ORIGINAL RESEARCH

Behavioral Predictive Analytics Towards Personalization for Self-management: a Use Case on Linking Health-Related Social Needs

Bon Sy[1,2,3] · Michael Wassil[3] · Helene Connelly[3] · Alisha Hassan[4]

Abstract

The objective of this research is to investigate the feasibility of applying behavioral predictive analytics to optimize patient engagement in diabetes self-management, and to gain insights on the potential of infusing a chatbot with NLP technology for discovering health-related social needs. In the U.S., less than 25% of patients actively engage in self-health management, even though self-health management has been reported to associate with improved health outcomes and reduced healthcare costs. The proposed behavioral predictive analytics relies on manifold clustering to identify subpopulations segmented by behavior readiness characteristics that exhibit non-linear properties. For each subpopulation, an individualized auto-regression model and a population-based model were developed to support self-management personalization in three areas: glucose self-monitoring, diet management, and exercise. The goal is to predict personalized activities that are most likely to achieve optimal engagement. In addition to actionable self-health management, this research also investigates the feasibility of detecting health-related social needs through unstructured conversational dialog. This paper reports the result of manifold clusters based on 148 subjects with type 2 diabetes and shows the preliminary result of personalization for 22 subjects under different scenarios, and the preliminary results on applying Latent Dirichlet Allocation to the conversational dialog of ten subjects for discovering social needs in five areas: food security, health (insurance coverage), transportation, employment, and housing.

Keywords Self-health management · Behavioral predictive analytics · Natural language processing (NLP) · Latent dirichlet allocation

---

This article is part of the topical collection "Biomedical Engineering Systems and Technologies" guest edited by Hugo Gamboa and Ana Fred.

✉ Bon Sy
bsy@sippasolutions.com

Michael Wassil
mike.wassil@sippasolutions.net

Helene Connelly
helene.connelly@sippasolutions.net

Alisha Hassan
alisha.hassan79@myhunter.cuny.edu

1 Queens College/City University of NY, 65-30 Kissena Blvd, Queens, NY 11367, USA

2 Graduate Center/City University of NY, 365 5th Ave, New York, NY 10016, USA

3 SIPPA Solutions, 42-06A Bell Blvd, Queens, NY 11361, USA

4 School of Public Health, Hunter College/City University of NY, New York, NY 10065, USA

Introduction

(Pre-)diabetes is a chronic disease that affects over 115 million Americans and over 440 million people worldwide. In the U.S., diabetes affects disproportionally among racial and ethnic minority and low-income populations. Due to the health disparity characterized by the Social Determinants of Health (SDoH), personalized intervention and engaging individuals in self-health management are healthcare strategies being actively discussed for achieving health equity (Haire-Joshu 2019).

Some of the risk factors of diabetes are mitigatable through behavior change towards a healthy lifestyle. It has been demonstrated elsewhere (Bollyky 2018) that behavior change can achieve a 10% or more improvement in diabetes symptoms if an individual is engaged in proactive self-management of diabetes.

Self-management is generally accepted as a viable intervention strategy (Hadjiconstantinou 2020). Self-management is the patient's ability to manage their chronic disease through their own activities, such as taking their blood glucose readings 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 evidenced in a survey conducted elsewhere (Volpp 2016), less than 25% of patients are considered 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 the technical feasibility of (1) behavioral predictive analytics, and (2) NLP technology for eliciting Social Determinants of Health (SDoH). The main goal is to optimize the effectiveness of self-management strategies using personalization based on predicting behavior readiness and its relationship to engagement outcomes. A second goal is to determine the feasibility of NLP technology for eliciting SDoH that could inform Health-Related Social Needs (HRSN). In this study, we aim to demonstrate a potential predictive system that delivers personalized actionable health based on the users' behavior readiness and to gain insights on the feasibility of applying (Blei 2003) Latent Dirichlet Allocation (LDA) to discover HRSN based on the SDoH revealed in a conversational dialog.

"Relationship to State-of-the-Art" contains a brief review of the state-of-the-art and the context of this research within it. We will first discuss the Theory of Planned Behavior, and the use of behavior constructs as an attribute vector of behavior readiness. We will also discuss different behavior theories that have been applied to explain health outcomes.

In "Predictive Analytics Foundation", the research results reported elsewhere will be restated as it is applied in this research. In "Predictive Analytics for Personalization" using either an auto-regression model or a population-based model will be discussed. The population-based model provides an alternative mechanism when the auto-regression model derivation fails. This could occur when there are insufficient data, or it fails the statistics test of the model selection process based on Bayesian/Akaike Information Criteria. The foundation of applying LDA for predicting health-related social needs will also be discussed. In "Preliminary Study", we will present the results of manifold clustering based on the attribute vector of behavior readiness of 148 subjects with type 2 diabetes. This will be followed by the results of a preliminary study involving 22 subjects who were in the intervention phase for personalization during the study period. We will then present and discuss the result of predicting health-related social needs based on the conversational dialog of ten subjects. In "Limitations", the limitations of this research will be discussed, followed by the "Conclusions" section sharing our thoughts on the future research plans.

Relationship to State-of-the-Art

In health psychology, behavior models have been developed and applied to address healthcare issues in different settings. Linden et al. [17] summarized several theory-based models that were applied in different health settings. For example, Theories of Organizational Change, and Community Coalition Action Theory, target disease management programs at the community level and focus on the planning and implementation of population-based interventions that influence social norms and structures. On the other hand, models such as Motivational Interviewing and Stages of Change Model focus on the motivation aspect as a driver for behavioral change. One of the characteristics among these models summarized by Linden is their focus on affecting behavioral change on the organizational level or community level. While these models were discussed in terms of the theories behind, applications, and limitations for disease management, their applicability to affect behavioral change on an individual level is less clear.

In contrast to the models focusing on community or organizational level, the Theory of Planned Behavior Model [1], Trans-theoretical Model of Behavior Change (Prochaska 1992), Health Belief Model (Strecher 1997), and IMB (Information Motivation and Behavior Skill) Model (Chandra 2010) are built upon theories of behavioral change on an individual level. These models have been applied to a specific intervention of chronic diseases, and have shown

clinical efficacy. One common thread is 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.

In this research, our behavior model is grounded on the Theory of Planned Behavior [1], with the incorporation of two additional behavior constructs—ownership, and motivation advocated in IMB. The Theory of Planned Behavior (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 (a 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).

The TPB has been applied to study a variety of health-related behaviors, with attitude and perceived behavioral control having the strongest association with intentions and behavior (Armitage 2001). Downs (2005) have reported the efficacy of the TPB to explain physical activity, while Conner (2002) and Sjoberg (2004) reported the effectiveness of TPB to explain diet activity. Blue [6], on the other hand, applies TPB to investigate the cognitive factors relevant to physical activity and healthy eating intentions or behaviors of the diabetic patient population.

In many applications discussed previously, the behavior constructs are often modeled qualitatively. This may suffice if the goal is to explain health outcome efficacy in terms of the behavior theory being applied. In our research, the application intent of the behavior theory is different. Instead of applying the behavior theory to explain the health outcome, our approach is to develop a quantitative behavior model to support the prediction of an individual's behavior readiness for self-health management. Towards this end, Structure Equation Modeling was applied to derive a quantitative behavior model. Further details are presented in the following section, as well as in a report elsewhere [25].

In line with the Theory of Planned Behavior, our research proposes targeting a user's behavioral beliefs to change their attitudes and intentions towards actionable health behaviors. One of the most important features of our approach is to track health activities that reveal information about appropriate health behaviors. In a review of the literature, Fry and Neff [19] found that frequent periodic prompts around: improving diet, increasing physical activity, and weight loss all led to positive results. Tailored prompts were found to be statistically significant in encouraging user engagement. However, for users who are already not engaged, these prompts do little to engage users [3]. Sawesi et al. [22] 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 is an engaging method for improving health behaviors and is cost-effective for behavioral change [27].

Predictive Analytics Foundation

SIPPA (Secure Information Processing with Privacy Assurance) predictive analytics relies on two foundational building blocks developed in the research reported elsewhere [25], [26]. 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. Furthermore, a pilot subject is invited to respond to a survey on social needs, participate in an interview, and optionally participate in interacting with a chatbot. 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 this research, health-related social needs are not included in segmenting subpopulations, because not all subjects agree/have a need to participate in the health-related social needs study. 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. In addition, LDA is applied to discover the keyword mixtures that define the categories of health-related social needs. The details on stage 3 will be presented in the following 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 Theory of Planned Behavior. Structural Equation Modeling [13] was employed to link questions of a survey instrument to the behavior constructs defined by a weighing factor 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 [25].

The second building block is an unsupervised learning approach for discovering manifold clusters. The novelty of manifold clustering is to induce patient subpopulation

clusters based on statistically significant association patterns. This approach is not restricted to only continuous data (number of Real). In other words, this approach could be applied to a data set of mixed-type of both continuous and discrete variables. 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—exceeds a pre-defined threshold according to the domain problem, and (2) the association among the observed values does not happen by chance as measured by the mutual information measure. There are two important results of the manifold clustering technique. First, each manifold cluster has a semantic interpretation characterized by statistically significant association patterns, i.e., the grouping criterion is based on behavior readiness. Second, the manifold clustering does not require linearity assumption as is required by the Principal Component Analysis (PCA). However, 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. While the behavior constructs are related to the Theory of Planned Behavior, variations exist as shown in the confirmative factor analysis regarding the assumption on linearity; i.e., the existence (and strength) of a linear relationship between the behavior constructs that quantify behavior readiness for self-management in a population.

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 "fulfillment" 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 (CR)

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 one's glucose once a day under the clinical recommendation commensurate to one's specific diabetic condition. 30 is the expected number of

self-monitoring measurements. Over this period, the user self-monitors 18 times. 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 engagement ratio is $(15 + 6 + 10)/90 = 31/90$.

Auto-regression and Maximum Likelihood for Prediction

To facilitate the discussion on predictive analytics for personalization, let P be a population consisting of n individuals; i.e., $|P| = n$. $C = \{C1, \dots Ck\}$ is the set of subpopulations obtained by applying the manifold clustering to P as described in "Predictive Analytics Foundation"; where $Ci \subseteq P$, $Ci \cap Cj = \phi$ if $i \neq j$ and $P = \cup_i Ci$. p^j_{Ci} is the jth 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 $\{ER^1, \dots, ER^T\}$. 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 one 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 are "*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 are *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 (Akaike Information Criteria) or BIC (Bayesian Information Criteria) that pass statistical tests.

Information-Theoretic Model Selection Approach

Bayes and Akaike Information Criteria are two common information-theoretic approaches for model selection as stated below:

Bayes Information Criterion (BIC) : $BIC(l) = ln(SSR(l)/T) + [(l+1)ln(T)]/T$ (1)

Akaike Information Criterion (AIC) $AIC(l) = Ln(SSR(l)/T) + 2/T,$ (2)

where l = number of lags,

T = total number of observations,

$SSR(l)$ = 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 $p\text{-}value < 0.05$, 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. Further technical details about manifold clustering based on statistically significant association patterns could be found elsewhere (Sy, 2019).

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

For each subpopulation Ci, derive the population statistical (joint) distribution of ER and ΔER based on the available engagement ratios of all individuals (p^j_{Ci}) in the subpopulation; for $j = 1, 2, \dots |Ci|$. In other words, the joint distribution characterized by $Pr(ER, \Delta ER)$ is derived from using the ER^t and ΔER^{t+1} ($t = 1 \dots T\text{-}1$) 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 ΔER^{T+1}_p based on auto-regression using $T, T\text{-}1, T\text{-}2 \dots T\text{-}l$. If the test statistics in (3) are reasonable (i.e., *p value < 0.05* and *threshold $\leq R^2$*), keep the predicted value ΔER^{T+1}_p and stop. Otherwise continue to step 5.
5. Determine the predicted value ΔER^{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 ΔER^{T+1}_p.

The steps above are repeated for predicting/recommending coaching agenda based on compliance ratio.

Predicting HRSN from LDA Result

Latent Dirichlet allocation (LDA) is a generative probabilistic approach suitable for topic modeling based on a given text corpus. LDA is a three-level hierarchical Bayesian model. When it is applied for topic discovery in natural language processing, a word could be conceived as a finite mixture over an underlying set of topics. Each topic is modeled as an infinite mixture over an underlying set of topic probabilities. In the context of topic modeling, the topic probabilities provide an explicit representation of a document. The classical approach for topic inference is based on variational methods and an EM algorithm for empirical Bayes parameter estimation. Mathematically, the probability distribution of topic (mixture) parameters θ, z_u (a topic), and w (set of words) is defined as below in LDA.

$$Pr(\theta, z_u, w|\alpha, \beta) = Pr(\theta|\alpha)\Pi_{n=1}^{N} Pr(z_u, |\theta) Pr(w_n|z_u, \beta) \quad (3)$$

where θ is a k-dimensional Dirichlet random variable;

$z = \{ z_u \ | \ u = 1..k\}$ is a set of k topics;

$D = \{d_1, d_2, \dots, d_M\}$ denotes a corpus of a collection of M documents;

N is the size of a document d ($\epsilon \ D$); i.e., a document d is represented by a sequence of N words denoted by $w = (w_1, w_2, \dots, w_N)$, where w_n is the nth word in the sequence;

V is vocabulary size (i.e., # of unique word tokens) that defines the index of a word represented by a $Vx1$ unit-basis vector (i.e., one entry of "1" and all others "0");

β is a hyperparameter represented by a kxV matrix parametrizing word probabilities with $\beta_{ij}=Pr(w_j=1|z_i=1)$;

α is a Dirichlet model hyperparameter of k dimensions for EM (Expectation-Maximization); i.e., $\theta \sim Dir(\alpha)$.

In this research, a conversational dialog obtained from a chatbot/an interview such as below are analyzed via LDA:

P1: "I asked the MTA bus driver for a ride."
P2: "My mother threatens to kick me out of the house if I do not find a job."
P3: "I search for events on campus that are serving food, so I can eat the leftovers."

The goal of applying LDA is to obtain topic probabilities that reveal the distributional mixtures of words appearing in a topic. For example, topical focus on disparity within the context of SDoH and health may include (topic 1) lack of transportation, (topic 2) housing insecurity, (topic 3) food insecurity, and (topic 4) employment. An example of distributional mixtures of the four topics is shown.

Topic 1: 0.04*train + 0.04*bus + 0.02*ride
Topic 2: 0.08*house + 0.05*campus
Topic 3: 0.08*food + 0.01*eat + 0.01*leftovers
Topic 4: 0.09*job + 0.01*events.

The keywords of each of the three sample conversations shown previously for *P1—P3* will be similarly derived via automatic keyword extraction methods such as YAKE (Campos 2020) or RAKE (Rose 2010). An example outcome for *P2* is shown below.

P2: 0.5*house + 0.5*job.

In this research, predicting social needs for each of the *Pi* (for *i = 1.0.3*) is based on calculating the product between an LDA-based inference matrix *M* and keyword distribution of *Pi* (for *i = 1...3*) discovered by YAKE (Campos 2020) from the conversational dialog obtained via an interview/chatbot interaction; where *M* represents the distributional mixtures defining the topics on disparity (i.e., *topic 1.. topic 4*).

Let $w = \{w_1, \ldots w_N\}$ be the set of unique word list. In this example, $N = 10$ with $W = \{train, bus, ride, house \ldots job, events\}$. *M* is a 4×10 matrix (*4 topics, 10 unique word tokens*) with the first row being *[0.04 0.04 0.02 0 0 …0]*, and the last row being *[0 0 … 0 0.09 0.01]*. Furthermore, the weights of the keyword distribution of *P2* can be represented by $TP_2 = [0\ 0\ 0\ 0.5\ 0\ \ldots\ 0.5\ 0]$. $M\ x\ TP_2^T = V$ returns a vector $[v1 \ldots v4]^T$ where vi ($i = 1...4$) is the dot product revealing the degree of match indicating health-related social needs; e.g., the vector *V* derived from $M\ x\ TP_2^T = [0, 0.04, 0, 0.045]^T$ shows that *P2* may experience employment and housing insecurity.

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/Caucasian	1.10%
Indian/Asian	1.10%
Mexican/Black	1.10%
Income (in U.S. $):	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-yr college degree	16.26%
4-yr 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%

Preliminary Study

The predictive analytics approach discussed in the previous sections was applied to the diabetes subjects of a self-health management pilot conducted under an IRB-approved study protocol (CUNY IRB #2018–1043). The objective was to investigate the impact of digital health solutions to affect individuals' behavior towards self-management of chronic diseases, particularly type 2 diabetes.

To be included in the study, the participants had to be at least 18 years old. They also 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
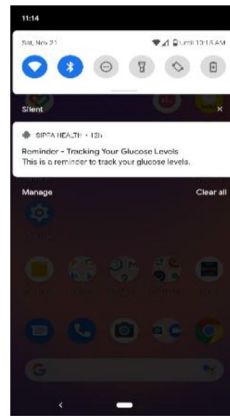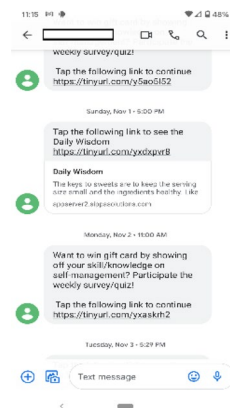
Fig. 1 Push notification

Fig. 2 SMS reminder

Fig. 3 In-app service

Fig. 4 Weekly survey

diagnosis at risk of diabetes or pre-diabetes. This means that participants also had a perceived risk of developing or had already 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 for predicting behavior readiness.

148 individuals with type 2 diabetes were involved in stage two of the preliminary study. These participants had a mean age of 49 and a mean H1AC of 7.89. The population characteristics are shown in Table 1.

These 148 individuals participated in this pilot on a rolling basis. During the reporting period of this research, some were still in a 1-month hold period for establishing a baseline without intervention; i.e., they have not entered the pilot phase for personalized intervention. Among the rest, 49 subjects completed/ entered the intervention phase of the study. During the intervention phase, the self-health management focused on the following three health coaching agenda items:
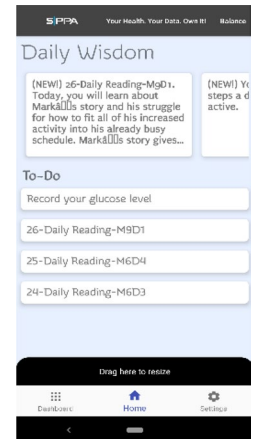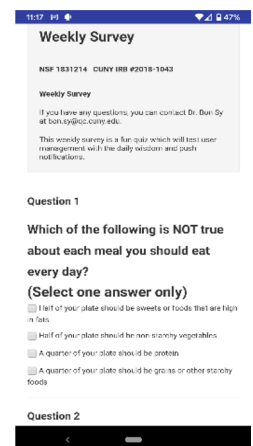
- Knowledge building and information gathering (through daily wisdom sent via SMS and/or push notifications).

- Discipline and skill development (through notifications and reminders).

- Awareness improvement (through a weekly survey).

The self-health management activities of this pilot included the delivery of (1) daily wisdom on diabetes management, (2) text messaging, and/or (in-app) notification reminders on the 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 of self-management. An example of each of these is shown in Figs. 1, 2, 3, 4. This study will focus on only a retrospective analysis based on compliance ratio, and a forward-looking prediction based on engagement ratio, 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 protocol, the SIPPA Health platform collects de-identified 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 data from the 148 subjects of this pilot were used for the manifold clustering to identify subpopulation characteristics defined by behavior readiness. The data from the 49 subjects who completed/entered the "intervention" phase were used to derive the population-based models (Sect. "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.

Using the behavior readiness of 148 subjects as training data, the 49 subjects in the intervention phase were covered by four manifold clusters. Each of them was assigned to a cluster based on the similarity between the individual's behavior readiness expressed as a pattern and the behavior patterns exhibiting statistically significant association that defines the cluster. Further details on the similarity distance function could be found elsewhere (Sy, 2019).

Within each cluster subpopulation, a normalized compliance ratio and an engagement ratio 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, 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 4 (Sect. "Predictive Analytics for Personlization") during the model selection process using *BIC/AIC*. Typically, this happens when a subject is in the intervention phase for less than 4 weeks.

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 *nxm* 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*.

Self-management Study

The subjects included in this study were distributed across four different clusters (subpopulations). On average, a participant spent 13 weeks in the intervention phase. 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 of 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 a forward-looking prediction of the engagement. The purpose of retrospective analysis is to establish a base reference for performance assessment based on historical results. The forward-looking prediction is for evaluating the prediction performance as a time-series on a rolling basis in real time.

Retrospective Analysis The predictive analytics will 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 Sect. "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.
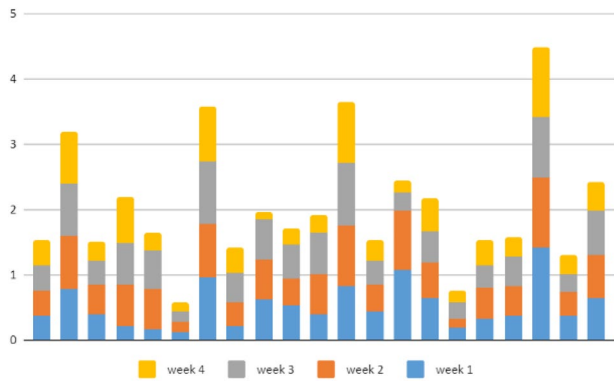
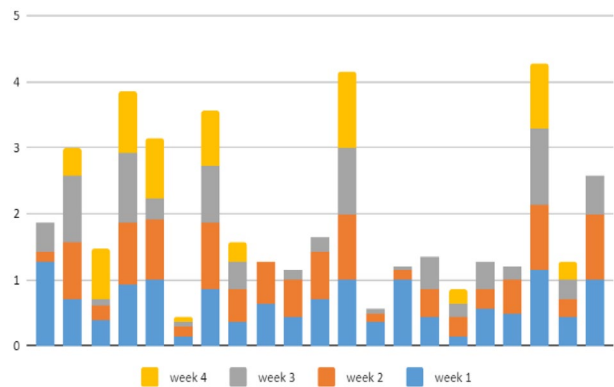**Fig. 5** Predicted compliance ratio for a subject

Fig. 7 Average predicted vs observed CR

Table 2 R and p values of (2.5 months) r 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	0.000045	0.0004

Table 3 Hypothesis testing results for each cluster

	t Statistic	p Value
All data without clustering	0.51758	.303733
Cluster 1	0.32971	0.372949
Cluster 2	1.79319	0.061554
Cluster 3	− 0.48247	0.319928
Cluster 4	− 0.10798	0.459604

Fig. 6 Observed compliance ratio for a subject

To gain insight into such scenario 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 (who completed or entered the intervention phase). Therefore, manifold clustering is required for this particular use case on applying the algorithm described in Sect. "Compliance Ratio (CR)".

The compliance ratio is computed on the 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 is used to predict the compliance ratio of the last 4 data points for evaluation purposes.
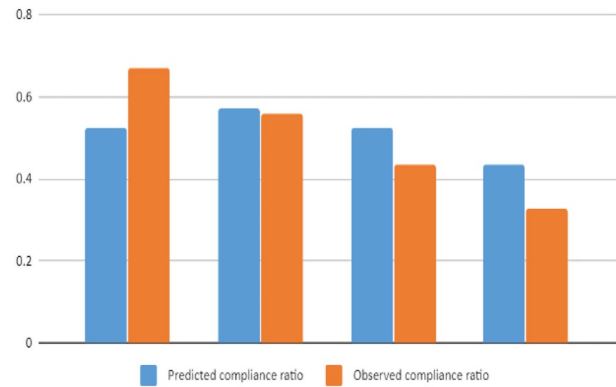
Forward-Looking Prediction In contrast to the retrospective analysis, forward-looking prediction involves only those subjects who were in the intervention phase during the study period. There are 22 of them out of the 49 subjects. Their engagement ratios were used for the predictive analytics task.

The engagement ratio of each active subject was computed on a weekly basis. Similar to the retrospective analysis, an estimated engagement ratio is derived for each week based on the predictive analytics technique described in Sect. "Predictive Analytics for Personalization". The prediction was performed by looking forward. For example, the prediction on engagement ratio 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 observed engagement ratio was recorded at week n. This forward-looking prediction process was repeated ten times in the 11-week study period.

Results of Predictive Analytics

Retrospective Analysis on Compliance Ratio

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 Fig. 7, there is a consistent pattern across the 4-week prediction period. 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 the 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).

Forward-Looking Prediction on Engagement Ratio

In the forward-looking prediction experiment for engagement ratio (instead of compliance ratio), the prediction is on actionable health recommendations based on the maximal posterior estimate as described in Sect. "Predictive Analytics for Personalization". In this study, the personalized actionable health recommendation would be either diet management or exercise. 22 subjects were in the intervention phase during this period of research.

Figures 5 through 7 show evidence of its accuracy and consistency for predicting compliance ratio. However, we are also interested in the effectiveness of the technique for predicting engagement ratio to improve self-efficacy. To evaluate its effectiveness for improving self-efficacy on
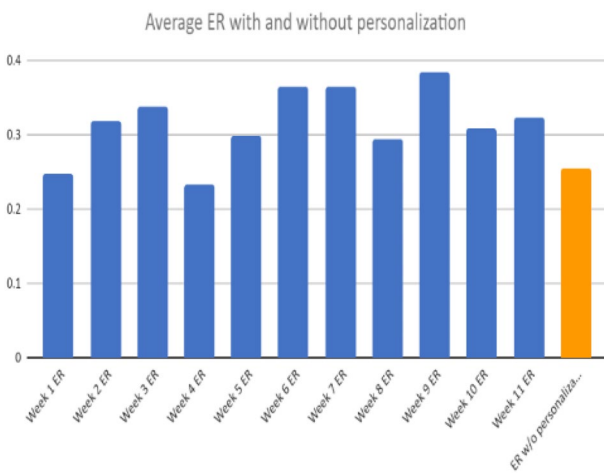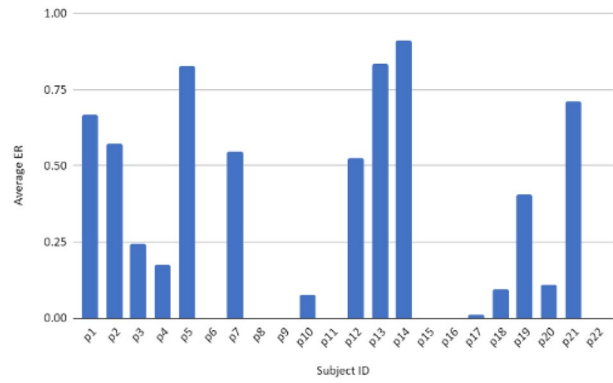

Fig. 9 Individual ER average (over 11 weeks)

health management, this study also attempts to show personalized actionable health (recommended by the behavioral predictive analytics) resulting in a more active engagement when it is compared to that without personalization.

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

In calculating the engagement ratio without personalization, the average engagement ratio of each subject overtime prior to personalization is first calculated, then the average over all the subjects. Note that the average engagement ratio of each subject over time prior to personalization spans over different periods and lengths, as well as the actionable health recommendations because of the rolling nature of the subject participation in the pilot.

Figure 9 shows the engagement ratio of each individual averaged over the participation period. There are half a dozen subjects with a low/zero engagement ratio. All of them received follow-up from this research team to
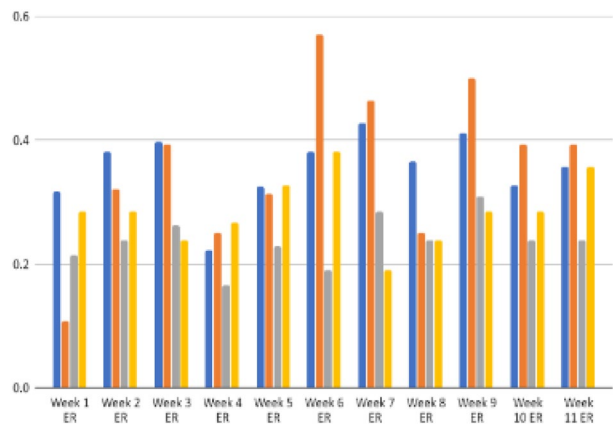

Fig. 8 Aggregated ER w(/o) personalization

Fig. 10 Observed ER by subpopulation clusters

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 Fig. 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.

Discussion

Experimental Results The results shown in Figs. 5 through 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.

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 Figs. 9 and 10 in the forward-looking experiment demonstrate the practical implementation feasibility. The results in Fig. 10 also reveal indirect evidence of the effectiveness of the manifold-based clustering technique for grouping subjects into subpopulations using 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, in particular, the second cluster subpopulation that is not as high performing at the beginning.

Finally, the overall average engagement ratio with personalization had a mean value of 0.31 with a standard deviation of 0.33. The 95% confidence interval around this was [0.17, 0.45]. By contrast, without personalization, the overall mean engagement ratio is 0.26 with a standard deviation of 0.31. The 95% confidence interval for this value was [0.13, 0.38]. These are overall promising results; however, with such large standard deviations, one of the next steps in the research would be to gather larger samples to mitigate this issue.

Hypothesis Testing Although the results shown in the previous figures are encouraging, it is necessary to conduct a hypothesis test analysis to understand the extent of improvement with clustering and personalization, as well as its statistical significance.

In reference to the results of the forward-looking prediction shown in Figs. 8, 9, 10, an analysis was conducted to understand the effect of the population size on the statistical power. In particular, is the change in engagement ratio reported in this study generalizable?

This question was approached by conducting a t test to compare the difference between the means of the engagement ratio with personalization and without personalization for the entire sample and within each cluster by investigating such change of each participant over the 11 weeks of the study.

While the t-statistic shows an overall improvement on engagement ratio when personalization is applied—irrespective of clustering, and a more significant improvement with clustering, none shown in Table 3 passes the p value test for the result to be generalizable. This suggests that the study will need a larger population to achieve a power that allows the result to be generalizable.

Health-Related Social Needs Study

Study Protocol

A smaller subset of the pilot participants with potential health-related social needs (HRSN) participated in the HRSN study. This smaller group of ten pilot participants was engaged in the following three activities:
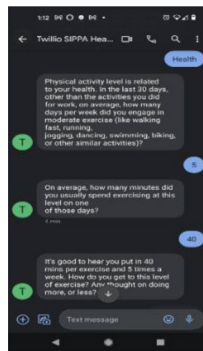
1. Online health-related social needs survey—SIPPA-HRSN.
2. One-on-one interview for eliciting social needs in five areas: health, employment, transportation, housing, and food security.
3. Chatbot interaction in an open-ended conversational dialog.

SIPPA-HRSN is a reduced set of survey questions from two validated surveys. The first is the Accountable Health Communities HRSN survey (Billioux 2017) developed by the Centers for the Medicaid and Medicare Services (CMS) in the U.S. The second is PRAPARE developed by the National Association of Community Health Centers NACHC-PRAPARE [18] (National Association of Community Health Centers website. https://www.nachc.org/research-and-data/prapare/). 2016. They both aim to capture the social determinants of health to inform health-related social needs. Both surveys are validated and comprehensive in terms of coverage areas ranging from the lack of transportation's impact on medication refill to housing security due to domestic violence. As CMS moves towards codified reimbursement based on the z-codes in ICD for social services, AHC-HRSN is appealing in that z-code findings from the survey response could be particularly useful for identifying providers for social service referrals. PRAPARE,

Fig. 11 In-app chatbot

Fig. 13 Topical keywords alluding health

Fig. 12 SMS chatbot

on the other hand, addresses the needs for SDoH screening in a clinical setup. In particular, the survey is designed for capturing SDoH through survey responses that could be encoded into an Electronic Medical Record (EMR) under an interoperable format.

Since this research is focused on specific areas of social needs mentioned before (i.e., health, employment, transportation, housing, and food security), it would not be appropriate to utilize the entire screening tool of AHC-HRSN or PRAPARE. Instead, only questions relevant to the areas of our focus are included in our survey SIPPA-HRSN rather than questions for identifying social needs due to, say, mental health. In doing so, SIPPA-HRSN could leverage the strengths of both surveys to not only capture the health-related social needs of pilot participants, but to (1) facilitate the encoding of the findings into an EMR, and (2) identify the z-code(s) for matching participants with providers through referral services.

Study Design

While validated surveys are effective in capturing the "what" part of the health-related social needs, structured responses coded as multiple choices for a survey question often are limited to reveal the "how" and "why" part of the needs. For example, a question in AHC-HRSN "*What is your living situation today?*", a response "*I have a place to live today, but*
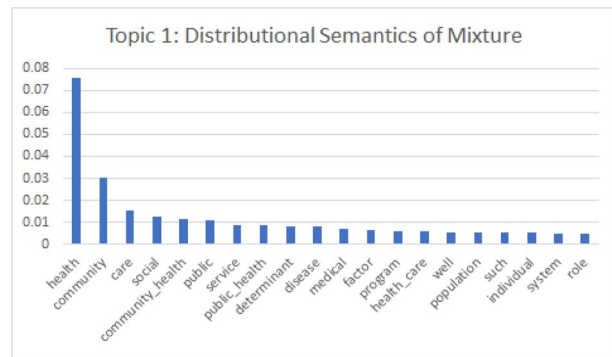
I am worried about losing it in the future." does not reveal why one is worried and how the living situation may change. Is it because of a rent hike that is no longer affordable? Is it because of aging or a health-related issue that becomes a problem living on a top floor of a multi-story building without an elevator, or something else? Understanding the underlying reason(s) could improve the quality and success rate of referral services.

The one-on-one interview is included in this study that attempts to better understand the "how" and "why" parts of the social needs. In addition, a pilot participant was also invited to interact with a chatbot available via in-app service or text messaging shown in Figs. 11 and 12.

By cross-referencing different sources including the survey response, one-on-one interview, and/or chatbot interaction, insights could be drawn from empirical observations to also better understand the concerns on the stigma associated with social needs inquiry—a factor often impedes the openness required for eliciting the information on health-related social needs. Furthermore, the unstructured conversational dialog during the one-on-one interview and chatbot interaction also provide data to better understand the feasibility of applying topic modeling via LDA to infer social needs and its consistency in relation to the survey responses.

LDA Implementation and Setting

Gensim is a Python NLP (Natural Language Processing) software library providing an implementation of LDA. The implementation of LDA in Gensim provides a set of arguments for pre-processing. The following shows the setting for the arguments in the calibration process that locally optimizes the topic coherence:

Data source for lemmatization: WordNet.

Occurrence frequency for bigram construction: 400.

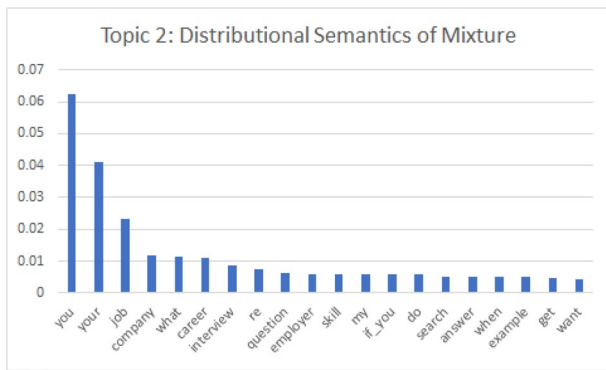Threshold for word filtering: $> 65\%$ of documents.

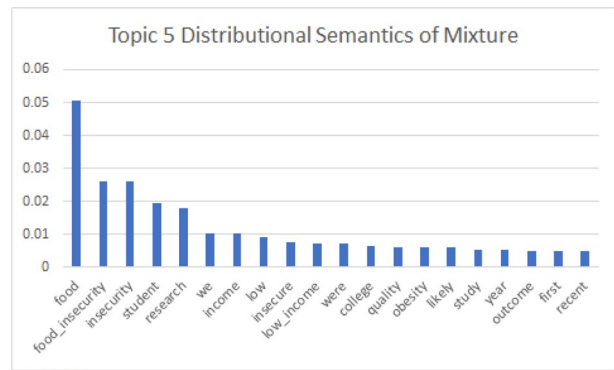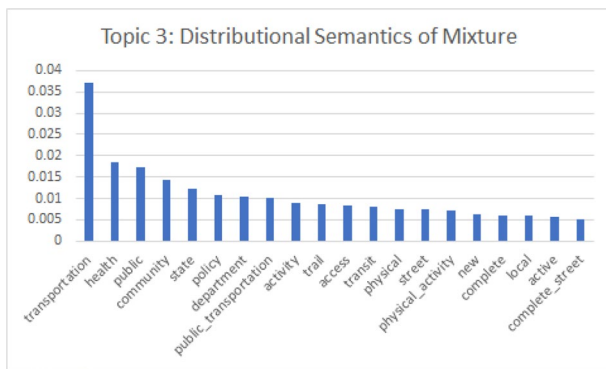**Fig. 14** Topical keywords alluding to employment

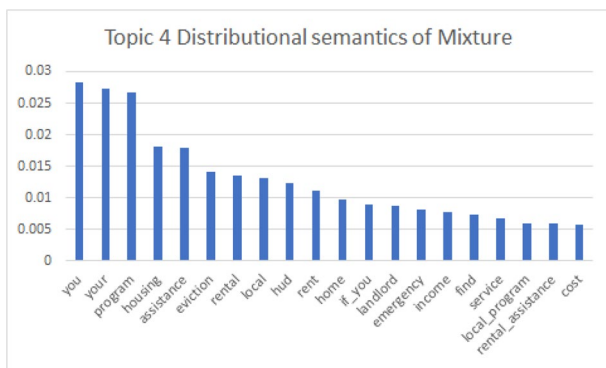**Fig. 15** Topical keywords alluding to transportation

Fig. 16 Topical keywords alluding to housing

To determine the setting for the arguments during the calibration, a corpus of documents focusing on health-related social needs was employed. The corpus consists of 55 documents with 2747 unique tokens. A token is either a word or a bigram. Each document targets a specific topical area of health-related social needs mentioned before. These documents were drawn from a pool comprised of information/

Fig. 17 Topical keywords alluding to food security

questions/comments common during the in-take process for social service referrals.

Results

Calibration and Parameter Tuning In the calibration process for the arguments of LDA implementation for pre-processing, the criteria for determining the setting for the arguments are:

1. Trial and error to optimize Topic Coherence score based on the conditional information (log of conditional probability) measure.
2. Interpretability of the mixture defined by Topic Coherence.

The basis of the calibration is the corpus mentioned in the previous section that targets the following areas of social needs: health, employment, transportation, housing, and food security.

In the calibration process, LDA was set to analyze with 1000 iterations per document, and 500 passes for the entire corpus. The topic coherence based on Umass that measures the conditional information content in form of the log of conditional probability is *-0.7745*. The optimized Dirichlet model hyperparameter α is *[0.06872154, 0.022918368, 0.018769965, 0.031925187, 0.053812448]*. Figures 13, 14, 15, 16, 17 show the top 20 distributional mixtures of the five topics.

Predicting Social Needs To predict the possible social needs of a participant in the pilot study, the distributions shown in Fig. 13, 14, 15, 16, 17 were used to formulate LDA-based inference matrix *M* described in Sect. "Predicting HRSN from LDA Result". In this study, *M* is a 5×100 *block diagonal* matrix where each row anchors on the 20 keywords of
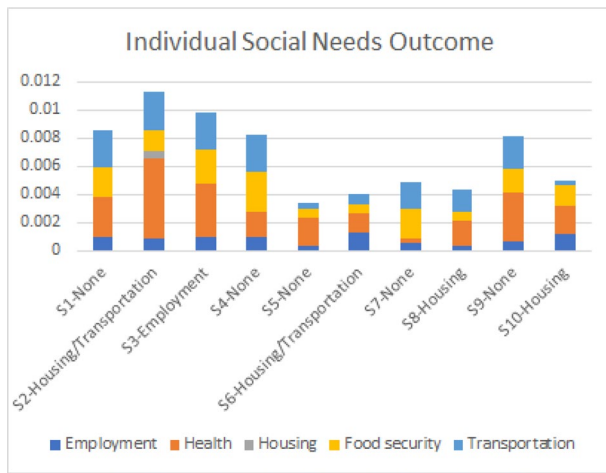
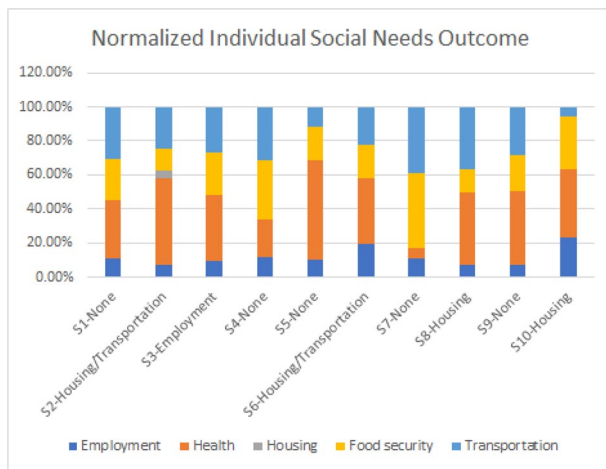**Fig. 18** Individual social needs distribution

Fig. 19 Individual social needs distribution

a topic with the corresponding weights discovered by LDA spanning over 20 entries in each of the five diagonal blocks.

The conversational dialog obtained from an interview/interaction with a chatbot is analyzed for keywords extraction using YAKE (Campos 2020). YAKE operates on a single document to identify keywords based on the text's statistical features. It should be noted that LDA requires multiple documents in a corpus for topic discovery; thus, it is inappropriate when the conversational dialog is transcribed to a single document per pilot participant.

For each pilot participant, a text document is generated based on the transcription of the conversational dialog. YAKE is applied to the text document to extract the top 20 keywords for capturing the topical focus of potential social needs. A 100×1 query vector is composed based on the matching keywords between the keywords returned by YAKE and the topical keywords derived by LDA.

Mathematically, the 100×1 query vector is constructed as below:

Let $H(i)$ be a mapping function that returns a keyword term derived by LDA for $i = 1...100$. There are five topics, and each topic consists of 20 mixture terms that could be a keyword or a *n-gram* word; thus, resulting in 100 keyword terms. The value at the jth entry of a query vector ($j = 1...100$) is either the weight of $H(j)$ derived by YAKE if $H(j)$ appears as a keyword term identified by YAKE, or zero if $H(j)$ is not identified as a keyword by YAKE.

The LDA-based inference matrix M (5×100) is then multiplied with the 100×1 query vector just mentioned to produce a 5×1 vector that predicts the potential needs in the five categories of this study; i.e., health, employment, transportation, housing, and food security. This vector is then normalized to show the relative needs in the five categories if there is at least one non-zero entry. Below shows one such example:

Topic/social needs categories	Needs-estimators	Normalized needs-estimator
Employment	0.000487231	6.49%
Health	0.002917556	38.84%
Housing	0	0.00%
Food security	0.002050444	27.30%
Transportation	0.002056469	27.38%

Figure 18 shows the distribution of the predicted social needs across the five categories of each pilot participant. The social needs indicated by each individual through the SIPPA-HRSN survey are encoded in the label for each participant. For example, S1-None means that in the survey the participant S1 did not indicate the need for any social support, while S2 indicated possible interest in the social service support for housing, and transportation. Figure 19 shows the normalized distribution of the social needs across five categories of each individual.

Discussion

Figure 18 shows the social needs predicted using LDA and YAKE. While the predicted overall needs for the ten participants vary, in all except three cases, the relative percentage of social needs related to transportation and health is higher than the other three categories. This is consistent with the distribution shown in Fig. 19 when we focus on the individual (as opposed to the population) normalized social needs. In addition, there is an interesting observation on this proposed approach. It correctly predicts virtually no transportation need for the subject S10 as is indicated in the survey by S10. However, a negligible

transportation need appears in Fig. 19. By tracing back to the conversational dialog, S10 has mentioned on one occasion that the subject's mother needed a ride from the subject's uncle to the hospital. It is because the subject's mother was infected by COVID-19 and all car services were reluctant to transport a COVID-19 patient.

In reference to the label that encodes the category of social needs indicated by each individual in the survey response, the original design of the study is to use the survey response to evaluate the usefulness and effectiveness of the prediction outcome based on the conversational dialog. While the areas of social needs indicated by the pilot participants were revealed by the prediction outcome based on the conversational dialog, it is also noted that in every case, the prediction outcomes include the social needs in categories not indicated in the survey response.

In reviewing the survey responses, two participants indicated not currently employed, while the third one was underemployed. Furthermore, all three did not indicate interest in employment assistance. Yet, in the conversational dialog, a significant portion of the conversations was centered on employment in all three cases, and two even indicated financial strains. The inconsistency between the conversational dialog and the survey response suggested the need for a statistical reliability test if the survey response and the conversational dialog are to be used for cross-validation.

In addition, in one case, the pilot participant mentioned a living condition issue in both the survey and the conversational dialog. However, the prediction outcome missed it. In reviewing the topical mixtures of LDA and the keyword extraction outcome of applying YAKE, it is noted that the topical mixtures of LDA are focused on housing assistance, while the context of the survey on housing is on living conditions. This suggests the need for broadening the scope of coverage (related to housing) during the model training phase of LDA, and including training materials to cover not just housing assistance but also living conditions.

Finally, it is not yet fully understood the effect of keyword extraction on the query vector composition. RAKE (Rose 2010) is another well-known keyword extraction method. While YAKE is unsupervised, RAKE allows granular control on the list and the frequency definition for a stop word. As a result, conversational stop words such as "yeah" could be filtered by RAKE through the definition of stop words, while YAKE will rely on the statistical distribution of the local features. In addition, when n-gram words $(n > 1)$ are allowed in a query vector, more n-gram words are extracted as keywords by RAKE in comparison to YAKE. The differences between various keyword extraction methods are topics subject to further research in the future.

Chatbot Acceptability Assessment

As discussed in the previous section, individuals may respond differently depending on the communication modalities. In the implementation of chatbot, an individual could interact with the chatbot either as an in-app service or via text messaging. Similarly, a survey is made available online via a web browser on a desktop or laptop, or as an in-app service. The goal is to remove the access barrier for an individual to engage in the process of discovering health-related social needs.

It is conceivable that one communication modality could be more effective than another. A technical assessment question of interest is chatbot acceptability, and more specifically, the feasibility of a chatbot eliciting user information regarding social determinants of health. Towards answering this assessment question, the conversational dialog of chatbot interaction of individuals with diabetes was used to predict the survey response. Four human evaluators participated in the prediction process based on the information extracted and synthesized from the chatbot interaction. Their predictions on the survey response were then compared against the actual survey response of an individual. Using the actual survey response as a reference, the metrics for assessing chatbot feasibility are the recall and precision measures:

Recall = # of questions with a predicted response/total # of questions.

Precision = # of questions with a correct prediction/ # of questions with a predicted response.

By averaging the recall measure of an evaluator over the number of subjects, the recall measure for each of the four evaluators falls into the range between 0.62 and 0.67; whereas the consistency across the evaluators P1 to P4 in terms of the R score and p value (in bracket) is shown below:

	P1　P2	P3	P4
P1	0.9857 (0)	0.9857 (0)	0.9725 (0)
P2		1 (0)	0.9738 (0)
P3			0.9732 (0)

It is noted that the recall [0.62, 0.67] is encouraging; i.e., the information elicited from the chatbot interaction is applicable to two-third of the survey questions on SDoH. The R scores and the zero p value indicate a consistency across the evaluators.

By averaging the precision measure of an evaluator over the number of subjects, the precision measure for each of the four evaluators falls into the range between 0.85 and 0.88; whereas the consistency across the evaluators P1–P4
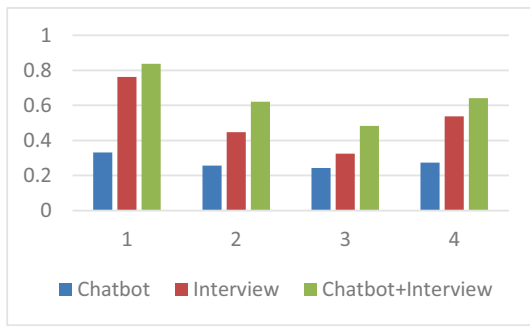
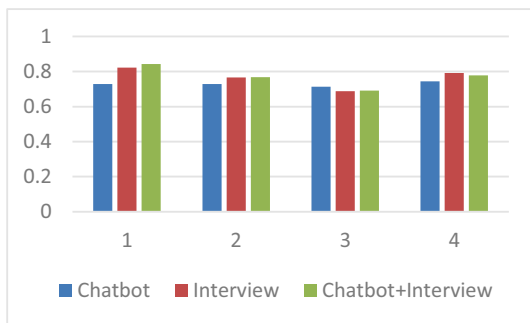**Fig. 20** Cross modalities recall rate

Fig. 21 Cross modalities precision rate

in terms of the R score and p value (in bracket) is shown below:

	P1	P2	P3	P4
P1		0.995 (0)	0.995 (0)	0.6752 (0.01598)
P2			1 (0)	0.9959 (0)
P3				0.9959 (0)

The precision ranges over [0.85, 0.88] is an evidence of the usefulness of the information collected via the chatbot. And the R scores towards 1 and the p values towards 0 is a strong indicator of the consistency among the evaluators in synthesizing the survey response from the chatbot interaction.

Assessment specific to HRSN When the analysis is repeated for predicting the survey response on not just SDoH but Health-Related Social Needs (HRSN), the results were less encouraging. The recall measure for each of the four evaluators falls into the range between 0.23 and 0.33; whereas the consistency across the evaluators P1 to P4 in terms of the R score and p value (in bracket) is shown below:

	P1	P2	P3	P4
P1		0.907 (0)	0.5496 (0.04175)	0.8219 (0.0031)
P2			0.6718 (0.0085)	0.933 (0)
P3				0.6568 (0.01072)

The precision measure for each of the four evaluators falls into the range between 0.71 and 0.76; whereas the consistency across the evaluators P1 to P4 in terms of the R score and p value (in bracket) is shown below:

	P1	P2	P3	P4
P1		0.2445 (0.3995)	0.1833 (0.53)	0.3911 (0.1668)
P2			0.7305 (0.003)	0.3354 (0.241)
P3				0.06814 (0.817)

Although the consistency across different evaluators of recall rate remains statistically valid, the recall rate on covering the HRSN [0.23, 0.33] is significantly lower than that of SDoH [0.62. 0.67]. On the other hand, when the precision measure for HRSN [0.71, 0.76] is comparable to that of SDoH [0.85, 0.88], it did not pass the p value test except one case (P2 and P3).

Assessment on Hybrid Approach To better understand the use case of chatbot, the analyses were repeated by comparing the result with the in-person interview; i.e., how does the recall/precision rate differ from each other when the user interacts with only the chatbot, with only the in-person interview, and that by combining the two? Figs. 20 and 21 show the differences among the three settings.

The results shown in Figs. 20 and 21 are insightful. Since it is expected that the interview will cover a more comprehensive scope as shown in the result of the recall measure, a main question is that whether the scope covered by chatbot is embedded in that of the in-person interview. If this is the case, then the recall measure utilizing the outcomes of both the chatbot and in-person interview will not improve the recall outcome. Similarly, how would the precision change by combining the information obtained from the chatbot interaction and the in-person interview? If the information contradicts each other, the precision performance will degrade.

Figure 20 shows that the scope of coverage for eliciting SDoH and HRSN was significantly improved when the process incorporates both the chatbot and in-person

interview. Furthermore, Fig. 21 shows that the precision measure did not degrade.

In other words, the qualitative analyses just shown suggest that the best practice is to integrate "human into the loop" for deploying NLP/data analytics technology for eliciting SDoH and HRSN.

Limitations

In this research, the preliminary results are limited by the sample population size. As a result, qualitative analysis was performed to gain insights for designing a larger scale study that could provide statistical evidence on the feasibility of the approach; i.e., enabling self-health management with personalization–at scale—for population health management.

There are also human factors that need to be explored in further analyses. These include time spent in the training period, level of proficiency with technology, and demographic features that can impact engagement such as gender and socioeconomic status.

Two technical factors related to the population-based model are also noteworthy. First, the population-based model approach is non-parametric and could potentially be sensitive to the additional data available over time that could change the behavior of the model as measured by information-theoretic entropy. Second, when a personalized recommendation is based on the population model, it should be noted that the prediction strategy is a "greedy" approach.

In reference to step 5 of the algorithm in Sect. "Predictive Analytics for Personalization" that determines the predicted value ΔER^{T+1}_p based on $Max\ Pr(\Delta ER^{T+1}_p| ER^T)$, a large ΔER^{T+1}_p is unlikely to come from a large ER^T. For example, if $ER^T = 0.9$, it is not possible for $\Delta ER^{T+1}_p > 0.1$; or $Pr(\Delta ER^{T+1}_p > 0.1| ER^T = 0.9) = 0$. Therefore, the "greedy" approach has an inherent bias to work better in personalization for those who are moderately active compared to others.

Conclusions

A behavioral predictive analytics approach was presented for self-management personalization. The personalized recommendation is based on the engagement outcomes that reveal the behavior readiness of an individual in self-management. Auto-regression and population models were derived to support the proposed predictive analytics approach for generating personalized recommendations. A limitation of this research is the requirement for a "wait" period to accumulate

sufficient data to derive a personalized auto-regression model. In this research, we adopt a 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 improvement on engagement. We do not yet know how this affects engagement and at what pace. Our future research will focus on understanding this aspect.

In regard to discovering health-related social needs based on social determinants of health via LDA, there are two future research goals. The first future research is to investigate the effect of employing different keyword extraction methods on discovering health-related social needs. The second future research goal is to develop an inference scheme for referral service generation that will match the social needs of an individual with the social service providers.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest involved with any affiliation not listed in this paper.

References

1. Ajzen I. Attitudes Personality Behav. Dorsey Press, 1988.
2. Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: a meta-analytic review. Br J Soc Psychol. 2001;40(4):471–99.
3. Bidargaddi N, Pituch T, Maaieh H, Short C, Strecher V. Predicting which type of push notification content motivates users to engage in a self-monitoring app. Prevent Med Rep. 2018;11:267–73. https://doi.org/10.1016/j.pmedr.2018.07.004.
4. Billioux A, Verlander K, Anthony S, Alley D. Standardized screening for health-related social needs in clinical settings: The accountable health communities screening tool. Discussion Paper, National Academy of Medicine, Washington, DC. 2017. https://nam.edu/wp-content/uploads/2017/05/Standardized-Screening-for-Health-Related-Social-Needsin-Clinical-Settings.pdf
5. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res. 2003; 33/1/2003: 993–1022. https://dl.acm.org/doi/pdf/https://doi.org/10.5555/944919.944937

6. Blue CL. Does the theory of planned behavior identify diabetes-related cognitions for intention to be physically active and eat a healthy diet? Public Health Nurs. 2007;24(2):141–50.

7. Bollyky JB, Bravata D, Yang J, Williamson M, Schneider J. 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;2018:3961730. Published 2018 May 16. https://doi.org/10.1155/2018/3961730

8. Campos R, Mangaravite V, Pasquali A, Jatowt A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. In Inform Sci J Elsevier. 2020;509:257–89.

9. CDC. National Diabetes Statistics Report. 2020. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

10. Chandra Y, Osborn K, Amico R, Fisher WA, Egede LE, Fisher JD. An information-motivation-behavioral skills analysis of diet and exercise behavior in Puerto Ricans with diabetes, J Health Psychol. 2010; 15(8): 1201–1213. https://doi.org/10.1177/1359105310364173.

11. Conner M, Norman P, Bell R. The theory of planned behavior and healthy eating. Health Psychol. 2002;21(2):194–201.

12. Downs DS, Hausenblas HA. The theories of reasoned action and planned behavior applied to exercise: a meta-analytic update. J Phys Act Health. 2005;2(1):76–97.

13. Duncan OD. Introduction to structural equation models. New York Academic Press. 1975..

14. Hadjiconstantinou M, Schreder S, Brough C, et al. Using intervention mapping to develop a digital self-management program for people with type 2 diabetes: tutorial on MyDESMOND. J Med Internet Res. 2020;22(5):e17316. Published 2020 May 11. https://doi.org/10.2196/17316

15. Haire-Joshu D, Hill-Briggs F. The next generation of diabetes translation: a path to health equity. In Annual Rev Public Health. 2019;2019(40):391–410.

16. Kan MPH, Fabrigar LR. Theory of planned behavior. In: Zeigler-Hill V., Shackelford T. (eds) Encyclopedia of personality and individual differences. Springer, Cham; 2017.

17. Linden A, Butterworth SW, Roberts N. Disease Management Interventions II: What Else Is in the Black Box?, Dis Manag. 2006; 9(2): 73–85.https://doi.org/10.1089/dis.2006.9.73

18. NACHC-PRAPARE. National Association of Community Health Centers website. 2016. https://www.nachc.org/research-and-data/prapare/.

19. Neff R, Fry J. Periodic prompts and reminders in health promotion and health behavior interventions: Syst Rev. J Med Int Res. 2009; 11(2). https://www.jmir.org/2009/2/e16 . https://doi.org/10.2196/jmir.1138

20. Prochaska J, DiClemente C, Norcross J. In search of how people change: applications to addictive behaviors. Am Psychol. 1992;47(9):1102–14.

21. Roblin N, Little M, Mcguire H. Diabetes Self-efficacy questionnaire (DSEQ) outcome measurement for diabetes education. 2004. https://www.semanticscholar.org/paper/DIABETES-SELF-EFFICACY-QUESTIONNAIRE-(DSEQ)-OUTCOME-Roblin-Little/b94747994e18f744b206788782fb9830d7d9543d?sort=relevance&citationIntent=methodology

22. Sawesi S, Rashrash M, Phalakornkule K, Carpenter JS, Jones JF. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR Med. Inform. 2016;4(1):e.1 https://doi.org/10.2196/medinform.4514

23. Sjoberg S, Kim K, Reicks M. Applying the theory of planned behavior to fruit and vegetable consumption by older adults. J Nutr Elder. 2004;23(4):35–46.

24. Strecher VJ, Champion VL, Rosenstock IM. The health belief model and health behavior, In D. S. Gochman (Ed.), Handbook of health behavior research I. Personal and social determinants (pp. 71–91). New York: Plenum Press; 1997.

25. Sy B (2017) SEM Approach for TPB: application to digital health software and self-health management, 2017 international conference on computational science and computational intelligence (CSCI), Las Vegas, NV, 2017;1660–1665, https://doi.org/10.1109/CSCI.2017.289.

26. Sy B, Chen J, Horowitz R. Incorporating association patterns into manifold clustering for enabling predictive analytics, 2019 international conference on computational science and computational intelligence (CSCI), Las Vegas, NV, USA, 2019; 1300–1305. https://doi.org/10.1109/CSCI49370.2019.00243.

27. Van Stee SK, Yang Q. The effectiveness and moderators of mobile applications for health behavior change. In: Technology and Health. pp 243–270. 2020 https://doi.org/10.1016/b978-0-12-816958-2.00011-3

28. Volpp KG, Mohta N. Insights report: patient engagement survey: improved engagement leads to better outcomes, but better tools are needed. NEJM Catalyst 2016. https://catalyst.nejm.org/patient-engagement-report-improved-engagement-leadsbetter-outcomes-better-tools-needed/

29. Rose S, Engel D, Cramer N, Cowley W. RAKE algorithm: Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley & Sons; 2010