Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

Building a Personalized Messaging System for Health Intervention in Underprivileged Regions Using Reinforcement Learning

Sarah Kinsey[1] , **Jack Wolf**[1] , **Nalini Saligram**[2] , **Varun Ramesan**[2] , **Meeta Walavalkar**[2] , **Nidhi Jaswal**[2] , **Sandhya Ramalingam**[2] , **Arunesh Sinha**[3] and **Thanh Nguyen**[1]
[1]University of Oregon
[2]Arogya World
[3]Rutgers University
sarahevekinsey@gmail.com, jwolf5@uoregon.edu, {nalini, varun, meeta, nidhi}@arogyaworld.org, ramalingamsandhya@gmail.com, arunesh.sinha@rutgers.edu, thanhhng@cs.uoregon.edu

Abstract

This work builds an effective AI-based *message generation system* for diabetes prevention in rural areas, where the diabetes rate has been increasing at an alarming rate. The messages contain information about diabetes causes and complications and the impact of nutrition and fitness on preventing diabetes. We propose to apply reinforcement learning (RL) to optimize our message selection policy over time, tailoring our messages to align with each individual participant's needs and preferences. We conduct an extensive field study in India which involves more than 1000 participants who are local villagers and they receive messages generated by our system, over a period of six months. Our analysis shows that with the use of AI, we can deliver significant improvements in the participants' diabetes-related knowledge, physical activity levels, and high-fat food avoidance, when compared to a static message set. Furthermore, we build a new neural network based behavior model to predict behavior changes of participants. By exploiting underlying characteristics of health-related behavior, we manage to significantly improve the prediction accuracy of our model compared to baselines.

1 Introduction

Non-communicable diseases (NCDs) such as cardiovascular disease, diabetes, and cancer, are among the top health challenges of the century. According to WHO, NCDs kill 41 million people each year, equivalent to 74% of all deaths globally. Notably, 85% of premature deaths from NCDs occur in low- and middle-income countries [Organization, 2018]. Fortunately, these serious diseases are largely preventable. According to WHO, NCDs are preventable with three lifestyle changes: eat healthy food, increase physical activity, and avoid tobacco. Yet barriers in underprivileged regions often prevent engagement in these activities, particularly poor education about the disease, lack of social support, and limited healthcare access.

We propose to overcome some of these barriers by building a new AI-based system which can help in improving diabetes risk behavior of people in such underprivileged regions. We build an effective messaging intervention system, that dynamically sends personalized messages to participants (through Whatsapp). These messages contain information about diabetes causes and complications, and the impact of nutrition and fitness on preventing diabetes. We chose Whatsapp to transmit our messages since mobile phone uptake is high in India, and Whatsapp is an essential communication channel that is accessible to almost everyone. Existing non-profit healthcare programs often pre-design fixed non-personalized messages [Pfammatter *et al.*, 2016; Ranjani *et al.*, 2020]. Our work seeks to improve this using AI, leveraging techniques in RL to optimize our message selection policy and tailoring sent messages to align with each individual participant's needs and preferences.

We provide four main contributions. First, we build an online diabetes-targeted intervention system that automatically sends out messages to participants in our study on a weekly basis. Each week, messages sent to each participant are determined based on behavior dynamics of the participant. In the same week, our system collects information about changes in behavior of participants through a question/answer mechanism, which is leveraged to improve our message generation in later weeks. Second, we model the problem of optimizing message generation as a RL problem and develop a RL algorithm for solving this problem, tackling concrete real-world challenges exhibited in our problem domain. Our algorithm is an extension of DQN [Mnih *et al.*, 2015], a well-known RL method, with adaptations to handle simultaneous interactions with multiple participants in an online learning fashion.

Third, we run an extensive field study that involves over 1000 participants from local villages that received messages generated by our system for more than six months. Our analysis shows that there are significant improvements in the participants' diabetes-related knowledge, physical activities, and high-fat food avoidance at the end of our field study. Finally, we build a new neural net-based behavior model to predict participants' behavior changes, leveraging the data collected during our study. We show that our model, which exploits inherit differences in behavior changes across different types (knowledge, physical activity, and food consumption) obtains

the best prediction accuracy compared to baselines.

2 Related Work

Reinforcement Learning in Healthcare. RL has been widely used in tackling various problems in healthcare. In particular, RL was used in developing effective personalized treatment plans which can be adaptive to the dynamic changes of clinical states. There are several works in this line of research including studies of chronic diseases such as cancers [Zhao *et al.*, 2009; Ahn and Park, 2011; Hassani and others, 2010; Padmanabhan *et al.*, 2017], diabetes [Bothe *et al.*, 2013; Daskalaki *et al.*, 2013; Noori *et al.*, 2017; Asoh *et al.*, 2013], anemia [Gaweda *et al.*, 2005; Gaweda *et al.*, 2006; Martín-Guerrero *et al.*, 2009], and HIV [Yu *et al.*, 2019; Parbhoo *et al.*, 2017]. In addition, there is an increasing number of studies that applied RL techniques to problems in critical care such as generating optimal sepsis treatment policies [Saria, 2018], and anesthesia control [Moore *et al.*, 2014; Sinzinger and Moore, 2005]. RL was also used in automated medical diagnosis [Sahba *et al.*, 2008; Ghesu *et al.*, 2017; Chu *et al.*, 2016; Kao *et al.*, 2018]. We refer readers to [Yu *et al.*, 2021] for a complete literature review.

The research topic that is closest to our work is the problem of health management. Specifically, there are works on using RL to optimize messages sent to users to improve their physical activities [Hochberg *et al.*, 2016; Yom-Tov *et al.*, 2017]. They essentially developed a mobile phone app that runs in the background of patients' smartphones and automatically collects data of physical activity performed by patients. They then run RL that utilizes the collected data to determine which SMS message is likely to increase the physical activity of the patient. This approach requires consensuses from patients to record all of their physical activities. Our work focuses on developing a personalized text message mechanism which targets not only physical activities but also diabetes-related knowledge and food consumption. This is accomplished through a automatic message/question/answer process in which participants can opt to respond or not.

Knowledge Tracing. Our behavior modeling of participants is related to knowledge tracing, an important research research areas for enhancing personalized education [Corbett and Anderson, 1994]. The main task is to build machine models of the knowledge of a student as they interact with coursework. Recently, given the rise of deep learning, there have been several works that utilized deep neural nets to model the student learning [Piech *et al.*, 2015; Yeung and Yeung, 2018; Xiong *et al.*, 2016; Ghosh *et al.*, 2020; Pandey and Karypis, 2019]. Our work, on the other hand, focuses on building a ML model of behavior change of participants in our study as they interact with our intervention system. We target not only diabetes-related knowledge improvement of participants, but also their physical activity and food consumption dynamics, as influenced by our messages sent to them on a weekly basis.

3 Personalized Message-Generation System

In this work, our goal is to optimize the impact of our message intervention program on participants' lifestyle behavior. The challenge is that participants have varied lifestyle and also may have various kinds of reaction to our messages. Therefore, it is important that we can personalize the message selection policy for each individual participant. We propose to apply techniques in reinforcement learning to serve this purpose. Overall, given a message bank as an input, our RL-based system runs for a number of rounds (i.e., weeks). In each round, our system selects a pair of messages for each participant and sends out these messages to the participants. In addition, the system sends out a separate pair of questions to each participant and collect answers from them. These questions collect information regarding the changes in the participants' lifestyle on a weekly basis. Our system then uses the participants' responses to these questions as feedback to update our message selection policy. The overview of our system is illustrated in Figure 1.

Our messaging process has three distinct phases. In phase one of participant sign-up, health workers visit villagers in person. The health workers (i) sign up people for receiving our health-related messages; (ii) provide detailed instructions of the program to the participant; and (iii) distribute initial questionnaire and collect answers from participants. The initial questionnaire includes questions about the participants' demographics, family health-related history, knowledge about diabetes, physical activities, and food intake, etc. We use participants' responses to this questionnaire to build the initial state of participants.

In more detail, each participant's *state* consists of "scores" for the five categories: (a) healthy food intake, (b) unhealthy food/tobacco/alcohol intake, (c) fitness/physical activity level, (d) diabetes cause knowledge, and (e) diabetes complication knowledge. The scores take value in $\{1, 2, 3\}$ where 1 means "Low", 2 means "Medium", and 3 represents "High". For example, the state for a participant can be $(1, 3, 2, 2, 3)$ where the score for the health food intake of this participant is 1, it means that this participant rarely consumes healthy food such as vegetables and fruits. Thus, this score implies that this participant should receive messages that encourage the participant to eat more healthy foods.

In phase two, participants receive two messages and two questions each week, each message or question targets one of the above five categories. The sets of messages and questions are carefully designed in local language by domain experts. In this phase, our message-based intervention system interacts with participants via Whatsapp. The flow of the weekly interactions with each participant is illustrated in Figure 2 and an example of weekly messages/questions is provided in Figure 3. As we show in Figure 3, the purpose of the questions asked in each week is to determine the impact of the messages sent out in the prior week on the participants' behavior change. We remark that our system sends out only two messages and two questions on two days per week (i.e., Tuesdays and Fridays and one message/question per day) for the sake of participants' comfort. Overloading the participants with messages/questions can potentially disrupt participants' daily activities, causing unnecessary burden, decreasing their engagement in our program. Finally, responses of participants to our questions is used to update the participants' state. Our system then leverages the participants' updated state to generate new messages and questions in the following week.

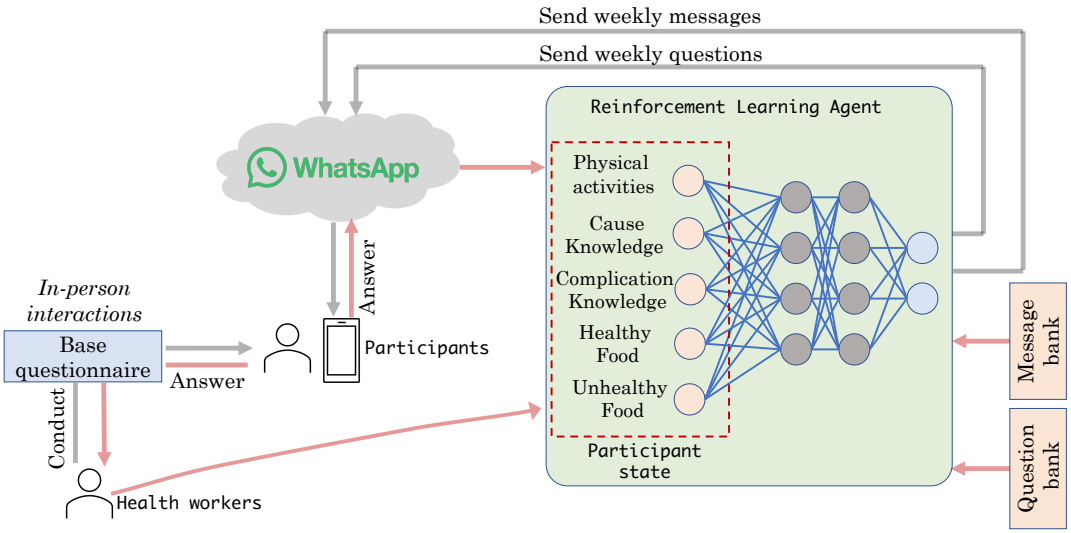

Figure 1: Personalized Message-based Intervention System Overview

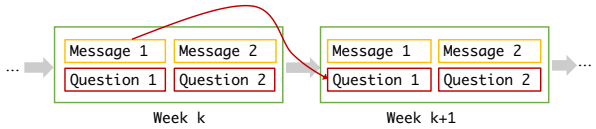

Figure 2: Weekly Message/Question Flow

Message example on week k: (Fitness) You can help avoid diabetes by being physically active. Walk to the temple or shops, climb stairs. Walk briskly or exercise for 30 mins daily.

Question example on week k+1: In last week, how often would you have done any form of exercise (Yoga/Running/Jogging/In Gym/Aerobics etc.)?
Answer format: 1. Daily days 3. Selective days/weekends
2. Alternate 4. Never

Figure 3: An Example of Weekly Messages and Questions

Finally, in phase three we perform a post-study analysis, in which our health workers meet participants in person again and conduct the same questionnaire as in phase one. Our goal is to have a complete comparison of the behavior changes in beneficiaries before and after participation in our program.

4 RL-based Message Generation Algorithm

The core of our message-based intervention system is the RL agent that actively selects weekly messages and questions for individual participants in a personalized manner based on their previous responses. We adopt ideas from Deep Q-Learning (DQN) [Mnih *et al.*, 2015], a well-known RL method in literature, to build our RL agent. However, we face the following challenges. *First*, RL methods are effective typically when they are pre-trained before the actual real-world deployment. However, in this project, there is no historical participant data that can be used for pre-training RL methods. Our RL agent has to be trained directly on the job during limited weekly interactions across only 25 weeks. *Second*, ideally, we can treat each participant as a separate Markov Decision Process and train a separate RL agent for each participant. However, our system can only obtain a single trajectory of state transition for each participant, which is extremely limited information for building an RL agent for each individual. Conversely, a single RL agent for the whole population may fail to capture the diverse behavior of participants. *Third*, traditionally, RL models are iteratively trained by sequentially acquiring different episodes of interactions with the environment. However, here our system interacts with multiple participants in parallel across a single episode.

Given these real-world deployment challenges, we create a new variant of DQN with the following revisions. *First*, we provide a warm-up stage in which we "pre-train" our model; we leverage the participants' responses to the initial questionnaire to initialize values of learnable parameters of our model. Intuitively, these values are determined such that our model generates messages for each participant that target state categories in which the participant has low scores. For example, if the participant has low physical activity level (i.e., score is 1), the model will likely select a message that encourages the participant to do more physical exercise. *Second*, based on initial state scores, we use a clustering method to divide participants into three groups such that participants in the same group have similar initial states. We aim to train a single RL agent for each group, anticipating similar in-group behavior. Note that this means the RL agent is trained based on data collected from all participants within a group; however, messages and questions selected by the RL agent for each participant are determined by their individual state. By doing so, we use enough data to train our model while still personalizing messages. *Third*, we modify the DQN training process to allow multiple updates of model training in each step (a week) and simultaneous state and message updates for all participants. The details are presented in Algorithm 1, which runs separately for each participant group.

Essentially, in each week t, each participant i is associated with a state $s_{t-1}^{(i)}$ which represents the latest status of

Algorithm 1: Adaptive Message Generation.

Calculate participant initial states based on their responses to the base questionnaire $\{s_0^{(i)}\}$;
Warm-up step: Pre-train action-value network Q with parameters θ based on initial states of participants;
Initialize target action-value network $\widehat{Q}$ with $\widehat{\theta} = \theta$;
for $t = 1 \rightarrow T$ **do**
  for $i = 1 \rightarrow N$ **do**
    With prob. ϵ, select a random action (i.e., a pair of messages) $a_t^{(i)}$ for participant i;
    With prob. $1 - \epsilon$, select $a_t^{(i)} = \text{argmax}_a\, Q(s_{t-1}^{(i)}, a, \theta)$;
    if $t > 1$ **then**
      Send questions $q_{t-1}^{(i)}$ and obtain answer$_{t-1}^{(i)}$;
      Update participant state: $s_t^{(i)} = \text{Update}\left(s_{t-1}^{(i)}, q_{t-1}^{(i)}, \text{answer}_{t-1}^{(i)}\right)$;
      Calculate reward $r_{t-1}^{(i)}$ based on state update and add transition $\left(s_{t-1}^{(i)}, a_{t-1}^{(i)}, s_t^{(i)}, r_{t-1}^{(i)}\right)$ to D;
  for $n = 1 \rightarrow$ *numUpdate* **do**
    Sample random minibatch of transitions (s_j, a_j, s_{j+1}, r_j) from D;
    Perform a gradient descent step to update θ on $\left[r_j + \gamma \max_{a'} \widehat{Q}\left(s_{j+1}, a', \widehat{\theta}\right) - Q\left(s_j, a_j, \theta\right)\right]^2$;
  if $t \mod step = 0$ **then** Update $\widehat{\theta} = \theta$;

the participant lifestyle behavior as we discussed previously in Section 3. The goal of Algorithm 1 is to train a neural net model with unknown parameters θ to predict the q-value $Q(s_{t-1}^{(i)}, a, \theta)$ of the state-action pair $(s_{t-1}^{(i)}, a)$. Here, an action a represents a pair of messages selected from the weekly message bank. Intuitively, the q-value $Q(s_{t-1}^{(i)}, a, \theta)$ is the total expected reward that we receive if the action a is chosen for participant i given the latest state value $s_{t-1}^{(i)}$. This total expected reward captures the long-term impact of the selected action on the behavior change of participant i in the future.

Warm-up. In the warm-up phase, we pre-train our neural net model by minimizing the following MSE:

$$\theta^{init} \in \underset{\theta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in A} \left[Q(s_0^{(i)}, a, \theta) - \text{init.value}(s_0^{(i)}, a)\right]^2$$

where init.value$(s_0^{(i)}, a)$ is the estimated importance score of action a for participant i given the participant initial state is $s_0^{(i)}$. Note that this initial state is estimated based on the participant's response to the questionnaire. The score init.value$(s_0^{(i)}, a)$ is determined such that this value will be high if the action a includes messages that target categories that the participant has low scores, and vice versa. For example, if the participant has score 1 for the physical activity category in $s_0^{(i)}$, the value of init.value$(s_0^{(i)}, a)$ is 3 if the action a has physical activity-related messages. Here, N is the total of participants in the considering group and A is the set of all possible actions (i.e., pairs of messages).

Weekly message/question selection. In the weekly message/question phase, at every week t, for every participant i, with a probability of $\epsilon > 0$, we select an action $a_t^{(i)}$ for participant i uniformly at random. And with a probability of $1 - \epsilon$, we select the optimal action $a_t^{(i)} = \text{argmax}_a\, Q(s_{t-1}^{(i)}, a, \theta)$. This ϵ-greedy approach allows us to balance between exploration and exploitation during the learning process. We then send a pair of questions $q_{t-1}^{(i)}$. These questions are used to measure the impact of messages the participants received in the last week $t - 1$. We remark that if we receive responses from participants for these questions quickly, we can immediately use these responses to update participants' state and then generate new messages for this week. However, this is not the case due to delayed updates, meaning that we have to send out messages for this week prior to receiving responses. Therefore, we have the following message/question procedure:

- Week $t = 1$: we send out messages $a_1^{(i)}$ based on $s_0^{(i)}$.
- Week $t = 2$: we send out messages $a_2^{(i)}$ based on state $s_1^{(i)} = s_0^{(i)}$. We then send out questions $q_1^{(i)}$ regarding impact of $a_1^{(i)}$ and receive responses answer$_1^{(i)}$. We update state $s_2^{(i)} = \text{Update}(s_1^{(i)}, q_1^{(i)}, \text{answer}_1^{(i)})$.
- Week $t = 3$: we send out messages $a_3^{(i)}$ based on state $s_2^{(i)}$. We send out questions $q_2^{(i)}$ regarding impact of $a_2^{(i)}$ and receive responses. We update state $s_3^{(i)} = \text{Update}(s_2^{(i)}, q_2^{(i)}, \text{answer}_2^{(i)})$, and so on.

Note that, in reality, participants may not respond every week. When we do not receive any answer from a participant, there will be no state change for that participant, i.e., $s_t^{(i)} = s_{t-1}^{(i)}$.

Model training update. Similar to DQN, we maintain a replay buffer D of historical interactions with participants. At every step t, transitions $\left(s_{t-1}^{(i)}, a_{t-1}^{(i)}, s_t^{(i)}, r_{t-1}^{(i)}\right)$ of every participant i will be added to D. Here, the reward $r_{t-1}^{(i)}$ is calculated based on the state change $(s_{t-1}^{(i)}, s_t^{(i)})$. For example, if the physical activity score part of the state is updated from a value of 1 in $s_{t-1}^{(i)}$ to a value of 2 in $s_t^{(i)}$, then the reward $r_{t-1}^{(i)} = 1$. This reward value indicates that the action $a_{t-1}^{(i)}$ had positive impact on the participant i's exercise behavior. If multiple categories in the state have their scores changed, then the reward is computed as the sum of rewards over all these categories. This buffer D is used to update the neural net parameters θ. In addition to D, we also maintain a target network $\widehat{Q}$ with target parameters $\widehat{\theta}$. The values of the parameters $\widehat{\theta}$ are updated periodically based on the network Q. The replay buffer D and target network $\widehat{Q}$ are the main ideas of DQN that help in stabilizing and improving the q-learning process. Finally, at each week t, after updating D with new transitions, we run a number of iterations

$numUpdate$ to perform gradient descent updates on θ based on the loss $\left[r_j + \gamma \max_{a'} \widehat{Q}\left(s_{j+1}, a', \widehat{\theta}\right) - Q\left(s_j, a_j, \theta\right)\right]^2$ which is computed based on a mini-batch of transitions (s_j, a_j, s_{j+1}, r_j) sampled from D.

5 Real-world Deployment

For real world deployment, our primary goals were to lay the groundwork for our message intervention program. These included completing all the required preparations, obtaining IRB approval, working with domain experts to design a bank of messages and questions, completing front-line worker training, collecting baseline data, and testing the transmission system. In addition, our critical objectives included recruiting beneficiaries and commencing the message transmission.

In January and February of 2022, we successfully set up the automated messaging pipeline by linking our Google Cloud VM to our partner's cloud storage, allowing for seamless and automatic transmission of messages and questions selected by our RL system to participants. We also completed Facebook business verification, obtained approval for WhatsApp message transmission, and developed the AI tool. Crucially, we ensured that key people from multiple project partners could seamlessly share de-identified data files and responses from the villagers each week. Finally, we completed training for 20 front-line workers on the project implementation.

In February and March of 2022, our front-line workers collected behavior surveys (questionnaire) from 1698 participants who are local villagers in India. We successfully completed a pilot test of the diabetes-related AI messaging bot to verify functionality of the overall system. In the end, we successfully recruited 1049 participants to opt-in to receive the diabetes messages. To compare with the existing static message program in which all participants received the same sequence of messages, we randomly divided participants into two groups: 548 participants joined our AI-based message program (we call this group the AI group) and 501 joined the baseline static message program (the non-AI group).

In March of 2022, we began the message transmission for the first batch of users with those in the AI cohort receiving two messages and two questions weekly on Fridays and Tuesdays. Participants in the non-AI group received two messages on Thursdays and Mondays each week. We remark that all participants did not opt-in at the very start but gradually joined over a couple of weeks. The villagers in the AI group responded to questions asked — engagement levels were around 35% past week 8 of the study.

The study ended in November 2022, at which time the post-study questionnaire was sent to the participants. This is the same questionnaire used at the beginning of our study. Based on responses of participants, we are able to evaluate the effectiveness of the AI system by comparing the performance of participants between the AI and non-AI groups.

6 Post-Study Intervention Result

We analyze behavior changes in both AI and non-AI participant groups based on their responses to the same questionnaire before/after joining our study. We divide questions into

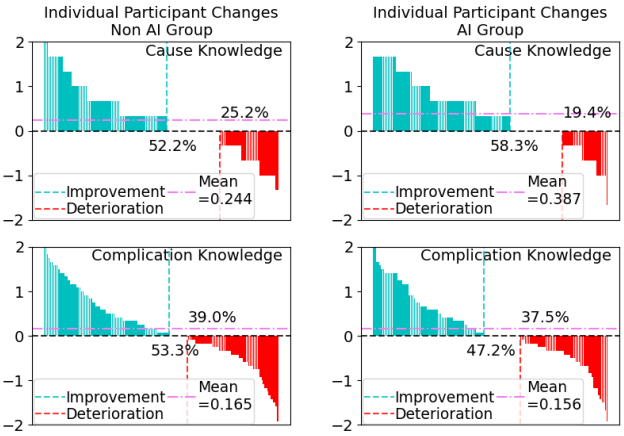

Figure 4: **Knowledge Comparison**. We plot results on *cause knowledge* and *complication knowledge* (of diabetes). Overall, AI group shows a substantial improvement in both cause and complication knowledge scores. In particular, AI group outperforms the non-AI group significantly in cause knowledge. In complication knowledge, non-AI group has a higher percentage of participants who have positive score changes but also a higher percentage of participants who have negative score changes — In the end, the AI-group obtains a higher mean score improvement compared to the non-AI group.

different types that focus on *knowledge*, *physical activity*, and *dietary*. Responses to each question are converted into three scores: 1 (low), 2 (medium), and 3 (high). The final score of a participant in each category is averaged over all questions in that category. We compute the score difference between the pre- and post-studies for each participant. Positive score changes imply participants improve their behavior at the end of our study. Our analysis results are shown in Figures 4, 5, 6 where the x-axis represents the participants and the y-axis represents the score change.

To compare our AI versus the non-AI intervention, we consider three statistics: (i) mean score change across each population; (ii) percentage of participants who improve their scores; and (iii) percentage of participants who decrease their scores. For example, in Figure 4 for the AI group, 58.3% of the participants have a positive score changes while 19.4% have a negative score changes in the diabetes-cause knowledge type. In addition, the mean score change is 0.387. For the AI group, we only consider participants who have response rates of at least 50%. Our rationale is that low-response rate participants do not engage in the study, and thus, their behavior will not be impacted by the AI messages.

7 Human Behavior Modeling

We aim to build a predictive model characterizing how participants behave in response to our message intervention program. Such predictive model could be used in the future to create simulated data to refine our message selection policy.

Feature Extraction. We use data collected from our field study for this modeling task. We divide the questionnaire and the weekly messages/questions into 17 finer categories. The

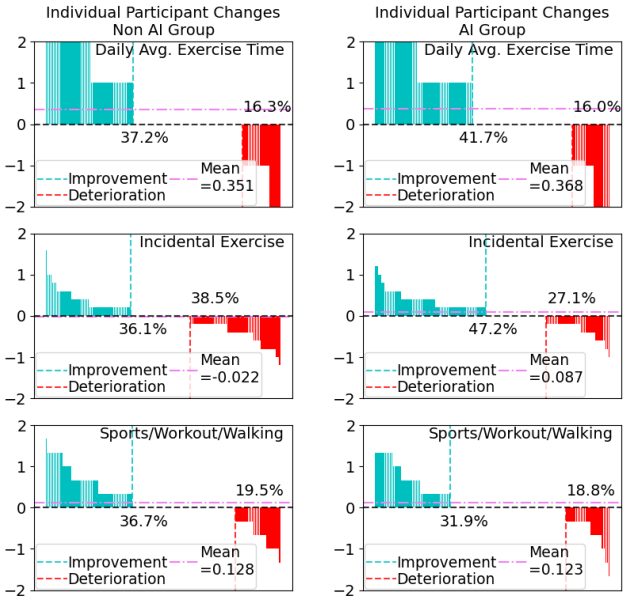

Figure 5: **Physical Activity Comparison**. We plot results on *daily average exercise time*, *incidental exercise* (this refers to exercises incurred throughout daily activities, such as choosing to walk for errands, walking around the house, and taking stairs instead of elevators, and *sport/workout/walking*). Overall, AI group shows a substantial improvement in both average exercise time and incidental exercise, which outperforms the non-AI group. In the sports/workout/walking type, non-AI group has a higher percentage of participants who have positive score changes; nevertheless, the mean score change is not much different.

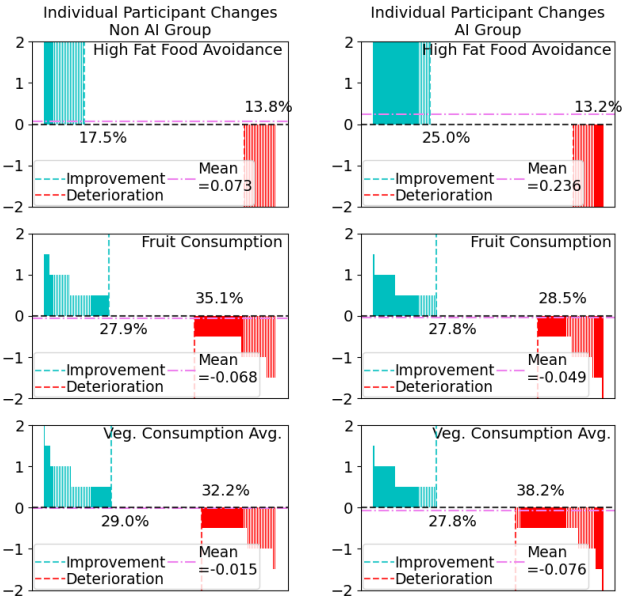

Figure 6: **Dietary Comparison**. We plot results on unhealthy food, fruit, and vegetable consumption. Overall, in the unhealthy food consumption category, the AI group shows a substantially more improvement compared to non-AI group — (25% improvement, 13.2% deterioration, 0.236 mean) versus (17.5% improvement, 13.8% deterioration, 0.073 mean). In the fruit/vegetable category, we observe somewhat negative changes in both non-AI and AI groups. That is, a higher percentage of participants had negative score changes compared to positive score changes. However, the mean score changes in both groups are very close to zero.

questionnaire responses are then used to compute scores in each of these categories for all participants, which are then used as features. Additionally, we include the message and question ID that the participant received, along with the category that they belong to. To provide the model with more context, we include the previous week's information for all these categories, along with the responses received from participants. Lastly, we also include a coarser categorization feature, indicating whether the question asked is regarding one of three types: knowledge, physical activity, or dietary.

Model Description. The prediction task, then, is to use the extracted features to predict participants' responses to the questions they receive. We cast this problem as a multi-classification problem. In each week, for each participant, we take each question and other features associated with that participant as an input to produce a prediction of the corresponding response of that participant (which is categorized into three levels 1 (low), 2 (medium), and 3 (high)).

We consider three different models as baselines for this task: (i) the classic logistic regression; (ii) a simple neural network (NN) with two fully-connected linear layers; and (iii) a Long Short Term Memory (LSTM) based model [Hochreiter and Schmidhuber, 1996] (i.e., a LSTM block followed by a simple linear layer). We remark that LSTM is commonly used in deep knowledge tracing [Piech *et al.*, 2015].

Importantly, we observe participants' behavior changes vary across different types of diabetes-related activities. As a result, a single behavior model may perform poorly in predicting responses to various types of questions. Therefore, we aim to build predictive models that can differentiate behavior changes of three different types: food consumption, physical activities, and diabetes-related knowledge. More specifically, we propose the three models: (i) type-trifecta logistic regression — this model consists of three separate logistic regression components, each produces predictions for responses in one of the aforementioned types; (ii) type-trifecta simple neural network — this model consists of three separate neural net components, each is a simple 2-layer neural net; and (iii) type-trifecta LSTM — this model has a shared LSTM block followed by three separate blocks of linear layers.

Accuracy Evaluation. All of our behavior prediction models were trained on dual Intel E5-2690v4 processors. All experiments were trained in PyTorch using cross entropy loss and the Adam optimizer. For our experiments, we collect all results over 30 random seeds (resulting in different model initializations and test/train splits) and report the mean along with the standard deviation. We remark that there are a lot of missing responses in our training data set (response rates of participants is less than 40%). Therefore, in our experiments, we examine two options: one is to simply encode

Model	Acc. Train	Acc. Test
Simple NN	0.533 ± 0.016	0.525 ± 0.022
LSTM	0.541 ± 0.022	0.533 ± 0.030
Logistic Regression	0.524 ± 0.009	0.515 ± 0.018
Simple NN Type Trifecta	0.59 ± 0.005	0.58 ± 0.019
LSTM Type Trifecta	0.583 ± 0.006	0.575 ± 0.019
Logistic Regression Type Trifecta	0.585 ± 0.005	0.578 ± 0.018

Table 1: Evaluation with no noise, no predicted feature insertion.

Model	Acc. Train	Acc. Test
Simple NN	0.536 ± 0.017	0.528 ± 0.022
LSTM	0.548 ± 0.023	0.542 ± 0.028
Logistic Regression	0.525 ± 0.009	0.516 ± 0.018
Simple NN Type Trifecta	0.59 ± 0.005	0.58 ± 0.019
LSTM Type Trifecta	0.584 ± 0.005	0.576 ± 0.019
Logistic Regression Type Trifecta	0.584 ± 0.005	0.578 ± 0.019

Table 2: Evaluation with noise, but no predicted feature insertion.

Model	Acc. Train	Acc. Test
Simple NN	0.531 ± 0.014	0.523 ± 0.019
LSTM	0.548 ± 0.023	0.535 ± 0.029
Logistic Regression	0.53 ± 0.010	0.522 ± 0.019
Simple NN Type Trifecta	0.592 ± 0.005	0.578 ± 0.019
LSTM Type Trifecta	0.579 ± 0.008	0.572 ± 0.022
Logistic Regression Type Trifecta	0.587 ± 0.006	0.577 ± 0.019

Table 3: Evaluation with noise and predicted feature insertion.

missing responses as "-1" and the another is to replace missing responses by our model predictions. In addition, we try adding noise (i.e., zero mean Gaussian noise) to participants' responses. The purpose is to examine if this noise helps in improving the robustness of our models or not. Our prediction accuracy results for all models are shown in Tables 1–3, corresponding to three settings: (i) no noise is added to participants' responses and missing responses are encoded as "-1" (Table 1); (ii) similar to (i) but Gaussian noise is added (Table 2); and Gaussian noise is added and missing responses are replaced with the model prediction (Table 3).

All three tables show that differentiating behavior changes according to three different categories of food consumption, physical activities, and knowledge significantly improves the prediction accuracy of our models compared to the baselines. For example, in Table 1, the type-trifecta simple NN model obtains an averaged prediction accuracy of 59% and 58% on the training and test sets, respectively. This is significantly higher than the prediction accuracy of the single simple NN model (i.e., 53.3% and 52.5%). We also observe the similar performance enhancement trend for the logistic regression and the LSTM-based models. In addition, interestingly, unlike in knowledge tracing [Piech *et al.*, 2015] where LSTM-based models are shown to be superior in predicting the knowledge of students, our results show that the type-trifecta simple NN model performs the best. This phenomenon perhaps comes from the limited data availability (with missing responses) in our domain that potentially deteriorates the performance of complex models like LSTM. Lastly, we do not observe a substantial changes in prediction accuracy of all models when we introduce Gaussian noise to responses or insert model predictions to replace missing responses.

8 Conclusion: Results and Learned Lessons

In this work, we developed an RL-based personalized messaging system for diabetes intervention tailored to people living in rural areas where access to healthcare, social support, and education are limited. We ran an extensive field study that involves more than 1000 local villagers participating in our messaging program. Our post-study analysis results show the significant benefit of our approach (compared to the existing system) in the participants' diabetes-related knowledge, physical activities, and high-fat food avoidance. Furthermore, our behavior models which leverage characteristics of participants' responses outperform baselines including the single LSTM model that is commonly used in knowledge tracing.

We would like to highlight some important lessons learned from this work. First, collaborations among different partners with different areas of expertise including NGOs, health domain experts, and academics are the key to the success of social impact projects. Second, real-world domains exhibit various challenges that we may not be able to anticipate when building our AI models. Continuing to improve our models and solutions with the adaptation to rising challenges is essential to the long-term impact of the project. For example, our current RL model does not directly account for missing responses from participants. Furthermore, answers about behavior are self-reported and could be misleading. We plan to address these limitations in future work. Third, available real-world data in this domain is extremely limited. Thus, simple models may work better than complex deep learning models.

Acknowledgements

This paper was partially supported by the Google AI for Social Good program and by grant W911NF-20-1-0344 from the US Army Research Office.

Contribution Statement

Sarah Kinsey and Jack Wolf contributed equally to the paper and thus are joint first authors.

References

[Ahn and Park, 2011] Inkyung Ahn and Jooyoung Park. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *BioSystems*, 106(2-3):121–129, 2011.

[Asoh *et al.*, 2013] Hideki Asoh, Masanori Shiro1 Shotaro Akaho, Toshihiro Kamishima, Koiti Hasida, Eiji Aramaki, and Takahide Kohro. An application of inverse reinforcement learning to medical records of diabetes treatment. In *ECMLPKDD2013 workshop on reinforcement learning with generalized feedback*, 2013.

[Bothe *et al.*, 2013] Melanie K Bothe, Luke Dickens, Katrin Reichel, Arn Tellmann, Björn Ellger, Martin Westphal, and Ahmed A Faisal. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert review of medical devices*, 10(5):661–673, 2013.

[Chu *et al.*, 2016] Tianshu Chu, Jie Wang, and Jiayu Chen. An adaptive online learning framework for practical breast cancer diagnosis. In *Medical imaging 2016: Computer-aided diagnosis*, volume 9785, pages 537–548. SPIE, 2016.

[Corbett and Anderson, 1994] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4:253–278, 1994.

[Daskalaki *et al.*, 2013] Elena Daskalaki, Peter Diem, and Stavroula G Mougiakakou. Personalized tuning of a reinforcement learning control algorithm for glucose regulation. In *2013 35th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3487–3490. IEEE, 2013.

[Gaweda *et al.*, 2005] Adam E Gaweda, Mehmet K Muezzinoglu, George R Aronoff, Alfred A Jacobs, Jacek M Zurada, and Michael E Brier. Individualization of pharmacological anemia management using reinforcement learning. *Neural Networks*, 18(5-6):826–834, 2005.

[Gaweda *et al.*, 2006] Adam E Gaweda, Mehmet K Muezzinoglu, Alfred A Jacobs, George R Aronoff, and Michael E Brier. Model predictive control with reinforcement learning for drug delivery in renal anemia management. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5177–5180. IEEE, 2006.

[Ghesu *et al.*, 2017] Florin-Cristian Ghesu, Bogdan Georgescu, Yefeng Zheng, Sasa Grbic, Andreas Maier, Joachim Hornegger, and Dorin Comaniciu. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):176–189, 2017.

[Ghosh *et al.*, 2020] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2330–2339, 2020.

[Hassani and others, 2010] Amin Hassani et al. Reinforcement learning based control of tumor growth with chemotherapy. In *2010 International Conference on System Science and Engineering*, pages 185–189. IEEE, 2010.

[Hochberg *et al.*, 2016] Irit Hochberg, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Elad Yom-Tov. A reinforcement learning system to encourage physical activity in diabetes patients. *arXiv preprint arXiv:1605.04070*, 2016.

[Hochreiter and Schmidhuber, 1996] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 473–479, Cambridge, MA, USA, 1996. MIT Press.

[Kao *et al.*, 2018] Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Martín-Guerrero *et al.*, 2009] José D Martín-Guerrero, Faustino Gomez, Emilio Soria-Olivas, Jürgen Schmidhuber, Mónica Climente-Martí, and N Víctor Jiménez-Torres. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Systems with Applications*, 36(6):9737–9742, 2009.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Moore *et al.*, 2014] Brett L Moore, Larry D Pyeatt, Vivekanand Kulkarni, Periklis Panousis, Kevin Padrez, and Anthony G Doufas. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *The journal of machine learning research*, 15(1):655–696, 2014.

[Noori *et al.*, 2017] Amin Noori, Mohammad Ali Sadrnia, et al. Glucose level control using temporal difference methods. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 895–900. IEEE, 2017.

[Organization, 2018] World Health Organization. *Noncommunicable diseases country profiles 2018*. World Health Organization, 2018.

[Padmanabhan *et al.*, 2017] Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical biosciences*, 293:11–20, 2017.

[Pandey and Karypis, 2019] Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837*, 2019.

[Parbhoo *et al.*, 2017] Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez.

Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

[Pfammatter *et al.*, 2016] Angela Pfammatter, Bonnie Spring, Nalini Saligram, Raj Davé, Arun Gowda, Linelle Blais, Monika Arora, Harish Ranjani, Om Ganda, Donald Hedeker, et al. mhealth intervention to improve diabetes risk behaviors in india: a prospective, parallel group cohort study. *Journal of medical Internet research*, 18(8):e207, 2016.

[Piech *et al.*, 2015] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

[Ranjani *et al.*, 2020] Harish Ranjani, Sharma Nitika, Ranjit Mohan Anjana, Sandhya Ramalingam, Viswanathan Mohan, Nalini Saligram, et al. Impact of noncommunicable disease text messages delivered via an app in preventing and managing lifestyle diseases: Results of the "myarogya" worksite-based effectiveness study from india. *Journal of Diabetology*, 11(2):90, 2020.

[Sahba *et al.*, 2008] Farhang Sahba, Hamid R Tizhoosh, and Magdy MA Salama. Application of reinforcement learning for segmentation of transrectal ultrasound images. *BMC medical imaging*, 8:1–10, 2008.

[Saria, 2018] Suchi Saria. Individualized sepsis treatment using reinforcement learning. *Nature medicine*, 24(11):1641–1642, 2018.

[Sinzinger and Moore, 2005] Eric D Sinzinger and Brett Moore. Sedation of simulated icu patients using reinforcement learning based control. *International Journal on Artificial Intelligence Tools*, 14(01n02):137–156, 2005.

[Xiong *et al.*, 2016] Xiaolu Xiong, Siyuan Zhao, Eric G Van Inwegen, and Joseph E Beck. Going deeper with deep knowledge tracing. *International Educational Data Mining Society*, 2016.

[Yeung and Yeung, 2018] Chun-Kit Yeung and Dit-Yan Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the fifth annual ACM conference on learning at scale*, pages 1–10, 2018.

[Yom-Tov *et al.*, 2017] Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

[Yu *et al.*, 2019] Chao Yu, Yinzhao Dong, Jiming Liu, and Guoqi Ren. Incorporating causal factors into reinforcement learning for dynamic treatment regimes in hiv. *BMC medical informatics and decision making*, 19(2):19–29, 2019.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[Zhao *et al.*, 2009] Yufan Zhao, Michael R Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315, 2009.