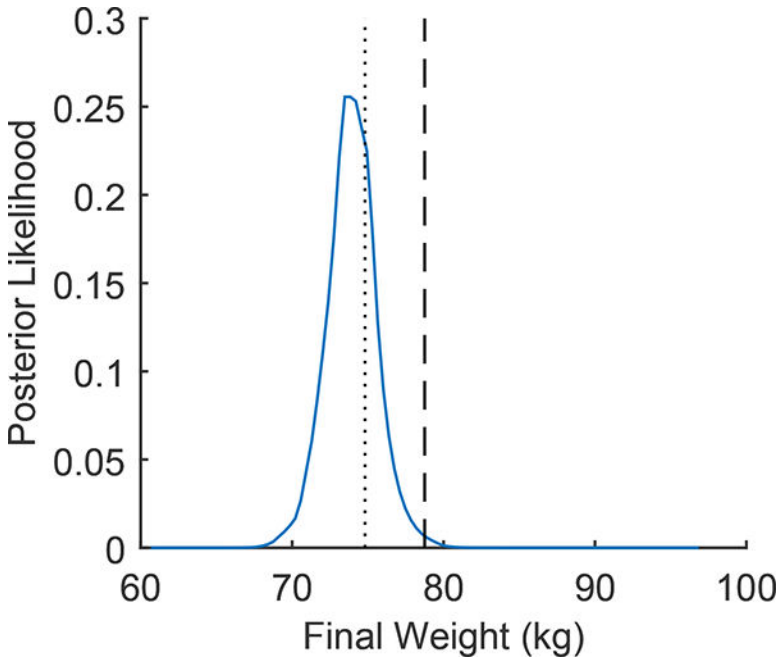

Figure 2:
Posterior likelihood of final weight conditioned on 30 days of data (solid) compare to intial weight (dashed) and final weight corresponding to a 5% weight loss (dotted).

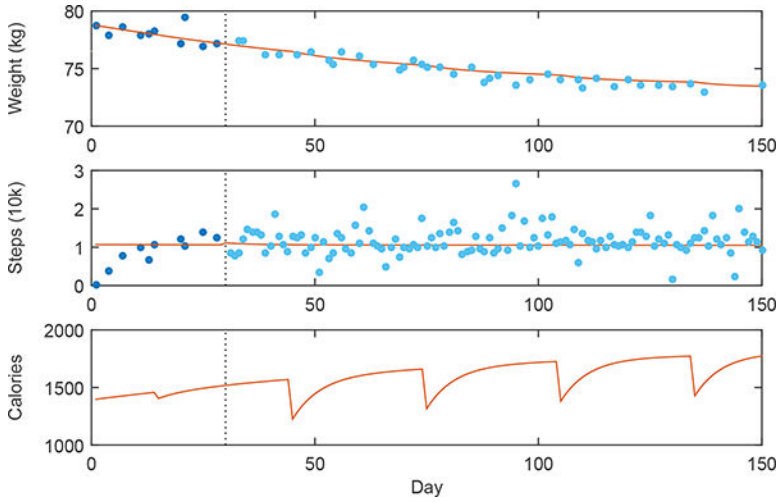

Figure 3:
Comparison of MAP estimates of weight, exercise, and caloric intake trajectories dark blue dots) with future data not used to computed estimates (light blue dots).

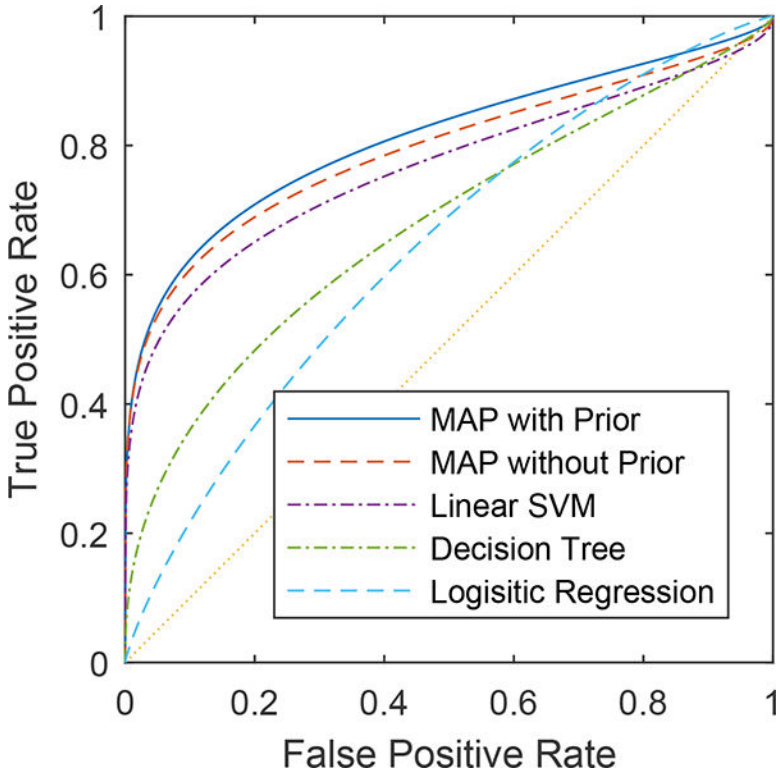

Figure 4:

ROC curves computed using leave-one-out cross-validation for our predictive model with an empirical Bayesian prior (blue solid), our predictive model without a Bayesian prior (red dashed), linear SVM model (purple dash dot), decision tree model (green dashed dot), and logistic regression (cyan dashed) are compared.

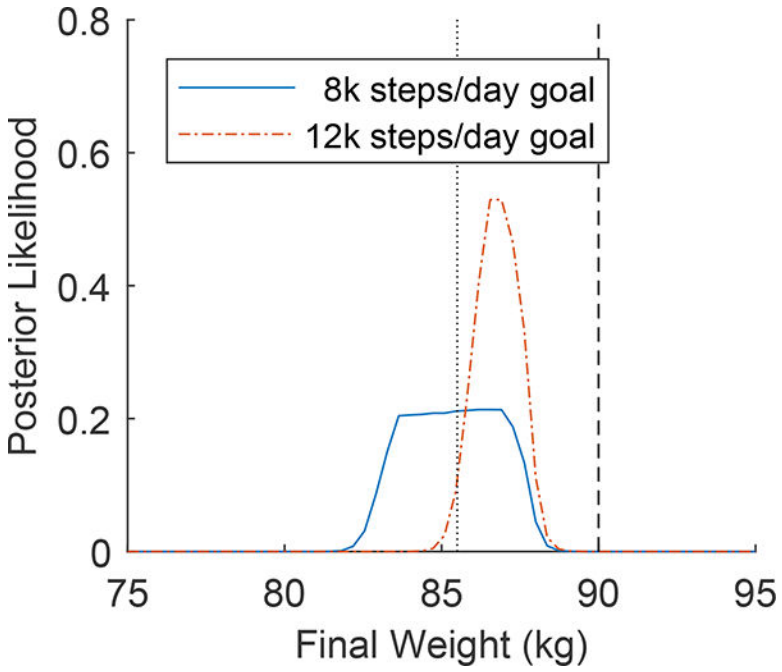

Figure 5:
Posterior likelihood of final weight of an individual conditioned on 50 days of data and conditioned on either having 12,000 steps/day goals after 50 days (dash dotted) or 8,000 steps/day goals after 50 days (solid), and compared to initial weight (dashed) and final weight corresponding to a 5% weight loss (dotted).

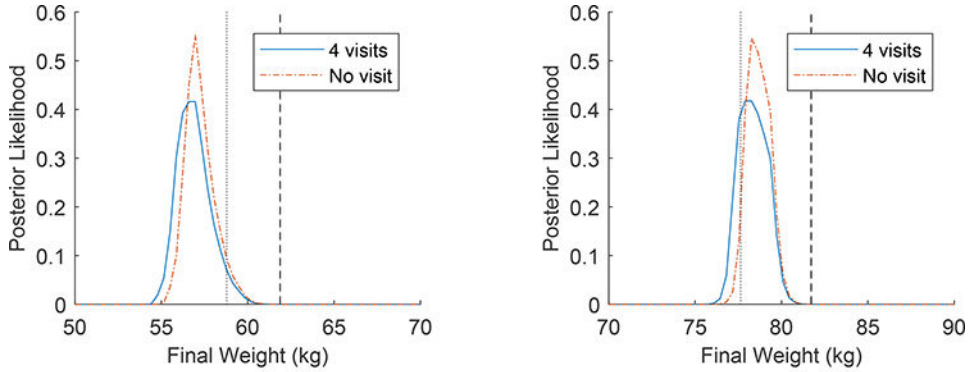

Figure 6:

Posterior likelihood of final weight of an individual conditioned on 50 days of data and conditioned on either having no office visits after 50 days (dash dotted) or having 4 office visits after 50 days (solid and final weight corresponding to a 5% weight loss (dotted).