Patterns

Personalizing self-management via behavioral predictive analytics with health education for improved self-efficacy

Highlights

- Assess behavior readiness via a proven psychology model: theory of planned behavior
- Segment patients into groups by behavior readiness via advanced manifold clustering
- Enable a closed loop via mobile app to support prediction for dynamic personalization
- Deliver asynchronous health education with demonstrated self-efficacy improvement

Authors

Bon Sy, Michael Wassil, Alisha Hassan, Jin Chen

Correspondence

bon.sy@qc.cuny.edu

In brief

Previous studies concluded that only 25% of patients are actively engaged in self-health management. In comparison to the engagement rate prior to the intervention, this study shows that, on average, the engagement rate among the pilot participants was improved by 20% when behavioral predictive analytics were applied to personalize actionable health activities. In addition, the improvement of diabetes self-efficacy was found statistically significant in the asynchronous setting of delivering health education through the SIPPA Health mobile app.

Sy et al., 2022, Patterns *3*, 100510
June 10, 2022
https://doi.org/10.1016/j.patter.2022.100510

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Personalizing self-management via behavioral predictive analytics with health education for improved self-efficacy

Bon Sy,[1,2,3,5,*] Michael Wassil,[3] Alisha Hassan,[4] and Jin Chen[3]
[1]Graduate Center/City University of NY, 365 5th Avenue, NY 10016, USA
[2]Queens College/City University of NY, 65-30 Kissena Boulevard, Queens, NY 11367, USA
[3]SIPPA Solutions, 42-06A Bell Boulevard, Queens, NY 11361, USA
[4]Hunter College/City University of NY, School of Public Health, New York, NY 10065, USA
[5]Lead contact
*Correspondence: bon.sy@qc.cuny.edu
https://doi.org/10.1016/j.patter.2022.100510

THE BIGGER PICTURE Type 2 and pre-diabetes is a chronic disease that affects over 115 million Americans and over 440 million people worldwide. Active patient self-management improves health outcome and lowers healthcare cost. Yet, less than 25% of patients are engaged in active self-health management. Behavioral predictive analytics was developed to improve patient engagement. It applies an advanced clustering technique in machine learning to segment patients into subpopulations by behavior readiness. It dynamically personalizes actionable health activities such as self-monitoring of glucose as well as health education based on one's behavior readiness. This paper reports (1) the practical feasibility of an engagement channel through an individual's mobile device to deliver health education for improving diabetes self-efficacy and (2) the validated outcomes of the behavioral predictive analytics to improve engagement in self-health management.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

The objective of this research is to investigate the feasibility of applying behavioral predictive analytics to optimize diabetes self-management. This research also presents a use case on the application of the anaytics technology platform to deliver an online diabetes prevention program developed by the CDC. The goal of personalized self-management is to affect individuals on behavior change toward actionable health activities on glucose self-monitoring, diet management, and exercise. In conjunction with personalizing self-management, the content of the CDC diabetes prevention program was delivered online directly to a mobile device. The proposed behavioral predictive analytics relies on manifold clustering to identify subpopulations by behavior readiness characteristics exhibiting non-linear properties. Utilizing behavior readiness data of 148 subjects, subpopulations are created using manifold clustering to target personalized actionable health activities. This paper reports the preliminary result of personalizing self-management for 22 subjects under different scenarios and the outcome on improving diabetes self-efficacy of 34 subjects.

INTRODUCTION

Type 2 and pre-diabetes is a chronic disease that affects over 115 million Americans and over 440 million people worldwide. Some of the risk factors are mitigatable through health education and behavior change toward a healthy lifestyle.

Health education programs, such as the in-person, year-long, Diabetes Prevention Program (DPP) of Centers for Disease Control and Prevention (CDC) in the US has shown results impacting patients from all walks of life; e.g., 6% among DPP participants compared with 11% in the placebo group developing diabetes across different gender, racial, and ethnicity groups.[1] It has also

been demonstrated elsewhere[2] that behavior change can achieve a 10% or more improvement in diabetes symptoms if an individual is engaged in pro-active self-management of diabetes.

Self-management is generally accepted as a viable intervention strategy.[3] Self-management is the patient's ability to manage their chronic disease through their own activities, such as taking their blood glucose and focusing on meeting diet and activity goals. However, we do not fully understand the relationship between the behavior readiness of an individual and the specific intervention strategy that could deliver optimal patient engagement in self-management activities. As demonstrated in a survey conducted elsewhere,[4] less than 25% of patients are considered as actively engaged in self-health management. Population health management will not be cost effective if self-management programs do not consider the readiness of the patient population. A contribution of this research is to provide an insight into (1) the technical feasibility of behavioral predictive analytics built upon the outcome of manifold clustering, and (2) the efficacy of delivering DPP online via the SIPPA Health platform in 3 months as opposed to the traditional in-person format over a 12-month period.

Our main goal is to optimize the effectiveness of self-management strategies by means of personalization based on predicting behavior readiness and its relationship to engagement outcomes. A second goal is to determine the feasibility and efficacy of delivering DPP health education online over a 3-month period. In this study, we aim to demonstrate a potential predictive system that delivers personalized content to the users based on their behavior readiness and user profile.

Relationship to state-of-the-art contains a brief review on the state-of-the-art, and the context of this research within it. We first discuss various theory-based behavior models including the theory of planned behavior (TPB), and their use in health applications. We put in context our application of TPB to model behavior readiness. We also briefly discuss DPP, and then the state-of-the-art on clustering techniques. In Predictive analytics foundation the research results reported elsewhere are restated as it is applied in this research. For completeness, the algorithmic steps for entropy-based discretization and manifold clustering are presented. In Predictive analytics for personalization we discuss predictive analytics for personalization using either an auto-regression model or a population-based model. The population-based model provides an alternative mechanism when the auto-regression model derivation fails. This could occur when there is insufficient data, or if it fails the statistics test of the model selection process based on Bayesian information criteria (BIC)/Akaike information criteria (AIC). In Personalized online health education we discuss the CDC DPP program and a validated assessment tool for diabetes health education—Diabetes Self-Efficacy Questionnaire (DSEQ) developed elsewhere.[5] In Preliminary study we present the results of manifold clustering based on the attribute vector of behavior readiness of 148 subjects with type 2 diabetes. This is followed by the results of a preliminary study involving 22 subjects who were in the intervention phase for personalization during the study period. An online delivery mechanism of DPP via the platform used in this research is also described. In Health education assessment using DSEQ we report the preliminary results of 34 participants receiving DPP online. This is followed by our final thoughts and future research in Final thoughts and open research questions. We then summarize this research in the Conclusions.

Relationship to state-of-the-art

In health psychology, behavior models have been developed and applied to address healthcare issues in different settings. As summarized in an article by Linden et al.[6] there are a number of theory-based behavior models—natural helper model, diffusion of innovations model, theories of organizational change, community coalition action theory, social marketing model, precede-proceed model, motivational interviewing, stages of change model, social learning interpersonal theory, consumer information processing model, implementation individual intentions models, and health belief model. Models, such as the theory of organizational change model, target disease management programs at the community level, and focus on the planning and the implementation of population-based interventions that influence social norms and structures.

On the other hand, the TPB model,[7] transtheoretical model of behavior change,[8] health belief model,[9] and IMB (information motivation and behavior skill) model[10] have been applied to interventions of chronic diseases, and have shown clinical efficacy. It was suggested that individuals perceiving the risk of a condition are more likely to engage in behavior to reduce risk. Thus perceived health risks, resulting in the change of attitude and behavior, are proponents for higher intentions to be physically active and to maintain a healthy diet.

The TPB provides a model to manifest the relationship among attitude, subjective norm, perceived behavioral control, intention, and behavior. TPB is modeled through expectancy value, and assumes that the best single predictor of an individual's behavior is an intention to perform that behavior. The intention in turn depends on the attitude of an individual (positive or negative evaluation of performing a behavior); the subjective norm (perception of whether relevant others think one should or should not perform the behavior); and perceived behavioral control (perception of the ease or difficulty of carrying out a behavior). These all work together or in opposition to fuel behavioral attitudes and beliefs in subjective norms, based on the importance the individual places on these attitudes and norms. This then decides one's intentions, which lead to the behaviors in question.[11]

In line with this theory, two additional behavior constructs—motivation and ownership, as advocated in the IMB—were incorporated into our behavior model. This allows us to target a user's behavioral beliefs to change their attitudes and intentions toward actionable health behaviors. One of the most important features of our approach is the use of frequent reminders to track health activities that reveal information about health behaviors.

In a review of the literature, Fry and Neff[12] found that frequent periodic prompts around improving diet, increasing physical activity, and weight loss all led to positive results for study participants. Tailored prompts were found statistically significant in encouraging user engagement. However, for users who are already not engaged, these prompts do little to engage users.[13] Sawesi et al.[14] found, in a systematic review of the literature, that digital methods such as text messages, web applications, and social media interventions all were good intervention tools. These tools can support behavioral change in users and usually improve patient engagement. Finally, the use of mobile health

interventions has been found to be an engaging method for improving health behaviors and is cost effective for the behavioral change.[15] This is particularly the case regarding the potential of mobile technology for delivering online health education content[1] on diabetes prevention, such as the DPP by CDC. When online health education via mobile technology can deliver similar efficacy, this reduces not only the operational cost of DPP, but the cost for patients in terms of the transportation and commuting time to a DPP.

On the technical side, this research intends to contribute to a better understanding of our manifold clustering approach that is applied to segmenting the diabetes population of this pilot study based on behavior readiness. Many researchers have proposed clustering algorithms to address the issue of linearity, but each comes with (dis)advantages.

k-Means[16] is one of the most popular algorithms due to its $O(n^2)$ complexity. The algorithm consists of selecting the first k data points to be the centers of k clusters and finding the minimum arithmetic mean between each data point and the k clusters. However, k-means breaks down in higher dimensions. Zhang and Kwok[17] suggested using an applied Nyström method to approximate the Eigen decomposition with low-rank kernel matrices. Alternatively, Wang et al.[18] suggested using local adaptive learning to perform graph embedding and k-means simultaneously, thereby reducing dimensionality. Both algorithms decrease the run-time of typical clustering methods but do not address the information lost in the dimension reduction process. Our clustering approach examines the effect of dimension reduction on information loss from an information-theoretic perspective, as well as from a reconstruction error perspective during the projection of a data point to a hyperplane of a cluster.

Recent clustering research focuses on minimizing dimensionality without losing meaning in the data. Ge et al.[19] suggested a geometrically local embedding (GLE) process that reduces dimensionality by assigning clusters according to geometric distance in the higher dimension. After finding optimal reconstruction weights, the algorithm filters for outliers, and the manifold is mapped to a lower dimension. Although GLE is effective, the procedure is computationally expensive; thus challenging for practical applications. Gong et al.[20] proposed using a structured sparse k-means algorithm to reduce the randomness of clusters. In doing so, they used Laplacian smoothing to exploit the correlation information among features, thereby improving clustering accuracy and retaining meaning. Faivishevsky and Goldberger[21] took a different approach by combining spectral clustering with a nonparametric information-theoretic clustering algorithm to retain information via mutual information measure. Their algorithm assumes that the conditional density of each cluster follows a Gaussian distribution. Our approach differs from that of Faivishevsky and Goldberger in that our approach does not assume Gaussian distribution, but rather an asymptotic convergence of mutual information measure toward chi-square. This was proven by Kullback[22] for the low dimension, and was extended to high dimension.[23]

Predictive analytics foundation

SIPPA (Secure Information Processing with Privacy Assurance) predictive analytics relies on two foundational building blocks developed in the research reported elsewhere.[24,25] The workflow process for the application of the proposed predictive analytics consists of three stages. In stage 1, an individual responds to a survey instrument linked to a behavior model for measuring readiness. In stage 2, the outcome measure of the behavior readiness determines the cluster/subpopulation that the individual is assigned to. The assignment is based on the similarity between the individual's behavior pattern and the statistically significant association patterns that characterize the cluster/subpopulation. In stage 3, the population-based model and individualized week-over-week engagement models are applied to predict personalized weekly activities that optimize the success rate of engagement in self-health management. The theoretical framework for manifold clustering that enables stage 2 and the details on stage 3 are presented in the next section.

The first building block of SIPPA predictive analytics is a behavior model to enable behavior readiness prediction. Behavior readiness is a 1×4 vector of continuous (real) numbers quantifying [*ownership*, *motivation*, *intention*, *attitudes*]. These behavior attributes of real are constructs of behavior modeling grounded on the TPB. Structural equation modeling[26] was employed to link questions of a survey instrument to the behavior constructs defined by weighing factors derived from the confirmatory factor analysis. The behavior model linking to the survey questions was statistically validated based on the responses from over 500 participants.[24]

The second building block is an unsupervised learning approach for discovering manifold clusters without the assumption of linearity. While the behavior constructs are related according to the TPB, variations exist as shown in the confirmative factor analysis regarding the assumption of linearity; i.e., the existence (and strength) of a linear relationship between the behavior constructs that quantifies behavior readiness for self-management in a population.

The concept of manifold clustering is to induce patient subpopulation clusters based on statistically significant association patterns on behavior readiness. This approach is not restricted to only continuous data (number of real). In other words, this approach could be applied to a dataset of mixed-type of both continuous and discrete variables.

Significant behavior patterns

A behavior pattern, which is manifested by the instantiation of finite discrete variables, is statistically significant if it survives two tests: (1) a support measure—as defined by normalized frequency occurrence, which exceeds a pre-defined threshold, and (2) the association among the observed values is not by chance as measured by the mutual information measure. The following shows the technical formulation of statistically significant association patterns:

Let $\boldsymbol{X} = \left\{ \boldsymbol{X}_i \middle| \boldsymbol{X}_i = \left[X_i^1 \quad , \ldots, \quad X_i^n \right]^T \in \mathbb{R}^n \text{ for } i = 1, \ldots, N \right\}$ be a dataset of real.

Let $Y = \left\{ Y_i \middle| Y_i = \left[Y_i^1 \quad , \ldots, \quad Y_i^n \right]^T \in \mathbb{Z}^n \text{ for } Y_i^j = 0, \ldots, K-1 < N;\ i = 1, \ldots, K \leq N \right\}$ be a dataset of discrete non-negative integers.

Let $\boldsymbol{M} = \{ M_k | M_k \subseteq \boldsymbol{X} \text{ for } k = 1, \ldots, |\boldsymbol{M}| \}$ be the set of $|\boldsymbol{M}|$ manifold clusters.

Let $F : X^q \rightarrow Y^q$ (for $q = 1, \ldots, n$) be a mapping function that defines the discretization of the multivariate dataset $\boldsymbol{X}$.

Let $S(M_k) = \{P_j^{k,o} | M_k, P_j^{k,o} = (val_{j,1}^{k,o}, ..., val_{j,o}^{k,o})$ for $j = 1, ..., |S(M_k)|\}$ $P_j^{k,o}$ is an o^{th} $(2 \leq o \leq n)$-order statistically significant association pattern[23] when $Pr(val_{j,1}^{k,o}, ..., val_{j,o}^{k,o}) > \alpha$ for some pre-defined threshold α, and $MI(val_{j,1}^{k,o}, ..., val_{j,o}^{k,o}) \rightarrow$ adjusted χ^2 as defined below:

$$MI\left(val_{j,1}^{k,o}, ..., val_{j,o}^{k,o}\right) \rightarrow \left(\frac{1}{Pr\left(val_{j,1}^{k,o}\ val_{j,2}^{k,o}\ ...\ val_{j,o}^{k,o}\right)}\right)\left(\frac{\chi^2}{2N}\right)\left(\frac{\widehat{E}}{E'}\right)^{o/2} \quad \text{(Equation 1)}$$

where $MI(val_{j,1}^{k,o}, ..., val_{j,o}^{k,o}) = Log_2 \frac{Pr(val_{j,1}^{k,o}, ..., val_{j,o}^{k,o})}{Pr(val_{j,1}^{k,o}) Pr(val_{j,2}^{k,o}) ... Pr(val_{j,o}^{k,o})}$, N is the sample size, χ^2 is Pearson chi-square defined as $(o_i - e_i)^2 / e_i$ with o_i as the observed count of $P_j^{k,o}$ and e_i the expected count under the assumption on independence.

$\widehat{E}$ is the expected entropy measure and E' is the maximum possible entropy.

Insights on significant patterns. Recall $S(M_k)$ represents a set of statistically significant association patterns that characterize the k^{th} cluster M_k. To illustrate using the example below:

Let $\boldsymbol{Y} = \{ [d1:0, d2:0, d3:0, d4:0],$

$$[d1:0, d2:0, d3:0, d4:1], \ldots$$

$$\ldots, [d1:1, d2:1, d3:1, d4:1]\}$$

$|\boldsymbol{Y}| = 16$

There are $4xC(4,2) + 8xC(4,3) + 16xC(4,4) = 72$ patterns.

Let's assume $S(M_2) = \{[d1{:}0, d3{:}1], [d2{:}1, d4{:}0], [d2{:}0, d3{:}1, d4{:}0], [d1{:}0, d2{:}0, d3{:}1, d4{:}0]\}$.

There are two second-order patterns, one third-order pattern, and one fourth-order pattern in $S(M_2)$ as shown below:

$$P_1^{2,2} = \left(val_{1,1}^{2,2}, val_{1,2}^{2,2}\right) = (\text{d1:0, d3:1})$$

$$P_2^{2,2} = \left(val_{2,1}^{2,2}, val_{2,2}^{2,2}\right) = (\text{d2:1, d4:0})$$

$$P_3^{2,3} = \left(val_{3,1}^{2,3}, val_{3,2}^{2,3}, val_{3,3}^{2,3}\right) = (\text{d2:0, d3:1, d4:0})$$

$$P_4^{2,4} = \left(val_{4,1}^{2,4}, val_{4,2}^{2,4}, val_{4,3}^{2,4}, val_{j,4}^{2,4}\right) = (\text{d1:0, d2:0, d3:1, d4:0})$$

Note that $MI(\bullet)$ is the mutual information measure. In brief, mutual information measure examines in a more granular level the "independence" property on the event level, and asymptotically converges toward χ^2 (proven by Kullback as mentioned in Relationship to state-of-the-art) adjusted for high order. In contrast to standard correlation coefficient analysis that examines whether two variables are independent of each other, mutual information measure could discover inter-dependency among multiple variables on event level, while such inter-dependency may be missed by techniques such as correlation analysis on the variable level.

Although we focus on only the independence test, it is noteworthy to point out a drawback of mutual information measure; i.e., its value is unbounded, making the interpretation on the strength of inter-dependency less clearer compared with, say, correlation coefficient analysis, which is bounded between −1 and 1.

Entropy-based discretization

Consider a discrete variable Y of N possible states, the Shannon entropy of a system defined by Y:

$$H_N(P_1 \ldots P_N) = \sum_{i=1}^{N} -Pr\,(Y = y_i) Log_2 Pr(Y = y_i) = \sum_{i=1}^{N} -P_i Log_2 P_i \quad \text{(Equation 2)}$$

It can be shown that the following equality holds:[23]

$$H_N(P_1 \ldots P_N) = H_{N-1}(P_1 + P_2, P_3 \ldots P_N) + (P_1 + P_2) H_2\left(\frac{P_1}{P_1 + P_2}, \frac{P_2}{P_1 + P_2}\right) \quad \text{(Equation 3)}$$

In the quantization process, combining two terms will reduce the number of terms by one, while resulting in an information loss amounting to the second term on the right-hand side of Equation 3. In other words, information loss is monotonic. The quantization of a dataset of real will utilize the above entropy equation to incrementally combine terms until it reaches the inflection point where there is a change of direction in the rate of change of information loss. The details of the algorithm are shown below:

Step 1: order X_i^j in ascending order. Create a bin for each term X^j. Treat each bin as a state of a discrete variable of $\boldsymbol{Y}$ and associate a value for a bin equal to the mean of its term(s). In other words, Y^j is a discrete variable of N states. If the values of X_i^j are all different, the initial distribution of Y^j is then even and the probability of Y_i^j is equal to $1/N$.

Step 2: initialize an iteration count $C = 1$. Derive the entropy $H_N(P_1 \ldots P_N)$ and record it as H_N^C.

Step 3: increment the iteration count by 1. Identify two adjacent bins, l and $l + 1$ in the ordered list where the difference between the mean of the terms in the l^{th} and $(l + 1)^{th}$ bins is the smallest. Merge the two adjacent l^{th} and $(l + 1)^{th}$ bins via arithmetic mean and update the probability distribution of Y^j. Re-derive the entropy H_{N-1}^{C+1}. Record the information loss I^{C+1} (i.e., the second term in Equation 3) from combining the terms in two bins.

Step 4: repeat step 3 until it reaches a pre-defined number of iterations, or the direction in the rate of change of I^k is changed. When this occurs, the following result is obtained:

$$Y = \left\{Y_i | Y_i = \begin{bmatrix} Y_i^1 & \cdots & Y_i^n \end{bmatrix}^T \in \mathbb{Z}^n \text{ for } Y_i^j = 0, \ldots, N - k - 1 < N;\ i = 1, \ldots, N - k \leq N\right\}$$

The mapping function mentioned before $F: X^j \rightarrow Y^j$ can then be defined for discretizing $\boldsymbol{X}$ to $\boldsymbol{Y}$.

Manifold clustering

Two important results of the manifold clustering technique previously reported elsewhere[25] are recited for completeness. First, each manifold cluster has a semantic interpretation characterized by statistically significant association patterns; i.e., grouping according to behavior readiness in this application. Second, the manifold clustering does not require linearity assumption as in principal-component analysis (PCA). But it will produce the same result as PCA if the linearity assumption holds, and the iteration is based on minimizing reconstruction errors; i.e., "phase 2" regrouping is skipped in the manifold clustering. Below, we describe the algorithmic steps of the manifold clustering:

Given $\boldsymbol{X}$, $\boldsymbol{Y}$, F, and a pre-defined error threshold δ, the algorithm for the manifold clustering based on statistically significant association patterns is shown below:

Step 1: based on $\boldsymbol{Y}$, derive the set of statistically significant association patterns $S(M_k)$.

Step 2: define $|\boldsymbol{M}|$ disjoint clusters such that initially each cluster has one and only one statistically significant association pattern (i.e., $|S(M_k)| = 1$ for $k = 1..\ |\boldsymbol{M}|$). Let W be the set of cluster reference "*holding*" the data points in $\boldsymbol{X}$; i.e., $W = \{\boldsymbol{X}^{n,j} | \boldsymbol{X} = \cup_j \boldsymbol{X}^{n,j} \text{ for } j = 1, \ldots, |M|\}$ In other words, $\boldsymbol{X}^{n,j}$ is a set reference to the data points of $\mathbb{R}^n$ in the cluster M_j, while P'') is the set of statistically significant association pattern(s) defining the cluster M_j.

Step 3: partition $\boldsymbol{X}$ by assigning each data point X_i to $\boldsymbol{X}^{n,k}$ if $ArgMax_{q,k}\, f(F(X_i), P_q^{k,o}) = k$; where $P_q^{k,o}$ is a pattern that defines the cluster M_k, thus $\boldsymbol{X}^{n,k}$. If $f(F(X_i), P_q^{k,o})$ is zero in all cases, X_i is assigned to a non-semantic cluster NS; where $f(F(X_i), P_j^{k,o}) \rightarrow [0,1]$ is a set membership function defined by the geometric mean measure below:

$$f\left(F(X_i), P_j^{k,o}\right) = \sqrt{\frac{\left|SC(F(X_i)) \cap SC\left(P_j^{k,o}\right)\right|}{|SC(F(X_i))|} \times \frac{\left|SC(F(X_i)) \cap SC\left(P_j^{k,o}\right)\right|}{\left|SC\left(P_j^{k,o}\right)\right|}}$$

The scope coverage $SC(P_j^{k,o})$, with respect to a set $\boldsymbol{Y}$, is defined as a subset of $\boldsymbol{Y}$ in which the semantic interpretation of the existence of $P_j^{k,o}$ is always true.

Example

Let $\boldsymbol{Y} = \{[d1:0, d2:0, d3:0, d4:0], [d1:0, d2:0, d3:0, d4:1], \ldots\ldots, [d1:1, d2:1, d3:1, d4:1]\}$

$|\boldsymbol{Y}|=16$.

Let $P_j^{k,o} = [d1:1, d3:0]$.

$$SC\left(P_j^{k,o}\right) = \{[d1:1, d2:0, d3:0, d4:0], [d1:1, d2:0, d3:0, d4:1], [d1:1, d2:1, d3:0, d4:0], [d1:1, d2:1, d3:0, d4:1]\}$$

Step 4: let $S = \{S_j \mid j = 1, \ldots, |\boldsymbol{M}|\}$ be the set of manifold subspaces corresponding to the clusters defined in step 2. Repeat the following for each j where the corresponding cluster has more than one element:

Let $\boldsymbol{D}^{n,j} = \{d_k^{n,j} | k = 1, \ldots, |\boldsymbol{X}^{n,j}|\}$ be the dataset of the cluster $\boldsymbol{X}^{n,j}$. The manifold subspace S_j corresponding to $\boldsymbol{X}^{n,j}$ is then derived based on the following:

Step 4.1: derive the mean vector $\mu^{n,j}$ and co-variance matrix $\boldsymbol{A}^{n,j}$ of $\boldsymbol{D}^{n,j}$ for each $j = 1, \ldots, |\boldsymbol{M}|$ i.e., $\boldsymbol{A}^{n,j} = \frac{1}{|\boldsymbol{D}^{n,j}|} \sum_{k=1}^{\boldsymbol{D}^{n,j}} (d_k^{n,j} - \mu^{n,j})(d_k^{n,j} - \mu^{n,j})^T$, where $\mu^{n,j} = \frac{1}{|\boldsymbol{D}^{n,j}|} \sum_{k=1}^{\boldsymbol{D}^{n,j}} (d_k^{n,j})$.

Step 4.2: conduct eigen decomposition on $\boldsymbol{A}^{n,j}$ to obtain the eigenvector matrix $\boldsymbol{Q}^{n,j}$ and the diagonal matrix of eigenvalue values $\wedge^{n,j}$ such that $\boldsymbol{A}^{n,j} = (\boldsymbol{Q}^{n,j})^{n,j} \wedge (\boldsymbol{Q}^{n,j})^{-1}$.

Step 4.3: let $P''(\leq n)$ be the number of non-zero eigenvalues obtained in step 4.2. Sort the P'' eigenvalues and define a cut-point based on some pre-defined criteria to split the corresponding eigenvectors into P'' leading and $n - P'$ remaining (zero and non-zero) eigenvectors.

Step 4.4: use the eigenvectors in $\boldsymbol{Q}^{n,j}$ that correspond to P'' leading eigenvalues in the sorted array to define the local coordinate frame for the subspace S_j, and rewrite $\boldsymbol{Q}^{n,j} = \left[W^{P',j}\ W^{n-P',j}\right]$

Step 4.5: the projection error of mapping a data point $d_k^{n,j}$ to the subspace S_j defined by the local coordinate frame is $e = (W^{n-P',j})^T (d_k^{n,j} - \mu^{n,j})$; where $W^{n-P',j}$ is an n by (n-P') matrix. Or the square-magnitude projection error of $d_k^{n,j}$ to the subspace S_j is then equal to $Err(d_k^{n,j}, S_j) = (d_k^{n,j} - \mu^{n,j})^T (W^{n-P',j})(W^{n-P',j})^T (d_k^{n,j} - \mu^{n,j})$.

Step 4.6: calculate the total error: $\sum_j \sum_k Err(d_k^{n,j}, S_j)$.

Step 4.7: repeat steps 4.4 and 4.5 with a new P'' (leading eigenvectors) that is one less; i.e., $P' - 1$. Record the total error.

Step 4.8: compute the total reconstruction error ratio of two successive rounds in step 4.6; i.e., (total reconstruction error using P'-q-*1* leading eigenvectors)/(total reconstruction error using P'-q leading eigenvector) where $q = 0, \ldots, P'$-*2*.

Step 4.9: finalize the local coordinate frame for the subspace S_j with a dimension P'-q when the error ratio in step 4.8 is the largest for the given q.

Step 5: merge two or more clusters that do not involve NS. If there are clusters with only one data point, these clusters will take the priority; then repeat step 4. Retain the solution with a lower total error.

Step 6: repeat step 5 until the total error is below the pre-defined error threshold δ, or the algorithm reaches the maximum number of iterations allowed.

It is noteworthy that step 5 of the manifold clustering algorithm above may result in a merged cluster characterized by possible multiple statistically significant association patterns; i.e., $|\boldsymbol{M}| \leq \sum_j |S(M_j)|$ in step 2 as iterations progress. The meaning of a data point will be its closeness to association patterns in a merged cluster in high dimension in terms of the semantic interpretation defined by the scope coverage and the membership function.

Furthermore, this manifold clustering technique is a two-phase optimization on grouping data. In phase 1, it groups data according to similarity to statistically significant association patterns that define the clusters. This is similar to using metrics, such as silhouette, to optimize clustering data in the same (1 × 4) dimensional space. However, in phase 2 it tries to find the most compact embedded subspace for a cluster according to the projection error and the reconstruction error when reducing the dimension of a cluster to an embedded lower dimension.

Predictive analytics for personalization

The *behavior* goal of personalization for self-management is to target specific user-directed activities that will be communicated to a user through a mobile app, and to inform "fulfilment" through feedback from the app. For example, when a personalized recommendation is to walk 10,000 steps a day, one would like to know whether a user follows through after the user received the recommendation from the mobile app. Two specific metrics are defined for this research to gain insights into the effectiveness of personalization:

Compliance ratio: over a period of time, compliance ratio is the ratio of the number of times a proposed health-related activity (i.e., actionable health) was acted on over the recommended/expected number of the related activity given the clinical condition/disease state of an individual.

Example: over a period of 30 days, a diabetes user is encouraged to self-monitor their glucose once a day under the clinical recommendation in commensurate to the user's diabetic condition. The expected number of self-monitoring measurements is 30. Over this period, the user self-monitors 18 times. Therefore, the compliance ratio is 0.6.

Engagement ratio (ER): over a given period, engagement ratio is defined as the total number of user interactions to the messages over the total number of messages sent. These messages are health tips or reminders for health actions, and are sent through text messaging, push notification, or as an in-app message.

Example: over a period of 30 days, three messages are sent daily: one healthy tip, one reminder to self-monitor, and one reminder on exercise. The total number of messages sent is 90. A diabetes user responds to half of the healthy tips (i.e., 15 out of 30), and 1/5 of the reminders on self-monitoring, and 1/3 of the reminders on exercise. The ER is (15 + 6 + 10)/90 = 31/90.

Prediction based on auto-regression and maximum likelihood

To facilitate the discussion on predictive analytics for personalization, let P be a population consisting of n individuals; i.e., $|P| = n$. $C = \{C1, \ldots Ck\}$ is the set of subpopulations obtained by applying manifold clustering described in Predictive analytics foundation to P; where $Ci \subseteq P$, $Ci \cap Cj = \phi$ if $i \neq j$, and $P = \cup_i Ci$. p^j_{Ci} is the j^{th} individual in the subpopulation cluster Ci. Recall each manifold cluster Ci is characterized by one or more statistically significant association patterns of behavior readiness attribute vector(s). For each p^j_{Ci} individual, there exists a set of engagement/compliance ratios over some period of time T. Let us denote the set of engagement ratios be $\{ER^1, \ldots, ER^T\}$. T could be different from one individual to another due to the rolling basis of the enrollment into the pilot. For example, one individual who just starts self-management may have $T = 2$ weekly engagement/compliance ratios, while another in the same subpopulation may have $T = 6$ weekly engagement/compliance ratios. Yet they both belong to the same subpopulation because of their behavior readiness.

This proposed predictive analytics is based on a two-pronged approach. First, individualized auto-regression will be applied for personalization when there is "*sufficient*" data on the engagement (compliance) ratio on a type of messages related to self-management; e.g., healthy diet. Second, a population-based model prediction for personalization will be applied when an individual does not (yet) have *sufficient* data on the engagement (compliance) ratio, or the individualized auto-regression model derivation fails on statistic validation. There is *sufficient* data for generating an individualized auto-regression model when $T \geq l$ for l being the order of the auto-regression model as discovered through model selection criteria, such as AIC or BIC, that pass statistical tests.

Information-theoretic model selection approach

BIC and AIC are two common information-theoretic approaches for model selection as stated below:

BIC: $BIC(l) =$

$$\ln(SSR(1)/T) + [(1+1)\ln(T)]/T. \quad \text{(Equation 4)}$$

AIC: $AIC(l) =$

$$\ln(SSR(1)/T) + 2/T \quad \text{(Equation 5)}$$

where l is the number of lags, T is the total number of observations, $SSR(l)$ is the sum of squared residual calculated from the difference between the estimated value derived from l^{th}-order auto-regression and the actual one.

Objective: choose l that minimizes BIC/AIC and $p < 0.05$, and R^2 correlation is "large."

Predictive analytics for personalization

Stage 1: the behavior readiness (a 1 × 4 vector of real [*ownership*, *motivation*, *intention*, *attitude*]) of each individual in a population is derived based on the user's response to a survey instrument.

Stage 2: the population is partitioned into subpopulations based on the result of manifold clustering; where each cluster is a subpopulation. In other words, the 1 × 4 behavior readiness vectors of real characterizing individuals in the population are the dataset for the manifold clustering technique described in Predictive analytics foundation.

Stage 3: repeat the following for each possible self-management activity (e.g., self-monitoring, exercise, diet management):

For each subpopulation Ci, derive the statistical (joint) distribution of ER and ΔER based on the available engagement ratios of all individuals (p^j_{Ci}) in the subpopulation; for $j = 1, 2, \ldots |Ci|$. In other words, the joint distribution characterized by $Pr(ER,$

ΔER) is derived from using the ER^t and ΔER^{t+1} ($t = 1 \ldots T\text{-}1$) of each individual p^j_{Ci} in the population who has participated in the study for a time period T. This is referred to as a population-based model to support predictive analytics specific to the subpopulation cluster Ci for the rest of the discussions in this paper.

For each individual p^j_{Ci} residing in a subpopulation (manifold cluster) Ci:

1. Perform l^{th}-order auto-regression (for $l = 1.. k \leq T$) on successive change in engagement ratio ΔER; in other words, $\Delta ER^{t+1}=ER^{t+1} - ER^t$, where $t = 1.. T\text{-}1$.
2. Perform *AIC* or *BIC* to determine the desirable lag l given the time series data that minimize *AIC/BIC.*
3. Note the p value and the correlation R^2 between the actual and the estimated based on some pre-selected threshold for R^2.
4. Predict the change in engagement ratio $\Delta ER^{T+1}{}_p$ based on auto-regression using T, $T\text{-}1$, $T\text{-}2 \ldots T\text{-}l$. If the test statistics in (3) are reasonable (i.e., $p < 0.05$ and *threshold* $\leq R^2$), keep the predicted value $\Delta ER^{T+1}{}_p$ and stop. Otherwise continue to step 5.
5. Determine the predicted value $\Delta ER^{T+1}{}_p$ based on $\Delta ER^{T+1}{}_p = \text{Arg}Max_{\Delta ER}\, Pr(\Delta ER|\, ER = ER^T{}_p)$.

Among the choices on the actionable health (e.g., self-monitoring, exercise, diet), determine the actionable health recommendation based on the one with the largest $\Delta ER^{T+1}{}_p$.

Predicting/recommending coaching agenda based on compliance ratio is similar by repeating the steps.

Personalized online health education

CDC's DPP[27] is a health education program targeting at individuals at high risk for type 2 diabetes. It was reported that participants in the lifestyle intervention introduced by DPP who lost 5%–7% of their bodyweight experienced a 58% lower incidence of type 2 diabetes than those who did not receive the lifestyle intervention.

The curriculum is designed as a year-long program. The delivery mode could be in-person, online, or a combination of the two; whereas online delivery refers to health coach-led teleconference format in synchronous mode. The program's aim[28] is to help a participant to achieve a modest weight loss in the range of 5%–7% of baseline body weight, a combination of 4% weight loss and 150 min of physical activity per week on average, or a reduction in H1AC of 0.2%. Strategies of the program focus on self-monitoring of diet and physical activity, building self-efficacy and social support for maintaining lifestyle changes.

The health education component of the pilot study in this research tests an alternative. Instead of the synchronous mode led by a health coach, health education is delivered asynchronously directly to the mobile device of an individual via the SIPPA Health app. This helps to improve the efficiency and the flexibility, and the opportunity to personalize DPP in terms of the delivery schedule, amount of health education content, and the rate of delivery. In this research, the strategy is to deliver DPP fully online asynchronously based on personalized programming of the content to be delivered in 3 months. In addition, pilot participants are reminded (on a daily basis) of the self-monitoring activities that include not only just the diet and physical activities but the glucose tracking, so that pilot participants can review the trend and the interaction relationship among diet, physical activity, and glucose level. Similar to compliance and engagement ratios, the metric used to gain an understanding on the effect of delivering DPP health education online via a mobile app is the change in diabetes self-efficacy. The validated survey instrument developed elsewhere[5] is adopted as a data collection instrument for assessment purposes.

Self-efficacy questionnaire

DSEQ[5] was developed to evaluate the outcome measure of the diabetes health education program delivered under the Rideau Valley Diabetes Services in ON, Canada. The development process of DSEQ follows psychometric design principle. It focuses on two aspects of diabetes self-efficacy covering a comprehensive range of diabetes self-management activities: belief and action. *Belief* refers to the perception on the importance of self-management activities, while *action* refers to the confidence on carrying out the self-management activities.

In the development process of the DSEQ, the main factors being considered include:

(1) Reliability in terms of test-retest and internal consistency.
(2) Validity in terms of the meaning/interpretation of the results and bias.
(3) Responsiveness in terms of sensitivity and stability; i.e., could the instrument detect change when there is a successful intervention, and could it show no change when there is a lack of (effective) intervention?
(4) Invariance in terms of the presentation order of the questions that may affect the outcome of the responses via Spearman coefficient on split-half and odd-even shuffling of the questions.

A total of 58 questions of a six-point Likert scale are included in DSEQ. Principal-component factor analysis (PCFA) was applied to group the questions into six scales:

Scale 1: managing social, emotional, and food-related aspects of diabetes
Scale 2: communicating with health professionals and planning
Scale 3: managing low blood sugars
Scale 4: managing diabetes related to exercise, blood glucose, and prevention
Scale 5: integrating knowledge and day to day care
Scale 6: managing insulin

Scale 6 was not included in PCFA as not all patients of Rideau Valley Diabetes Services were insulin dependent. The following were the result of PCFA reported by Roblin et al. based on analyzing the responses by 478 individuals:

Scale	No. of questions	% coverage of variance
1	17	17.8
2	8	10.61
3	4	6.42
4	13	13.69
5	10	10.45

Table 1. Participant demographic information

Ethnicity	Distribution (%)
Caucasian	41.40
African American	30.90
African American/Hispanic	3.10
Asian	13.80
Hispanic	7.50
Hispanic/White	1.10
Indian/Asian	1.10
Mexican/Black	1.10
Income (in US $)	**Distribution (%)**
0–24,999	27.50
25,000–49,000	23.33
50,000–99,999	28.33
100,000–150,000	12.50
150,000–199,999	4.17
>200,000	4.17
Education level	**Distribution (%)**
High school diploma	17.89
Some college—no degree	21.95
2-year college degree	16.26
4-year college degree	26.83
Some graduate work	5.69
Graduate-level degree	11.38
Self-perceived health	**Distribution (%)**
Poor	8.13
Fair	28.46
Good	43.09
Very good	16.26
Excellent	4.06
Sex	**Distribution (%)**
Female	51
Male	49

Incorporating DSEQ for pilot

In this research, a pilot participant was asked to respond to DSEQ—referred to as *pre-survey*—during the enrollment for establishing a baseline. The pilot participant was then placed in a 1-month "*hold*" period under the assumption that the participant would not receive any intervention involving a change in lifestyle or medication, as well as that the participant would not remember the response to DSEQ after 1 month. After the 1-month *hold*, the pilot participant was invited to respond to the same DSEQ again—referred to as *post-survey*. The post-survey is required no later than the orientation for on-boarding. The orientation includes instructions on the use of the monitoring devices, such as glucose meter and lancing device, as well as the mobile app. After the orientation, the participant enters the intervention phase, which lasts on average about 3 months. Further details on the actionable health activities during the intervention phase are described in the next section. After the intervention phase, the participant was asked to repeat the DSEQ again—referred to as *exit-survey*.

Under the assumption of no intervention, the difference in pre-survey and post-survey provides a baseline on the intra-variations indicating the (in)consistency of the survey responses. By comparing the responses of pre-survey and post-survey, we could better understand the stability and any possible change in self-efficacy when there is no intervention. Similarly, by comparing the responses of pre-survey and exit-survey, the effect of behavioral predictive analytics and the mobile app on self-efficacy could be examined. To understand the effect, ANOVA with repeated measures (ANOVA-RM) was applied to analyze the pre-survey, post-survey, and exit-survey responses.

Preliminary study

The proposed approach was applied to the diabetes subjects of a self-health management pilot conducted under an IRB-approved study protocol (CUNY IRB no. 2018-1043). The objective was to investigate the impact of digital health solutions to affect individuals' behavior toward self-management of chronic diseases, particularly type 2 diabetes.

To be included in the study, the participants had to be at least 18 years of age. They needed a minimum education level of a high school diploma. An additional criterion was that the participants had to have an H1AC of 6.0, or a diagnosis of diabetes or pre-diabetes. This means that participants also had a perceived risk of developing, or had developed diabetes and other associated chronic illnesses.

The behavior model developed under previous research for predicting behavior readiness was based on a population of over 500 individuals. The population consisted of both healthy individuals as well as individuals with chronic diseases. The statistically validated model was applied in stage 1 of the proposed predictive analytics for personalization.

This pilot strives for an equitable recruitment: 148 individuals with type 2 diabetes were involved in stage 2 of the preliminary study. These participants had a mean age of 49 years and a mean H1AC of 7.89. The population characteristics are shown below (Table 1).

The survey responses of these 148 individuals were used to derive behavior readiness vectors, which were then subsequently used to identify manifold clusters (subpopulations). Individuals were grouped into a cluster when their behavior readiness patterns were close to the statistically significant association patterns characterizing the cluster.

Among the 148 individuals participating in this pilot on a rolling basis, some were still in a 1-month hold period for establishing a baseline without intervention; i.e., they had not entered the pilot phase for personalized intervention. Among the rest, 49 subjects with type 2 diabetes were included in deriving the population-based models for personalized intervention. These were the subjects who entered/were in the intervention phase of the study during this research. The self-health management focused on three health coaching agenda:

- knowledge building and information gathering through daily wisdom sent via SMS and/or (in-app) push notifications
- discipline and skill development (through notifications and reminders)
- awareness improvement (through weekly survey)

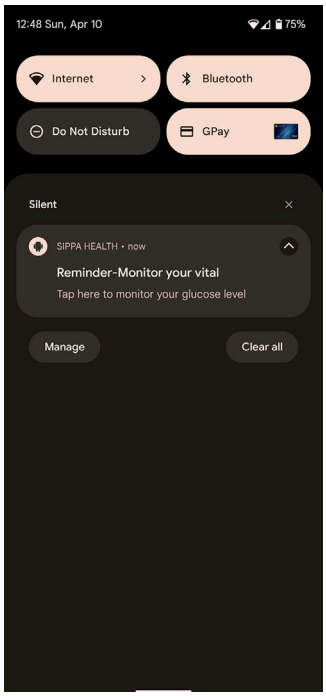

Figure 1. Push notification

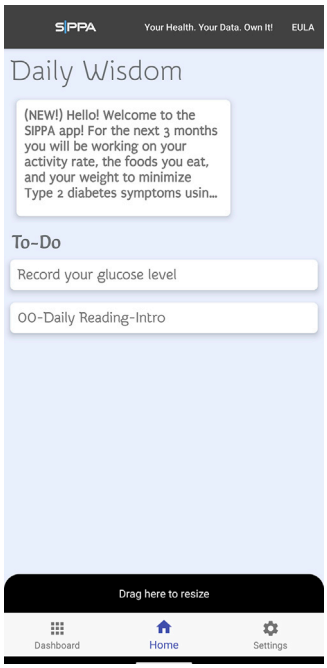

Figure 3. In-app service

The self-health management activities of this pilot included the delivery of (1) daily wisdom on diabetes management, (2) text messaging, and/or notification reminders on diet, physical exercise, and self-monitoring, and (3) in-app services to track self-monitoring, diet and steps. This is followed by weekly online surveys to improve awareness on self-management. Examples of each of these are shown in Figures 1, 2, 3, and 4. This study focuses on only a retrospective analysis based on compliance

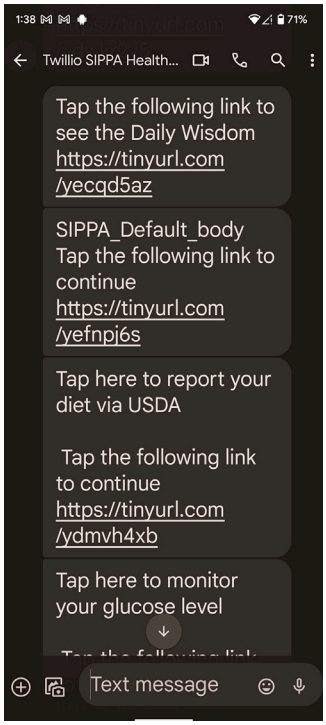

Figure 2. SMS reminder

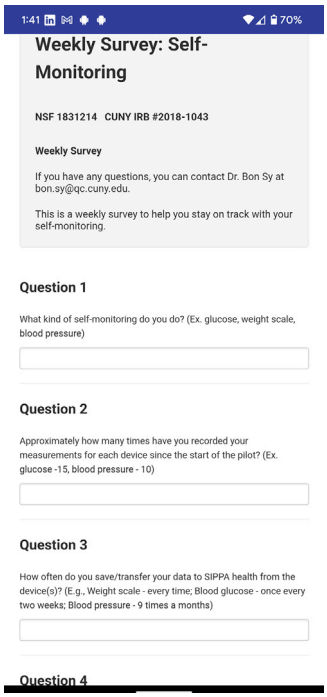

Figure 4. Weekly survey

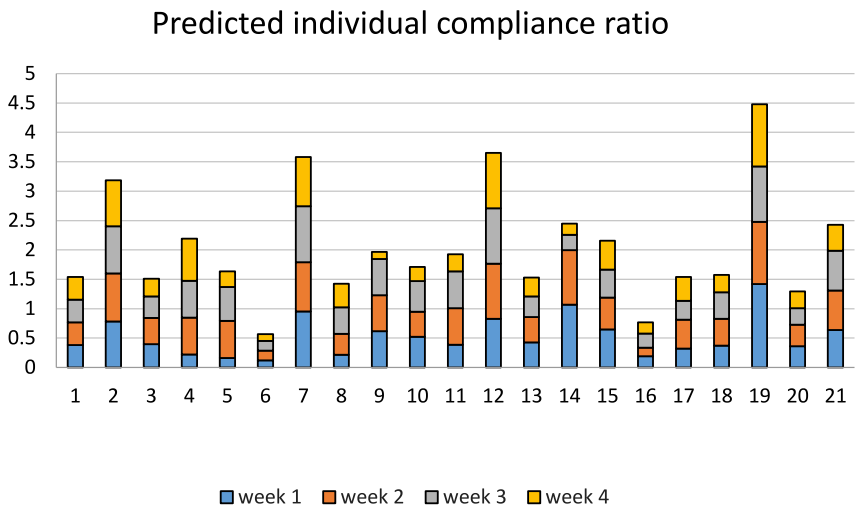

Figure 5. Predicted compliance ratio for a subject

ratio, and a forward-looking prediction based on ER, for evaluation purposes.

Data-driven model development

The data collected and used for this preliminary study are a subset of our pilot sample. When a subject enters the "intervention" phase of the study, the SIPPA Health platform collects activity meta-data on user interactions with the SIPPA Health mobile app. This allows us to infer adherence and engagement in certain activities; e.g., using the app to conduct medication research or schedule medication reminders.

The survey response data of 148 subjects were used to derive individuals' behavior readiness. Among the 148 subjects, 49 of them had either completed the study or were in the intervention phase during the study period.

The data from all 148 subjects were used for the manifold clustering to identify subpopulation characteristics defined by behavior readiness. The 49 subjects fell into four of the manifold clusters. Each of the 49 subjects was assigned to a subpopulation cluster based on the similarity between the behavior readiness measure of the individual and behavior patterns exhibiting statistically significant association that define the cluster. The data from the subjects in each cluster were used to derive the population-based models (Predictive analytics for personalization, stage 3) to support the behavioral predictive analytics for personalization.

The personalization results reported in this paper are based on 22 subjects who were in the intervention phase during the study period of this research. A subject in the intervention phase of the study receives a recommendation on a weekly basis about the activities on diet management, physical activities, and self-monitoring of glucose and other vital signs. Personalization for each subject is performed on a weekly basis to recommend one activity to focus on during a week.

Within each cluster subpopulation, a normalized compliance ratio and an ER of each subject, as well as the change on a weekly basis, are derived for each one of the activities: diet management, physical activities, and self-monitoring. Each ratio is normalized to account for the different starting times of the participants. For each subject, an auto-regression model is derived for each activity for each ratio.

It is noted that developing an auto-regression model is not always feasible. For example, there may not be sufficient data because in an early stage of the participation an individual may have only activity data in one category (such as self-monitoring) but not the others (such as physical activities). Furthermore, the data may not yield a valid auto-regression model because it fails the statistical test in step 3.2 during the model selection process using *BIC*/*AIC*. Typically, this happens when a subject is in the intervention phase for less than 4 weeks.

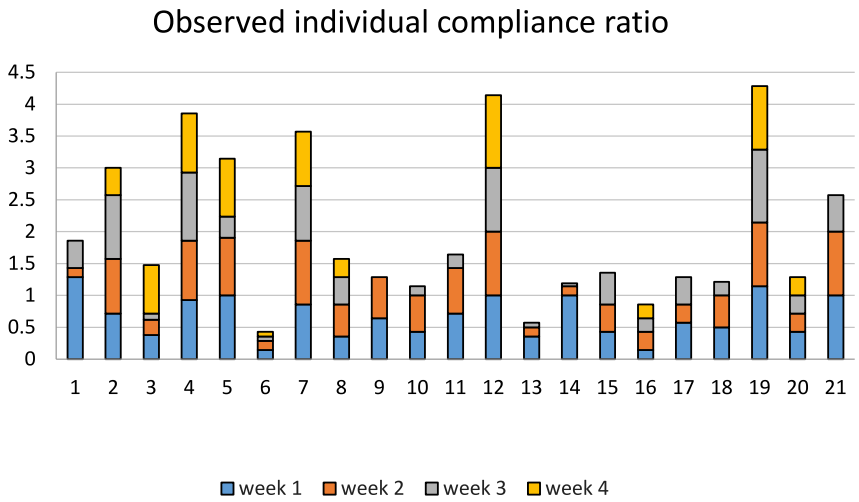

Figure 6. Observed compliance ratio for a subject

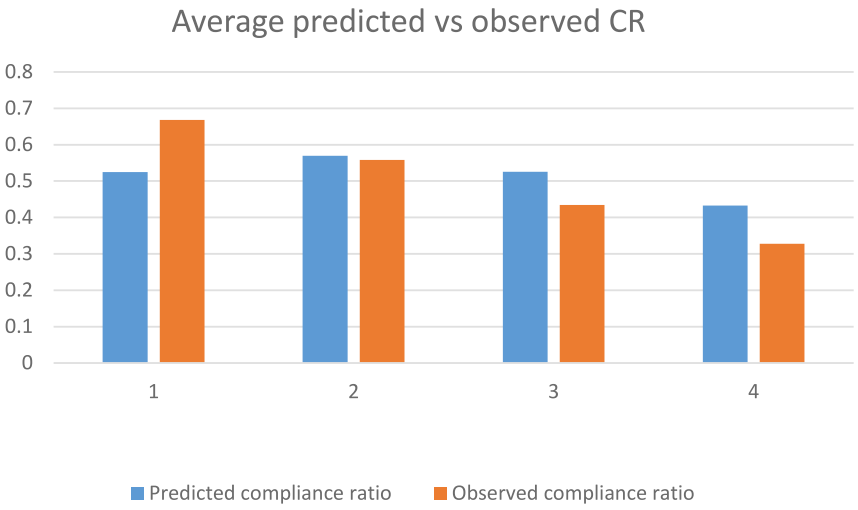

Figure 7. Average predicted versus observed CR

In a scenario where an individual auto-regression model is not feasible, prediction for personalization for the individual will rely on the population-based model. For each cluster subpopulation, we derive a population-based model—one for each activity—defined by the distribution of the compliance/engagement ratio and the amount of change using the data of all the subjects in the cluster subpopulation. In other words, there are $n \times m$ such models to capture engagement (compliance) ratios; where n is the number of clusters and m is the number of activity categories. For example, $m = 3$ if there are three categories of activities, such as diet management, physical exercise, and self-monitoring. A population-based model developed for an activity category Aj (where $j = 1..\ m$) in a cluster Ci (where $i = 1..\ n$) is used to predict an engagement (compliance) ratio for an individual in Ci when an individual auto-regression model is not available for the activity category Aj.

Preliminary study

The subjects included in this study were distributed across four different clusters (subpopulations). The results reported in this paper are based on an 11-week (2.5 months) study of personalization. In other words, the activity data of each subject since participating in this pilot, leading up to the week of personalization, were used to develop the prediction models for the self-management activities. Then for each subject a recommendation (either exercise or diet management) was derived using the prediction algorithm described in the previous section.

Feasibility assessment. To determine the feasibility on the real-world application of the proposed behavioral predictive analytic technique, the design of the preliminary study consists of two parts. The first part is a retrospective analysis using the data related to compliance. The second part is looking forward prediction on the engagement. The purpose of retrospective analysis is to establish a base reference for performance assessment based on historical results. The looking forward prediction is for evaluating the prediction performance as a time series on a rolling basis in real-time.

Retrospective analysis. The predictive analytics would be greatly simplified if personalization could be based on only the time series (engagement/compliance) data. That is, for each subject, it is possible to derive an auto-regression model that is also statistically valid according to the information-theoretic model selection criteria described in Information-theoretic model selection approach. In such a case, manifold-based clustering could be completely skipped because a population-based model to support personalization would not be necessary.

To gain insight into such a scenario as just described, an attempt was made to derive an auto-regression model for each subject who completed/entered the intervention phase. Out of the 49 subjects, the auto-regression model derivation was successful for 21 subjects. Therefore, manifold clustering is required for this particular use case on applying the algorithm described in Predictive analytics for personalization.

The compliance ratio is computed on a weekly basis for each subject. A subject has n data points of compliance ratio; where n is the number of weeks of participation in the intervention phase. For deriving the auto-regression model for a subject, n-4 data points were used to derive/train the auto-regression model, and the model was used to predict the compliance ratio of the last four data points for evaluation purposes.

Forward-looking prediction. In contrast to the retrospective analysis, forward-looking prediction involves only those subjects who were in the intervention phase during the study period. Out of the 49 subjects mentioned earlier, 22 of them were involved.

The ER of each active subject was computed on a weekly basis. Similar to the retrospective analysis, an estimated ER is derived for each week based on the predictive analytics technique described in Predictive analytics for personalization. The prediction was performed forward looking. For example, the prediction on ER for week n ($n = 2 \ldots 11$) of the 11-week study period for a subject would be conducted at week n-1. Then the actual observed ER was recorded at week n. This forward-looking prediction process was repeated 10 times in the 11-week study period.

Table 2. *R* and p values for the tests

	Week 1	Week 2	Week 3	Week 4
R	0.5178	0.6673	0.7698	0.7008
p value	0.0162	0.00095	4.5×10^{-5}	0.0004

RESULTS AND DISCUSSION

Retrospective analysis

Figures 5 and 6 show the predicted and observed compliance ratios of the 21 subjects for whom a statistically valid auto-regression model could be derived. The result shows the predicted and observed compliance ratios for each week on each of the 21 subjects; whereas a compliance ratio is derived based on a 7-day average. As shown in Figure 7, there is a consistent pattern across the 4-week prediction period.

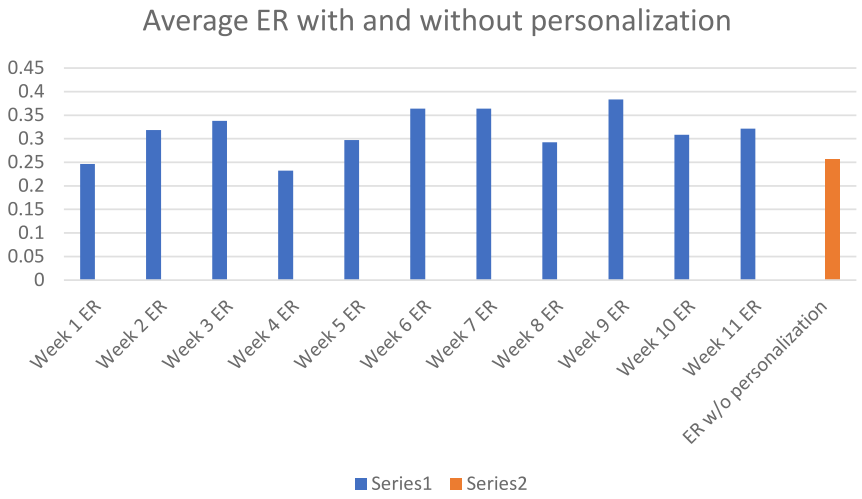

Figure 8. Aggregated ER w(/o) personalization

Table 2 shows the *R* and the p value of the 4 weeks; whereas *R* is the correlation coefficient measuring the strength and direction of a linear relationship between the predicted and observed compliance ratio, and p value is a probability measure on the value of *R* that have occurred just by random chance (which is typically compared against the gold standard requiring it to be less than 0.05).

While auto-regression based on the AIC/BIC criteria for model selection is intuitively reasonable, it is desirable to obtain evidence from retrospective analysis (i.e., looking backward) that auto-regression is indeed reasonable. When auto-regression is applied to predict the compliance ratio using the weekly data of a subject, one could compare this against the observed actual compliance ratio. If a linear relationship exists between the observed compliance ratio and that predicted by employing auto-regression across the 21 subjects, this provides evidence to support the appropriateness of auto-regression. To determine the strength of the linear relationship, correlation coefficient analysis is conducted. Correlation coefficient *R* is bounded between −1 and 1. Generally speaking, there is a strong linear relationship between the predicted and the observed values when *R* is greater than 0.5, as shown in Table 2.

Forward-looking prediction

In the forward-looking prediction experiment, the prediction is on actionable health recommendations based on the maximal posterior estimate as described in Predictive analytics for personalization. In this study, the personalized actionable health recommendation would be in either diet management or exercise. Twenty-two subjects were in the intervention phase during this period of research.

Figures 5, 6, and 7 show evidence of its accuracy and consistency. But we are also interested in the effectiveness of the prediction technique for personalization. To evaluate its effectiveness for improving self-efficacy on health management, this study attempts to show personalized actionable health (recommended by behavioral predictive analytics) resulting in a more active engagement when it is compared with that of without personalization.

To understand the effect of personalization on engagement, the weekly average ER without personalization is compared against the ER with personalization. Figure 8 shows the aggregated weekly engagement average, disregarding subpopulations, for comparison purposes.

In calculating the ER without personalization, the average ER of each subject over time prior to personalization is first

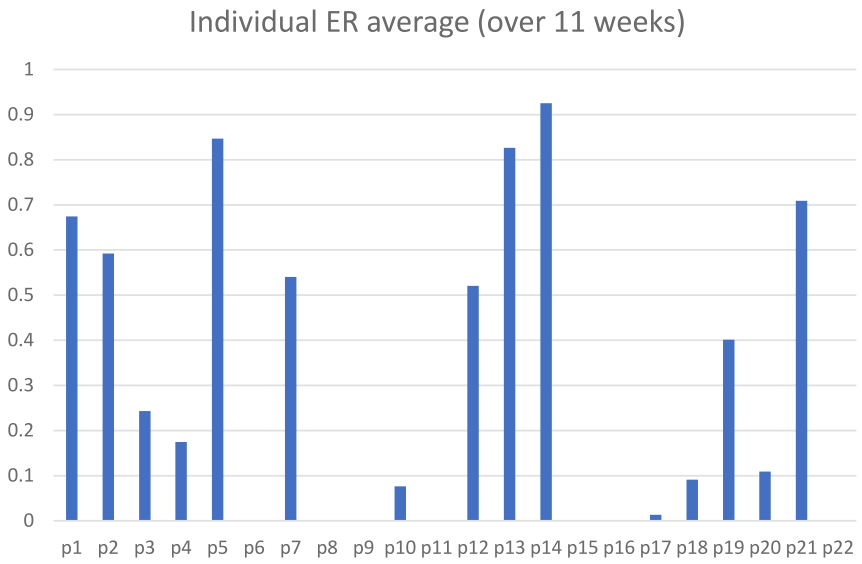

Figure 9. Individual ER average (over 11 weeks)

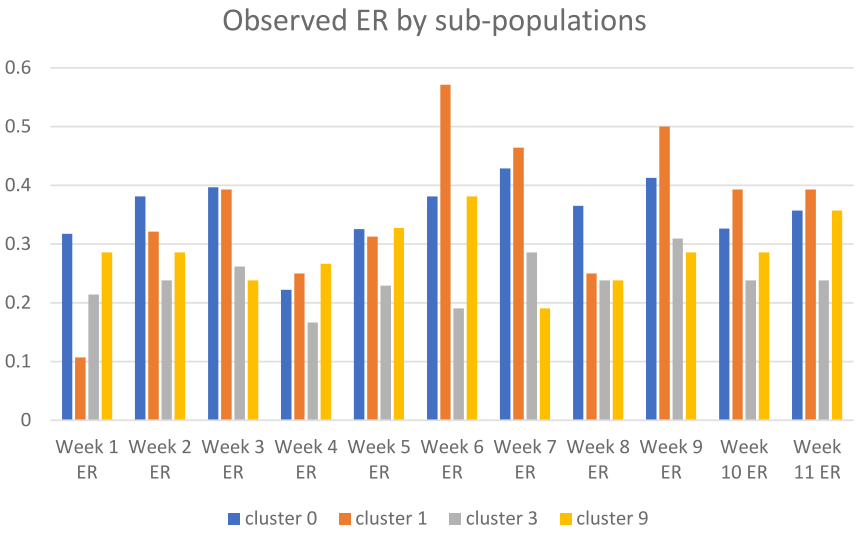

Figure 10. Observed ER by subpopulation clusters

calculated, then the average over all the subjects. Note that the average ER of each subject over time prior to personalization spans over different time periods and lengths, as do the actionable health recommendations because of the rolling nature of the subject participation in the pilot.

Figure 9 shows the ER of each individual averaged over the participation period. There are six subjects with low/zero ER in forward-looking prediction. All of them received follow-up from this research team to understand these unusual outcomes. One withdrew from the study, and two were unreachable during the study period. Among the rest, one has limited technology proficiency, and one other older adult subject relies on her daughter to assist her on certain self-management activities at a time convenient to her daughter. Furthermore, one subject (participant 15 in Figure 9) was active until he damaged his phone during the study period of this research.

Figure 10 shows the aggregated engagement average of 22 subjects (with personalization) for each week during the study period distributed across four cluster subpopulations.

Experimental results and discussion

The results shown in Figures 5, 6, and 7 in the retrospective analysis show evidence of the feasibility of behavioral predictive analytics in terms of computational efficacy, as measured by accuracy and consistency.

Table 3. Factor loading of DSEQ questions excluded

Scale 1	factor loading max: 0.737, min: 0.413
Missing questions: 38 (0.737), 48 (0.678), 36 (0.634), 43 (0.551), 18 (0.506), 40 (0.413)	
Scale 2	factor loading max: 0.799, min 0.477
Missing questions: 46 (0.647), 41 (0.477)	
Scale 3	factor loading max: 0.814, min: 0.551
Missing questions: none	
Scale 4	factor loading max: 0.65, min: 0.447
Missing questions: 14 (0.502), 3 (0.447)	
Scale 5	factor loading max: 0.693, min: 0.393
Missing questions: 20 (0.619), 29 (0.56), 12 (0.393)	

Figure 8 shows the evidence of the applicability of the approach in terms of health efficacy. It shows that engagement level with personalization is better than that without personalization.

The results shown in Figures 9 and 10 in the forward-looking experiment demonstrate the practical implementation feasibility. The results shown in Figure 10 also reveal indirect evidence of the effectiveness of the manifold-based clustering technique for grouping subjects into subpopulations by means of behavior readiness. In particular, subpopulation clusters 1 and 2 are the more engaged patient subpopulations reflected in the behavior readiness characteristics of the clusters. Furthermore, personalization with strategies tailored for a cluster seems to show an effect over time for improving the engagement, e.g., the second cluster subpopulation is not as high performing at the beginning.

Finally, the overall average ER with personalization had a mean value of 0.31 with a an SD of 0.33. The 95% confidence interval was [0.17, 0.45]. By contrast, without personalization, the overall mean ER is 0.26 with an SD of 0.31. The 95% confidence interval for this value was [0.13, 0.38].

Health education assessment using DSEQ

The effectiveness of delivering online health education through the SIPPA Health platform was evaluated. The health education content is based on the diabetes prevention program developed by CDC. In this pilot study, a total of 34 individuals with type 2 diabetes completed the health education.

DSEQ assessment instrument

Similar to the populations that were based on the psychometric design of the DSEQ by Roblin et al., only a fraction of subjects in this pilot population are insulin dependent. Therefore, questions on scale 6 relating to insulin management were not included.

In addition, the original set of 52 questions in DSEQ[5] covering the five scales were further reduced to 39 questions after considering the fatigue factor that impacts the participants in our pilot. Due to the pilot constraints, a subject was asked to complete all 78 responses (one on belief and one on action to 39 questions) in one session. The 13 questions excluded from the original DSEQ, and the corresponding factor loading, are listed in Table 3.

Experimental results

In this pilot study, 34 participants completed the online health education via the SIPPA Health mobile app. A participant is considered to having completed the health education if the following conditions were met:

1. Completed the DSEQ—refer to as the *pre-survey* defined in Incorporating DSEQ for pilot—as soon as the participant was enrolled into the self-health management pilot.

Table 4. ANOVA-RM on belief disregarding scales

	F statistic	Significance	Greenhouse-Geisser
Sphericity assumed	4.185	0.019	
Mauchly's test of sphericity		0.795	0.986
Within subject contrast	7.272	0.011	

2. Completed the DSEQ again—refer to as the *post-survey* defined in Incorporating DSEQ for pilot—soon after 30 days since the pre-survey.
3. Completed the DSEQ one last time—refer to as the *exit-survey* defined in Incorporating DSEQ for pilot—after the completion of the self-health management intervention phase.

ANOVA-RM was performed on the belief responses provided by the 34 participants during the pre-survey, post-survey, and exit-survey—without concerning the scales. ANOVA-RM was then repeated on each scale. This was then repeated on the action responses.

In this research, health education introduced during the self-management intervention phase occurs after the post-survey. When the pilot subject population did not receive intervention elsewhere during the *hold* period (i.e., between pre-survey and post-survey), the expected outcome of applying ANOVA, or t-test, to the responses of pre-survey and post-survey should show no change. In other words, the null hypothesis stating no difference in the means of the pre-survey and post-survey responses cannot be rejected using the gold standard of p-value < 0.05.

On the other hand, if SIPPA Health is effective in delivering online health education in terms of evidence-based outcome, one would expect a statistically significant difference between the mean of the pre-survey and exit-survey; i.e., the null hypothesis stating no change in the means of the pre-/exit surveys is expected to be rejected using the gold standard of p-value < 0.05.

Therefore, positive outcomes attributed to SIPPA Health can be formulated as:

(1) Null hypothesis is rejected using the gold standard of p-value < 0.05 in considering the pre-survey and the exit-survey; i.e., there is a change in the self-efficacy.
(2) Null hypothesis is not rejected due to insufficient statistical evidence using the gold standard of p-value < 0.05 in considering the pre-survey and the post-survey; i.e., there is no change in the self-efficacy.

The strategy is to perform ANOVA-RM on the responses to the pre-post-exit surveys. It does not need to analyze further if the null hypothesis stating no change could not be rejected. On the other hand, if the null hypothesis is rejected, then further analysis will be conducted to confirm the change is from responses to the pre-exit, but NOT pre-post, surveys.

ANOVA-RM results on belief. ANOVA-RM was performed on the responses by the 34 participants to the belief aspect of the DSEQ, without concerning the scales. Similar to the psychometric evaluation of the DSEQ, the responses to the belief aspect of the 39 questions by each participant in a survey were averaged. This resulted in three datasets—pre-survey, post-survey, and exit-survey. Each dataset consists of 34 responses. These three datasets are the basis of the ANOVA-RM leading to the results shown in Table 4. The lower and upper bounds of the estimate with 95% confidence based on the sample mean are shown in Table 5.

The process is then repeated for each scale. On the scale level, the average is performed using the responses to only the questions on that scale. For example, there are four questions for scale 3. The average for the responses by an individual to these four questions was derived and used in the ANOVA-RM when the focus was on scale 3. The p-value results of the analysis for each scale are shown in Table 6.

It is noted that the p-values of ANOVA-RM analysis on scales 4 and 5 are less than the gold standard 0.05 under the assumption of sphericity—suggesting a change in self-efficacy. The analysis result on each scale was further examined.

Further details on the ANOVA-RM analysis for scales 4 and 5 are shown in Tables 7, 8, 9, and 10. In this research, Mauchly's test was applied to assess the sphericity assumption. The Greenhouse-Geisser F statistic and significance are also included in Tables 7 and 9 should there be statistical evidence on the violation of sphericity assumption (see Tables 8 and 10).

ANOVA-RM results on action. The analyses described in ANOVA-RM results on belief were repeated on the responses to the action aspect of the DSEQ.

The results of ANOVA-RM analysis are shown in Tables 11, 12, 13, 14, and 15. In Table 11, two p-values (0.234 and 0.236) were reported on the first row because the p-value of Mauchly's test for sphericity assumption is less than 0.05; i.e., the null hypothesis of Mauchly's test on sphericity assumption is rejected. Therefore, the adjustment based on Greenhouse-Geisser (0.236) was included. Similar adjustments were also made for scale 3 and scale 5, as shown in Tables 12 and 13.

Table 6. ANOVA-RM on belief for each scale

Sig	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
Sphericity assumed	0.072	0.503	0.098	0.017	0.007
Mauchly's test of sphericity	0.039	0.865	0.058	0.369	0.336
Within subject contrast	0.042	0.284	0.606	0.019	0.003

Table 5. Mean estimate on belief disregarding scales

	Sample mean	95% confidence Lower	Upper
Pre	4.396	4.204	4.589
Post	4.417	4.193	4.641
Exit	4.59	4.429	4.752

Table 7. Details on ANOVA-RM for belief scale 4

	F statistic	Significance
Sphericity assumed	4.361	0.017
Mauchly's test of sphericity		0.369
Greenhouse-Geisser	4.361	0.019
Within subject contrast	6.074	0.019

Table 8. Mean estimate on belief for scale 4

	Sample mean	95% confidence Lower	Upper
Pre	4.464	4.269	4.66
Post	4.437	4.202	4.672
Exit	4.674	4.513	4.835

Table 10. Mean estimate on belief for scale 5

	Sample mean	95% confidence Lower	Upper
Pre	3.946	3.585	4.308
Post	4.109	3.784	4.435
Exit	4.332	4.05	4.615

Review on results and limitations

Result review and discussion. By comparing the effect of delivering health education via SIPPA Health on affecting the belief and action aspects of self-efficacy, the following results are noted:

1. Without concerning the scales, ANOVA-RM indicates a change in the belief aspects of the diabetes self-efficacy, as shown in Table 4. Null hypothesis of Mauchly's test of sphericity could not be rejected since the p-value was 0.951. Therefore, sphericity assumption is valid. In other words, no adjustment on the significance value ($p = 0.023$) is required. Furthermore, the within-subject pre-exit contrast is significant ($p = 0.017$), as well as the upward trend on the mean shown in Table 5, confirming the positive overall effect on the belief aspects of diabetes self-efficacy improvement.
2. To better understand the change in self-efficacy at the scale level, ANOVA-RM was conducted on each level. Table 6 shows that the change in scales 4 and 5 is of significance.

Scale 4: managing diabetes related to exercise, blood glucose, and prevention

Scale 5: integrating knowledge and day-to-day care

Note: the results of Mauchly's test on scales 4 and 5 analysis show no adjustment required for the p values derived during the ANOVA-RM process.

3. In contrast to the belief aspects, the result of ANOVA-RM did not indicate a change in the overall action aspects of self-efficacy, as shown in Tables 11 and 12.
4. ANOVA-RM conducted on the scale level revealed action scale 5 is significant, as shown in Table 13. With further details are shown in Tables 14 and 15.

Note that the null hypothesis of Mauchly's test of sphericity for scale 5 was rejected, as shown in Table 13. Therefore, the p-value significance should be adjusted to 0.04 from 0.032 according to the Greenhouse-Geisser correction.

Overall, the result shown above is a validation of the SIPPA Health (mobile app) platform for delivering online health education programs grounded on the DPP of CDC. Although the result did not show improvement in self-efficacy on all scales, this is not a surprise.

Table 9. Details on ANOVA-RM for belief scale 5

	F statistic	Significance
Sphericity assumed	5.406	0.007
Mauchly's test of sphericity		0.336
Greenhouse-Geisser	5.406	0.008
Within subject contrast	10.383	0.003

The behavioral predictive analytics is focused on personalization toward optimizing engagement in three areas: self-awareness of health conditions, knowledge and skill-building through health education, and discipline on actionable health activities, including self-care and self-monitoring of glucose, diet, and physical activities. These areas fall into scales 1, 4, and 5.

Limitations of the study

The outcome as shown in DSEQ indicates improvement on self-efficacy in both belief and action aspects of scale 5: integrating knowledge and day to day care. It also indicates improvement on self-efficacy in belief scale 4: managing diabetes related to exercise, blood glucose, and prevention. That is, increased awareness on health and skill building on self-management. However, action scale 4 did not appear to be significant. We suspect that the 3-month duration is not sufficient for pilot participants to feel confident on carrying out actionable self-care activities.

Although this pilot covers diet self-reporting and vital sign measurement, it did not cover the full scope of scale 1: managing social, emotional, and food-related aspects of diabetes. This is due to at least in part the exclusion of the mental health support. Nonetheless, the p-value of 0.072 for belief scale 1 is encouraging since it is close to the gold standard 0.05.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Sy (bon.sy@qc.cuny.edu).

Materials availability

The mobile devices and/or vital sign monitoring devices used in this study are not available to the readers. However, readers could contact the lead contact for information about the models and vendors for procuring the devices.

Data and code availability

Due to the sensitivity of the personal data and the re-identification risk, it was concluded, after discussion with the administration and the editors of *Patterns*, that data access shall be requested through the lead contact Dr. Sy. Arrangements may be made for readers interested in the data. This will require seeking approval by the CUNY IRB, and the individuals interested in the data are subjected to the same standard and criteria as the personnel of this research; i.e., completing and maintaining currency of CITI training on research, ethics,

Table 11. ANOVA-RM on action disregarding scales

	F statistic	Significance	Greenhouse-Geisser
Sphericity assumed	1.483	0.234/0.236	
Mauchly's test of sphericity		0	0.707
Within subject contrast	1.482	0.232	

Table 12. Mean estimate on action disregarding scales

	Sample mean	95% confidence Lower	Upper
Pre	4.000	3.736	4.264
Post	4.005	3.74	4.271
Exit	4.143	3.859	4.426

compliance, and safety training (https://about.citiprogram.org/) for human subject protection.

Currently the Android SIPPA Health mobile app is available in the Google Play Store closed testing group. Readers may send a request via email to info@sippasolutions.com. Once one is being added to the closed testing group, one will be able to find and download the SIPPA Health in the Google Play Store.

The software for developing a model for behavior readiness is based on the commercially available (https://ssicentral.com/index.php/products/lisrel/) application LISREL. Software code developed for manifold clustering is based on Java servlet hosted in a Tomcat server. It requires periodic security update, but could be arranged for non-commercial use upon request made to the lead author. Software code written on Python/R and SQL for model selection and predictive analytics are tightly integrated into the application and database infrastructure. Readers interested in replicating the environment are encouraged to contact the lead author. The analysis result of the diabetes self-efficacy was conducted using SPSS, and ANOVA-RM could be carried out on Excel as well. A tutorial explaining the steps for carrying out ANOVA-RM closely resembling that of this research could be found in YouTube (https://www.youtube.com/watch?v=6T6dvrwDe_U). Readers interested in replicating the infrastructure of this research environment for their own use could contact the lead author for a further discussion.

Experimental procedures: Further details

In this research the experimental procedure for the study followed the CUNY IRB-approved protocol (CUNY IRB no. 2018-1043). Due to the rolling basis of the subject recruitment and participation, only a subset, rather than the entire population, participated in the intervention phase that involves behavioral predictive analytics at any given time point. In reference to Table 1, as well as the inclusion criteria approved by the institutional review board for the pilot study, our study did not include patients with diabetes from all walks of life. In particular, the vulnerable population defined by individuals younger than 18 years is excluded from the study sample. In addition, the health education material for the DPP developed by the US CDC was developed for readers with a sixth-grade reading level and is available in both English and Spanish. In our pilot, the inclusion criteria stipulate a minimum high school reading level. The minimum high school reading level is related to the survey instrument for deriving behavior readiness. While the survey was validated with sufficient statistical power, it has not been assessed for its appropriateness regarding the required minimum reading level.

Final thoughts and open research questions

Type 2 diabetes generally requires 3–6 months before short-term health outcome improvement could be observed. There may be a relapse in the health outcome improvement over time. A conclusive health outcome improvement such as that shown in the DPP (of the US CDC) entails a large-scale, long-term study that lasts more than 5–10 years.

In this research, most pilot participants who completed the study were engaged for a 3-month period. Since behavior predictive analytics based on auto-regression is conducted on a weekly basis for each participant, transient changes within a week are smoothed via the average in deriving the weekly engagement level. Patterns, such as the gradual change or trajectory change over time in the period of weeks, generally could be incorporated in the auto-regression due to its nature to make a prediction based on the observations made available over time. In other words, the 3-month study period of a subject is assumed stationary. Beyond the short-term 3-month study, a subject may relapse, as observed in this and other studies. The best practice to address this in the real-world environment is to engage the subject for "re-training." The ultimate goal is to empower an individual to develop long-lasting discipline, skill, and knowledge to better self-manage their health and their chronic diseases. Some of the critical open questions for future research are:

1. If the longitudinal study is feasible, could short-term outcome data from studies such as this be useful to support (statistical-based) change-point detection to obtain evidence on the effect of behavior change toward pro-active/relapse of engagement in self-management to inform change in long-term health outcomes?
2. If a large patient population is available, could this research be repeated by applying (manifold) clustering to segment the patient population by behavior readiness and social determinants of health (SDoH), including race/ethnicity, gender, and social-economic status, to inform findings that are specific to population subgroups by the categories defined through the social determinants of health? In doing so, we could achieve a better understanding of any hidden bias embedded in the research result when the distribution of the subjects by SDoH is skewed.
3. Furthering on (2), could health-related social needs be incorporated into a predictive model to improve engagement in self-management, as well as to translate health-related social needs into actionable social services, for enhancing the likelihood of improving long-term health outcome improvement?

Conclusions

A behavioral predictive analytics approach was presented for self-health management with personalization. The personalized recommendation is based on population segmentation via manifold clustering, and the engagement outcomes that reveal the behavior readiness of an individual in self-management.

Auto-regression and population models were derived to support the predictive analytics approach for generating personalized recommendations. A limitation is the requirement for a "wait" period to accumulate sufficient data to derive a personalized auto-regression model. In this research we adopt a

Table 13. ANOVA-RM on action for each scale

Sig	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
Sphericity assumed	0.96	0.18	0.807/ 0.73	0.234	0.032/ 0.04
Mauchly's test of sphericity	0.212	0.256	0	0.07	0.028
Within subject contrast	0.899	0.328	0.681	0.226	0.022

Table 14. Details on ANOVA-RM for action scale 5

	F statistic	Significance
Sphericity assumed	3.645	0.032
Mauchly's test of sphericity		0.028
Greenhouse-Geisser	3.645	0.04
Within subject contrast	5.802	0.022

Table 15. Mean estimate on action for scale 5

	Sample mean	95% confidence Lower	Upper
Pre	3.714	3.277	4.150
Post	3.904	3.556	4.242
Exit	4.081	3.735	4.447

strategy that aims to prioritize personalization based on the greatest improvement possible on engagement in a self-management area. This has an inherent bias that may negatively impact individuals with limited potential for improvement on engagement.

The health education delivered through SIPPA Health mobile app shows evidence on improving diabetes self-efficacy. However, the pilot study is focused on only the English version of the DPP of CDC. We do not yet know how health education delivered in a different language, and the SDoH, may affect engagement and at what pace.

Our future research will focus on understanding these aspects. In addition, our future research goal will also aim to develop partnerships for collecting larger samples to gain insights into statistical significance for generalizability.

ACKNOWLEDGMENTS

This research team is grateful to the reviewers for their suggestions to this research and for the guidance and patience of the editors. This research is conducted under the support of US NSF phase 2 grant 1831214. The pilot study was approved by CUNY IRB no. 2018-1043. All participants have completed the Informed Consent Form approved by CUNY IRB. Magdalen Beiting-Parrish contributed to preparing the CDC DPP materials and the self-efficacy evaluation. Michael Van der Gaag leads the usability study of the mobile app used in this research. The pilot team consists of Arora Ashima, Connor Brown, Brandon Huang, Rebecca Horowitz, Sumaita Hussain, Pan Lin, and Deniz Turgut. Dr. Catherine Benedict advised on this research regarding patient self-efficacy. Dr. Adebola Orafidiya (MD) helped this pilot team by sharing clinical best practice on recommending self-monitoring. This pilot team also benefited from discussions with Dr. Joseph Tibaldi (MD) and Caterina Trovato (CDE) on patient engagement. Dr. Moritz Boettger helped proofread this paper. J.C. performed this work while with SIPPA Solutions. A portion of this paper appeared in the 14th International Conference on Health Informatics, February 2021, BIOSTEC.

AUTHOR CONTRIBUTIONS

B.S. directed the technical research and the technical write-up. M.W. led the pilot study coordination and the pilot participant recruitment. A.H. contributed to the study by keeping track of participants' completion of surveys and interviews and prepared the diabetes self-efficacy data using SPSS tools. J.C. contributed to the implementation of the manifold clustering technique and data collection for grouping pilot participants into subpopulations according to their behavior readiness to support predictive analytics.

DECLARATION OF INTERESTS

B.S. is the Founder of SIPPA Solutions as well as the lead principal investigator on behalf of the City University of New York on this research, which is funded by the US National Science Foundation under grant no. 1831214. M.W. is the lead principal investigator on behalf of SIPPA Solutions on this research. A.H. declares no competing interests. J.C. is a co-inventor of the manifold clustering listed in a patent application. The manifold clustering recited in this paper is patent pending on the (US) national and (PCT) international stage.

Received: November 9, 2021
Revised: February 10, 2022
Accepted: April 22, 2022
Published: May 17, 2022

REFERENCES

1. Diabetes Prevention Program Research Group (2012). The 10-year cost-effectiveness of lifestyle intervention or metformin for diabetes prevention: an intent-to-treat analysis of the DPP/DPPOS. Diabetes Care *35*, 723–730. https://doi.org/10.2337/dc11-1468.
2. Bollyky, J.B., Bravata, D., Yang, J., Williamson, M., and Schneider, J. (2018). Remote lifestyle coaching plus a connected glucose meter with certified diabetes educator support improves glucose and weight loss for people with type 2 diabetes. J. Diabetes Res. *2018*, 3961730. https://doi.org/10.1155/2018/3961730.
3. Hadjiconstantinou, M., Schreder, S., Brough, C., Northern, A., Stribling, B., Khunti, K., and Davies, M.J. (2020). Using intervention mapping to develop a digital self-management program for people with type 2 diabetes: tutorial on MyDESMOND. J. Med. Internet Res. *22*, e17316. https://doi.org/10.2196/17316.
4. Volpp, K.G., and Mohta, N. (2016). Insights report: patient engagement survey: improved engagement leads to better outcomes, but better tools are needed. NEJM Catalyst *2*. Notes. https://catalyst.nejm.org/patient-engagement-report-improved-engagement-leadsbetter-outcomes-better-tools-needed/.
5. Roblin, N., Little, M., and Mcguire, H. (2004). Diabetes self-efficacy questionnaire (dseq) outcome measurement for diabetes education, Oct 2004. https://www.semanticscholar.org/paper/DIABETES-SELF-EFFICACY-QUESTIONNAIRE-(DSEQ)-OUTCOME-Roblin-Little/b94747994e18f744b206788782fb9830d7d9543d?sort=relevance&citationIntent=methodology.
6. Linden, A., Butterworth, S.W., and Roberts, N. (2006). Disease management interventions II: what else is in the black box? Dis. Manag. *9*, 73–85. https://doi.org/10.1089/dis.2006.9.73.
7. Ajzen, I. (1988). Attitudes, Personality, and Behavior*1988* (Dorsey Press).
8. Prochaska, J.O., DiClemente, C.C., and Norcross, J.C. (1992). In search of how people change: applications to addictive behaviors. Am. Psychol. *47*, 1102–1114. https://doi.org/10.1037/0003-066x.47.9.1102.
9. Strecher, V.J., Champion, V.L., and Rosenstock, I.M. (1997). The health belief model and health behavior. In Handbook of health behavior research I. Personal and social determinants, *1997*, D.S. Gochman, ed. (New York: Plenum Press), pp. 71–91.
10. Osborn, C.Y., Rivet Amico, K., Fisher, W.A., Egede, L.E., and Fisher, J.D. (2010). An information-motivation-behavioral skills analysis of diet and exercise behavior in Puerto Ricans with diabetes. J. Health Psychol. *15*, 1201–1213. https://doi.org/10.1177/1359105310364173.
11. Kan, M.P.H., and Fabrigar, L.R. (2017). Theory of planned behavior. In Encyclopedia of Personality and Individual Differences, V. Zeigler-Hill and T. Shackelford, eds. (Springer) https://doi.org/10.1007/978-3-319-28099-8_1191-1.
12. Fry, J.P., and Neff, R.A. (2009). Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. J. Med. Internet Res. *11*, e16. https://doi.org/10.2196/jmir.1138.
13. Bidargaddi, N., Pituch, T., Maaieh, H., Short, C., and Strecher, V. (2018). Predicting which type of push notification content motivates users to engage in a self-monitoring app. Prev. Med. Rep. *11*, 267–273. https://doi.org/10.1016/j.pmedr.2018.07.004.
14. Sawesi, S., Rashrash, M., Phalakornkule, K., Carpenter, J.S., and Jones, J.F. (2016). The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR Med. Inform. *4*, e1. https://doi.org/10.2196/medinform.4514.
15. Van Stee, S.K., and Yang, Q. (2020). The effectiveness and moderators of mobile applications for health behavior change. In Technology and Health, pp. 243–270. https://doi.org/10.1016/b978-0-12-816958-2.00011-3.
16. Macqueen, J.B., and Lee, H.B. (1980). A K-means cluster Analysis computer program with cross-tabulations and next-nearest-neighbor analysis. Educ. Psychol. Meas. *40*, 133–138. https://doi.org/10.1177/001316448004000118.
17. Zhang, K., and Kwok, J.T. (2010). Clustered Nyström method for large scale manifold learning and dimension reduction. IEEE Trans. Neural Network. *21*, 1576–1587. https://doi.org/10.1109/tnn.2010.2064786.
18. Wang, X.-D., Chen, R.-C., Zeng, Z.-Q., Hong, C.-Q., and Yan, F. (2019). Robust dimension reduction for clustering with local adaptive learning. IEEE Trans. Neural Network. Learn. Syst. *30*, 657–669. https://doi.org/10.1109/tnnls.2018.2850823.
19. Ge, S.S., He, H., and Shen, C. (2012). Geometrically local embedding in manifolds for dimension reduction. Pattern Recogn. *45*, 1455–1470. https://doi.org/10.1016/j.patcog.2011.09.022.

20. Gong, W., Zhao, R., and Grünewald, S. (2018). Structured sparse K-means clustering via Laplacian smoothing. Pattern Recogn. Lett. *112*, 63–69. https://doi.org/10.1016/j.patrec.2018.06.006.

21. Faivishevsky, L., and Goldberger, J. (2012). An unsupervised data projection that preserves the cluster structure. Pattern Recogn. Lett. *33*, 256–262. https://doi.org/10.1016/j.patrec.2011.10.012.

22. Kullback, S. (1959). Information Theory and Statistics (Wiley and Sons).

23. Sy, B., and Gupta, A. (2004). Information-Statistical Data Mining: Warehouse Integration with Examples of Oracle Basics. eBook ISBN: 978-1-4419-9001-3 (Springer). https://doi.org/10.1007/978-1-4419-9001-3.

24. Sy, B. (2017). SEM Approach for TPB: Application to Digital Health Software and Self-Health Management. In 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2017, pp. 1660–1665. https://doi.org/10.1109/CSCI.2017.289.

25. Sy, B., Chen, J., and Horowitz, R. (2019). Incorporating association patterns into manifold clustering for enabling predictive analytics. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 1300–1305. https://doi.org/10.1109/CSCI49370.2019.00243.

26. Duncan, Otis Dudley (1975). Introduction to Structural Equation Models (Academic Press).

27. Diabetes Prevention Program (DPP) Research Group (2002). The Diabetes Prevention Program (DPP): description of lifestyle intervention. Diabetes Care *25*, 2165–2171. https://doi.org/10.2337/diacare.25.12.2165.

28. Knowler, W.C., Fowler, S.E., Hamman, R.F., Christophi, C.A., Hoffman, H.J., Brenneman, A.T., Brown-Friday, J.O., Goldberg, R., Venditti, E., and Nathan, D.M. (2009). 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet *374*, 1677–1686. https://doi.org/10.1016/S0140-6736(09)61457-4.