



# Recommender System for Responsive Engagement of Senior Adults in Daily Activities

Igor Kulev<sup>1</sup>  · Carlijn Valk<sup>2</sup> · Yuan Lu<sup>2</sup> · Pearl Pu<sup>1</sup>

Received: 30 September 2019 / Accepted: 16 January 2020 / Published online: 06 February 2020  
© The Author(s) 2020

## Abstract

Understanding and predicting how people change their behavior after an intervention from time series data is an important task for health recommender systems. This task is especially challenging when the time series data is frequently sampled. In this paper, we develop and propose a novel recommender system that aims to promote physical activeness in elderly people. The main novelty of our recommender system is that it learns how senior adults with different lifestyle change their activeness after a digital health intervention from minute-by-minute fitness data in an automated way. We trained the system and validated the recommendations using data from senior adults. We demonstrated that the low-level information contained in time series data is an important predictor of behavior change. The insights generated by our recommender system could help senior adults to engage more in daily activities.

**Keywords** recommender system · behavior change · representation learning · predictive modeling.

---

✉ Igor Kulev  
igor.kulev@epfl.ch

Carlijn Valk  
c.a.l.valk@tue.nl

Yuan Lu  
y.lu@tue.nl

Pearl Pu  
pearl.pu@epfl.ch

<sup>1</sup> School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

<sup>2</sup> Department of Industrial Design, Eindhoven University of Technology, Eindhoven, the Netherlands

## Introduction

The proportion of the global population aged 60 years or over increases rapidly: from 8.5% in 1980 to 12.7% in 2017 (United Nations, Department of Economic and Social Affairs 2017). It is projected to rise to 21.3% in 2050 (United Nations, Department of Economic and Social Affairs 2017). One of the main reasons for this is the increasing trend of the global average life expectancy: it increased by 5.5 years between 2000 and 2016 (as indicated by the World Health Organization), and is projected to increase by 4.4 years by 2040 (Foreman et al. 2018). The health of the elderly population is an enormous challenge for the health and social care services (The Lancet Public Health 2017). Increased longevity is associated with increases in the number of chronic diseases in the elderly population (Ory and Cox 1994). The leading contributors to disease burden in older people are cardiovascular diseases (30.3% of the total burden) (Prince et al. 2015).

Demographic, social, and environmental factors, including physical activity and dietary habits, play a major role in the health and functioning of older adults (Drewnowski and Evans 2001). Increased physical activity is associated with a lower risk of developing cardiovascular disease when compared to less physical activity (Carnethon 2009). Moreover, it is associated with a clear decrease in the risk of mortality from any cause (Leitzmann et al. 2007). Physical activities that improve muscular strength, endurance, and flexibility also improve the ability to perform the tasks of daily living (Drewnowski and Evans 2001). In this way, physical activity could enhance the quality of life of the older population.

There is evidence that cognitive and behavioral interventions designed to improve physical activity behavior are effective, both for the general (Conn et al. 2011) and the elderly population (Park et al. 2014). By intervention, we mean an explicit and pro-active recommendation for the purpose of changing the current behaviors of a user. Successful strategies include goal-setting, self-monitoring, feedback, rewards, social support, and coaching (Sullivan and Lachman 2017). As technology advances, it is becoming easier to integrate new and emerging platforms, software, and devices into behavioral interventions to improve physical activity (Lewis et al. 2017). Wearable devices can measure heart rate, number of steps, distance, and sleep duration with very high accuracy (Xie et al. 2018). Also, activity trackers are becoming more comfortable to wear: this is an important factor for user acceptance. The number of health and fitness apps on the market is growing. Over 318,000 health apps are now available on top app stores worldwide with more than 200 health apps being added each day (Aitken et al. 2017). These apps help people to change their behavior using different behavior change strategies (Direito et al. 2014).

It is less investigated how behavioral interventions actually result in responses at the individual levels. We are mainly interested in knowing whether a given intervention can lead to positive and engaged responses and avoid giving harmful interventions. We believe it is possible to scientifically analyze and assess the intervention effect before it is given. The best intervention for the general population is not likely to be equally effective for each individual (Kulev et al. 2018). Interventions should be tailored to the individual needs, account for personal levels of fitness, allow for personal control of the activity and its outcomes, and provide for social support by family, peers, and communities (Seefeldt et al. 2002). Tailoring variables include time-invariant predictors (e.g., sex), time-variant predictors (e.g., stress) and contextual factors (e.g., weather, day of the week) (Phatak et al. 2018).

In this paper, we describe our effort in creating a recommender system that predicts the personalized effects of different behavior interventions on the same user to select one intervention over another. We call it the behavior recommender system. Behavior recommender systems that promote physical activities would be most beneficial for the elderly population. We have identified seven main challenges for developing a recommender system whose main purpose is to select those interventions that are most likely to work:

- **Interventions should be based on successful behavior change strategies.** For all types of interventions, the development process benefits from applying evidence-based theories and techniques because they indicate under which conditions the interventions are effective (Moller et al. 2017).
- **Predictive models should be trained on data from senior adults.** The aging process leads to a reduction in physical activity level and functional fitness (Milanović et al., 2013). Thus, the distribution of the fitness data collected from young people and senior adults is different. This is why models trained on data from the general population might not work for the elderly subpopulation.
- **Data collection should not interfere with the normal functioning of the elderly.** This happens when the sensor devices are obtrusive or when the data collection is performed in a laboratory environment. Sasaki et al. (2016) have demonstrated that the algorithms developed on free-living accelerometer data are more accurate in classifying activity type in free-living older adults than the algorithms developed on laboratory accelerometer data.
- **Intervention bias in the data should be minimized.** Intervention bias exists when people who receive different interventions are not drawn from the same population. Predictive models trained on a dataset with large intervention bias could underestimate or overestimate the true intervention effect on the target population. Ideally, training data should be collected from a randomized controlled trial (RCT). If this is not possible, bias-reducing techniques (Shalit et al. 2017) should be applied before the predictive model is built.
- **Data should contain repeated time-varying measurements of the individual's behavior.** Most of the existing works do not consider temporal dynamics as a predictor of behavior change. Human actions vary over time, for example, based on time of day (Kurashima et al. 2018). Kurashima et al. (2018) have demonstrated that the time-varying action propensities can be useful to predict the next user action and when this action will occur.
- **Predictive models should be able to learn features in an automated way.** Manual feature engineering is both difficult and expensive. This process could generate a large number of features out of which only some are relevant for the predictive task. Also, during manual feature engineering, some important information from the data could be missed, resulting in trained models that have poor prediction performance. This is why features should be learned in an automated way that takes into account the machine learning task of interest.
- **Recommendations should be feasible and effective.** Very ambitious recommendations might cause injuries to the sensitive elderly population. This is why the recommender system should provide evidence that the interventions are likely to cause positive behavior change in the individual before they are suggested to him or

her. One of the ways to do this is by ensuring that the predictions for the target user are comparable with the actual responses of existing users that are similar to the target user. If similar people who already received the intervention did not respond to it, this intervention should not be recommended to the target user.

Existing work focuses only on a subset of these challenges. For example, Phatak et al. (2018) developed a system that generates recommendations based on the median value of steps/day from the baseline period. Their system does not take into account the distribution of the physical activity throughout the day as a predictor of the behavior change. In contrast, we propose a novel recommender system that could be used to engage senior adults in daily activities while addressing all the challenges discussed above. The main novelty of our recommender system is that it uses time series fitness data to predict intervention effect.

In this paper, we consider two different mobile app interventions that aim to promote physical activeness in senior adults. Each mobile application incorporates one of two motivational strategies: self-reflection and social reflection. Under the first intervention, users were able to see real-time step count information only about themselves. Under the second intervention, users were paired up and were able to see real-time step count information about each other. In previous studies, it has been shown that self-monitoring and social support are associated with increased physical activity (Greaves et al. 2011; Park et al. 2014).

Our system requires pre- and post-intervention fitness data from real users. For this purpose, we designed a randomized trial so that each participant received either a self-reflection mobile app or a social reflection mobile app. Participants were wearing a fitness tracker for three weeks before and five weeks after they received the intervention. We used this data to train the recommender system and perform an offline evaluation. More specifically, we built machine learning methods to predict the change of the physical activity levels after each intervention for new users. This allows the system to decide which intervention should be recommended. The quality of the recommendations could be estimated in online evaluation where one part of users are served by the derived recommender system, and another part of users are given an intervention without taking into account their current behavior. However, in our work we focus on the offline evaluation — in the future we plan to perform an online evaluation as well.

## Related Work

Different machine learning methods have been used to gain meaningful insights from health data. Supervised methods have been focused on either detection of health conditions or prediction of health conditions. The former refers to the process of analyzing information to understand the health condition better. The latter refers to the process of analyzing information to predict a health outcome of interest. For example, deciding whether a patient has heart arrhythmia from electrocardiograms is a detection task, but guessing whether the patient will develop a heart arrhythmia in the next year from electrocardiograms and fitness data is a prediction (or *forecasting*) task. Unsupervised methods have generally been used to discover the dominant patterns in

the data that explain people's behavior. For example, going to bed early vs. going to bed late. However, they can also be combined with the supervised algorithms and used in the pre-processing step.

## Monitoring and Tracking

Health and behavioral data, which is usually provided by sensor devices, contains valuable information that could help people improve their well-being. This data may be high-dimensional, making it challenging to extract relevant patterns from it. Most existing work has focused on processing this data to recognize various health and behavioral states. For example, classifying activity types from acceleration (Sasaki et al. 2016) and body tags data (Luštrek and Kaluža 2009), detecting falls from acceleration (Albert et al. 2012) and body tags data (Luštrek and Kaluža 2009), detecting anxiety and depression from socio-demographic and health-related data (Sau and Bhakta 2017), estimating physical activity levels from questionnaire data (Park et al. 2014), estimating body fat from accelerometer data (Swartz et al. 2012), estimating mental health burden from self-reported physical activity data (Chekroud et al. 2018), detecting heart arrhythmia from electrocardiograms (Rajpurkar et al. 2017), detecting multiple medical conditions, including diabetes, high cholesterol, high blood pressure, and sleep apnea, from heart rate sensor data (Ballinger et al. 2018), etc. Feature engineering was common in most of this works. For example, Sasaki et al. (2016) extracted time- and frequency- domain features from acceleration signals in 20-s windows. This process requires a lot of effort and may often result with features that are not predictive of the target variable. Andrew Ng stated that “coming up with features is difficult, time-consuming, requires expert knowledge” (Ng 2013). Although in our work, we have no classes or labels available and we are not interested in detecting particular health or behavioral state, the methods described in this section are relevant because of the way they process the data and extract useful features.

Deep learning techniques are able to automatically extract relevant features and they are especially useful when the data has a complex nature. Rajpurkar et al. (2017) used convolutional neural networks to detect irregular heart rhythms from electrocardiograms better than a cardiologist. In terms of data and feature extraction, our work is most similar to the work of Ballinger et al. (2018). They used sequence-to-sequence autoencoder to learn features from step count and heart rate time series data. These features were more useful than hand-engineered biomarkers derived from the medical literature to detect diabetes, high cholesterol, high blood pressure, and sleep apnea. In our task, we also use sequence-to-sequence autoencoder to learn features from step count time series data. Our work differs from Ballinger et al. (2018) in that we use longer user history to make a prediction and we have a limited amount of data to train the model. Also, we try to predict (forecast) intervention effect. This makes the data collection phase more challenging.

## Prediction of Health Conditions

Predicting health conditions is an important task because it can turn data into actionable insights. Much work has focused on early prediction of future health behavior that could be used in preventive healthcare. These works include: predicting future

cognitive impairment in elderly people from variables, which are commonly collected in community health care institutions (Na 2019), predicting mortality in elderly people from medical history, diet, exercises and lifestyle activity (Leitzmann et al. 2007), predicting mortality in older women from mean daily step counts (Lee et al. 2019), predicting changes in exercise behavior from historical physical activity data (Kotsev et al. 2014), predicting daily blood pressure levels from historical blood pressure and health behavior (Chiang and Dey 2018), predicting in-hospital mortality, readmission, prolonged length of stay and final discharge diagnoses from electronic health records data (Rajkomar et al. 2018), predict future actions from past activities (Kurashima et al. 2018), etc. The early prediction information can be useful for a health recommender system to decide *when* it needs to act, however, it is not sufficient to decide *how* to act. Expert knowledge is usually needed to choose a suitable preventive intervention based on these predictions.

Another line of work has applied data-driven approach to understand and predict the effect of interventions on people's health (Sloan et al. 2018; Robinson et al. 2019; Zeevi et al. 2015; Phatak et al. 2018). The most traditional method estimates the population-level intervention effect from randomized controlled trial (RCT) data (Sloan et al. 2018; Robinson et al. 2019). The problem with this method is that not all people respond to the same intervention in the same way. The optimal intervention depends on the individual's characteristics — this aligns with the goals of personalized medicine (Bates 2010; Kulev et al. 2018). Zeevi et al. (2015) demonstrated that people eating identical meals present high variability in post-meal blood glucose response. They developed a machine learning algorithm that uses information about blood parameters, dietary habits, anthropometrics, physical activity, and gut microbiota to predict personalized blood glucose response to real-life meals. These predictions were used to design personalized diets composed of the meals predicted by the algorithm to have low post-meal blood glucose responses. They performed both offline and online evaluation and showed that their dietary interventions improve multiple aspects of glucose metabolism.

In another work, Phatak et al. (2018) developed a system to deliver personalized daily step goals that aimed to improve people's physical activity levels. For this purpose, they performed a data collection study with people wearing activity tracker for 14 weeks. Baseline physical activity measured in weeks 1–2 has been used to inform personalized daily step goals delivered in weeks 3–14. They learned a regression model that predicts daily step count given few different variables including people's baseline median daily step count and current daily step goal. In contrast to Phatak et al. (2018) and Zeevi et al. (2015), we are interested to generate predictions based on much more fine-grained baseline data, such as the minute-by-minute step count measurements. This requires machine learning methods that are able to extract temporal patterns from time series data which are relevant to predict intervention effect. Also, our work differs from Phatak et al. (2018) in that our work focuses on improving physical activity levels in elderly people — there is no other group in our society that can benefit more from physical activities (Evans 1999).

Building a recommender system for responsive engagement of senior adults in daily activities is a challenging problem. One of the main reasons is that it is very costly to perform a study providing evidence about intervention effectiveness. Another reason is that existing personalized models for predicting intervention effect cannot learn relevant insights when they are directly applied to frequently-sampled fitness data from a

limited number of people. In our research, we provide a complete solution to this problem — from data collection to recommendation generation. We show the benefit of learning representations from time series data in predicting intervention effect. This allows the recommendation system to select better one intervention over another.

## Recommender System

The main research objective of our work is to develop a recommender system that would suggest personalized interventions likely to improve the physical activity levels of senior adults. Our recommender system pipeline consists of four phases: data collection, representation learning, predictive modeling and recommendation generation. The first phase provides time series sensor data containing evidence of the intervention effectiveness. The second phase reduces the dimensionality of each time series while preserving information about its underlying temporal dynamics. The third phase builds machine learning models that are able to predict how an elderly person would respond if he or she was given an intervention. These models take as input the representations learned in the previous phase. The fourth phase selects and recommends an optimal personalized intervention based on the predictions generated in the third phase.

## Data Collection

The first step towards building a recommender system was to collect sensor data from senior adults who received an intervention. The system would then use this data to decide whether the intervention would be effective for a new senior adult. To collect data, we conducted an experiment that included 55 senior adults aged 65+ years wearing a Fitbit Flex 2 wristband for eight weeks. The Fitbit device recorded the number of steps performed in each minute. After the first three weeks, each participant received one of two mobile app interventions by random assignment. Under the first intervention (*self-reflection*), users were able to see real-time step count information only about themselves. Under the second intervention (*peer-to-peer*), users were paired up and were able to see real-time step count information about each other. The app had been used for five weeks, until the end of the trial. The user interface of this mobile app is given in Fig. 1. We were interested to extract insights about which behavior patterns are associated with a larger improvement of physical activity levels under each intervention. These insights would be used by the system to recommend feasible and effective interventions to new senior adults.

Participants were instructed to wear the Fitbit at all times during the trial. However, it was not possible to know for sure whether they were always wearing the device or not. Only the days with a positive number of steps were counted as valid days of data. We filtered out participants who didn't have at least one valid day of data for each different day of the week, both before and after the intervention. Also, we filtered out one participant who had an increase of her or his average daily step count by 145% after the intervention (an outlier). 49 participants remained and the data associated with them were included in our analysis. Out of these people, 14 received the first intervention, “self-reflection”, and 35 received the second intervention, “peer-to-peer”. The average daily step count per day in the trial is given in Fig. 2. It can be observed that

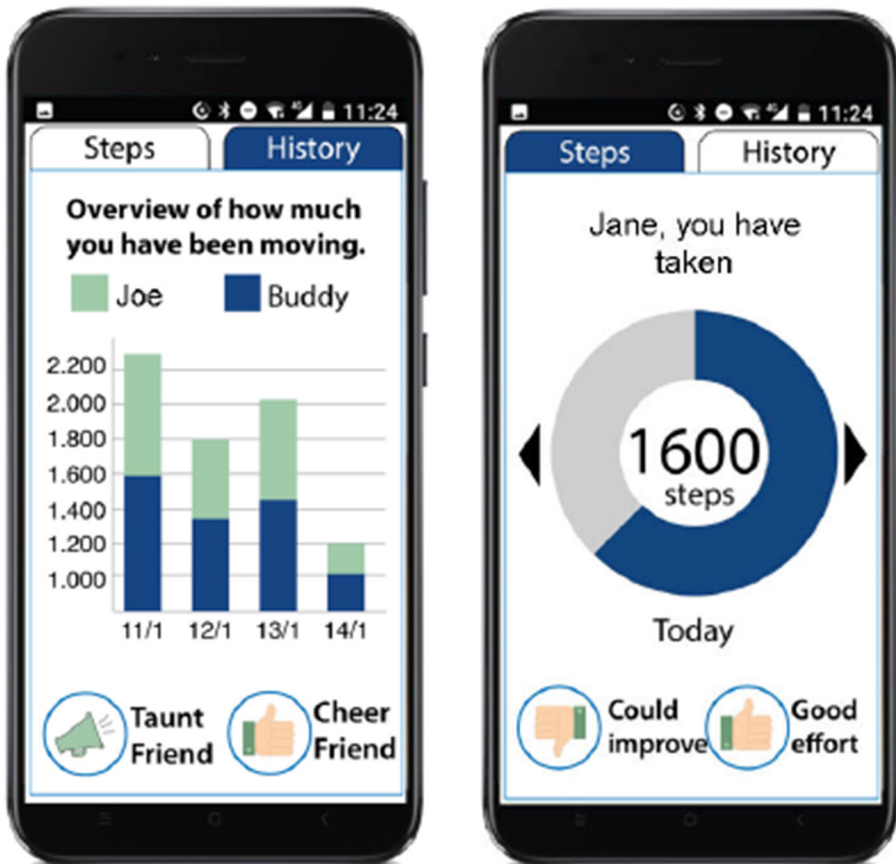


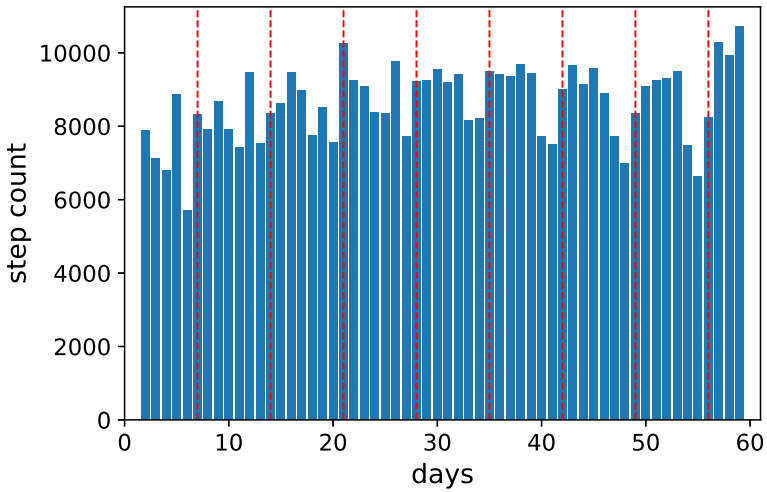
Fig. 1 Mobile app interface

people manifested periodic weekly behavior and they were the least active during the weekends. In 18.5% of the minutes there was at least one step performed. There was no significant difference between the two intervention groups in terms of their pre-intervention average daily step counts (two-sample t-test for the null hypothesis that the two samples have identical average values,  $p$  value = 0.838245). Figure 3 shows that users varied a lot in terms of their average pre-intervention (post-intervention) daily step count: the least active ones performed 2500 steps, and the most active ones performed 18,000 steps per day on average.

We define the *absolute improvement* of the user  $i$  as the difference between his or her post-intervention average daily step count  $post_i$  and his or her pre-intervention average daily step count  $pre_i$ . We define the *relative improvement* of the user  $i$  as the relative increase of his or her post-intervention average daily step count compared to his or her pre-intervention average daily step count:

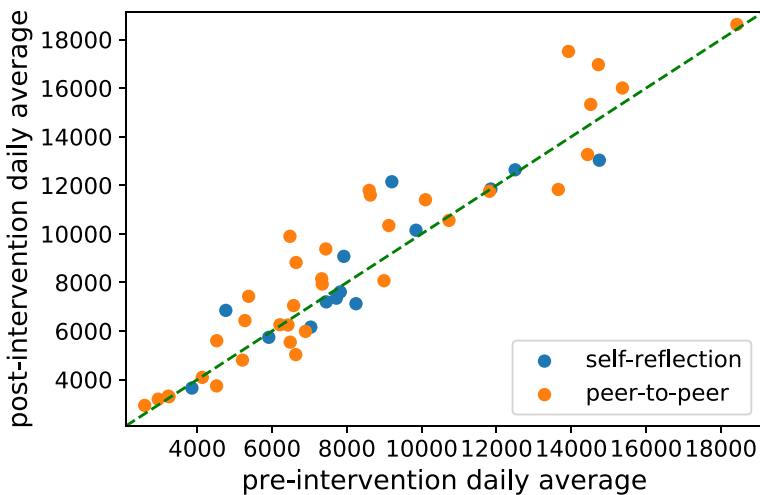
$$\text{relative improvement}_i = \frac{\text{post}_i - \text{pre}_i}{\text{pre}_i}$$





**Fig. 2** Average daily step count per day in the trial (only valid days of data were included in the estimate). Red dashed lines indicate the beginning of each week (Monday)

Increase of 1000 steps might not be much for someone who performs 18,000 steps per day, but might be for someone who performs 2500 steps per day. This is why we are more interested in relative improvement. We assume that the relative improvement is a proxy for the individual intervention effect. The peer-to-peer group improved more on average than the self-reflection group. The improvement was significant for the peer-to-peer group (8.1%, one-sample t-test for the null hypothesis that the mean is positive,  $p$  value = 0.005409), but not for the self-reflection group (2.1%, one-sample t-test for the null hypothesis that the mean is positive,  $p$  value = 0.277266). The main task of the recommender system is to predict the individual’s relative improvement under each of the two interventions given his or her pre-intervention data. This allows the system to select and recommend the optimal intervention for each individual.



**Fig. 3** Pre- vs post-intervention average daily step count per user (only valid days of data were included in the estimate)

## Representation Learning

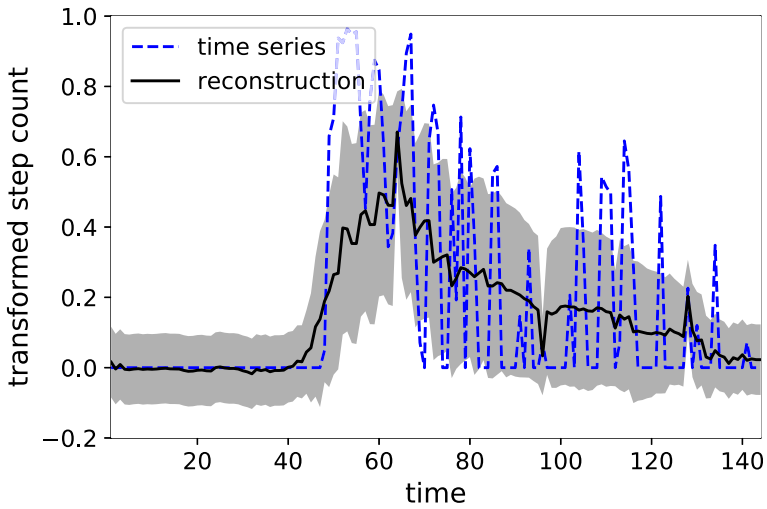
Fitbit provides 1440 step count measurements per day. This granularity makes it difficult for predictive models to gather insights from data. Most prior works (Swartz et al. 2012; Phatak et al. 2018; Sloan et al. 2018; Lee et al. 2019) used aggregated step count data in their analysis. In contrast, we were interested to extract much more information explaining the time series dynamics that could be further used for the prediction task. We used RNN (Recurrent Neural Network) autoencoder to reduce the dimensionality of daily time series and generate embeddings (representing each time series as a lower-dimensional data point) that preserve the low-level information as much as possible. These embeddings were used directly by the predictive models described in the next section.

The autoencoder consists of two parts: an Encoder and a Decoder. The Encoder processes the input time series and produces a low-dimensional embedding. The Decoder tries to reconstruct the original time series given its embedding as input. We model both the Encoder and the Decoder using RNN. The most popular kind of RNN is built using LSTM (Long short-term memory) units (Gers et al. 1999). However, in our model, we use Clockwork RNN because LSTM RNN performs worse than Clockwork RNN in time series reconstruction (Koutnik et al. 2014). Clockwork RNN is an architecture in which the hidden layer is partitioned into separate modules, each processing inputs at its own temporal granularity, making computations only at its prescribed clock rate (Koutnik et al. 2014). As a consequence, long-term information propagates faster through the network.

853 time series collected before the intervention and associated to valid days of data were used to train the model. First, we pre-processed the data using aggregation and data transformation. The aggregation step included segmenting each sequence into 10-min non-overlapping sliding windows and summing up the minute-level step counts belonging to the same window. In this way, we reduced the length of the time series from 1440 to 144, without losing much information about the distribution of the physical activities during the day. Also, in this way, we improved the balance between the observations indicating no activity and the observations with a positive step count. The distribution of the observations with a positive step count was skewed to the right. Thus, we used box-cox transformation (Sakia 1992) to transform these data into a more normal distribution.

Our autoencoder tries to learn parameters of a Normal distribution (mean and standard deviation) for each time step of the Decoder so that it is more likely that the observations are generated from the associated distributions. We specified a minimum value for the standard deviation and we used weight regularization to reduce overfitting. Figure 4 shows a reconstruction of a sample time series. It can be seen that the Decoder was able to capture the high activity in the middle and the low activity in the second part of the time series. We used an embedding space of 10 dimensions.

For visualization purposes, we further reduced these embeddings into two dimensions using t-SNE (Maaten and Hinton 2008). This allowed us to see each time series as a point in a two-dimensional space and to visually validate the embeddings generated by the autoencoder. In Fig. 5, we see that embeddings that are closer to each other represent time series that have similar step counts. In addition, these time series have a similar distribution of the physical activities during the day. This demonstrates that the Encoder has learned to embed similar time series into similar vectors.

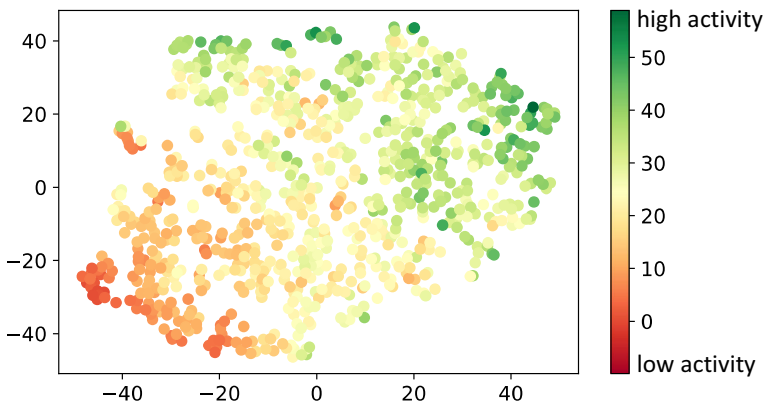


**Fig. 4** Mean and standard deviation of a time series reconstruction generated by the RNN autoencoder. The dashed blue line represents a sample daily time series given as input

**Predictive Modeling**

The main component of our recommender system utilizes the pre-intervention sensor data to predict how a new elderly person would respond if he or she was given the intervention. For this purpose, we trained supervised machine learning models separately on the data from the people that received the self- reflection intervention and the people that received the peer-to-peer intervention. Our models used the features extracted from the pre-intervention time series data to predict his or her *relative improvement*. The main idea behind predictive modeling is that people who behave similarly will respond to the same intervention in a similar way.

Before we apply predictive modeling, we needed to deal with missing data; 17% of the days of pre-intervention data were invalid i.e. did not contain any activity at all. Most machine learning models do not support missing data as input, thus we decided to



**Fig. 5** 2D visualization of the time series embeddings generated by the encoder. The embeddings are visualized using t-SNE. Each point represents a time series. Closer points indicate similar time series

replace the missing values with an estimate. There are correlations in the data that could help us choose a more relevant estimate. An important observation is that users differ in terms of their activity levels, but maintain consistent and periodic behavior from one day to another. We used an imputation method that replaces the missing data by a random time series from the valid days of data generated by the same user on the same day of the week. Alternative data imputation methods are to replace the missing time series with user average, or to generate time series using deep learning. User average is a simple, but unsuitable method because it produces smoothed time series whose distribution is different from the distribution of raw sensor time series data. Deep learning techniques are unsuitable as well because we don't have enough data to learn to generate realistic time series from the conditional distribution.

Using the whole pre-intervention data as a predictor in our machine learning models means that when we deploy the recommender system, it needs to observe a new user for a few weeks before it decides which intervention is better for him or her. Ideally, the user should be observed for as short period as possible. Thus, we were interested in the minimum amount of pre-intervention data that we could use to predict an individual's relative improvement under each intervention. In our experiments, we applied models that take as input either one day or one week of pre-intervention data. Since we had two weeks of pre-intervention data available for each user,<sup>1</sup> we split it into multiple data samples. Each data sample consists of either a day or a week of user's pre-intervention data and his or her relative improvement. Thus, the number of data samples per user was 14 in the first case (one for each day) and 2 in the second case (one for each week).

We scaled our output variable (relative improvement) so that its variance was one. We used root-mean-square error (RMSE) to measure the error of our models in predicting the scaled relative improvement:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{prediction}_i - \text{relative improvement}_i)^2}{N}}$$

where  $N$  is the number of data samples,  $\text{relative improvement}_i$  is the scaled relative improvement of the user associated to the  $i$ -th data sample, and  $\text{prediction}_i$  is a prediction generated by our method, based on the pre-intervention data associated to the  $i$ -th data sample. The generalization ability of our methods was estimated using 10-fold cross-validation. We ensured that data from the same user belonged to the same fold. In this way, the model was tested on users whose data was not used in the training process. We repeated each cross-validation 10 times with a different random partition each time to obtain the mean and the variance of the test error.

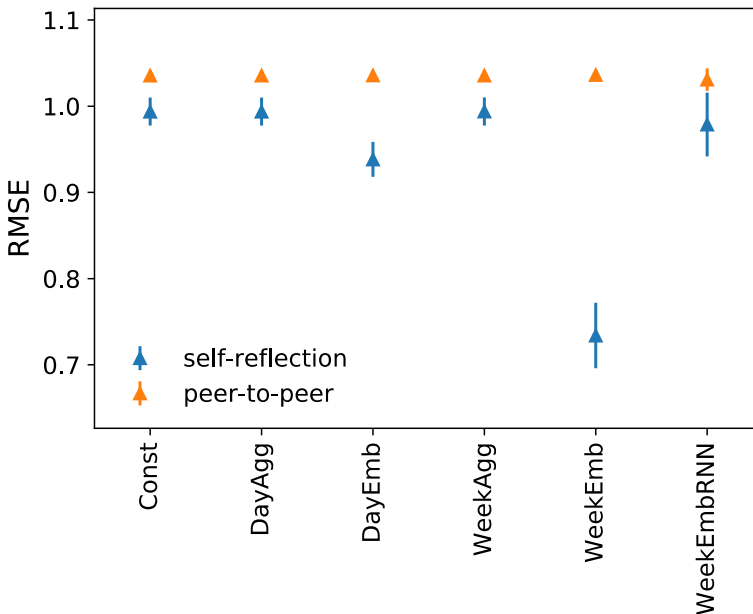
We used six different models to predict the relative improvement under each intervention:

<sup>1</sup> In the experiment, the pre-intervention phase lasted for three weeks, but it started and ended in the middle of the week. Also, some of the participants started with the experiment one or two days later than the others. For more relevant results, in our analysis, we decided to use the pre-intervention data from the two complete weeks, from Monday to Sunday, when all the participants belonged in the pre-intervention phase of the trial.

- **Const.** The simplest model that could be used is to predict that new people would improve according to the mean improvement of the existing people who already received the intervention. Thus, this model does not base its predictions on the pre-intervention data at all. It was expected that this model would score  $RMSE \approx 1$  because the standard deviation of the output variable is 1.
- **DayAgg.** This model is a simple ridge regression that uses one day of data to generate predictions. It does not use the whole time series as a predictor, but only the total (or aggregated) daily step count. The model has a hyper-parameter  $\lambda$  that indicates the regularization strength. Larger  $\lambda$  forces weights to decay more towards zero, so  $\lambda = 0$  means that we do not do any regularization at all.
- **DayEmb.** This model also uses one day of data to generate predictions. In contrast to model DayAgg, it uses the whole time series as a predictor. More specifically, the feature vector contains the time series embeddings provided by the RNN autoencoder in addition to the total step count. Ridge regression is used to generate predictions.
- **WeekAgg.** This model is a ridge regression that uses one week of data to generate predictions. It takes as input a set of features that represent the total daily step counts for each different day of the week.
- **WeekEmb.** This model is a ridge regression that also uses one week of data to generate predictions. It considers the input data as a set of embeddings (plus the total daily step count associated with each embedding).
- **WeekEmbRNN.** This model is RNN that takes the same input as the model WeekEmb. It considers the input data as a sequence of embeddings (plus the total daily step count associated with each embedding). In this way, it tries to extract relevant predictive information from both the temporal dynamics within a single day and the temporal dynamics from one day to another.

The test errors for each model applied separately on the data from the self-reflection and the peer-to-peer group are given in Fig. 6. It can be seen that personalized models (these are all except model Const) applied on the data from the peer-to-peer group performed the same as the sample mean estimator (model Const). This means the either (1) one week of minute-by-minute data does not contain enough information to explain the individual response to the peer-to-peer intervention, or (2) we don't have enough data to learn the individual response. On the other hand, two of the personalized models (model DayEmb and model WeekEmb) applied to the data from the self-reflection group performed much better than the sample mean estimator (model Const). Both models were applied to fine-grained time series data. This means that the low-level information contained in the daily time series data is an important predictor of the intervention effect. Model WeekEmb performs better than model DayEmb. This suggests that the information about the higher-level human behavior in different days of the week contains relevant predictive information as well. It is interesting that WeekEmbRNN does not perform better than WeekEmb although it takes the same input as WeekEmb, but is more complex. This can be explained by the fact that WeekEmbRNN has a large number of parameters, but there is a small amount of training data available.

We continued the analysis by inspecting the predictions generated by the optimal WeekEmb for the self-reflection group on the test set (see Fig. 7). There is a significant positive correlation between the true relative improvements and the predictions



**Fig. 6** Comparison of the test error of different models

(Pearson correlation,  $r = 0.644734$ ,  $p$  value  $< 0.000001$ ). We were also interested to know how accurate would be the continuous predictions if we used just their sign (positive or negative) to predict whether the user will increase or decrease his physical activities after the intervention (a binary prediction). Since 64.29% of the participants who received the self-reflection intervention did not improve their physical activities, the simplest baseline method would predict that every user would not improve. In this way the accuracy of the method would be 64.29%. If we generated predictions about the improvement using our optimal model, but just care about the sign of the predictions, then we would obtain accuracy of 65.36% – not much different than the baseline method. However, if we define a threshold, and we predict the direction (or sign) of the behavior change only if the absolute value of the prediction is larger than this threshold, the accuracy improves. This can be seen in Fig. 8. When the threshold is 0.6 and above, we obtain accuracy of more than 87% – much better than the baseline method. In other words, when the predictions have higher absolute value, we are more certain whether the user will increase or decrease his activities after the intervention. In practice, this means that we could choose a subset of people that are more likely to benefit from the intervention and give the intervention only to those people. However, there is a trade-off: larger subset means lower certainty in the sign of the improvement.

### Generating Recommendations

The best predictive model (model WeekEmb) is used to predict the potential improvement of a new user under each intervention. The recommender system chooses the intervention that is associated with a higher improvement and recommends it to the user. To evaluate the recommendations, we analyzed whether the improvement of a new user that received one of the two interventions is better predicted by the model that

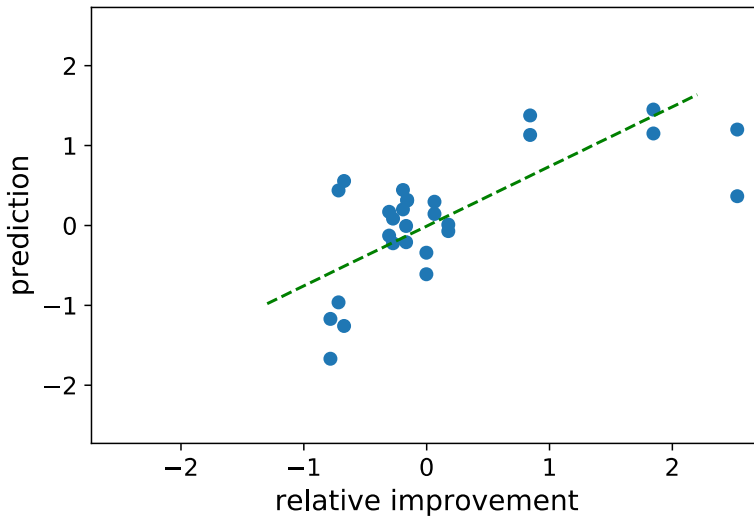


Fig. 7 True relative improvement vs. predicted improvement for the self-reflection group. There are two data points per each user. The dashed green line shows the trend

is trained on existing users that received the same intervention as the target user. We discovered that the RMSE is smaller when we apply the correct model. This observation strengthens the results from the previous section. Our recommender system would give the peer-to-peer intervention to 75% of the people that actually received the self-reflection intervention and the self-reflection intervention to 25.7% of the people that actually received the peer-to-peer intervention. This means that, although the peer-to-peer intervention is more beneficial for the general population, it is likely that 25% of the people would still benefit more from the self-reflection intervention.

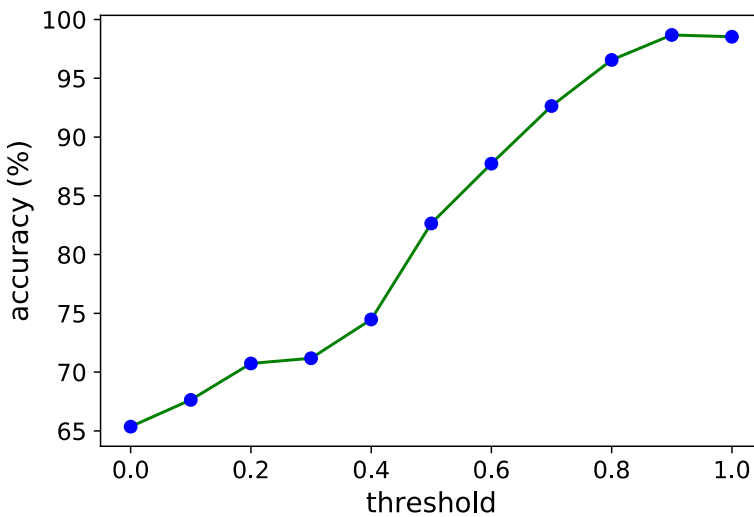


Fig. 8 Percentage of people in the self-reflection group for whom we would predict correctly the direction of the improvement (positive or negative) if we apply the predictions only when their absolute value is larger than a threshold

The recommender system could be designed not to generate any recommendation at all if it predicted that the target user would worsen his or her activity levels under any available intervention. It could also be designed to consider only recommendations that are more likely to cause positive improvement i.e. interventions that are associated with predictions larger than a threshold (Fig. 8). This is especially important for the vulnerable elderly population. The proposed design could also be useful when there is a limited number of interventions and a large population, because in this case the interventions could be given only to the people that are most likely to benefit from them.

## Conclusion

In this paper, we proposed a novel recommender system that aims to promote physical activeness in senior adults. The main novelty of our recommender system is that it uses time series fitness data to predict the intervention effect. These predictions allow the system to select better one intervention over another. We trained the system using data from real senior adults. In our experiment, we used two different mobile app interventions. The first intervention (self-reflection) allowed the user to see a real-time step count information about himself or herself. The second intervention (peer-to-peer) allowed the user to see a real-time step count information about himself or herself and his or her partner. Although the peer-to-peer intervention was more beneficial for the general elderly population, we demonstrated that 25% of the senior adults were still likely to benefit more from the self-reflection intervention. Our personalized predictive models were able to discover who are these people based on their pre-intervention behavior. We showed that models that utilize fine-grained sensor data from a longer period (one week) perform better. This suggests that both the lower-level human behavior within a single day and the higher-level human behavior from one day of the week to another are important predictors of behavior change in senior adults.

The following use case demonstrates the use of our recommender system. Consider an elderly person who wants to become more physically active, but doesn't know how to achieve that. He or she installs our recommendation app and starts wearing an unobtrusive fitness tracker. After one week, the recommendation app learns his or her behavior patterns and recommends him or her to start using either the mobile app that incorporates self-reflection or the mobile app that incorporates social reflection to promote physical activeness. The target user installs the recommended app and improves his or her physical activity levels over time.

In the beginning of our research, we identified several main challenges for developing a recommender system whose main purpose is to select those interventions that are most likely to work for senior adults. In our recommender system design and implementation, we ensured that our system addresses all these challenges. First, we used mobile app interventions based on two motivational strategies that were shown to be successful in the scientific literature: self-reflection and social reflection. Second, we trained our predictive models using only data from senior adults. In this way, the models learned behavioral patterns that are characteristic of this vulnerable subpopulation. Third, we collected data from a trial in which participants were wearing a smart wristband that tracked their activities without interfering with their normal functioning. Fourth, our machine learning models utilized randomized trial data that allowed them to



make a more relevant comparison between the different interventions. Fifth, the generated predictions are based on how existing people that are similar to the target user responded to the same intervention, thus, there is evidence about the effectiveness of the recommendations. Finally, our recommender system utilizes fully the frequently-sampled time series data and learns relevant predictive information from it in an automated way, without human interference.

Our recommender system is scalable and fast to train. It also supports multi-variate time series and multiple interventions. For example, the same physical activities may result in different heart rates in different people. Thus, heart rate time series could be an important predictor for behavior change besides step count time series. Other time-invariant predictors (e.g., sex) and contextual factors (e.g., weather) could also be used to explain the behavior change. Our recommendations system supports both simpler and more complex predictive models, such as deep learning. The latter could generate more accurate predictions. However, deep learning methods require a large amount of data. We showed that LSTM performs much worse than ridge regression on our dataset from a limited number of users. In this paper, we trained our recommendation system to choose one out of two different interventions. However, the system could be easily extended if data for more than two interventions were available. In this case, we would train a separate predictive model on the data associated with each different intervention.

Personalized recommendations for increased physical activity are of great practical value for senior adults. We believe that our system for personalized recommendations is an important contribution in this field because it learns relevant predictive information from unlabeled time series sensor data that is easy to collect, in an automated way. In our future work, we plan to perform an online evaluation of our recommender system.

**Acknowledgements** The research leading to these results presented in this article has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690425.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aitken M, Clancy B, Nass D (2017) The growing value of digital health: Evidence and impact on human health and the healthcare system. IQVIA Institute for Human Data Science.
- Albert, M. V., Kording, K., Herrmann, M., & Jayaraman, A. (2012). Fall classification by machine learning using mobile phones. *PLoS One*, 7(5), e36556.
- Ballinger, B., Hsieh, J., Singh, A., Sohoni, N., Wang, J., Tison, G.H., Marcus, G.M., Sanchez, J.M., Maguire, C., Olgin, J.E., et al. (2018) Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction. In: Thirty-Second AAAI Conference on Artificial Intelligence.
- Bates, S. (2010). Progress towards personalized medicine. *Drug Discovery Today*, 15(3–4), 115–120.
- Carnethon, M. R. (2009). Physical activity and cardiovascular disease: How much is enough? *American Journal of Lifestyle Medicine*, 3(1 suppl), 44S–49S.

- Chekroud, S. R., Gueorguieva, R., Zheutlin, A. B., Paulus, M., Krumholz, H. M., Krystal, J. H., & Chekroud, A. M. (2018). Association between physical exercise and mental health in 1.2 million individuals in the USA between 2011 and 2015: A cross-sectional study. *The Lancet Psychiatry*, 5(9), 739–746.
- Chiang, P.H., Dey, S., (2018) Personalized effect of health behavior on blood pressure: Machine learning based prediction and recommendation. In: 2018 IEEE 20th international conference on e-health networking, Applications and Services (Healthcom), IEEE, pp 1–6.
- Conn, V. S., Hafidahl, A. R., & Mehr, D. R. (2011). Interventions to increase physical activity among healthy adults: Meta-analysis of outcomes. *American Journal of Public Health*, 101(4), 751–758.
- Direito, A., Dale, L. P., Shields, E., Dobson, R., Whittaker, R., & Maddison, R. (2014). Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? *BMC Public Health*, 14(1), 646.
- Drewnowski, A., & Evans, W. J. (2001). Nutrition, physical activity, and quality of life in older adults: Summary. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(suppl 2), 89–94.
- Evans, W. J. (1999). Exercise training guidelines for the elderly. *Medicine and Science in Sports and Exercise*, 31(1), 12–17.
- Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M. A., Smith, A. E., Tang, K., Yuan, C. W., et al. (2018). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet*, 392(10159), 2052–2090.
- Gers, F.A., Schmidhuber, J., Cummins, F., (1999) Learning to forget: Continual prediction with LSTM. In: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), vol 2, pp 850–855 vol.2, DOI <https://doi.org/10.1049/cp:19991218>.
- Greaves, C. J., Sheppard, K. E., Abraham, C., Hardeman, W., Roden, M., Evans, P. H., & Schwarz, P. (2011). Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. *BMC Public Health*, 11(1), 119.
- Kotsev, G., Nguyen, L.T., Zeng, M., Zhang, J., (2014) User exercise pattern prediction through mobile sensing. In: 6th international conference on Mobile computing, Applications and Services, IEEE, pp 182–188.
- Koutnik, J., Greff, K., Gomez, F., Schmidhuber, J., (2014) A clockwork RNN. arXiv preprint arXiv:14023511.
- Kulev, I., Pu, P., & Faltings, B. (2018). A Bayesian approach to intervention-based clustering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(4), 44.
- Kurashima, T., Althoff, T., Leskovec, J., (2018) Modeling interdependent and periodic real-world action sequences. In: *Proceedings of the 2018 World wide web conference*. International WWW Conference, NIH Public Access, vol 2018, p 803.
- Lee, I.M., Shiroma, E.J., Kamada, M., Bassett, D.R., Matthews, C.E., Buring, J.E., (2019) Association of step volume and intensity with all-cause mortality in older women. *JAMA internal medicine*.
- Leitzmann, M. F., Park, Y., Blair, A., Ballard-Barbash, R., Mouw, T., Hollenbeck, A. R., & Schatzkin, A. (2007). Physical activity recommendations and decreased risk of mortality. *Archives of Internal Medicine*, 167(22), 2453–2460.
- Lewis, B. A., Napolitano, M. A., Buman, M. P., Williams, D. M., & Nigg, C. R. (2017). Future directions in physical activity intervention research: Expanding our focus to sedentary behaviors, technology, and dissemination. *Journal of Behavioral Medicine*, 40(1), 112–126.
- Luštrek, M., & Kaluža, B. (2009). Fall detection and activity recognition with machine learning. *Informatica*, 33(2).
- Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.
- Milanović, Z., Pantelić, S., Trajković, N., Sporiš, G., Kostić, R., & James, N. (2013). Age-related decrease in physical activity and functional fitness among elderly men and women. *Clinical Interventions in Aging*, 8, 549.
- Moller, A. C., Merchant, G., Conroy, D. E., West, R., Hekler, E., Kugler, K. C., & Michie, S. (2017). Applying and advancing behavior change theories and techniques in the context of a digital health revolution: Proposals for more effectively realizing untapped potential. *Journal of Behavioral Medicine*, 40(1), 85–98.
- Na, K. S. (2019). Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Scientific Reports*, 9(1), 3335.
- Ng, A., (2013) Machine learning and AI via brain simulations. Accessed: May 3:2018.
- Ory, M. G., & Cox, D. M. (1994). Forging ahead: Linking health and behavior to improve quality of life in older people. *Social Indicators Research*, 33(1–3), 89–120.

- Park, C. H., Elavsky, S., & Koo, K. M. (2014). Factors influencing physical activity in older adults. *Journal of exercise rehabilitation*, *10*(1), 45–52.
- Phatak, S. S., Freigoun, M. T., Martín, C. A., Rivera, D. E., Korinek, E. V., Adams, M. A., Buman, M. P., Klasnja, P., & Hekler, E. B. (2018). Modeling individual differences: A case study of the application of system identification for personalizing a physical activity intervention. *Journal of Biomedical Informatics*, *79*, 82–97.
- Prince, M. J., Wu, F., Guo, Y., Robledo, L. M. G., O'Donnell, M., Sullivan, R., & Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. *The Lancet*, *385*(9967), 549–562.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 18.
- Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y., (2017) Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:170701836.
- Robinson, S. A., Bisson, A. N., Hughes, M. L., Ebert, J., & Lachman, M. E. (2019). Time for change: Using implementation intentions to promote physical activity in a randomised pilot trial. *Psychology & Health*, *34*(2), 232–254.
- Sakia, R. (1992). The box-cox transformation technique: A review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(2), 169–178.
- Sasaki, J. E., Hickey, A., Staudenmayer, J., John, D., Kent, J. A., & Freedson, P. S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine and Science in Sports and Exercise*, *48*(5), 941–950.
- Sau, A., & Bhakta, I. (2017). Predicting anxiety and depression in elderly patients using machine learning technology. *Healthcare Technology Letters*, *4*(6), 238–243.
- Seefeldt, V., Malina, R. M., & Clark, M. A. (2002). Factors affecting levels of physical activity in adults. *Sports Medicine*, *32*(3), 143–168.
- Shalit, U., Johansson, F.D., Sontag, D., (2017) Estimating individual treatment effect: Generalization bounds and algorithms. In: *Proceedings of the 34th International Conference on Machine Learning*-volume 70, JMLR. Org, pp 3076–3085.
- Sloan, R. A., Kim, Y., Sahasranaman, A., Müller-Riemenschneider, F., Biddle, S. J., & Finkelstein, E. A. (2018). The influence of a consumer-wearable activity tracker on sedentary time and prolonged sedentary bouts: Secondary analysis of a randomized controlled trial. *BMC Research Notes*, *11*(1), 189.
- Sullivan, A. N., & Lachman, M. E. (2017). Behavior change with fitness technology in sedentary adults: A review of the evidence for increasing physical activity. *Frontiers in Public Health*, *4*, 289.
- Swartz, A. M., Tarima, S., Miller, N. E., Hart, T. L., Grimm, E. K., Rote, A. E., & Strath, S. J. (2012). Prediction of body fat in older adults by time spent in sedentary behavior. *Journal of Aging and Physical Activity*, *20*(3), 332–344.
- The Lancet Public Health. (2017). Ageing: A 21st century public health challenge? *The Lancet Public Health*, *2*(7), e297.
- United Nations, Department of Economic and Social Affairs (2017) World population ageing 2017: highlights.
- Xie, J., Wen, D., Liang, L., Jia, Y., Gao, L., & Lei, J. (2018). Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study. *JMIR mHealth and uHealth*, *6*(4), e94.
- Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E., & Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, *163*(5), 1079–1094.