Original Paper

Personalized Glucose Management With AI: Pilot Study Using a Multiarmed Bandit Approach

Shinji Hotta[1,2], MInf; Mikko Kytö[3,4], PhD; Saila Koivusalo[5], PhD; Seppo Heinonen[5], PhD; Pekka Marttinen[2], PhD

[1]Fujitsu Limited, Kawasaki, Kanagawa, Japan

[2]Department of Computer Science, Aalto University, Espoo, Uusimaa, Finland

[3]Development and Strategy Unit, Helsinki University Hospital, University of Helsinki, Helsinki, Uusimaa, Finland

[4]Department of Computer Science, University of Helsinki, Helsinki, Uusimaa, Finland

[5]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Uusimaa, Finland

Corresponding Author:

Shinji Hotta, MInf
Fujitsu Limited
4-1-1 Kamikodanaka, Nakahara-ku
Kawasaki, Kanagawa 211-8588
Japan
Phone: 81 44 777 1111
Email: hotta_s@fujitsu.com

Abstract

Background: Personalized behavioral recommendations through mobile apps have proven effective in preventing serious chronic diseases such as diabetes. Recent studies have primarily focused on optimizing personalized recommendations using reinforcement learning. However, the main problem with these approaches is that they focus on behavioral changes and overlook clinical outcomes.

Objective: This study aimed to propose a method for online planning of dietary and exercise recommendations to optimize postprandial glucose levels through behavioral changes directly.

Methods: The proposed method is a multiarmed bandit based on a two-stage reward prediction model, where an action is a combination of the total carbohydrate intake and postprandial walking duration, and the reward is the reduction in postprandial glucose levels. We implemented the prediction of the reward for each action based on the predicted behavioral responses to an action, and subsequently, the postprandial glycemic response.

Results: In a simulation experiment, we demonstrated that the proposed online algorithm can significantly improve postprandial glucose levels with personalized recommendations, compared to the randomized policy. Furthermore, we conducted a small real-world experiment with a simplified proposed method involving a single update of the recommendation policy into a personalized one. For 6 participants, compared to the randomized policy, we observed a 23% improvement, on average, in actual glucose responses along with the behavioral adherence to the recommendations concerning carbohydrate intake and postprandial walking.

Conclusions: The preliminary effectiveness of the proposed method was demonstrated from both the simulation experiment and the small real-world experiment. However, further longitudinal real-world experiments in patients with diabetes are needed to validate and generalize the findings.

JMIR Form Res 2026;10:e70826; doi: 10.2196/70826

Keywords: diabetes; mobile intervention; personalization; dietary and exercise recommendation; glucose management; multiarmed bandit

Introduction

Background

The global incidence of diabetes has increased considerably in recent years, accompanied by escalating disease severity. This progression has detrimental ramifications, including compromised quality of life, multifarious complications, and expensive surgical treatment. The medical expenses for diabetes are expected to increase to approximately US $2.5 trillion worldwide by 2030 [1]. This has facilitated the urgent need to reduce the severity of diabetes. The recent national clinical practice guidelines [2] state that the basic principle of this preventive treatment is to maintain glucose levels within the normal range. The central strategy to manage glucose levels is to maintain an appropriate lifestyle of eating and exercising [3], in addition to insulin or oral medication. In recent years, technological advances in continuous glucose monitoring devices have helped patients manage their glucose levels with mobile apps at any given time and place, thus promoting self-management [4,5]. However, it is difficult for patients to learn the optimal behavioral action to maintain glucose within a normal range [6]. Therefore, a personalized system that recommends individualized optimal behavioral actions is required to ensure that glucose levels do not deviate from the normal range in the future.

Over the past decade, it has become possible for patients with prediabetes to self-monitor their glucose levels, along with their lifestyle habits, such as diet and exercise, by visualizing them through an integrated mobile app [4,7]. While this has made it easier for patients to manage their own glucose levels daily, and maintaining high user adherence to such monitoring is achievable, as demonstrated by recent studies reporting over 90% daily adherence [8], learning the best behaviors to avoid abnormal glucose levels remains a significant hurdle [4]. For instance, the total grams of carbohydrates to be consumed or the total minutes of walking exercise to be performed vary from person to person. Therefore, the behavioral recommendations should be automatically generated for each patient by analyzing personal datasets collected daily.

Literature Review

Thus far, there has been significant research on personalized behavioral recommendations through mobile apps for the prevention of chronic diseases, including diabetes [9], and a framework called just-in-time adaptive intervention has been established. In just-in-time adaptive intervention, to improve both the short-term and long-term outcomes, decision rules regarding intervention conditions, such as content, frequency, and timing of recommendations, are personally optimized [10]. For instance, in [11], to promote physical activity in patients with hypertension, many interventions with different conditions were allocated and performed on each patient's timeline, and the best-fitting intervention conditions were identified based on the results of the interventional experiments. While this approach can robustly identify personally appropriate intervention conditions, its practical application for practitioners is constrained by the significant burden associated with prolonged experimental periods and the necessity for patients to undergo multiple interventions that may prove to be suboptimal. This underscores the need to develop more efficient and effective personalization methods.

With regard to the aforementioned factors, recent studies have focused on sequentially optimizing personal intervention conditions using reinforcement learning [12]. Yom-Tov et al [13] proposed a contextual bandit that optimized the message content of exercise interventions for patients with diabetes. In this study, an action is the message content, and a reward is the amount of activity after the recommendation. The individual optimization of message content can be adaptively performed by selecting an action based on a reward using Boltzmann sampling. Liao et al [14] proposed a framework that applies a multistep Markov decision process by introducing a user state, which represents how much intervention to increase daily walking has been received so far. This helps adjust both the intervention content and the intervention frequency to avoid user dropout caused by frequent interventions. However, while the focus of these existing methods is on patient behavior change, it is still inconclusive whether these methods can offer optimal interventions for improving clinical outcomes. For instance, in the case of diabetes, the extent to which the intervention lowers glucose levels is highly significant. In other words, further research is needed to identify the intervention conditions that can lower glucose levels most effectively to the target range.

Some studies have proposed optimizing clinical insulin interventions for people with more severe diabetes using reinforcement learning with an artificial pancreas to directly control their glucose levels [15,16]. When insulin is injected, the glucose level drops immediately. Therefore, the primary goal of such research is to determine the appropriate insulin dose and timing for each patient to ensure that the glucose level remains within the normal range. The action is the insulin dosage, the state is the user's glucose level, and the reward is the normality of the glucose levels after insulin injection. The integration of reinforcement learning helped obtain the optimal insulin dosage for each patient [17]. However, in this type of study, the variations in action were limited to clinical and direct interventions, such as insulin injection, which are not directly controlled by the patient. Incorporating behavioral interventions, such as diet and exercise recommendations, in addition to clinical interventions, is currently regarded as future work [15,16].

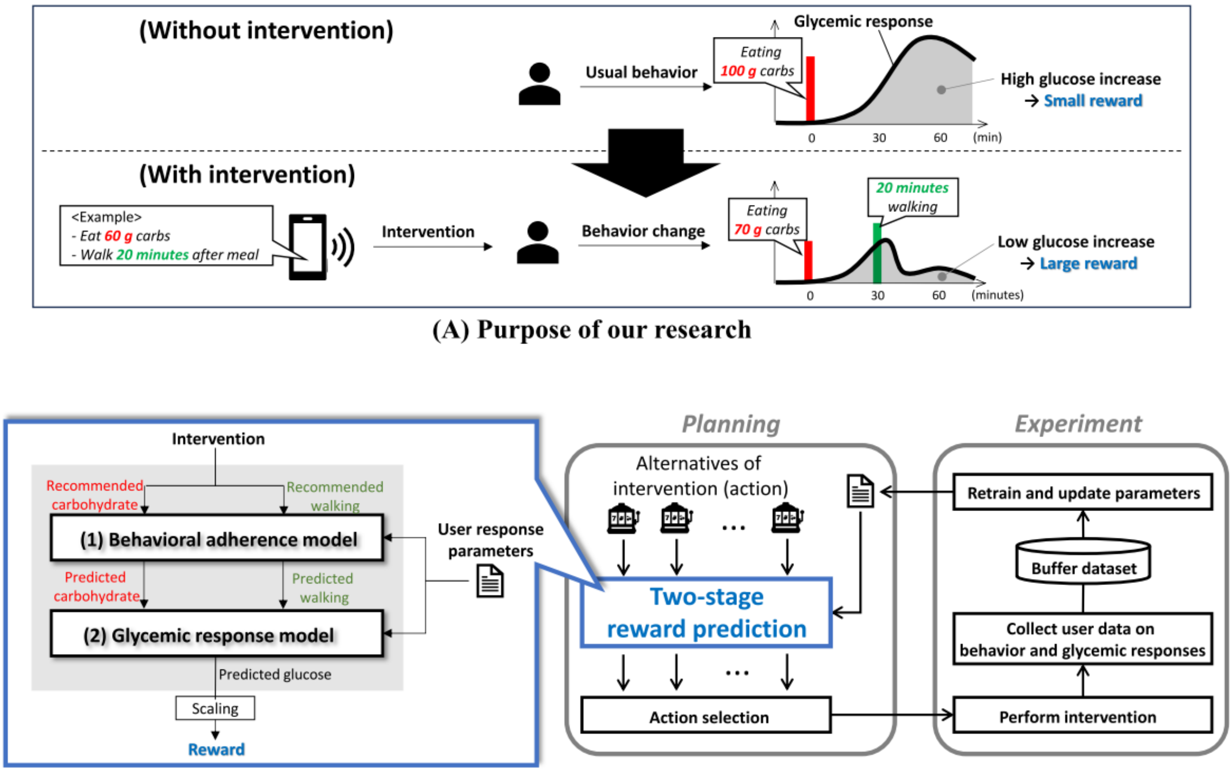
Objectives

Altogether, there is a lack of approaches that individually optimize behavioral recommendations such as diet and exercise to control glucose levels. In this paper, we propose a multiarmed bandit algorithm that can individually plan mobile behavioral interventions, making it easy for users to perform and reducing glucose levels after eating (ie, postprandial glucose, which is most likely to become abnormal in the daily lives of people with diabetes). The most novel aspect of the proposed method is the introduction of a two-stage reward prediction mechanism for each intervention: (1) prediction of the actual user behaviors that occur after the intervention, and (2) prediction of the postprandial glucose trajectory from these predicted behaviors.

An overview of the proposed method is provided in Figure 1. The aim is to lower postprandial glucose levels with the intervention (Figure 1A). Here, the action (or arm) is a mobile behavioral intervention that combines recommended amounts of carbohydrate intake and postprandial walking exercise, which have been proven to have specific effects on glucose levels [18,19]. Multiple alternative candidates for these recommended amounts are prepared in advance, and the optimal action for each user is selected based on the reward predicted in a two-stage manner, as provided in Figure 1B. In this reward prediction, the actual amounts of both carbohydrate intake and postprandial walking are predicted for each action candidate. For example, for carbohydrate intake, when a user receives a recommendation to take 60 g of carbohydrates, it is predicted that the participant will take 70 g because the recommended amount is small according to their dietary preference. Second, based on the predicted dietary and exercise behaviors, the temporal trajectory of postprandial glucose levels is predicted. Third, according to the predicted glucose trajectory, batched actions are selected using the ε-greedy method, based on the assumption that the greater the suppression of the increase in postprandial glucose, the higher the reward obtained by a user. Finally, after performing the selected actions using a real intervention, the behavioral adherence and glycemic response models are retrained using the buffer dataset of actual behaviors and glucose, and the parameters are updated accordingly. By repeating this experiential learning cycle, which is composed of action planning and real experiments, it is possible to fill the gap between each user’s real response and the virtually predicted response, which yields a more optimal action selection for lowering postprandial glucose levels. Furthermore, to accelerate this cycle, we suggest incorporating prior knowledge into the prediction models for retraining.

We demonstrate the effectiveness of the proposed method through simulation experiments and clarify the conditions under which the proposed method works best. Then, with a simplified algorithm based on these conditions, we conducted a small-scale mobile intervention experiment with healthy participants. The results provide preliminary evidence that the proposed method improves both behavior and glucose levels. To the best of our knowledge, this is the first attempt to (1) clarify how to optimize behavioral interventions for improving glucose levels and (2) demonstrate the effectiveness of the method through simulation experiments and real-world intervention experiments.

Figure 1. Purpose of our research (A) and overview of the proposed method (B).



(A) Purpose of our research

(B) Overview of proposed method

Methods

Overview

In the following subsections, we describe the problem setting, the details of the proposed method, the simulation experiment setup, and the real-world feasibility study setup. The proposed method is formulated as a model-based multiarmed bandit framework with a two-stage reward prediction model that integrates behavioral adherence modeling and glycemic response modeling to enable data-efficient personalization of dietary and exercise recommendations. Its effectiveness is evaluated through both simulation experiments and a real-world feasibility study.

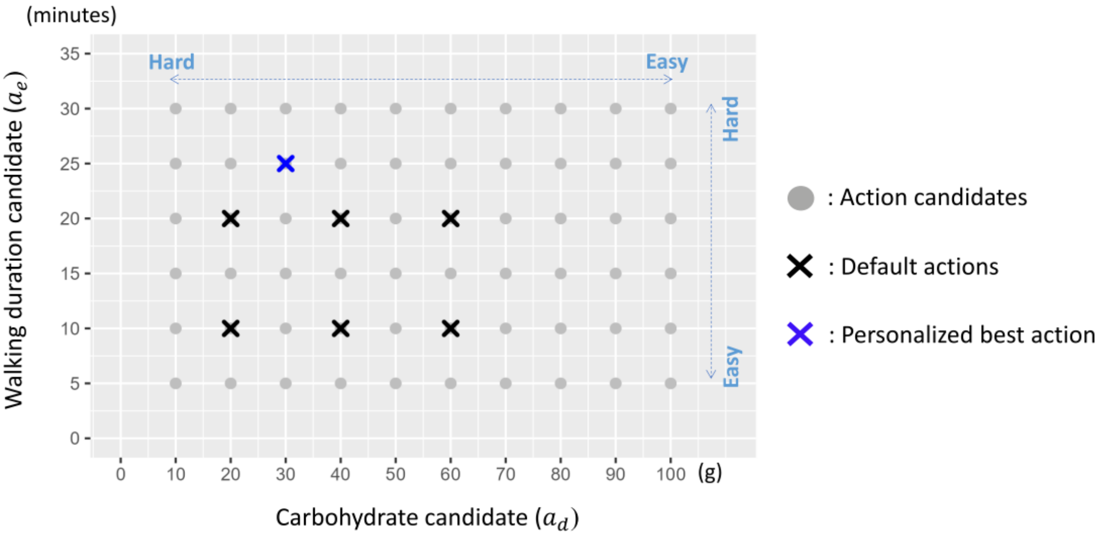
Problem Setting

The purpose of our method is to select and recommend the best action candidate for each user that is easy to perform and reduces postprandial glucose levels the most through interaction with the user. For this purpose, a multiarmed bandit algorithm is often used, as in previous literature [13,20,21]. The multiarmed bandit is one of the reinforcement learning methods. It continuously addresses the dilemma of selecting actions to gather more information (exploration) vs selecting actions to maximize immediate reward based on current knowledge (exploitation). Thus, a bandit algorithm can adjust this exploration-exploitation trade-off [22] by estimating the rewards associated with different actions.

In our method, the reward r can be defined as the function of the total increase in postprandial glucose levels after the target diet, which occurs after recommending an action A to a certain user. We mainly focus on postprandial glucose, given that glucose levels always rise after a meal and that deviations from the normal range of glucose levels occur most frequently after a meal [23,24].

An action in our method $A(\in \mathcal{A})$ can be represented as a combination of $A = \{a_d, a_e\}$, where a_d means a recommendation for "how much carbohydrates in grams a user should take at maximum in target diet," and a_e means a recommendation for "the minimum duration a user should walk just after target diet." For clarity, users do not directly estimate carbohydrate content; instead, it is automatically calculated by a mobile app such as MyFitnessPal (MyFitnessPal, Inc) based on inputted food items. The accuracy of MyFitnessPal's carbohydrate calculations has been rigorously validated, showing strong correlations (r=0.90) with established food composition databases [25]. Based on the assumption that each user has their own appropriate amounts for both actions, multiple action candidates are prepared in advance within a clinically plausible range. Regarding the amount of carbohydrate intake (a_d), we prepared 10 candidates in the range of 10-100 g, referring to the low-carbohydrate diet recommendations [26]. In addition, regarding the amount of postprandial walking exercise (a_e), we prepared 6 candidates between 5 and 30 minutes based on the experimental results of the effect of walking exercise on glucose levels [19,27-29]. Therefore, there are 60 action candidates in the action set $\mathcal{A}$ based on previous research [19,26-29], as shown in Figure 2.

Figure 2. Illustration of action candidates. For default actions (×), the amount of carbohydrate intake is set within the range of 20-60 g, and postprandial walking duration is set within the range of 10-20 minutes. Although it is easy to adhere to taking a large amount of carbohydrate and walking for a few minutes, this leads to higher postprandial glucose levels. The main purpose is to find the best personalized action (×) that can achieve lower postprandial glucose levels while maintaining adherence for each user.



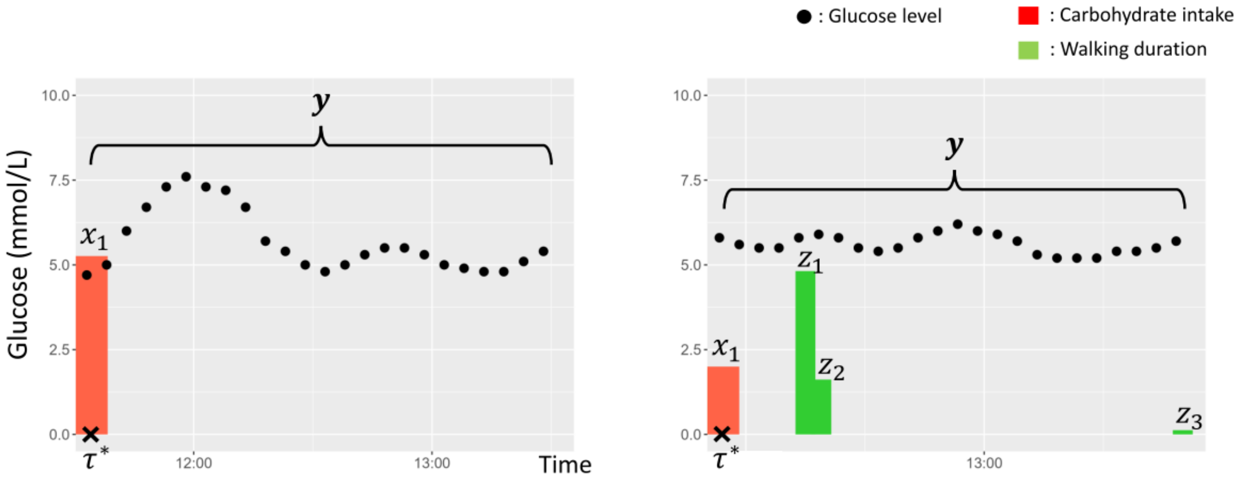
Here, there are 2 approaches for action selection in a multiarmed bandit, that is, the experience- and the model-based approaches. The former (eg, Upper Confidence Bound) evaluates each action based on the actual observed rewards, which enables robust evaluation but requires extensive exploration for all actions. In contrast, the latter (eg, Thompson Sampling) can reduce exploratory trials because it can estimate the reward (or value) for each action by learning the parameters of the reward distribution. Returning to our setup (Figure 1A), each observation of the reward (glucose level change) requires users to wear a glucose monitoring device, imposing a recurring physical burden. Therefore, it is important to reduce the burden of exploration trials on users for real applications, and we adopted this model-based approach to predict the reward for untried actions efficiently from limited observations.

Now, the reward prediction model is represented as follows:

$$r = R(A \mid \boldsymbol{\Theta})$$

The input is an action A, and $\boldsymbol{\Theta}$ is a model parameter set of a target user. Because $\boldsymbol{\Theta}$ is unknown at the initial trial, it is sequentially updated when observing user responses to previously recommended actions. Here, the user response is composed of 3 variables, $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ as shown in Figure 3; $\boldsymbol{x}$ is the actual amount vector of each carbohydrate intake at the target diet, $\boldsymbol{z}$ is the actual walking duration vector of each walking event occurring after starting the target diet, and $\boldsymbol{y}$ is the time series of postprandial glucose levels after the start time of the target diet as τ^*. Because the glucose level within 1 hour after a diet is of clinical importance, we focused on user responses within 90 minutes of starting the diet. As a result, for each action, the observed data are represented as $\boldsymbol{\mathcal{D}} = \{A, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ and are added to the buffered dataset $\mathfrak{B}$ after observation.

Figure 3. Two examples of observed data with a small reward (A) and a large reward (B). The greater the postprandial increase in glucose levels, the smaller the reward.



(A) Example with small reward

(B) Example with large reward

Here, glucose levels and behaviors in 1 observation "$\boldsymbol{\mathcal{D}}$" would involve high uncertainty because they may be affected by many factors. Thus, a parameter set "$\boldsymbol{\Theta}$" is updated in a Bayesian fashion when obtaining batched observation data containing "M" actions and "M" user responses. That is, as shown on the right side of Figure 1B, parameter learning is executed after recommending "M" actions to the users. This can be considered as a batched multiarmed bandits setting [30]. Hereafter, the method of selecting "M" actions is referred to as the policy.

Algorithm of the Proposed Method

In this subsection, we describe the proposed multiarmed bandit algorithm based on two-stage reward prediction. The pseudocode for the proposed algorithm is shown in Textbox 1. In this algorithm, the following processes are executed in the given order: (1) actions are recommended to users according to the predetermined policy, and batch observation data are obtained; (2) the parameters of the behavioral adherence and glycemic response models are retrained using these observation data; (3) the reward for each action candidate is predicted using a two-stage reward model with retrained parameters; and (4) finally, the next batch actions are selected based on the reward prediction result using the ε-greedy method.

These processes are repeated to improve the rewards and model parameters. This predictive model–based approach allows the algorithm to estimate potential outcomes of all possible actions based on learned models, rather than requiring direct, burdensome observation for every action candidate. In the first iteration, default actions are recommended for pure parameter exploration, aiming to gather diverse data for initial model parameter estimation. Afterward, as the number of iterations increases, more robust parameter learning can be performed with more observational data, resulting in accurate reward prediction and action selection. However, if the convergence of this parameter learning requires long-term (eg, several months of) observation data owing to the uncertainty contained in the observation, the user would drop out due to the burden of such time-consuming experiments for collecting data. To prevent this, we propose introducing prior knowledge for each parameter and performing Bayesian learning.

Textbox 1. Algorithm 1: multiarmed bandit with two-stage reward prediction.

```
INPUT: maximum iteration number $I$, batch size $M$, default policy $\pi_1$, and action set $\mathcal{A}$.
// Repeat experimental learning cycle of intervention experiment and action planning for a user
For $i$ in 1:$I$
// Perform intervention
    For $j$ in 1:$M$
        Recommend the action $A_j = \pi_i(j)$ for the user
        Observe the user responses $\boldsymbol{y}_{i,j}, \boldsymbol{x}_{i,j}, \boldsymbol{z}_{i,j}$ ........................................................ (†)
    End
    Update buffered dataset $\mathfrak{B}_i$ by adding new batch data$\mathcal{D}_i = \bigcup_{j=1}^{M}\{A_{i,j}, \boldsymbol{y}_{i,j}, \boldsymbol{x}_{i,j}, \boldsymbol{z}_{i,j}\}$
    // Train each user model from subset of buffered dataset
    Train the parameters $\theta_d(i)$ in dietary adherence model from the subset data$\mathfrak{B}_i^{(d)}$
    Train the parameters $\theta_e(i)$ in exercise adherence model from the subset data
    Train the parameters $\theta_g(i)$ in glycemic response model from the subset data
    // Two-stage reward prediction for each action
    For $A_k \in \mathcal{A}$
        Predict dietary behavior $x_k^*$ with $\theta_d(i)$ under action $A_k$
        Predict exercise behavior $z_k^*$ with $\theta_e(i)$ under action $A_k$
        Predict glucose increase trajectory $\boldsymbol{y}_k^*$ with $\theta_g(i)$ under predicted behaviors $x_k^*, z_k^*$,
        Obtain reward $r_k$ from glucose increase trajectory $\boldsymbol{y}_k^*$
    End
    // Action selection for next intervention
    If Rand() < ε
        Select actions randomly as $\pi_{i+1} = \pi_{random} = \bigcup_{m=1}^{M} Random(\mathcal{A})$ // for exploration
    Else
        Select actions greedy as $\pi_{i+1} = \pi^* = \bigcup_{m=1}^{M}\left\{\underset{k}{\operatorname{argmax}}\, r_k\right\}$ // for exploitation
    End
End
```

Two-Stage Reward Prediction Model

Here, we first focus on describing a two-stage reward prediction model that integrates behavioral adherence models into a glycemic response model. The starting point for modeling is based on previous clinical findings, which state that the changes in postprandial glucose are highly dependent on the amount of carbohydrates an individual takes [31] and the amount of exercise performed after a meal [19,27-29]. Furthermore, these dietary and exercise behaviors are assumed to depend on the kind of actions recommended immediately beforehand. Based on these premises, the reward prediction model is described as follows:

$$R(A \mid \boldsymbol{\Theta}) = f(\boldsymbol{y}^*(A \mid \boldsymbol{\Theta})) \quad (1)$$

$$\boldsymbol{y}^*(A \mid \boldsymbol{\Theta}) = \mathbf{g}(x^*, z^* \mid \theta_g) \quad (2)$$

$$x^* = h_d(a_d \mid \theta_d), \quad z^* = h_e(a_e \mid \theta_e) \quad (3)$$

where $R(A \mid \boldsymbol{\Theta})$ is the reward, $\boldsymbol{y}^*(A \mid \boldsymbol{\Theta})$ is the glucose time series, x^* is the amount of carbohydrate intake, and z^* is the walking duration. Further, $\boldsymbol{\Theta} = \{\theta_g, \theta_d, \theta_e\}$ and $\theta_g, \theta_d, \theta_e$ represent parameters in a glycemic response model, a carbohydrate intake model, and a postprandial exercise model, respectively. These parameters are defined for each user. In actual reward prediction, the first stage is to predict the dietary and exercise behaviors using Equation 3 with behavior adherence models $h_d(\cdot)$ and $h_e(\cdot)$, and the second stage predicts the trajectory of the postprandial glucose increase $\boldsymbol{y}^*$ using Equation 2 with the predicted behaviors (x^*, z^*). Finally, the glucose increase trajectory is converted into a reward scale using Equation 1.

Behavior Adherence Model

The first stage is behavior prediction. Because each user receives a specific action immediately beforehand, we can assume that they act according to this action. However, certain actions are difficult to execute in real-world scenarios. For example, most users would find it easy to keep their carbohydrate intake less than 100 g, but it is difficult to consume less than 10 g. In addition, the actual reaction to reducing carbohydrate intake to 50 g may vary for each user. This is consistent with the claim in the theory of planned behavior [32], which states that behavioral intentions may be influenced by "perceived behavioral control." Inspired by this, in the behavior adherence model, we hypothesize that there is a perceptual critical point specific to each user for each action candidate, as indicated in Figure 2, and that the actual behavioral tendency changes depending on this critical point. Specifically, as shown in Figure S1 (left) in Multimedia Appendix 1, we assume that if the recommended amount of carbohydrate a_d is higher than the critical point X_{th}, the user is likely to try to behave in accordance with the action. In contrast, if a_d is lower than X_{th}, the user is likely to ignore the action and behave normally.

Based on the aforementioned assumptions, the behavior adherence model of carbohydrate intake is represented by the following linear model:

$$h_d(a_d \mid \theta_d) = \begin{cases} a_d + e_{x1} & (if \;\; a_d \geq X_{th}) \\ a_d + X_{norm}\left(\frac{X_{th} - a_d}{X_{th}}\right) + e_{x2} & (if \;\; a_d < X_{th}) \end{cases} \quad (4)$$

where $\theta_d = \{X_{th}, X_{norm}\}$ and X_{th} is the perceptual critical point for a target user, X_{norm} is a parameter that represents the normal carbohydrate intake at the target diet. We also introduce Gaussian noises $e_{x1} \sim N(0, \sigma_{x1})$ and $e_{x2} \sim N(0, \sigma_{x2})$ to represent the uncertainty of carbohydrate intake from other factors. Here, different variance parameters $(\sigma_{x1}, \sigma_{x2})$ are set because the level of behavioral uncertainty may depend on whether the critical point is exceeded.

Subsequently, the adherence model for postprandial exercise is based on the same concept, as described for carbohydrate intake above. As shown in Figure S1 (right) in Multimedia Appendix 1, if the recommended walking duration a_e included in the action is shorter than the perceptual critical point of each user, it is expected to be adhered to. Otherwise, the recommended action is not adhered to, that is, the actual walking duration tends to be zero. Specifically, the exercise adherence model is represented using the following linear model:

$$h_e(a_e \mid \theta_e) = \begin{cases} a_e + e_{z1} & (if \;\; a_e \leq Z_{th}) \\ a_e - Z_{th}\left(\frac{a_e - Z_{th}}{Z_{max} - Z_{th}}\right) + e_{z2} & (if \;\; a_e > Z_{th}) \end{cases} \quad (5)$$

where $\theta_e = \{Z_{th}, Z_{max}\}$ and Z_{th} is the perceptual critical point and Z_{max} is the parameter that represents the upper bound of walking duration in terms of the user's physical strength. In this model, it was assumed that the walking time is zero at the upper bound. Further, we introduce Gaussian noises $e_{z1} \sim N(0, \sigma_{z1})$ and $e_{z2} \sim N(0, \sigma_{z2})$ to represent the uncertainty of exercise behavior from other factors.

Based on these behavior adherence models, each user's dietary behavior parameters θ_d and exercise behavior parameters θ_e are inferred according to the buffered observation data $\boldsymbol{\mathfrak{B}}_i$ accumulated for each user at an iterative time point i. Then, by deploying the inferred parameters $\theta_d(i), \theta_e(i)$ into each corresponding model $h_d(\cdot), h_e(\cdot)$, the behaviors $x^*(i)$, $z^*(i)$ in response to each action candidate can be predicted.

Glycemic Response Model

Next, the postprandial glucose trajectory $\boldsymbol{y}^*$ is predicted based on each behavior x^*, z^* predicted for action A. Our approach models this trajectory based on the clinical findings. First, it is well known that glucose levels increase with carbohydrate intake. For example, Ashrafi et al [33] showed that the actual postprandial glucose trajectory could be fitted by a response curve to carbohydrate intake. Second, it has been proven that exercises, such as walking, immediately lower glucose levels, as previously mentioned in this "two-stage reward prediction model" subheading [19,27-29]. The mechanism is that energy expenditure in muscle tissues due to exercise causes glucose absorption for energy supply [34]. For example, Jankovic et al [35] and Xie and Wang [36] adopted a linear additive model to predict the glucose trajectory and showed that predictive performance was improved by adding the exercise effect on glucose as a factor, in addition to the carbohydrate intake effect. Based on this glucose modeling, we hypothesized that the observed postprandial glucose trajectory $\boldsymbol{y}$ can be represented based on the summation of the response curve $\boldsymbol{\mathcal{R}}_d$ to carbohydrate intake and the response curve R_e to exercise to the baseline preprandial glucose level y_{base}, as follows [37]:

$$\boldsymbol{y} = y_{base} + \boldsymbol{\mathcal{R}}_d(\boldsymbol{x}, \boldsymbol{\tau}_x) + \boldsymbol{\mathcal{R}}_e(\boldsymbol{z}, \boldsymbol{\tau}_z) + \boldsymbol{e}_{\boldsymbol{y}} \quad (6)$$

Because the time range of the glucose trajectory that we focus on here is short-term, we assume that the baseline value y_{base} is constant and substitute the average of the glucose history from 15 minutes before starting the target diet. Here, $\boldsymbol{e}_{\boldsymbol{y}}$ is the vector of Gaussian noise following $N\left(\boldsymbol{0}, \boldsymbol{\sigma}_y\right)$. The response curves $\boldsymbol{\mathcal{R}}_d, \boldsymbol{\mathcal{R}}_e$ are represented by the following equations:

$$\boldsymbol{\mathcal{R}}_d(\boldsymbol{x}, \boldsymbol{\tau}_x) = \sum_{i=1}^{P} h_{d_i} exp\left(\frac{-0.5(\boldsymbol{\Delta}_i - 3\alpha_d)^2}{\alpha_d^2}\right), \; h_{d_i} = \beta_d x_i, \; \boldsymbol{\Delta}_i = \boldsymbol{t} - \tau_{\boldsymbol{x}, i} \quad (7)$$

$$\boldsymbol{\mathcal{R}}_e(\boldsymbol{z}, \boldsymbol{\tau}_z) = \sum_{j=1}^{Q} h_{e_j} exp\left(\frac{-0.5(\boldsymbol{\Delta}_j - 3\alpha_e)^2}{\alpha_e^2}\right), \; h_{e_j} = \beta_e z_j, \; \boldsymbol{\Delta}_j = \boldsymbol{t} - \tau_{\boldsymbol{z}, j} \quad (8)$$

where $\tau_{\boldsymbol{x}, i}, \tau_{\boldsymbol{z}, j}$ represents the time of event occurrence of the ith carbohydrate intake or the jth walking event, respectively. Further, P represents the number of carbohydrates consumed within the target diet, usually $P = 1$. Also, Q represents the number of walking events that occurred within 90 minutes of starting the target diet. In these equations, we adopted a bell-shaped function as the response curve, similar to that in previous works [33,38], owing to its interpretability and the small number of parameters. Some specific examples of each response curve, $\boldsymbol{\mathcal{R}}_d, \boldsymbol{\mathcal{R}}_e$ are shown in Figure S2 in Multimedia Appendix 1. This response curve starts at the timing of event occurrence $\tau_{\boldsymbol{x}, i}, \tau_{\boldsymbol{z}, j}$. The response curve is then amplified based on the amount of carbohydrate intake $\boldsymbol{x}_i$ or the duration of each walk $\boldsymbol{z}_j$. Here β_d, β_e is a parameter that represents the degree of amplification with respect to the volume of each behavior. Further, α_d, α_e is a parameter that represents the degree of time delay of the response. A hyperparameter of the prior is introduced for each parameter $\beta_d, \alpha_d, \beta_e, \alpha_e$, and we describe how this can be obtained in the subsequent "Parameter Learning" subheading.

Based on the above setting, specifically defined by Equations 6-8, it can be seen that the parameters of our glycemic response model $\theta_g = \{\beta_d, \alpha_d, \beta_e, \alpha_e\}$ for each user are estimated from the observed behavior sequence $(\boldsymbol{x}, \boldsymbol{z})$ and glucose trajectory $(\boldsymbol{y})$, at each iteration point (i). In this parameter learning, we use the observed behaviors $(\boldsymbol{x}, \boldsymbol{z})$ rather than the predicted behaviors (x^*, z^*) as training data to identify the parameters more accurately, reflecting the real relationship. In contrast, the predicted behaviors (x^*, z^*) are used as inputs for reward prediction for the action candidate A. Accordingly, the predicted postprandial trajectory is represented as follows:

$$\mathbf{g}(x^*, z^* | \theta_g) = \mathcal{R}_d(x^*, \tau_{x^*} | \beta_d, \alpha_d) + \mathcal{R}_e(z^*, \tau_{z^*} | \beta_e, \alpha_e) \qquad (9)$$

where the baseline term is removed compared to Equation 6. This is because the baseline term does not contribute to the glucose increase and, hence, the reward. Then, $\boldsymbol{y}^*(A | \boldsymbol{\Theta}) = \mathbf{g}(x^*, z^* | \theta_g)$ represents the series of pure increase of postprandial glucose affected by behaviors. It should be noted that for each iteration (i), the estimated parameter values $\theta_g(i)$ are used as θ_g in these equations.

Finally, the glucose increase trajectory, $\boldsymbol{y}^*$, is converted into a reward scale. The reward should be designed such that the smaller the degree of glucose increase, the higher the reward. We use the incremental area under the curve (iAUC), a traditional glucose metric [39], as the degree of glucose increase. Based on these factors, we perform reward conversion using the sigmoid function, as shown below:

$$f(\boldsymbol{y}^*) = \frac{1}{1 + exp\{a(iAUC - b)\}} \qquad (10)$$

$$iAUC = \sum_{t = \tau^*}^{\tau^* + T} y_t^*$$

where iAUC is the total value of the trajectory of glucose increase. As described in the "Problem Setting" subheading, τ^* denotes the start time of the diet, and T=90 minutes. a, b are parameters for the conversion from the iAUC to the reward scale. We set $a = -3$ and $b = 1$ empirically, based on the distribution of the iAUC. Consequently, the reward value ranges between 0 and 1.

Thus, for each action candidate A, a reward calculated from postprandial glucose increase $R(A | \boldsymbol{\Theta}) = f(\boldsymbol{y}^*(A | \boldsymbol{\Theta}))$ is predicted based on our original two-stage prediction.

Action Selection

After the reward is predicted for each action candidate, the policy for selecting batched actions for the real intervention is selected. Specifically, the action that maximizes the predicted reward is selected. However, the predicted reward is not always accurate due to overfitting caused by scarce training data at the initial stage. Moreover, if the action is always selected to maximize that reward, the real action for the user becomes fixed, and the observed user response subsequently exhibits a small variation, which also leads to overfitting. Therefore, we adopt the ε-greedy method for balancing exploration and exploitation. For exploitation, assuming that the prediction model is accurate, actions that maximize the reward (π) should be selected. For exploration, assuming that the prediction model needs to be trained more, actions are selected randomly (π_{random}). Such policies π^* and π_{random} are taken with probabilities of $1 - \varepsilon$ and ε, respectively. Because parameter learning is performed in a batched manner, M actions are selected simultaneously for each policy. Therefore, π^* and π_{random} are represented as follows:

$$\pi^* = \bigcup_{m=1}^{M} \left\{ \underset{k}{\operatorname{argmax}} R(A_k | \boldsymbol{\Theta}) \right\}, \pi_{random} = \bigcup_{m=1}^{M} Random(\boldsymbol{\mathcal{A}})$$

where, $Random(\boldsymbol{\mathcal{A}})$ represents an operation for extracting an element randomly from a set $\boldsymbol{\mathcal{A}}$. Specifically, in π_{random}, this operation is repeated M times independently.

Parameter Learning

The learning of the parameter set $\boldsymbol{\Theta} = \{\theta_g, \theta_d, \theta_e\}$ in the reward model is performed at each iteration time just after M-intervention experiments are finished, and the batched observation data are accumulated in the buffered dataset. In this learning, all the buffered datasets accumulated up to the ith iteration point are used. If the ith batch of data $\boldsymbol{\mathcal{D}}_i$ are observed, the training dataset at that time becomes $\mathfrak{B}_i = \bigcup_{j=1}{}^{i} \boldsymbol{\mathcal{D}}_j$. The batched data includes a history of 4 types of data (action, postprandial glucose level, amount of carbohydrate intake, and postprandial walking duration) for M target meals. This is represented as $\boldsymbol{\mathcal{D}}_j = \bigcup_{m=1}{}^{M} \{A_{j,m}, \boldsymbol{y}_{j,m}, \boldsymbol{x}_{j,m}, \boldsymbol{z}_{j,m}\}$. In the observation data, carbohydrate intake and postprandial walking can occur multiple times per target diet. Therefore, they are represented as vectors (recall Figure 3B). In practice, $\theta_g, \theta_d, \theta_e$ are separately trained from the following subsets of the buffer data $\mathfrak{B}_i^{(g)}, \mathfrak{B}_i^{(d)}, \mathfrak{B}_i^{(e)}$ for each parameter.

$$\mathfrak{B}_i^{(g)} = \bigcup_{j=1}^{i} \left[\bigcup_{m=1}^{M} \{\boldsymbol{y}_{j,m}, \boldsymbol{x}_{j,m}, \boldsymbol{z}_{j,m}\} \right],$$

$$\mathfrak{B}_i^{(d)} = \bigcup_{j=1}^{i} \left[\bigcup_{m=1}^{M} \{A_{j,m}, \boldsymbol{x}_{j,m}\} \right],$$

$$\mathfrak{B}_i^{(e)} = \bigcup_{j=1}^{i} \left[\bigcup_{m=1}^{M} \{A_{j,m}, \boldsymbol{z}_{j,m}\} \right].$$

Here, $\mathfrak{B}_i^{(g)}, \mathfrak{B}_i^{(d)}, \mathfrak{B}_i^{(e)}$ are the training datasets for the glycemic response, dietary adherence, and exercise adherence models, respectively.

To improve the efficiency of this learning process, we introduce prior distributions for each parameter. Specifically, for obtaining prior information for the behavioral parameter sets θ_d, θ_e, we ask each user to answer questions corresponding to each parameter beforehand and use the responses to define the prior distribution. For example, the question for the critical point X_{th} is "If you were to reduce carbohydrates in your dinner, what would you say is the minimum amount of carbohydrates?" Then, the user's answer is set to the mean $\mu_{X_{th}}$ of the Gaussian prior distribution $N\left(\mu_{X_{th}}, \sigma_{X_{th}}\right)$. For the other parameters $X_{norm}, Z_{th}, Z_{max}$, each prior distribution is set similarly.

On the other hand, the glycemic parameter set θ_g cannot be clarified through user interviews because each parameter is related to a complex individual biological response. Therefore, we adopt a data-driven approach wherein the prior distribution is pretrained through the original upstream task. Before starting the first intervention experiment, we recruited additional participants belonging to the same disease group as the target users. Second, we collect a glucose dataset, $\mathcal{D}^{(g)} = \bigcup_{l=1}{}^{L}\{\boldsymbol{y}_l, \boldsymbol{x}_l, \boldsymbol{z}_l\}$ for L diet events from each participant. Third, this dataset is used to infer glycemic parameters θ_g for each participant. Finally, the distribution of these parameters θ_g among all participants is used as the Gaussian prior distribution of θ_g. Specifically, the mean $\tilde{\boldsymbol{\mu}}_g$ and the variance $\tilde{\boldsymbol{\sigma}}_g$ of these parameters θ_g are set as $N\left(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\sigma}}_g\right)$. Here, each parameter θ_g is assumed to be generated independently.

Through this process, each prior distribution of the parameter set $\theta_g, \theta_d, \theta_e$ is obtained before the first intervention experiment, and this prior is used for Bayesian parameter learning at every iteration time point. This parameter learning is performed by executing a Markov Chain Monte Carlo simulation with the No U-Turn Sampler implemented in RStan (Stan Development Team) [40].

Simulation Experiment Setup

We studied in a simulation experiment whether our online algorithm can improve postprandial glucose levels with personalized recommendations. In addition, to carry out real experiments with real personalized recommendations, we investigated the best setting for the proposed method to maximize performance. Specifically, we aim to clarify the following questions through simulation experiments:

- Q1: In action selection, how should the trade-off point between exploration and exploitation be set to maximize the cumulative reward?
- Q2: How fast does parameter learning converge when introducing a prior distribution into the reward prediction model?

To answer these questions, we set up several virtual users who exhibited different behavioral and glycemic responses, performed repeated action planning and pseudo-intervention experiments using the proposed method for each user, and evaluated these experimental results to address the above questions. At the same time, we also show that by introducing the proposed method, the reward performance can be improved compared to selecting default actions uniformly.

Virtual User Setting

In the real world, behavioral tendencies and glycemic responses differ from user to user. Therefore, in this simulation, a parameter set $\theta_g, \theta_d, \theta_e$ for each user was set to ensure that virtual users exhibit response patterns as diverse as possible. Table 1 shows the actual settings of parameter values for each user. The total number of virtual users was 10.

Table 1. User settings for each parameter.

	Dietary adherence model		Exercise adherence model		Glycemic response model			
Virtual user	X_{th}(g)[a]	X_{norm}(g)[b]	Z_{th}(min)[a]	Z_{max}(min)[c]	β_d[d]	α_d[e]	β_e[d]	α_e[e]
*ID*1	50	100	20	60	0.05	20	–0.15	6.0
*ID*2	30	60	10	40	0.05	20	–0.20	8.0
*ID*3	70	120	30	80	0.05	20	–0.15	6.0
*ID*4	90	120	5	40	0.05	20	–0.20	8.0
*ID*5	50	100	20	60	0.04	15	–0.15	6.0
*ID*6	70	120	10	60	0.04	15	–0.20	8.0
*ID*7	30	60	30	80	0.04	15	–0.15	6.0
*ID*8	90	120	5	30	0.04	15	–0.20	8.0
*ID*9	10	40	30	80	0.04	15	–0.20	6.0
*ID*10	10	40	5	30	0.05	20	–0.20	8.0

[a] X_{th} and Z_{th}: perceptual critical points for carbohydrate intake and postprandial walking duration (Equations 4 and 5).

[b] X_{norm}: normal carbohydrate intake (Equation 4).

[c] Z_{max}: upper bound of walking duration (Equation 5).

[d] β_d and β_e: the degree of glycemic amplification of carbohydrate intake and postprandial walking duration (Equations 7 and 8).

[e] α_d and α_e: the degree of time delay of the glycemic response to carbohydrate intake and postprandial walking duration (Equations 7 and 8).

As an example of user diversity, it is difficult for user ID4 to reduce carbohydrate intake because of $X_{th} = 90$, whereas user ID9, with $X_{th} = 10$ finds it easy to reduce carbohydrate intake. User ID4 also has $\beta_d = 0.05$, indicating a greater glycemic response to carbohydrate intake, while user ID5 has a lower glycemic response due to $\beta_d = 0.04$.

Thus, each parameter value was set differently for different users. The value ranges of the behavior parameters θ_d, θ_e were within the range of the action set shown in Figure 2. These ranges were defined following established dietary and exercise guidelines. Specifically, the minimum carbohydrate intake (X_{th}) for virtual users was set with reference to "very low-carbohydrate" diets (20-50 g/d) and "low-carbohydrate" diets (<130 g/d) as defined by Oh et al [41], while the normal carbohydrate intake (X_{norm}) was guided by typical daily carbohydrate intake levels (200-325 g/d) as described in the literature [26]. For the parameters of postprandial walking durations (Z_{th}, Z_{max}), typical ranges in prior experimental studies (5-30 minutes) [19,27,29] and recommendations for at least 10 minutes of walking [27] informed the setting of Z_{th}, and cases of walking up to 90 minutes [28] guided Z_{max}. To introduce a diverse range of realistic individual differences among virtual users, slight deviations from these foundational literature values were sometimes permitted. Nevertheless, the chosen parameter ranges for virtual users in Table 1 do not drastically diverge from the prior information collected from real participants (in Table 2), suggesting that the simulated settings are not unrealistic. In addition, the value ranges of the glycemic parameters θ_g were set based on the prior knowledge obtained in the upstream task.

Table 2. Prior information on behavioral parameters.

Parameter	Questionnaire	ID1	ID2	ID3	ID4	ID5	ID6
X_{th}	"If you were to reduce carbohydrates in your target diet, what would you say is the minimum amount of carbohydrates?"	50 g	50 g	30 g	30 g	50 g	50 g
X_{norm}	"How much carbohydrates in grams do you typically consume in your target diet?"	110 g	90 g	50 g	50 g	70 g	150 g
Z_{th}	"If you were to walk after your target meal, how long do you think you could walk each day?"	30 minutes	15 minutes	15 minutes	20 minutes	25 minutes	5 minutes
Z_{max}	"Based on your physical fitness, how long do you think you can continue walking?"	60 minutes	60 minutes	180 minutes	60 minutes	40 minutes	120 minutes

User Response Simulator

Next, to perform the simulation experiment, it is necessary to generate a pseudo-response to any action according to each user setting. Therefore, we developed a user-response simulator that outputs 3 responses (dietary, exercise, and glycemic) to 2 inputs (an action and each parameter value), as shown in Figure S3 in Multimedia Appendix 1.

In this simulator, the simulated dietary behavior and exercise behavior for an action are the first outputs, and a simulated glucose trajectory is generated based on these behaviors. First, the simulated behaviors are pre-estimated according to the corresponding behavior adherence model in Eqs 4 and 5 by setting the input parameter values of the virtual user and input action $A = \{a_d, a_e\}$. Subsequently, by perturbing the simulated behaviors by adding Gaussian noise, we obtain the final simulated behaviors $\check{x}, \check{z}$. Here, the perturbation noise for carbohydrate intake followed $N(0, 10)$, and the noise for postprandial walking duration followed $N(0, 5)$. Subsequently, these simulated behaviors, $\check{x}, \check{z}$ are input into the glycemic response model (Equation 9), and the obtained trajectory is then perturbed by adding Gaussian noise to generate the final simulated glucose trajectory $\check{\boldsymbol{y}}$. In this perturbation, a time-independent noise following $N(0, 0.02)$ was added. Figure S4 in Multimedia Appendix 1 shows a representative example of these simulated responses.

By substituting the part (†) of the pseudocode in Algorithm 1 (Textbox 1) with the above simulator, it is possible to simulate repeated experiential learning cycles of action planning and intervention experiments, as shown in Figure 1B, for each virtual user.

Simulation Evaluation Settings

To answer the question (Q1), we tracked how the actual cumulative reward changes depending on the settings of the proposed method while changing them. One of the most significant settings is the balance of exploration and exploitation in action selection, which is determined according to the set value of ε, a hyperparameter of the proposed method. Therefore, we set multiple patterns of ε values and investigated which setting yielded the highest cumulative reward. For these value patterns, we considered $\varepsilon = 1.0$ for random selection, $\varepsilon = 0.5$ for high, $\varepsilon = 0.2$ for medium, and $\varepsilon = 0.01$ for low condition of ε.

Another significant setting is the introduction of prior distributions for each model parameter. Thus, we also investigated how the cumulative reward varied depending on whether the prior distribution was used in parameter learning. In this simulation, prior knowledge of the behavioral parameters $\{X_{th}, X_{norm}, Z_{th}, Z_{max}\}$ was pseudo-generated by adding Gaussian noise to the true parameter values. This may be attributed to the fact that the results of the survey questionnaire are subject to the human cognitive errors of each user regarding their true behavior, and these cognitive errors are assumed to follow Gaussian noise. On the other hand, to set the prior distribution of the glycemic parameters $\{\beta_d, \alpha_d, \beta_e, \alpha_e\}$, the mean and the variance of the true values among all users were substituted. The initial batch of 6 actions was based on past clinical findings (refer to Figure 2). From the second iteration ($i \geq 2$), action selection was performed by integrating the proposed method under each setting.

Metrics

Specifically, the cumulative reward of the virtual user (n) at iteration of the experiential learning cycle (i) can be represented as follows:

$$G_n(i) = \frac{1}{i \times D} \sum_{j=1}^{i} \sum_{d=1}^{D} r_{j,d,n}$$

where $r_{j,d,n}$ represents the actual reward transformed by Equation 10 from the glucose trajectory $\check{\boldsymbol{y}}$ of the virtual user n after the dth target diet in the jth intervention experiment. Further, D is the number of target diets included in each experiment; and $D = 6$ was used in the simulation. The average cumulative reward of N virtual users, which is used as an evaluation metric, can be represented as follows:

$$J(i) = \frac{1}{N} \sum_{n=1}^{N} G_n(i)$$

By increasing i to the maximum iteration time I, we can track how the metric $J(i)$ changes depending on each condition of ε. Considering the burden on users in real experiments, I was set to 10. This means that each user must continue wearing the necessary sensors to collect user responses for a maximum of 60 meals when $D = 6$.

In this simulation, the true value of each parameter for each user was known, as listed in Table 1. This enables the investigation of the estimation error of each parameter for any setting of our method. To answer the question (Q2), we investigated how the estimation error changes as the iteration time increases (ie, the amount of training data), with and without prior distribution. The estimation error was defined as follows:

$$\epsilon_p(i) = \frac{\left|\hat{\theta}_p(i) - \theta_p\right|}{|\theta_p|} \times 100$$

This corresponds to the standard error (%). p is the parameter number, θ_p is the true value of the parameter in Table 1, and $\hat{\theta}_p(i)$ is the estimated value of the parameter in the ith iteration time.

Real-World Feasibility Study Setup

Additionally, we conducted a preliminary set of short intervention experiments with real users using the proposed method, based on the simulation results. This experiment was positioned as a pilot study before conducting a full-scale study. The purpose of this experiment was to explore the feasibility of whether postprandial glucose levels can be improved in real users after recommending actions updated by the proposed method compared to default actions. Accordingly, in this study, we compared the performance when default actions were selected in the first experiment (referred to as the "default policy") with the performance when actions were selected by the proposed method in the second experiment (referred to as the "optimized policy") using a within-participant study design, to evaluate the feasibility of the proposed method. The optimized policy was updated from the observed participant responses in the first experiment, following the proposed method.

Experimental Protocol

First, we recruited 6 healthy participants (4 men and 2 women) as real users from April 2023 to July 2023. Because this experiment was the first feasibility study, the target group was healthy rather than diabetic. All participants were researchers at Aalto University and the University of Helsinki.

Next, each participant was assigned the goal of reducing postprandial glucose levels to prevent diabetes. To achieve this goal, the participants received daily recommendations for the target diet, including total carbohydrate intake and postprandial walking duration, via a dedicated smartphone app. The target diet was either dinner or lunch each day, and the participants chose according to their lifestyle preferences. While the recommendations in the first experiment were preset to default actions, those in the subsequent experiment were preset to actions selected by the proposed method. Participants were free to choose whether to follow the recommendations and to act accordingly. To collect a participant's behavioral and glycemic responses to the recommendation actions, they wore a continuous glucose monitoring device and an activity tracker.

This process of mobile intervention and data collection for each participant was carried out for 6 days in the first experiment and for 3 days in the second experiment. Because the first experiment corresponded primarily to the exploration phase and the second experiment corresponded to the exploitation phase, the number of days in the second experiment was set to be short.

Recommendation of Action

An action was delivered as a recommendation in a notification app on a dedicated smartphone (iPhone 12), which was provided to each participant. In the notification app, a dedicated message including the action was delivered along with a notification sound, a few hours before the target diet. The frequency and duration of this recommendation were predetermined according to participants' preferences.

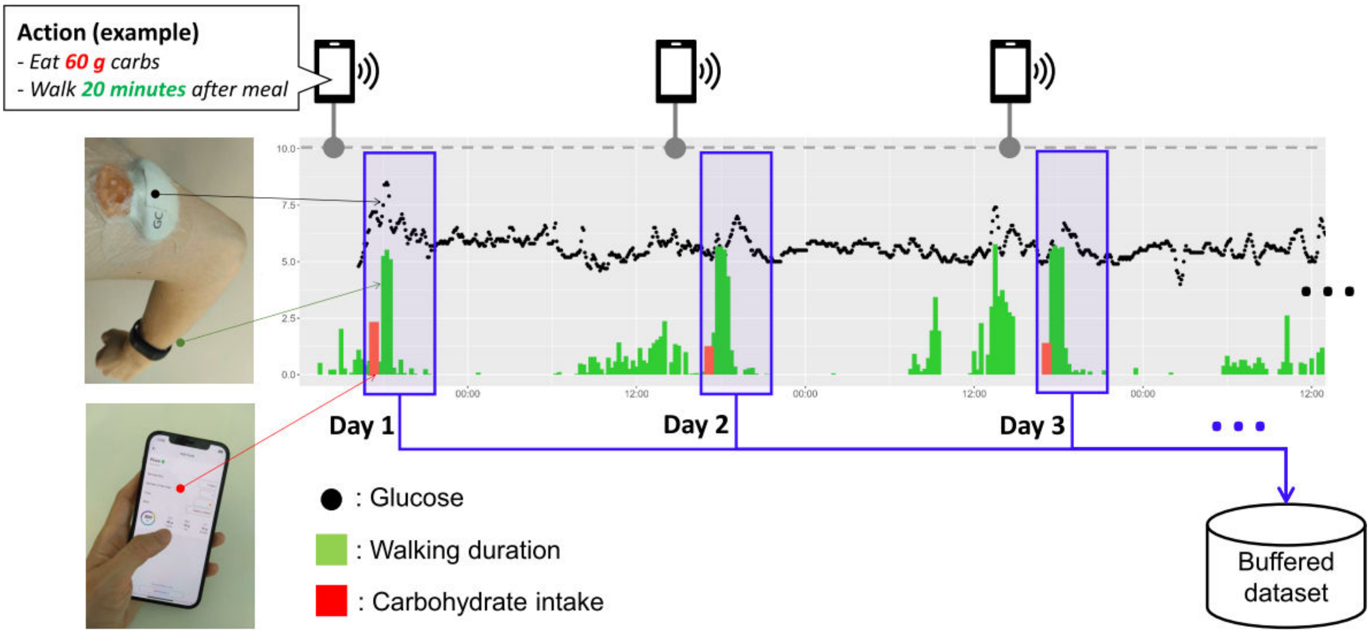
An example of this message appearing on the smartphone screen is "Eat less than 60 g of carbohydrates at dinner, and walk for more than 20 minutes just after dinner." Depending on the selected action, the amount of carbohydrate and postprandial walking duration in the message changed. For default actions in the first experiment, carbohydrate intake (a_d) was within the range of 20-80 g, and the postprandial walking duration (a_e) was within the range of 5-20 minutes, considering the participants' preferences and previous literature [19,26-29]. For example, in the second experiment, actions were selected based on the two-stage reward prediction results of each action candidate; therefore, they were not always within this range. This reward prediction was attained by setting the parameters learned from the user response data obtained in the first experiment.

Response Data Collection

For user response data, we collected data related to the participant's continuous glucose levels, physical activity, and food records. Figure 4 illustrates how the data were collected. Continuous glucose measurements were collected every 5 minutes using a Guardian Connect medical device (Medtronic Inc), which was inserted on the participant's skin. At the same time, physical activity data were collected using the activity tracker Vivosmart 3 (Garmin Ltd), which was worn on the participant's arm. To mitigate erroneous detection of walking from various daily movements, the device identifies a walking event when characteristic arm movements indicative of continuous walking are detected for a duration exceeding a certain minimum threshold. The physical activity data included the start time and duration of each walking event that was automatically detected by the device. A dedicated fitness tracker was chosen over participants' personal smartphones to ensure data consistency and reliability across participants, as smartphone usage and carrying patterns (eg, in a pocket or bag) can vary significantly and lead to inconsistent activity detection. This approach also minimizes issues related to battery drain on personal devices and addresses potential privacy concerns by avoiding continuous access to personal smartphone data. The glucose and activity data were continuously sent to a dedicated iPhone through a Bluetooth connection with each device. In addition, a food diary, including the amount of carbohydrates and the start time of each target diet, was recorded manually by the participant. The carbohydrate content was automatically calculated using the mobile app MyFitnessPal by inputting all food contents and amounts.

Finally, because we focus on the improvement of postprandial glucose levels, we extracted postprandial segment data from the longitudinal time series as user response data for each action. Specifically, the segment data were defined from the start of each target diet until 90 minutes later, as shown by the blue frames in Figure 4. Each data segment was added to a buffered dataset.

Figure 4. Illustration of response data collection.



Metrics

The evaluation was based on whether the reward related to postprandial glucose increased in the second experiment compared to the first. Therefore, we set the following metric as the actual reward average $J'(i)$ for all diets of all N participants in the ith experiment, and this metric was compared between the first and second experiments.

$$J'(i) = \frac{1}{N}\sum_{n=1}^{N} G'_n(i), G'_n(i) = \frac{1}{D}\sum_{d=1}^{D} r_{i,d,n}$$

It should be noted that the actual reward $r_{i,d,k}$ is transformed by Equation 10 from the glucose increase trajectory, which is obtained by subtracting its baseline derived from preprandial glucose levels from the observed glucose trajectory $\boldsymbol{y}$.

Acquisition of Prior Knowledge

To introduce the prior distribution of the behavioral parameters, each participant was asked to answer a questionnaire before starting the experiment. As described in the "Parameter Learning" subheading, the result was set as the mean value of the prior distribution of the corresponding parameter. Table 2 shows the actual answers for each parameter together with the corresponding question. It can be observed that the values vary according to the participant. As for glycemic parameters, we estimated the prior distribution $N(\tilde{\mu}_g, \tilde{\sigma}_g)$, which is common to target participants, using the glucose dataset from other participants. To obtain this, we recruited 4 additional healthy participants and collected continuous glucose levels, physical activity, and food records around lunchtime over 6 days for each participant in the same manner, as described previously. Consequently, a total of 6 segment data were obtained for each participant, as shown in Figure 4. Then, after estimating the glycemic parameters $\boldsymbol{\theta}_g$ for each participant using his segment data, the parameters $\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\sigma}}_g$ were calculated among 4 participants, as described in the "Parameter Learning" subheading in this section. As a result of this calculation, we obtained $\tilde{\boldsymbol{\mu}}_g = \{\tilde{\mu}_{\beta_d}, \tilde{\mu}_{\beta_e}, \tilde{\mu}_{\alpha_d}, \tilde{\mu}_{\alpha_e}\} = \{0.047, -0.148, 19.07, 6.22\}$ and $\tilde{\boldsymbol{\sigma}}_g = \{\tilde{\sigma}_{\beta_d}, \tilde{\sigma}_{\beta_e}, \tilde{\sigma}_{\alpha_d}, \tilde{\sigma}_{\alpha_e}\} = \{0.028, 0.035, 1.91, 0.39\}$.

Ethical Considerations

The study was approved by the Aalto University Research Ethics Committee (Treatment Planning Project/08.06/2023). Written informed consent was obtained from all participants before the start of the experiment. Direct identifiers were removed, and data were pseudonymized and stored securely. No personal data were shared with third parties. Additionally, to ensure participant safety, participants were informed orally and via the consent form that they could withdraw at any time if they experienced discomfort or unease.

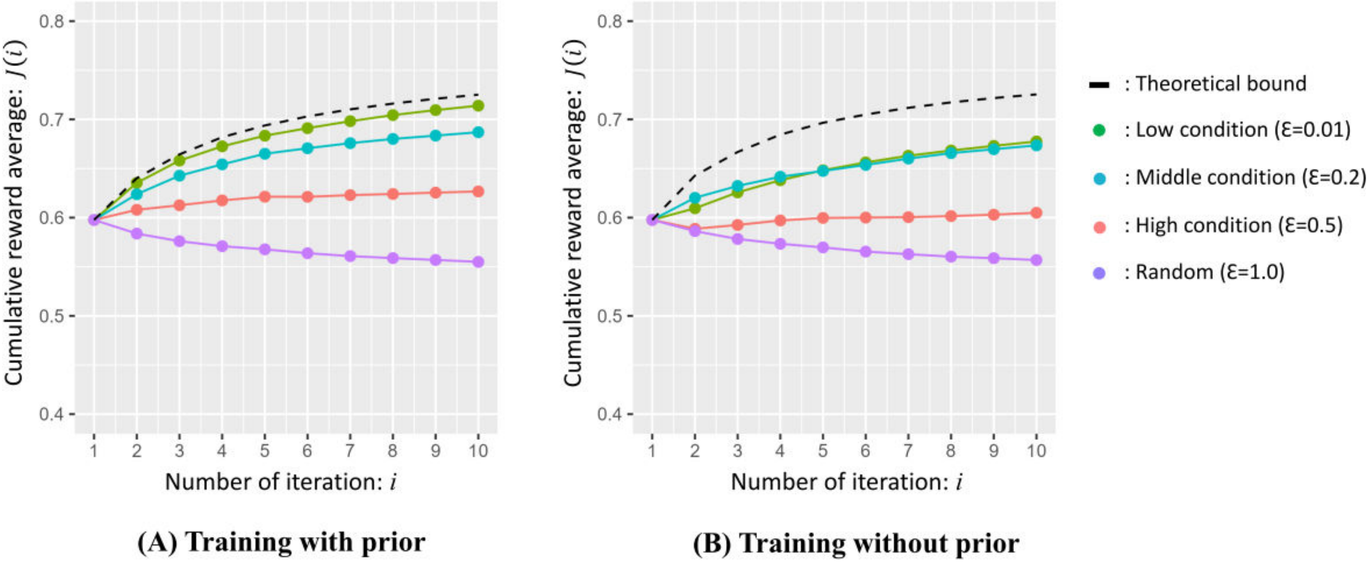
Results

Simulation Experiment Results

We present the simulated results of the average cumulative rewards ($J(i)$) with and without prior distribution in Figure 5. First, in any case, when an action selection by the proposed method is adopted (ie, when ε is other than 1.0), we can confirm that the cumulative reward gradually increases from the second iteration ($i = 2$) after executing default actions at the first iteration ($i = 1$). This demonstrates the effectiveness of action selection by the proposed method in the second and subsequent iterations ($i \geq 2$). Regarding question Q1, by focusing on the variation of ε, we can also confirm that the smaller the value of ε, the greater the accumulated reward, and the maximum cumulative reward is obtained when $\varepsilon = 0.01$.

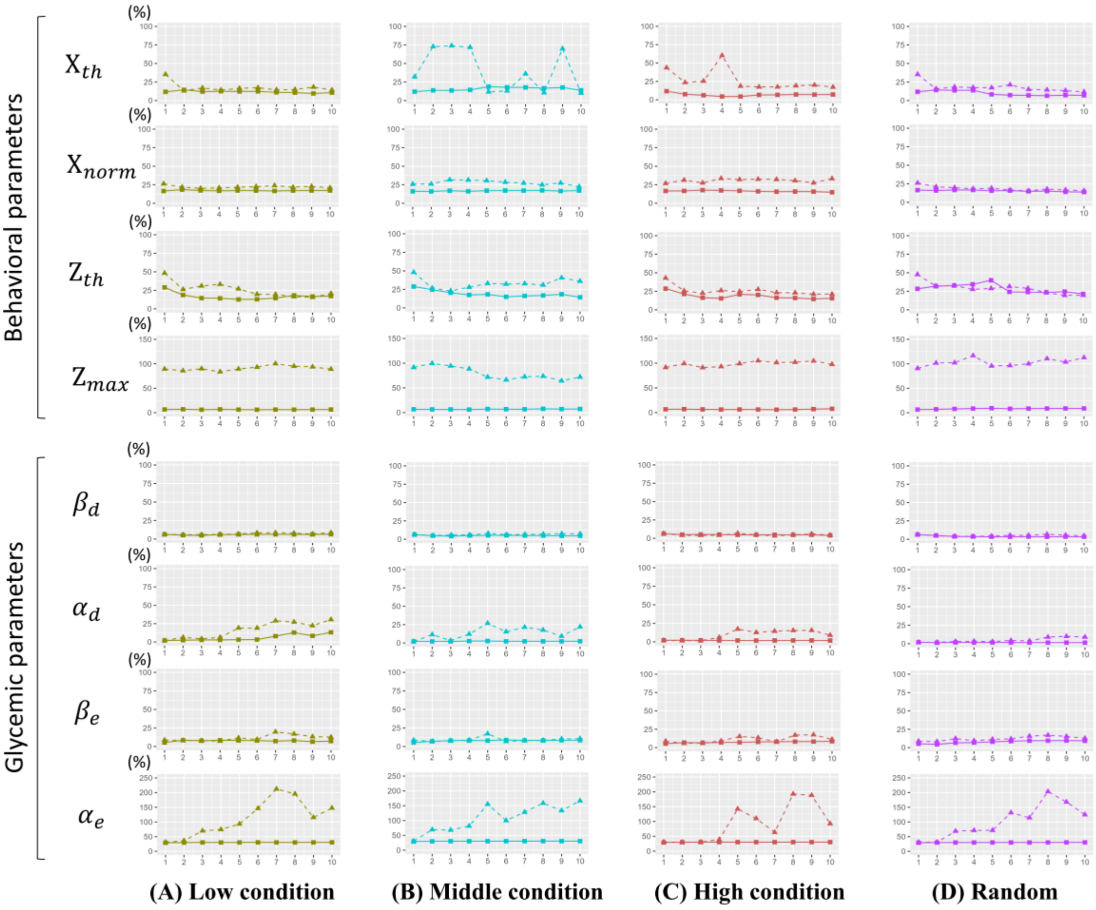
The theoretical boundary of the cumulative reward is shown by the black dotted line in Figure 5, and it is calculated as an ideal condition, wherein the true parameter values are known and used for action selection from the second iteration. The cumulative reward for $\varepsilon = 0.01$ with prior knowledge shows almost the same value as that in this theoretical bound. In other words, when prior knowledge is introduced, exploration with the policy π_{random} is unnecessary, and the best performance can be attained by full exploitation with the policy π that maximizes the predicted reward from the beginning. On the other hand, in the case without prior knowledge in Figure 5B, the cumulative reward decreases significantly when $\varepsilon = 0.01$ and 0.2 are compared with the case with prior knowledge in Figure 5A. In particular, when $\varepsilon = 0.01$, the decrease in the second iteration is remarkable.

Figure 5. Simulation results of average cumulative reward with prior distribution "(A)" and without prior distribution "(B)" of the proposed method.



Next, regarding question Q2, we show the actual changes in the parameter estimation error ($\epsilon_p(i)$) of the behavioral adherence and glycemic response models in Figure 6. First, it can be seen that the case without prior knowledge (▲) has a larger estimation error than the case with prior knowledge (■) for the first iteration. This large error leads to a prediction error of the reward for each action candidate and therefore makes it easier to select worse actions with the policy π^*. This may be the reason why the cumulative reward decreases immediately in case $\varepsilon = 0.01$ and 0.2 due to the erroneous action selection at $i = 2$ in Figure 5B. Furthermore, in the case without prior knowledge, it can be seen that the estimation errors of some parameters (such as Z_{max}, α_e) are not improved as iteration time of experiential learning cycle increases. This indicates the potential risk in real experiments, showing that it is impossible to collect the necessary training data for stable parameter learning, even when imposing heavy burdens of repeated interventions on users. However, when introducing prior knowledge, we can confirm that the parameter errors are minimized from the first iteration for most parameters. This suggests that introducing prior knowledge can eliminate the burden on users and maximize their cumulative rewards.

Figure 6. Average estimation error of behavioral parameters and glycemic parameters across different ε-greedy conditions: (A) low condition, (B) middle condition, (C) high condition, and (D) random. ■ represents results with prior, and ▲ represents results without prior. The vertical axis of each graph represents the average estimation error ($\varepsilon_p(i)$) of the pth parameter among all 10 users, and the horizontal axis represents the iteration time (i) of the experiential learning.



(A) Low condition (B) Middle condition (C) High condition (D) Random

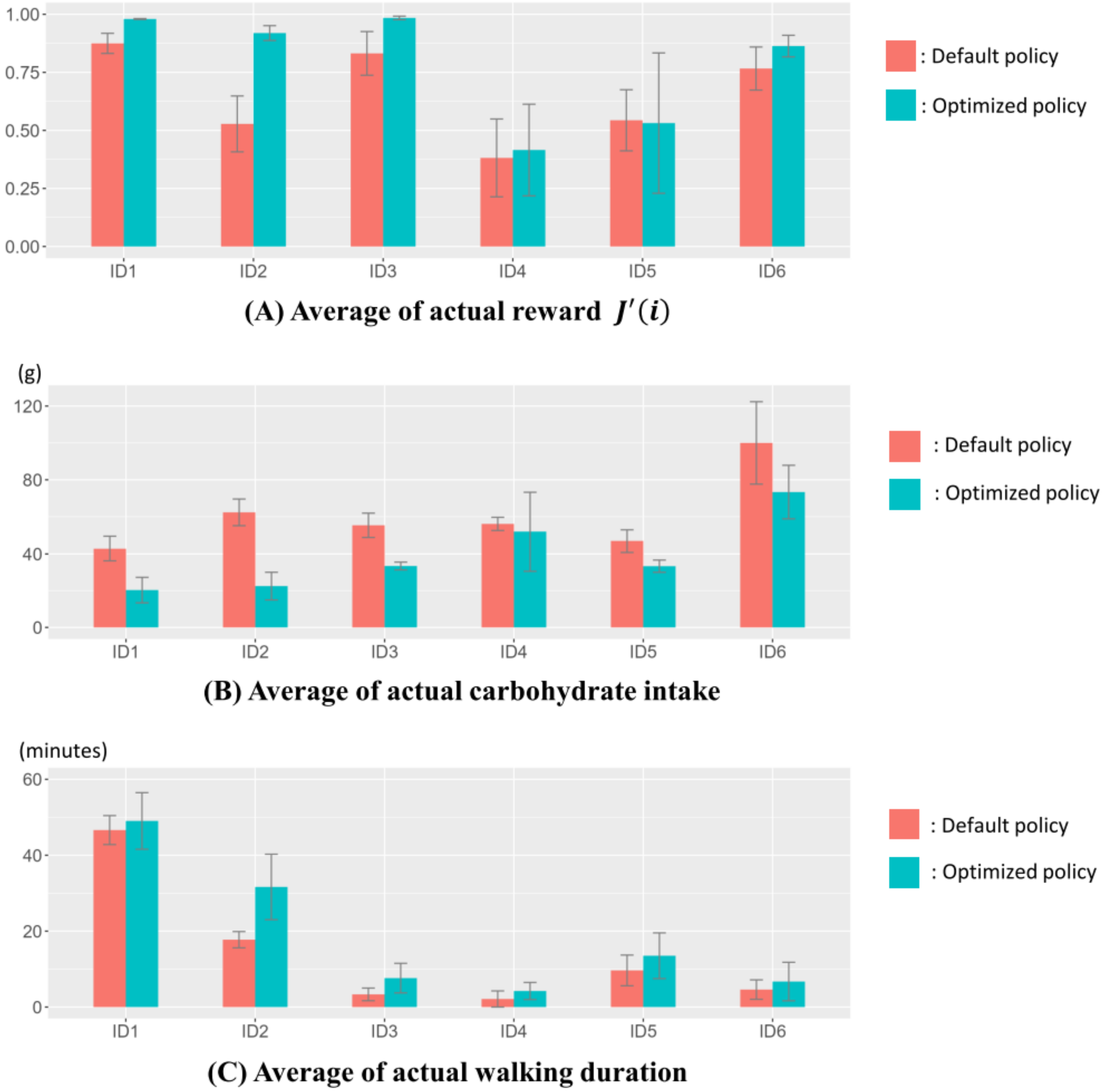
Real-World Experiment Results

Based on the simulation results, we conducted a real-world experiment with a simplified algorithm involving a single update of the policy. According to the setting derived in the simulation experiment, we introduced prior knowledge for parameter learning, and we took the greedy policy with $\varepsilon = 0$ for action selection.

The obtained metric values for each participant from the real experiments conducted on the within-participant design are shown in Figure 7A. For all participants except ID5, the simple reward average increased in the second experiment using the proposed method. On average, for all participants, this metric increased from 0.65 in the first experiment to 0.80 in the second experiment, representing a 23% increase in reward. To statistically evaluate this improvement, we performed a paired *t* test comparing the average rewards between the first (default policy, R1) and second (optimized policy, R2) experiments across the 6 participants. The analysis revealed a statistically significant increase in reward following the optimized policy (*P*=.04). This result indicates that action selection using the proposed method is effective for real-world applications, consistent with the simulation results.

Moreover, we show how carbohydrate intake and postprandial walking duration changed in the first and second experiments in Figures 7B and 7C. All participants showed decreased carbohydrate intake and increased postprandial walking duration. For all participants, the average carbohydrate intake decreased by 35.3% from 60.5 to 39.1 g, and postprandial walking duration increased by 34.3% from 14.0 to 18.8 minutes. To further quantify these behavioral improvements, we conducted paired *t* tests for carbohydrate intake and postprandial walking duration. For carbohydrate intake, the mean difference (R2 – R1) was −21.425 g, and the paired *t* test showed a statistically significant reduction (*P*=.003). Similarly, for postprandial walking duration, the mean difference (R2 – R1) was 4.780 minutes, and the paired *t* test indicated a statistically significant increase (*P*=.03). These statistically significant behavioral changes underscore the effectiveness of the proposed method in guiding users toward healthier actions.

Figure 7. Changes in average reward (A), average carbohydrate intake (B), and average postprandial walking duration (C) for each participant in real experiments. Standard errors are shown as error bars. Statistically significant improvements (*P*<.05) were observed in each of reward, carbohydrate intake, and postprandial walking duration from the first experiment (default policy) to the second experiment (optimized policy). On average, reward increased from 0.65 (default policy) to 0.80 (optimized policy), carbohydrate intake decreased from 60.5 g (default policy) to 39.1 g (optimized policy), and postprandial walking duration increased from 14 minutes (default policy) to 18.8 minutes (optimized policy).



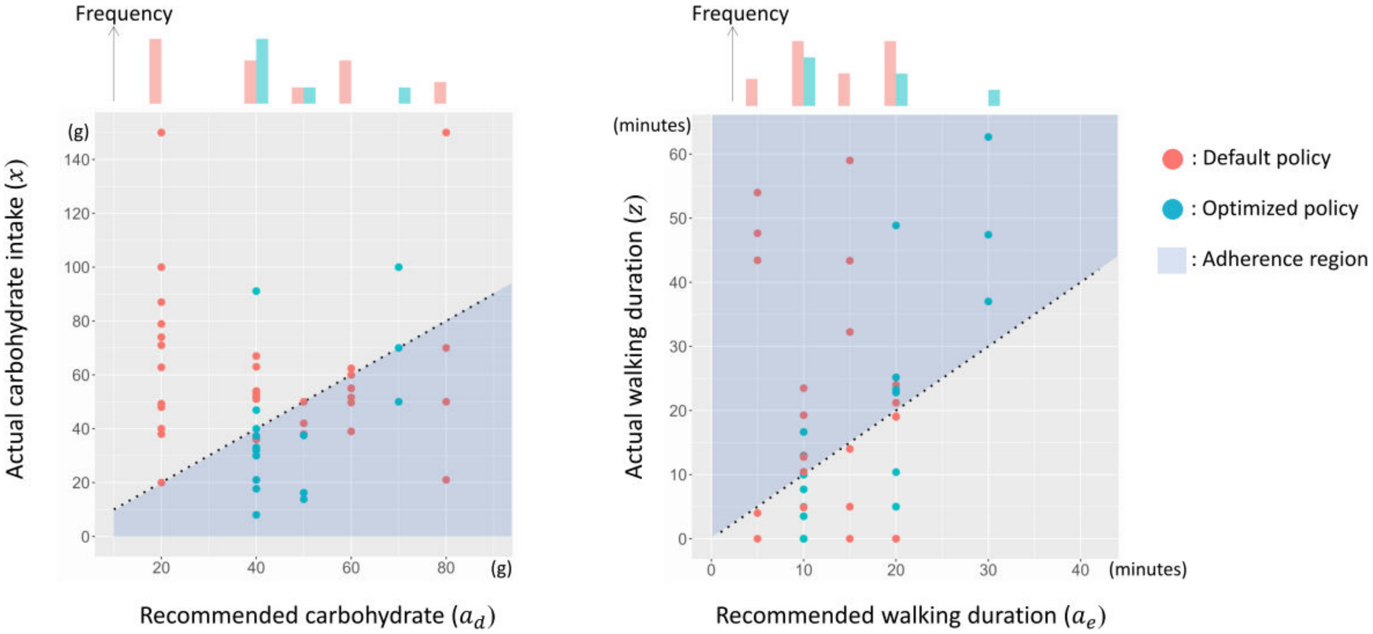
(A) Average of actual reward $J'(i)$

(B) Average of actual carbohydrate intake

(C) Average of actual walking duration

Furthermore, to investigate the impact on behavioral adherence, we show the actual behaviors of all participants in response to the recommended actions in Figure 8. The adherence to carbohydrate intake improved from the first experiment to the second experiment. In fact, as a result of calculating the adherence rate for each experiment, the adherence rate for carbohydrate intake increased from 45.7% to 83.3% and that for postprandial walking duration increased from 40% to 55.6%. These behavioral improvements support the hypothesis that action selection using the proposed method is appropriate for participants from the viewpoint of ease of behavior change.

Figure 8. Recommended action vs actual behaviors in real experiments (left: carbohydrate intake and right: postprandial walking). The horizontal and vertical axes show the values of recommended action (a_d, a_e) and actual behavior (x, z), respectively. Each data point denotes a pair of one action and one behavior for any participant. A participant is considered to adhere to the recommended action in cases $x \leq a_d$or $z \geq a_e$ shown in the blue region. For each policy, the frequency of action value is shown on top of each graph. The distribution differs between policies.



Discussion

Principal Findings

In this paper, we propose a multiarmed bandit algorithm based on the two-stage reward prediction mechanism, which can individually plan mobile behavioral interventions and make it easier for users to perform actions and reduce glucose levels after eating.

From the simulation experiment results, first, we confirm that a larger cumulative reward can be obtained by adopting the proposed method than by selecting the default action. This demonstrates the effectiveness of our online algorithm for personalizing recommendations. Second, it is also demonstrated that the parameter learning converged immediately at the first iteration $(i = 1)$ by introducing prior knowledge. This indicates it is not always necessary to repeat the intervention experiment many times for obtaining accurate parameters, emphasizing the importance of the initial single update we conducted in the real-world feasibility study.

From the short-term real-world feasibility study, with a simplified proposed method involving a single update of the policy into a personalized one, it is found that actual postprandial glucose levels can be improved using the proposed method through behavioral improvements in both carbohydrate intake and postprandial walking. Specifically, we observed an average 23% improvement in actual glucose responses, which was statistically significant (P=.04), along with statistically significant improvements in behavioral adherence to the recommendations concerning carbohydrate intake (P=.003) and postprandial walking (P=.03). The adherence rates also increased notably (from 45.7% to 83.3% for carbohydrate intake and 40% to 55.6% for postprandial walking). While promising, these results are based on a small-scale pilot study with healthy participants and a simplified intervention protocol. Therefore, they should be viewed as preliminary evidence rather than definitive conclusions, primarily serving as guidance for the design of future, more comprehensive clinical trials. This demonstrates the potential effectiveness of the proposed method in a real-world setting. Crucially, this pilot demonstrated that even a single update, informed by our novel two-stage reward prediction model and based on initial user-specific data, can yield statistically significant improvements in clinical outcomes (glucose levels) and behavioral adherence in a real-world setting. This offers compelling initial evidence for the practical viability and immediate impact of our personalized approach, serving as a vital proof-of-concept for its core mechanism in a complex human context.

While simulation experiments elucidated the algorithm's behavior across diverse iteration numbers (i), the real-world feasibility study, designed as a pilot, adopted a simplified approach: a single policy update (from default to optimized after the initial experiment), prioritizing feasibility and logistical tractability. This design inherently meant that the full extent of continuous learning and adaptive capabilities, central to multiarmed bandit frameworks over extended iterations, was not fully explored. Consequently, this study primarily demonstrated the potential of personalized recommendations derived from an initial learning phase, rather than evaluating continuous adaptive optimization over a prolonged period. This outcome underscores the critical need for future investigations to deploy the full online algorithm, addressing the challenges and benefits of continuous policy updates, particularly concerning user engagement and dynamic model accuracy.

To the best of our knowledge, this is the first attempt to (1) clarify how to optimize behavioral interventions for improving glucose levels and (2) demonstrate the effectiveness of the method through simulation experiments and real interventional experiments.

Limitation

While this preliminary study offers promising insights, it is important to acknowledge several limitations that warrant future investigation to ensure comprehensive validation and generalization. First, the real experiment was short-term and small-scale. The findings from this preliminary study, which involved only 6 healthy participants over a short duration, should therefore be interpreted with caution and are not readily generalizable to a broader population, particularly including patients with diabetes who represent the primary target group for this intervention. The exclusion of patients with diabetes in this initial feasibility study naturally limits the immediate clinical applicability of these specific real-world results. While this real-world feasibility study offers encouraging initial evidence that effective personalized policy updates could be established within a practical, short-term data collection period (eg, 6 days), future research will include a long-term and massive intervention study. This is needed to fully evaluate the sustained effects and long-term stability of parameter convergence and policy updates beyond the initial phase (ie, after the third iteration [$i \geq 3$]) in diverse real-world settings and patient populations. Also, as our target in this study was limited to healthy participants, the interventional experiments in patients with diabetes should be performed for a comprehensive evaluation, adhering to robust statistical principles, including multiple comparison adjustments.

Second, this study primarily compared the proposed online algorithm against a randomized (default) policy. While this comparison effectively demonstrated the benefit of personalization over a nonpersonalized baseline, it did not include a direct comparison with other existing reinforcement learning-based or contextual bandit algorithms from the literature. Such comparisons would provide a more comprehensive understanding of the relative performance advantages and disadvantages of our proposed method against alternative state-of-the-art approaches. Future work could aim to include these comparisons, potentially through simulation studies based on established benchmarks or by incorporating other algorithms into real-world trials. Furthermore, a limitation exists in the parameterization of our simulation experiments. While parameter ranges for virtual users (Table 1) were based on literature, specific values were partially informed by our small-scale feasibility study (Table 2). This limited sample of 6 healthy participants may not be fully representative, potentially affecting the generalizability of simulation results. Consequently, these simulations primarily served to qualitatively explore the algorithm's behavior and identify optimal operational settings (eg, epsilon values and importance of prior knowledge) for future large-scale trials, acknowledging they do not provide definitive quantitative predictions. Future research should prioritize parameter validation and derivation from larger, more diverse datasets, possibly through crowdsourcing or extensive clinical trials, to enhance simulation robustness.

Third, the predictive models, key to our two-stage reward prediction mechanism, possess inherent limitations, particularly regarding data requirements and accuracy. Furthermore, the two-stage prediction itself introduces a compounding effect of prediction errors: inaccuracies in behavioral adherence predictions can propagate and amplify errors in subsequent glycemic response prediction. This “multiplication” of noise impacts reward estimation, especially with limited data. While our preliminary study showed improvements in both glucose and behavior, the potential impact of such compounded prediction errors on long-term reward estimation and intervention effectiveness needs to be thoroughly evaluated in future large-scale studies. Finally, our behavior adherence model is simple, and user responses are also affected by external contextual factors such as weather or daily schedules [14], and adherence levels may decrease with engagement time [42,43]. Therefore, future development should focus on robustly handling limited initial data, efficient data collection strategies, precise quantification or mitigation of propagated uncertainty, and incorporating psychological and contextual factors into sophisticated behavioral models. Crucially, improving behavioral adherence through this future development is also key to minimizing the compounding effects of prediction errors within our two-stage model.

Fourth, the behavioral parameters’ prior knowledge, derived from self-reported questionnaires (Table 2), may not perfectly reflect true user tendencies due to cognitive biases. Inaccuracies in these initial priors could affect the model’s early parameter estimation accuracy and the speed of convergence to optimal recommendations. While our multiarmed bandit naturally adapts over time through experiential learning, more robust methods for acquiring these priors, such as integrating objective behavioral data, could further enhance initial performance and reduce the burden of exploration.

Finally, while our current action candidates are tailored around carbohydrate intake and postprandial walking, expanding the scope to include other types of exercise or dietary components would enhance personalization. Also, redesigning the reward function to align more closely with clinical guidelines for diabetes would improve its clinical relevance. Incorporating advanced behavioral change techniques [44], such as gamification, could further boost user engagement and usability.

Future Directions

Beyond the immediate limitations, our study highlights several important directions for future research and practical implementation. Foremost among these, addressing the ethical considerations and practical challenges associated with long-term behavior modification interventions is crucial for successful clinical deployment. Factors such as user fatigue from continuous data collection and potential disengagement with recommendations over extended periods need careful consideration. Moreover, the current framework does not explicitly incorporate mechanisms to prevent the recommendation of potentially clinically damaging actions. While action candidates are prepared within a clinically plausible range, individual-specific contraindications or extreme physiological states could render even a seemingly "plausible" recommendation detrimental (eg, recommending a very low carbohydrate intake to someone with specific metabolic conditions). Future work will also include empirically establishing practical timelines for robust parameter convergence and policy updates, considering user engagement and burden.

Another critical area concerns the data requirements and accuracy for personalized recommendations. A significant challenge, especially for initial deployment, is the need for sufficient user-specific data to effectively learn and personalize recommendations. While our method uses prior knowledge to accelerate initial learning, acquiring this initial information (eg, via questionnaires or pretraining data) can be difficult and prone to cognitive biases, as self-reported behavioral priors (Table 2) may not perfectly reflect true user tendencies. However, recent advancements in mobile health apps, such as eMOM (Helsinki University Hospital) [8], are enabling the comprehensive collection and visualization of vast amounts of user behavioral and physiological data, including glucose levels. This digital self-tracking, by visualizing relevant data, has shown a positive impact on user understanding of their own health characteristics and offers a promising avenue for acquiring more robust and objective priors. For future clinical real-world applications, directly providing recommendations without proper context carries medical risks. Therefore, a clinician-in-the-loop approach is crucial. This involves an initial phase focused on data collection and visualization (eg, through mobile apps such as eMOM [8]), followed by a collaborative process where clinicians work with users to understand individual characteristics based on the visualized data. This systematic approach ensures the quantity and quality of prior information for each user before recommendations are made. Future work will integrate such an eMOM-like mobile app within a clinician-in-the-loop framework, enabling large-scale validation of the entire process from data visualization to personalized recommendation.

To achieve full clinical deployment, it is paramount to integrate safety constraints into the action selection process, developing methods to filter out adverse clinical outcomes (eg, through real-time physiological data integration or clinician-in-the-loop systems). Future work should proactively investigate strategies to mitigate these issues, such as dynamic adjustment of intervention intensity, gamification elements to maintain sustained engagement, and robust ethical frameworks ensuring user well-being, privacy, and safety in longitudinal deployments.

Conclusions

In this study, we proposed a method for optimizing the planning of dietary and exercise recommendations to improve postprandial glucose levels through behavioral changes. The proposed method is a multiarmed bandit based on a two-stage reward prediction model, where an action is a combination of the amount of carbohydrate intake and postprandial walking duration, and the reward is the postprandial glucose level. We specifically realized the reward prediction for each action by predicting the behavioral responses to an action and then predicting the postprandial glycemic response using the predicted responses. From the simulation experiment, it was demonstrated that action selection using the proposed method significantly increased the cumulative reward compared to the default action selection. Simultaneously, we found that the most beneficial setting of the proposed method is adopting a policy to maximize the predicted rewards from the beginning, together with robust parameter learning using prior distributions for the prediction models. Finally, based on this finding, an intervention experiment was conducted on 6 healthy participants, and it was shown that the application of the proposed method provided preliminary evidence that it could improve postprandial glucose levels along with the behaviors of carbohydrate intake and postprandial walking. This initial real-world investigation, while valuable for understanding the practical considerations and informing the design of larger trials, underscores the need for extensive long-term validation in diverse patient populations, ensuring user safety and promoting sustained engagement.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets analyzed during this study are not publicly available due to ethical considerations and the privacy of personal data but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional material on the two-stage reward prediction model and the simulation experiment setup.
[DOCX File (Microsoft Word File), 658 KB-Multimedia Appendix 1]

References

1. Bommer C, Sagalova V, Heesemann E, et al. Global economic burden of diabetes in adults: projections from 2015 to 2030. Diabetes Care. May 2018;41(5):963-970. [doi: 10.2337/dc17-1962] [Medline: 29475843]

2. Yu J, Lee SH, Kim MK. Recent updates to clinical practice guidelines for diabetes mellitus. Endocrinol Metab (Seoul). Feb 2022;37(1):26-37. [doi: 10.3803/EnM.2022.105] [Medline: 35255599]
3. Asif M. The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. J Educ Health Promot. 2014;3:1. [doi: 10.4103/2277-9531.127541] [Medline: 24741641]
4. Kytö M, Koivusalo S, Ruonala A, et al. Behavior change app for self-management of gestational diabetes: design and evaluation of desirable features. JMIR Hum Factors. Oct 12, 2022;9(4):e36987. [doi: 10.2196/36987] [Medline: 36222806]
5. Ramakrishnan P, Yan K, Balijepalli C, Druyts E. Changing face of healthcare: digital therapeutics in the management of diabetes. Curr Med Res Opin. Dec 2021;37(12):2089-2091. [doi: 10.1080/03007995.2021.1976737] [Medline: 34511002]
6. Kytö M, Markussen LT, Marttinen P, et al. Comprehensive self-tracking of blood glucose and lifestyle with a mobile application in the management of gestational diabetes: a study protocol for a randomised controlled trial (eMOM GDM study). BMJ Open. Nov 7, 2022;12(11):e066292. [doi: 10.1136/bmjopen-2022-066292] [Medline: 36344008]
7. Everett E, Kane B, Yoo A, Dobs A, Mathioudakis N. A novel approach for fully automated, personalized health coaching for adults with prediabetes: pilot clinical trial. J Med Internet Res. Feb 27, 2018;20(2):e72. [doi: 10.2196/jmir.9723] [Medline: 29487046]
8. Kytö M, Hotta S, Niinistö S, et al. Periodic mobile application (eMOM) with self-tracking of glucose and lifestyle improves treatment of diet-controlled gestational diabetes without human guidance: a randomized controlled trial. Am J Obstet Gynecol. Nov 2024;231(5):541. [doi: 10.1016/j.ajog.2024.02.303] [Medline: 38432415]
9. Debon R, Coleone JD, Bellei EA, De Marchi ACB. Mobile health applications for chronic diseases: a systematic review of features for lifestyle improvement. Diabetes Metab Syndr. 2019;13(4):2507-2512. [doi: 10.1016/j.dsx.2019.07.016] [Medline: 31405669]
10. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med. May 18, 2018;52(6):446-462. [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
11. Klasnja P, Smith S, Seewald NJ, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med. May 3, 2019;53(6):573-582. [doi: 10.1093/abm/kay067] [Medline: 30192907]
12. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv. Jan 31, 2023;55(1):1-36. [doi: 10.1145/3477600]
13. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. J Med Internet Res. Oct 10, 2017;19(10):e338. [doi: 10.2196/jmir.7994] [Medline: 29017988]
14. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 2020;4(1):18. [doi: 10.1145/3381007] [Medline: 34527853]
15. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: a systematic review. Artif Intell Med. Apr 2020;104:101836. [doi: 10.1016/j.artmed.2020.101836] [Medline: 32499004]
16. Yau KLA, Chong YW, Fan X, Wu C, Saleem Y, Lim PC. Reinforcement learning models and algorithms for diabetes management. IEEE Access. 2023;11:28391-28415. [doi: 10.1109/ACCESS.2023.3259425]
17. Oroojeni Mohammad Javad M, Agboola SO, Jethwani K, Zeid A, Kamarthi S. A reinforcement learning-based method for management of type 1 diabetes: exploratory study. JMIR Diabetes. Aug 28, 2019;4(3):e12905. [doi: 10.2196/12905] [Medline: 31464196]
18. Shepherd E, Gomersall JC, Tieu J, Han S, Crowther CA, Middleton P. Combined diet and exercise interventions for preventing gestational diabetes mellitus. Cochrane Database Syst Rev. Nov 13, 2017;2017(11):CD010443. [doi: 10.1002/14651858.CD010443.pub3]
19. Borror A, Zieff G, Battaglini C, Stoner L. The effects of postprandial exercise on glucose control in individuals with type 2 diabetes: a systematic review. Sports Med. Jun 2018;48(6):1479-1491. [doi: 10.1007/s40279-018-0864-x] [Medline: 29396781]
20. Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. Presented at: UbiComp ’15: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 7-11, 2015; Osaka, Japan. [doi: 10.1145/2750858.2805840]
21. Ameko MK, Beltzer ML, Cai L, Boukhechba M, Teachman BA, Barnes LE. Offline contextual multi-armed bandits for mobile health interventions: a case study on emotion regulation. Presented at: RecSys ’20: Proceedings of the 14th ACM Conference on Recommender Systems; Sep 22-26, 2020; Virtual Event Brazil. [doi: 10.1145/3383313.3412244]

22. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. MIT press; 2018. URL: https://mitpress.mit.edu/9780262039246/reinforcement-learning/ [Accessed 2026-03-07]

23. Pustozerov EA, Tkachuk AS, Vasukova EA, et al. Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus. IEEE Access. 2020;8:219308-219321. [doi: 10.1109/ACCESS.2020.3042483]

24. Gallwitz B. Implications of postprandial glucose and weight control in people with type 2 diabetes: understanding and implementing the International Diabetes Federation guidelines. Diabetes Care. Nov 2009;32 Suppl 2(Suppl 2):S322-5. [doi: 10.2337/dc09-S331] [Medline: 19875573]

25. Evenepoel C, Clevers E, Deroover L, Van Loo W, Matthys C, Verbeke K. Accuracy of nutrient calculations using the consumer-focused online app MyFitnessPal: validation study. J Med Internet Res. Oct 21, 2020;22(10):e18237. [doi: 10.2196/18237] [Medline: 33084583]

26. Wylie-Rosett J, Aebersold K, Conlon B, Isasi CR, Ostrovsky NW. Health effects of low-carbohydrate diets: where should new research go? Curr Diab Rep. Apr 2013;13(2):271-278. [doi: 10.1007/s11892-012-0357-5] [Medline: 23266565]

27. Reynolds AN, Mann JI, Williams S, Venn BJ. Advice to walk after meals is more effective for lowering postprandial glycaemia in type 2 diabetes mellitus than advice that does not specify timing: a randomised crossover study. Diabetologia. Dec 2016;59(12):2572-2578. [doi: 10.1007/s00125-016-4085-2] [Medline: 27747394]

28. Aqeel M, Forster A, Richards EA, et al. The effect of timing of exercise and eating on postprandial response in adults: a systematic review. Nutrients. Jan 15, 2020;12(1):221. [doi: 10.3390/nu12010221] [Medline: 31952250]

29. Andersen MB, Fuglsang J, Ostenfeld EB, Poulsen CW, Daugaard M, Ovesen PG. Postprandial interval walking-effect on blood glucose in pregnant women with gestational diabetes. Am J Obstet Gynecol MFM. Nov 2021;3(6):100440. [doi: 10.1016/j.ajogmf.2021.100440] [Medline: 34216833]

30. Gao Z, Han Y, Ren Z, Zhou Z. Batched multi-armed bandits problem. Presented at: Proceedings of the 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019:503; Vancouver, BC, Canada. URL: https://dl.acm.org/doi/10.5555/3454287.3454333 [Accessed 2026-03-07]

31. Wolever TM, Bolognesi C. Source and amount of carbohydrate affect postprandial glucose and insulin in normal subjects. J Nutr. Nov 1996;126(11):2798-2806. [doi: 10.1093/jn/126.11.2798] [Medline: 8914951]

32. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process. Dec 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

33. Ashrafi RA, Ahola AJ, Rosengård-Bärlund M, et al. Computational modelling of self-reported dietary carbohydrate intake on glucose concentrations in patients undergoing Roux-en-Y gastric bypass versus one-anastomosis gastric bypass. Ann Med. Dec 2021;53(1):1885-1895. [doi: 10.1080/07853890.2021.1964035] [Medline: 34714211]

34. Richter EA, Derave W, Wojtaszewski JF. Glucose, exercise and insulin: emerging concepts. J Physiol. Sep 1, 2001;535(Pt 2):313-322. [doi: 10.1111/j.1469-7793.2001.t01-2-00313.x] [Medline: 11533125]

35. Jankovic MV, Mosimann S, Bally L, Stettler C, Mougiakakou S. Deep prediction model: the case of online adaptive prediction of subcutaneous glucose. Presented at: 2016 13th Symposium on Neural Networks and Applications (NEUREL); Nov 22-24, 2016; Belgrade, Serbia. [doi: 10.1109/NEUREL.2016.7800095]

36. Xie J, Wang Q. A data-driven personalized model of glucose dynamics taking account of the effects of physical activity for type 1 diabetes: an in silico study. J Biomech Eng. Jan 1, 2019;141(1). [doi: 10.1115/1.4041522] [Medline: 30458503]

37. Hotta S, Kytö M, Koivusalo S, Heinonen S, Marttinen P. Optimizing postprandial glucose prediction through integration of diet and exercise: leveraging transfer learning with imbalanced patient data. PLoS ONE. 2024;19(8):e0298506. [doi: 10.1371/journal.pone.0298506] [Medline: 39088422]

38. Zhang G, Ashrafi RA, Juuti A, Pietilainen K, Marttinen P. Errors-in-variables modeling of personalized treatment-response trajectories. IEEE J Biomed Health Inform. Jan 2021;25(1):201-208. [doi: 10.1109/JBHI.2020.2987323] [Medline: 32324579]

39. Le Floch JP, Escuyer P, Baudin E, Baudon D, Perlemuter L. Blood glucose area under the curve. Methodological aspects. Diabetes Care. Feb 1990;13(2):172-175. [doi: 10.2337/diacare.13.2.172] [Medline: 2351014]

40. RStan: the r interface to stan. R package version 2218. 2023. URL: https://mc-stan.org [Accessed 2026-02-20]

41. Oh R, Gilani B, Uppaluri KR. Low-carbohydrate diet. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023. [Medline: 30725769]

42. Gershkowitz BD, Hillert CJ, Crotty BH. Digital coaching strategies to facilitate behavioral change in type 2 diabetes: a systematic review. J Clin Endocrinol Metab. Mar 25, 2021;106(4):e1513-e1520. [doi: 10.1210/clinem/dgaa850] [Medline: 33206975]

43. Van Rhoon L, Byrne M, Morrissey E, Murphy J, McSharry J. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. Digit Health. 2020;6:2055207620914427. [doi: 10.1177/2055207620914427] [Medline: 32269830]

44. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. mHealth apps using behavior change techniques to self-report data: systematic review. JMIR Mhealth Uhealth. Sep 9, 2022;10(9):e33247. [doi: 10.2196/33247] [Medline: 36083606]

Abbreviations

iAUC: incremental area under the curve

Edited by Javad Sarvestan; peer-reviewed by Elad Yom-Tov, Vijay A Kamble; submitted 14.Jan.2025; final revised version received 24.Jan.2026; accepted 28.Jan.2026; published 19.Mar.2026

Please cite as:
Hotta S, Kytö M, Koivusalo S, Heinonen S, Marttinen P
Personalized Glucose Management With AI: Pilot Study Using a Multiarmed Bandit Approach
JMIR Form Res 2026;10:e70826
URL: https://formative.jmir.org/2026/1/e70826
doi: 10.2196/70826

© Shinji Hotta, Mikko Kytö, Saila Koivusalo, Seppo Heinonen, Pekka Marttinen. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.