A Machine Learning Approach to Classifying Self-Reported Health Status in a cohort of Patients with Heart Disease using Activity Tracker Data

Yiwen Meng, William Speier, *Member*, Chrisandra Shufelt, Sandy Joung, Jennifer E Van Eyk, C. Noel Bairey Merz, Mayra Lopez, Brennan Spiegel and Corey W. Arnold

***Abstract*— Constructing statistical models using personal sensor data could allow for tracking health status over time, thereby enabling the possibility of early intervention. The goal of this study was to use machine learning algorithms to classify patient-reported outcomes (PROs) using activity tracker data in a cohort of patients with stable ischemic heart disease (SIHD). A population of 182 patients with SIHD were monitored over a period of 12 weeks. Each subject received a Fitbit Charge 2 device to record daily activity data, and each subject completed eight Patient-Reported Outcomes Measurement Information Systems (PROMIS®) short form at the end of each week as a self-assessment of their health status. Two models were built to classify PRO scores using activity tracker data. The first model treated each week independently, while the second used a Hidden Markov model (HMM) to take advantage of correlations between successive weeks. Retrospective analysis compared the classification accuracy of the two models and the importance of each feature. In the independent model, a random forest classifier achieved a mean area under curve (AUC) of 0.76 for classifying the Physical Function PRO. The HMM model achieved significantly better AUCs for all PROs ($p<0.05$) other than Fatigue and Sleep Disturbance, with a highest mean AUC of 0.79 for the Physical Function-short form 10a. Our study demonstrates the ability of activity tracker data to classify health status over time. These results suggest that patient outcomes can be monitored in real time using activity trackers.**

***Index Terms*—Clinical diagnosis, machine learning, patient monitoring, telemedicine, wearable sensors.**

This work was supported by the California Initiative to Advance Precision Medicine (CIAPM) (BS, NBM, and JVE); as well as the National Heart, Lung, and Blood Institute (NIH/NHLBI R56HL135425, R01HL141773 CWA; K23HL127262, CS); the National Center for Research Resources (NIH/NCRR UL1RR033176); the National Center for Advancing Translational Sciences (NCATS) and UCLA Clinical Translational Science Institute (CTSI) (NIH/NCATS UL1TR000124); the Advanced Clinical Biosystems Research Institute (JVE); the Erika Glazer Endowed Chair in Women's Heart Health (NBM and JVE); and the Barbra Streisand Women's Cardiovascular Research and Education Program.

Y. Meng, W. Speier and C.W. Arnold are with the Computational Integrated Diagnostics Lab, the Department of Bioengineering, the Department of Radiology, and the Department of Pathology at the University of California Los Angeles, 924 Westwood Blvd, Suite 420, CA 90024 USA (e-mails: lanyexiaosa@ucla.edu, speier@ucla.edu, cwarnold@ucla.edu).

C. Shufelt, S. Joung and JEV Eyk are with the Barbra Streisand Women's Heart Center, Smidt Heart Institute, Los Angeles, California, USA, Los Angeles, CA 90048 USA.

M. Lopez and B. Spiegel are with the Centers for Outcomes Research and Education, Cedars-Sinai Medical Center, Los Angeles, California, USA.

I. Introduction

There has been significant effort in developing monitoring devices and protocols to diagnose patients remotely. However, device fatigue has been shown to be a barrier to adherence [1]–[3]. Commercially available devices, such as passive accelerometry, have been shown to overcome this barrier by reducing the burden of human intervention [4], and activity tracker accuracy has been demonstrated to be sufficient for documenting health indicators in real-time [5]–[7]. With wireless connections to portable electronics, such as smartphones or tablets, monitoring by activity tracker is an easy-to-use, accessible means of providing personalized information to peoples' health and daily activities [8]. This approach creates a feedback loop that is capable of positively impacting health interventions with the goal of lifestyle change [9], [10]. However, analysis of this data has largely been limited to simple correlations, and the ability to use this information to classify patient health status has not been explored [11], [12].

Patient-reported outcome (PRO) questionnaires are designed to capture a patient's perspective and experience of their own health and to provide valid and reliable data [13]–[15]. However, response fatigue is a common problem that can result in missing data due to an incomplete response from the subject, resulting in misclassification [16], [17]. Response fatigue can be common when an administered survey is too long, or when a survey is short, but administered too frequently. Because of these drawbacks, data collected from a less invasive method could potentially provide more reliable estimates of patient health status over time. Previous studies have shown high compliance in activity trackers, indicating that they may be more reliable methods for tracking continuous patient data [4], [18].

In this study, we explore the use of machine learning methods to classify PRO scores over time [14]. Machine learning algorithms have been widely used in biomedical research for tasks such as disease detection [19] and outcome prediction [20]. These methods traditionally use a set of demographic variables and baseline data as a feature vector to make a classification using a machine learning algorithm such as gradient boosting regression tree (GBRT) [21], AdaBoost [22], or random forests (RF) [20], [23]. However, traditional machine learning methods are effective for making single decisions, but do not allow for adjusting as more information

is learned. Temporal models are appropriate in the case of sequential observations where the value of the outcome may need to be adjusted over time. In particular, hidden Markov models (HMMs) are well-established temporal models that use sequential data to predict events such as patient state changes, such as estimating mean sojourn time of lung cancer patients using screening images [24], detecting homologous protein sequences [25], and gene finding [26].

The goal of this study was to investigate the feasibility of using machine learning models to classify PRO scores based on data collected using one type of activity tracker, the Fitbit Charge 2. In this study, we tested this goal within a population of patients with stable ischemic heart disease (SIHD). The rest of this article describes an approach for data preprocessing and constructing a model that treats weeks independently, as well as an HMM that takes temporal information into account. Performance of the classification algorithms is then evaluated for each PRO measure and feature importance in classification is analyzed. Finally, we provide a discussion and analysis of the results and suggest future directions for implementing such a classifier in a patient surveillance application.

II. Data Description

A set of 200 patients with SIHD were recruited for a feasibility study conducted by Cedars-Sinai Medical Center from 2017 to 2018 to predict surrogate markers of major adverse cardiac events (MACE), including myocardial infarction, arrhythmia, and hospitalization due to heart failure, using biometrics, wearable sensors, patient-reported surveys, and other biochemical markers. This study population size is similar to several previous studies that used activity trackers for patient monitoring [27], [28]. The desired monitoring period was 12 weeks for each subject, during which time subjects wore personal activity trackers to record their physiological indices, including steps, heart rate, calories burned, and distance traveled. At the end of each week, they were asked to fill out eight PROMIS short forms as a self-report assessment of their health status [4].

TABLE I
SUMMARY OF 17 TYPES OF FEATURE COLLECTED FROM FITBIT PER DAY

Type (units)	Mean ± Std
Steps (#)	6138 ± 4031
Total Distance (kilometers)	4.18 ± 3.00
Tracker Distance* (kilometers)	4.18 ± 3.00
Logged Activity Distance* (kilometers)	0.02 ± 0.56
Very Active Distance (kilometers)	0.71 ± 1.49
Moderate Active Distance (kilometers)	0.36 ± 0.60
Light Active Distance (kilometers)	2.69 ± 1.90
Sedentary Active Distance* (kilometers)	0.01 ± 0.08
Very Active Minutes	12.21 ± 22.29
Fairly Active Minutes	12.78 ± 21.89
Light Active Minutes	176.81 ± 99.73
Sedentary Minutes	823.24 ± 323.90
Calories	2032 ± 610
Floor (#)	5.1 ± 11.8
Calories BMR (basal metabolic rate)	1428 ± 254
Marginal Calories	372 ± 317
Resting Heart Rate (BPM)	61.81 ± 7.45

*means that feature was eliminated for model input because it was highly sparse or redundant.

A. Activity Data

The Fitbit Charge 2 (Fitbit Inc., San Francisco, CA, USA) is a popular commercially available activity tracker that can record a person's daily activities and health indices like heart rate, steps, and sleep (Table I). Previous work has validated the accuracy of heart rate monitoring specifically in the Fitbit Charge 2 [29]. The Fitbit hardware and its computational algorithms for calculating step counts and physical activity have been validated using other Fitbit devices [30], [31]. The Fitbit Charge 2 estimates activity using metabolic equivalents (METs), which are calculated based on heart rate and distance traveled [32]. Heart rate during activity is also provided, however it has been shown to be inaccurate during activity [33]. Data quality was assured by verifying that there were no extreme outliers based on subject-specific inter-quartile range [34]. We aggregated the data for each day to compensate for noise and redundancy. After data preprocessing, tracker distance was eliminated because it was identical to total distance, and logged activity distance and sedentary active distance were also deleted because of high sparsity. As a result, there were 14 features per day for each patient in our model.

B. Patient-Reported Outcome Measures

Patient-Reported Outcomes Measurement Information Systems (PROMIS®) questionnaires are a library of instruments developed and validated to measure many domains of physical and mental health [15]. This analysis uses data from eight PROMIS instruments: Global Physical Health and Global Mental Health, which are two composite scores from the Global-10 short form [35]; Fatigue-Short Form 4a; Physical Function-Short Form 10a; Emotional Distress-Anxiety-Short Form 6a; Depression-Short Form 4a; Social Isolation-Short Form 4a; and Sleep Disturbance-Short Form 4a. Each questionnaire either asks about current health or has a recall period of the previous seven days, so they are appropriate for weekly administration. The T metric method was used to standardize scores for each type to a mean of 50 and a standard deviation of 10, with a range between 0 and 100 [15], [36]. Symptom (i.e., Fatigue, Anxiety, Depression, Social Isolation, and Sleep Disturbance) scores of 60 or higher

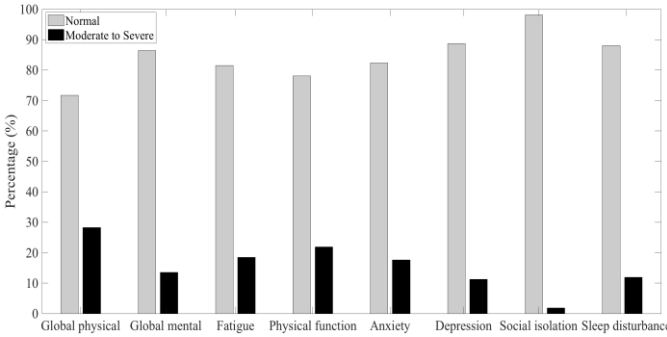

Fig. 1. Distribution of normal and abnormal (moderate to severe) class for each PRO measure.

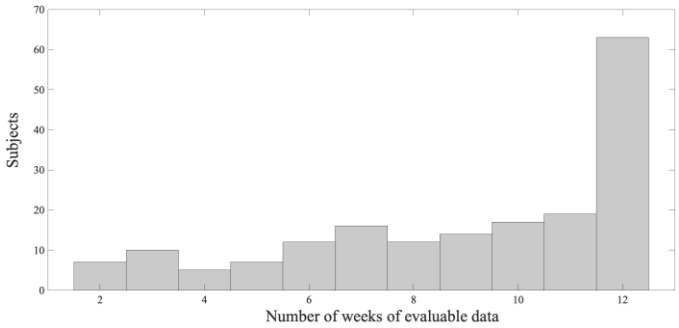

Fig. 2. Histogram of number of weeks of evaluable data for the 182 subjects used in the dataset.

are one standard deviation above the average, which is defined as moderate to severe symptom severity. For function (i.e., Global Physical Health, Global Mental Health, and Physical Function), scores less than 40 are classified as moderate to severe, meaning less functional ability than normal. For this study, PRO scores were predicted in two ways: regression was used to predict PRO scores from patient activity tracker data, and classification was used to determine whether subjects' PRO scores were above the threshold for at least moderate severity. The distributions of PRO scores are shown in Fig. 1. Because of a lack of moderate or severe cases for social isolation (<2%), this variable was eliminated for analysis in our model.

III. METHODS

Missing data is a common concern when dealing with activity tracker data and can result from subjects either forgetting to wear their devices or removing them for charging. Patients were asked to fill out eight PROMIS questionnaires at the end of each week for a 12-week monitoring period. In total, 19.1 percent of weeks had missing PRO data and 16.6 percent of weeks had missing values from the activity tracker in four or more days. If data was available for at least four days in a week, missing values were permuted by using the average value of the rest of the week for steps or resting heart rate. Weeks with missing survey scores, as well as those without step and resting heart rate data for more than three days, were removed from the analysis.

A correlation analysis between subjects' missing Fitbit data and their average Global Physical Health and Global Mental Health scores shows a slight negative relationship (-0.11 and -0.09, respectively) that was not statistically significant (p=0.13 and p=0.23, respectively). The correlation coefficient between number of missing PROs and the average global health scores are -0.17 (p=0.018) to -0.14 (p=0.048), respectively, indicating that the missing PROs are significantly related to patient health. Another correlation analysis was performed between subject's age and number of missing values with r^2 <0.001, which demonstrates no trend of more missing values for elder subjects. Finally, subjects with only one week of data were eliminated in order to ensure the continuity of transition of states from week to week when building the HMM model. After adopting this data preprocessing approach and using the classification criteria above, a total number of 182 subjects with a total of 1,640 weeks were collected, where the number of weeks of evaluable data for each patient ranged from two to 12 weeks as shown in Fig. 2.

A. Independent per week Model by Machine Learning Algorithms

Since survey scores were generated per week, a naive approach for using this data is to treat each week independently. The left plot in Fig. 3 illustrates the idea of the independent model as an example for one subject with a number of weeks of evaluable data of 12 weeks. The features for each of the seven days were appended into a single feature vector, which was then used as the input for binary classification of each PRO score. Ensemble methods like Adaboost, GBRT (gradient boosting regression tree) and Random Forest (RF) are relatively robust over unbalanced dataset and is capable of generating better classification accuracy than other types of machine learning algorithms [37]. Each of these methods was applied to the dataset using ten-fold cross-validation across subjects in conjunction with grid search to find the optimal parameters for each model. T-tests were applied to validate the statistical significance for each comparison of the result with different p values: 0.05, 0.01 as for different levels of significance. A sensitivity analysis was completed to investigate the model performance against missing values in feature vector by randomly withholding values from one to six days within a week.

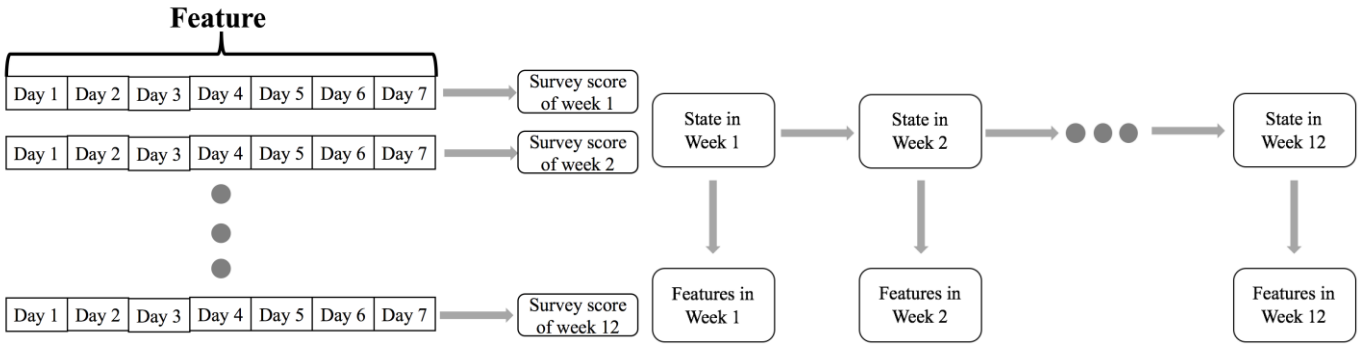

Fig. 3. Illustration of independent week model (left) and Hidden Markov Model (right). For HMM, feature in each week was observed while the state of health status transits from week to week.

B. Hidden Markov Model (HMM) with Forward Algorithm

In order to track changes in PRO responses over time, a model was built to incorporate temporal correlations of PRO scores across weeks. As shown in the right part of Figure 3, an HMM was used and formalized such that the state at each time point corresponded to the PRO score for that week, with the features collected for that week treated as observations. The transition matrix was derived by counting the s state transitions from week to week. The original number of states for each PRO was found by number of unique responses, ranging from 15 to 36. In order to make the transition matrix less sparse, we defined 10 states for all types of health status based on the score distribution of each PRO. The Forward algorithm computed the probability across states at time t, with the maximum probability representing the classified state,

$$S(y_t|y_{t-1},\dots,y_1,x_t,\dots,x_1) = P(x_t|y_t) * \sum \mathrm{P}(\mathrm{y_t}|\mathrm{y_{t-1}}) * \mathrm{S}(\mathrm{y_{t-1}}|\mathrm{y_{t-2}},\dots,\mathrm{x_{t-1}},\dots) \quad (1)$$

where the weekly PRO score was treated as state y_t, with observation of features x_t. The emission probability, $P(x_t|y_t)$, computed the probability of the observed feature vector x_t given state y_t, computed from the random forest classifier and $P(y_t)$:

$$P(x_t|y_t) \propto \frac{P(y_t|x_t)}{P(y_t)} \quad (2)$$

At the first-time step, the transition probability distribution is undefined, so the state probability was:

$$S(y_1|x_1) \propto P(x_1|y_1)P(y_1) \quad (3)$$

For analysis, states were binarized according to the criteria defined above. Because dichotomizing PRO score values loses some information and precision, a regression analysis was conducted between the median value of HMM stages and actual scores for the HMM. This method of predicting PRO scores was compared against multinomial logistic regression to evaluate the accuracy of predicting PRO scores over time.

TABLE II
MEAN AND STANDARD DEVIATION ROCAUC OF DIFFERENCE ALGORITHMS. BOLD VALUES ARE THE HIGHEST FOR A GIVEN PRO

Type	AdaBoost	GBRT	Random Forest
Global physical health	0.72 (0.03)	0.69 (0.04)	**0.73 (0.01)***
Global mental health	0.53 (0.03)	0.51 (0.03)	**0.55 (0.03) ***
Fatigue	0.59 (0.04)	0.60 (0.04)	**0.61 (0.03)**
Physical function	0.74 (0.03)	0.75 (0.03)	**0.75 (0.01)**
Anxiety	0.48 (0.03)	0.50 (0.03)	**0.54 (0.02)** †
Depression	0.47 (0.04)	0.50 (0.03)	**0.53 (0.02)** †
Sleep Disturbance	0.55 (0.06)	0.59 (0.05)	**0.61 (0.03)**

* Significant improvement over GBRT.
† Significant improvement over both GBRT and AdaBoost.

IV. RESULTS

Table. II shows the mean AUC for binary classification of PRO scores for the seven PROMIS measures using GBRT, AdaBoost and RF. The highest mean AUC was 0.75 using RF for classifying Physical Function, while the lowest was 0.47 using AdaBoost for Depression. The results indicated that RF significantly outperformed other models in classification of Anxiety and Depression ($p<0.05$), and it was also significantly better than GBRT for Global Physical Health and Mental Health ($p=0.01$ and $p=0.01$, respectively). The RF model was selected for the remaining analyses because its

TABLE III
IMPORTANCE FACTOR OF EACH FEATURE FOR CLASSIFYING VARIOUS HEALTH STATUS

Type	Global physical health	Global mental health	Fatigue	Physical Function	Anxiety	Depression	Sleep Disturbance
Step	**0.138 (0.010)**	**0.088 (0.006)**	**0.111 (0.009)**	**0.138 (0.007)**	0.075 (0.005)	**0.080 (0.006)**	**0.104 (0.005)**
Total Distance	**0.122 (0.009)**	**0.081 (0.003)**	**0.100 (0.009)**	**0.123 (0.011)**	**0.079 (0.004)**	0.075 (0.004)	**0.088 (0.005)**
Very Active Distance	0.062 (0.005)	0.066 (0.006)	0.061 (0.004)	0.068 (0.005)	0.065 (0.003)	0.056 (0.002)	0.053 (0.004)
Moderately Active Distance	0.059 (0.007)	0.058 (0.002)	0.052 (0.003)	0.051 (0.003)	0.052 (0.002)	0.054 (0.002)	0.0573 (0.003)
Light Active Distance	**0.088 (0.010)**	0.071 (0.002)	0.074 (0.001)	**0.086 (0.010)**	0.070 (0.005)	0.066 (0.003)	0.069 (0.004)
Very Active Minutes	0.054 (0.007)	0.060 (0.006)	0.052 (0.004)	0.060 (0.006)	0.057 (0.004)	0.053 (0.002)	0.055 (0.004)
Fairly Active Minutes	0.055 (0.009)	0.054 (0.002)	0.050 (0.007)	0.050 (0.006)	0.0530 (0.003)	0.051 (0.003)	0.063 (0.004)
Light Active Minutes	0.068 (0.006)	0.072 (0.005)	0.064 (0.003)	0.061 (0.005)	0.066 (0.002)	0.073 (0.002)	0.067 (0.004)
Sedentary Minutes	0.046 (0.003)	0.066 (0.005)	0.055 (0.002)	0.043 (0.001)	0.065 (0.007)	0.072 (0.007)	0.060 (0.005)
Calories	0.059 (0.006)	**0.088 (0.008)**	0.074 (0.007)	0.053 (0.005)	**0.098 (0.007)**	**0.094 (0.004)**	**0.082 (0.003)**
Floors	0.055 (0.008)	0.049 (0.001)	0.056 (0.010)	0.076 (0.019)	0.048 (0.003)	0.053 (0.003)	0.051 (0.005)
Calories BMR	0.073 (0.004)	**0.110 (0.016)**	**0.105 (0.007)**	0.071 (0.009)	**0.117 (0.006)**	**0.128 (0.009)**	**0.105 (0.012)**
Marginal Calories	0.066 (0.004)	0.071 (0.003)	0.060 (0.003)	0.058 (0.005)	0.070 (0.006)	0.069 (0.003)	0.077 (0.006)
Resting Heart Rate	0.058 (0.012)	0.067 (0.008)	0.085 (0.012)	0.063 (0.008)	**0.085 (0.003)**	**0.076 (0.002)**	0.069 (0.016)

Value in parentheses is the standard deviation. Bold values are significantly higher ($p<0.05$) than the average value for a feature (1/14 = 0.0714).

performance was equivalent to or better than the other methods for classifying all PRO scores. Additionally, it was notable that the AUC related to self-reported physical health PROs such as Global Physical Health, Fatigue, and Physical Function were higher than those related to mental health such as Global Mental Health, Anxiety, and Depression.

We then looked at the importance factor of each feature contributing to the classification in the RF model. Table III displays the importance factor for the 14 feature types summed over seven days. Features that were significantly higher ($p<0.05$) than the average value for each classification were determined. Steps, total distance, calories, and calories BMR contributed to most of the PRO scores. The importance factor of light active distance was significantly better than other features for classifying Global Physical Health and Physical Function, which were both related to a subject's physical health. On the other hand, resting heart rate contributed significantly more than other features for classification of mental health PROs such as Anxiety and Depression, while its importance factor was not significantly higher than other features in classification of PROs related to physical health.

TABLE IV
MEAN AND STANDARD DEVIATION ROCAUC OF DIFFERENT FEATURE SELECTION STRATEGY. BOLD VALUES ARE THE HIGHEST FOR A GIVEN PRO

Type	Steps Only	All Feature	Selected Feature
Global physical health	0.73 (0.03)	0.73 (0.01)	**0.73 (0.02)**
Global mental health	0.52 (0.02)	0.55 (0.03)†	**0.58 (0.02)***
Fatigue	0.60 (0.05)	0.61 (0.03)	**0.64 (0.03)***
Physical function	0.76 (0.03)	0.75 (0.01)	**0.76 (0.01)***
Anxiety	0.50 (0.04)	0.54 (0.02)†	**0.57 (0.02)***
Depression	0.51 (0.02)	0.53 (0.02)†	**0.56 (0.02)***
Sleep Disturbance	0.59 (0.03)	0.61 (0.03)	**0.64 (0.03)***

* Significant improvement from Selected Feature over All Feature.
† Significant improvement from All Feature over Steps Only.

The analysis was repeated using the RF classifier and only the significant features from Table III and are shown in Table IV. Because some studies such as [38] only used steps data to assess user's health status, we also compared the model performance in the same manner. The result suggested that the RF model can generate significantly better classification accuracy with the selected features than all features from Fitbit for all PROMIS short form survey scores except for Global Mental Health ($p=0.37$), with the highest AUC of 0.76 for classification of Physical Function.

Fig. 4 illustrates the results of sensitivity analysis on missing feature data on RF classification by randomly censoring data from one day to six days per a week. The results show that ROCAUC decreased monotonically as days were removed. For Global Physical Health, the value at missing four days drops significantly compared to no missing data ($p=0.03$), while the difference at missing three days was not significant ($p=0.11$). This was why that cutoff was chosen for inclusion in our analysis.

Table V displayed the comparison of means and standard deviations of the AUC for each PRO measure using the

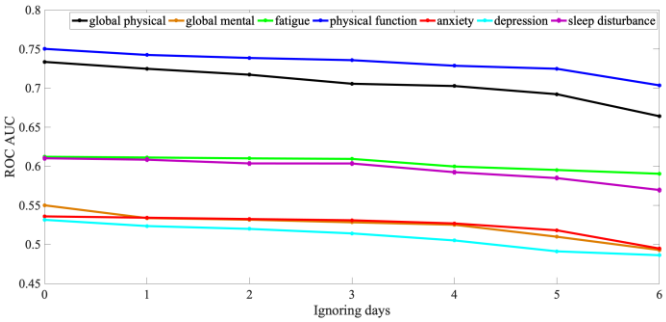

Fig. 4. Plot of ROCAUC for each type of PRO after randomly withhold feature values from one day to six days within a week.

independent model and HMM. AUCs derived using the HMM were significantly higher than those from the independent model in all domains other than Fatigue and Sleep Disturbance. Depression achieved the highest increase from 0.57 to 0.61. We also compared the R^2 value of the regression analysis between HMM and multinomial logistic regression. The value were 0.079 and 0.1526 from HMM in Global Physical Health and Physical Function. They were significantly better than the values achieved by the multinomial logit model (0.0016 and 0.0026, respectively; $p<0.001$ for both). This result suggested that HMM could also track the minor change of PRO score with higher precision over time than baseline models like multinomial logistic regression.

TABLE V
MEAN AND STANDARD DEVIATION OF AUC VALUES BETWEEN THE INDEPENDENT WEEK MODEL AND THE HIDDEN MARKOV MODEL. BOLD VALUES ARE THE HIGHEST AUC FOR A GIVEN PRO

Type	Independent model	HMM
Global physical health	0.73 (0.02)	**0.76 (0.02)***
Global mental health	0.58 (0.01)	**0.61 (0.02)***
Fatigue	0.64 (0.03)	**0.65 (0.03)**
Physical function	0.76 (0.01)	**0.79 (0.02)***
Anxiety	0.57 (0.02)	**0.61 (0.04)***
Depression	0.56 (0.02)	**0.59 (0.02)***
Sleep Disturbance	0.64 (0.03)	**0.66 (0.05)**

* Significant improvement over the independent model

V. DISCUSSION

In general, the AUCs related to classifying physical health were relatively higher than mental health PROs, such as Global Mental Health, Anxiety and Depression. This result makes intuitive sense, as collected data, such as steps, total distance, and calorie expenditure, are more directly related to physical health than mental health. It might therefore be useful to develop hardware to record data more related to mental health for future studies. For instance, there has been effort to develop non-invasive and continuous blood pressure tracking [39] using wearable devices, which may improve the performance of classifying mental health [40]. Also, only Anxiety and Depression measured by PROMIS instruments were used in this study, which lacks precision as mental health is a broad and complicated field. More thorough evaluations of subjects' mental states could provide more descriptive labels for training machine learning models, which could

further improve performance in predicting mental health status.

Our highest AUC was 0.79 from classification of Physical Function, which demonstrated the correlation between data collected from Fitbit and PROs. However, the AUC values also indicated that PROs cannot be completely determined by activity tracker data alone, suggesting that PROs, particularly those pertaining to mental health such as depression, contain additional information that was not captured in the tracking devices. While the current study demonstrates the use of activity trackers to capture information about patient's health status, in some cases PROs could be a preferable method. Internet access enables PRO data collection to be done outside of clinic through web or mobile apps, which provides convenience and reduces time commitment for patients.

According to Table III, steps and total distance have significantly higher importance for classifying the majority of survey scores, while calories BMR significantly contributes to mental health scores, like Anxiety and Depression. Their importance factor may due to data quality, as previous studies [5]–[7] have validated the data accuracy for step counts, distance travelled, and energy expenditure for activity trackers, while other features have not been validated in scientific work. As Fitbits are not sold as medical devices, many of their features are not validated or regulated like other medical devices. In our study, we found inconsistency in sleep data and sleeping stages for subjects. It was likely that Fitbit was taken off for charging during nights. Therefore, future studies should notice user not always charge it during nights to collect sleep data. Moreover, the data elements that have been validated are generally only tested in specific devices, rather than across all activity trackers, so it is not clear how these validation results translate to other devices. Future studies should be conducted to validate these features.

As indicated by the correlation between subject's average PRO scores and the number of missing PRO values, patients with moderate to severe health status were less likely to complete PRO questionnaires routinely, which may have introduced bias for data collection in this study. Future studies could try to provide incentives for continued participation, which may mitigate study attrition. Eight PROMIS instruments were used in this study, and some redundancy existed between the specific short forms such as Fatigue or Anxiety to the general Global-10 short form. Our current approach treated each score independently without considering this overlap. A possible future study could predict PRO scores simultaneously in a joint model such as Bayesian network, which considers the correlations between PRO scores.

In our dataset of patients with SIHD based on adjudicated clinical data, HMMs achieved significantly higher classification accuracy than treating weeks independently because they took advantage of correlations in subjects' survey scores from week to week. In our data-driven approach, the model states were determined based on the distribution of PRO scores in the clinical study [41]. Score bins for the states were defined to limit the sparsity during training and make the number or states consistent across all of the PROMIS PROs tested. However, this may not be the optimal way to define the number of states for clinical representation of health status. Future studies could conduct some analysis to find out the optimum number of states, which may further increase the classification accuracy. In addition, another future direction would approach this as a regression problem to predict actual PRO scores with high precision over time.

Sequential deep learning models, such as recurrent neural networks (RNNs) and long-short-term-memory (LSTM) networks have also demonstrated strong performance when dealing with sequential data [42], [43]. Therefore, these techniques may hold potential for applications to sensor data to classify or predict health status. However, such methods generally require a large amount of training data, which was not available in the current study. In future studies, deep learning methods could be explored if a sufficiently large data set were collected.

While activity trackers are able to produce patient information within seconds or minutes, the sampling periods for PROs like PROMIS [13] are on the order of weeks, requiring down-sampling of the Fitbit data for comparison. Given that the PROs measured in this study are unlikely to vary significantly from day to day, this temporal resolution is appropriate for the application of PRO prediction. However, predicting more acute events might require more temporal resolution, which could be addressed by using the activity tracker data at a finer time scale. Long term follow-up with patients including recordings of clinical events such as rehospitalizations could also allow us to evaluate the effect of mHealth monitoring on clinical outcome, an important step in determining the efficacy of such an intervention.

VI. Conclusion

A temporal machine learning model can be used to classify self-reported physical health in patients with SIHD using physiological indices measured by activity trackers. By constructing an HMM with feature selection and an RF classifier, the resulting model can achieve an AUC of 0.79 for classifying Physical Function. Our result indicates data generated from activity trackers may be used in a machine learning framework to classify validated self-reported health status variables. These techniques could play a future role in larger frameworks for remotely monitoring a patient's health state in a clinically meaningful manner.

References

[1] M. K. Ong *et al.*, "Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure the better effectiveness after transition-heart failure (BEAT-HF) randomized clinical trial," *JAMA Intern. Med.*, vol. 176, no. 3, pp. 310–318, 2016.

[2] R. J. Shaw *et al.*, "Mobile health devices: Will patients actually use them?," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 3, pp. 462–466, 2016.

[3] J. J. T. Black *et al.*, "A remote monitoring and telephone nurse coaching intervention to reduce readmissions among patients with heart failure: study protocol for the Better Effectiveness After Transition-Heart Failure (BEAT-HF) randomized controlled trial,"

Trials, vol. 15, no. 1, pp. 124–135, 2014.

[4] W. Speier *et al.*, "Evaluating utility and compliance in a patient-based eHealth study using continuous-time heart rate and activity trackers," *J. Am. Med. Informatics Assoc.*, 2018.

[5] J. Meyer and A. Hein, "Live long and prosper: Potentials of low-cost consumer devices for the prevention of cardiovascular diseases," *J. Med. Internet Res.*, vol. 15, no. 8, pp. 1–9, 2013.

[6] M. Alharbi, N. Straiton, and R. Gallagher, "Harnessing the Potential of Wearable Activity Trackers for Heart Failure Self-Care," *Curr. Heart Fail. Rep.*, vol. 14, no. 1, pp. 23–29, Feb. 2017.

[7] T. Ferguson, A. V. Rowlands, T. Olds, and C. Maher, "The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study," *Int. J. Behav. Nutr. Phys. Act.*, vol. 12, no. 1, pp. 1–9, 2015.

[8] N. C. Franklin, C. J. Lavie, and R. A. Arena, "Personal health technology: A new era in cardiovascular disease prevention," *Postgrad. Med.*, vol. 127, no. 2, pp. 150–158, 2015.

[9] C. Smith-spangler, A. L. Gienger, N. Lin, R. Lewis, C. D. Stave, and I. Olkin, "CLINICIAN ' S CORNER Using Pedometers to Increase Physical Activity A Systematic Review," *Clin. Corner*, vol. 298, no. 19, p. 2296, 2014.

[10] S. L. Shuger *et al.*, "Electronic feedback in a diet- and physical activity-based lifestyle intervention for weight loss: a randomized controlled trial," *Int. J. Behav. Nutr. Phys. Act.*, vol. 8, no. 1, p. 41, May 2011.

[11] S. Zan, S. Agboola, S. A. Moore, K. A. Parks, J. C. Kvedar, and K. Jethwani, "Patient engagement with a mobile web-based telemonitoring system for heart failure self-management: a pilot study.," *JMIR mHealth uHealth*, vol. 3, no. 2, p. e33, Apr. 2015.

[12] T. M. Hale, K. Jethwani, M. S. Kandola, F. Saldana, and J. C. Kvedar, "A Remote Medication Monitoring System for Chronic Heart Failure Patients to Reduce Readmissions: A Two-Arm Randomized Pilot Study.," *J. Med. Internet Res.*, vol. 18, no. 5, p. e91, Apr. 2016.

[13] P. A. Pilkonis, L. Yu, N. E. Dodds, K. L. Johnston, C. C. Maihoefer, and S. M. Lawrence, "Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study," *J. Psychiatr. Res.*, vol. 56, no. 1, pp. 112–119, 2014.

[14] D. Cella *et al.*, "Initial Adult Health Item Banks and First Wave Testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) Network: 2005-2008," *J. Clin. Epidemiol.*, vol. 63, no. 11, pp. 1179–1194, 2011.

[15] H. Liu *et al.*, "Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel," *J. Clin. Epidemiol.*, vol. 63, no. 11, pp. 1169–1178, 2010.

[16] B. L. Egleston, S. M. Miller, and N. J. Meropol, "The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects," *Stat. Med.*, vol. 30, no. 30, pp. 3560–3572, 2011.

[17] S. R. Porter, M. E. Whitcomb, and W. H. Weitzer, "Multiple Surveys of Students and Survey Fatigue," *New Dir. Institutional Res.*, no. 121, 2004.

[18] S. Hermsen, J. Moons, P. Kerkhof, C. Wiekens, and M. De Groot, "Determinants for Sustained Use of an Activity Tracker: Observational Study," *JMIR mHealth uHealth*, vol. 5, no. 10, p. e164, 2017.

[19] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[20] M. Eileen Hsich, MD, Eiran Z. Gorodeski, MD, MPH, Eugene H. Blackstone, MD, Hemant Ishwaran, PhD, and Michael S. Lauer, E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, "Identifying Important Risk Factors for Survival in Systolic Heart Failure Patients Using Random Survival Forests," *Circ. Cardiovasc. Qual. Outcomes*, vol. 4, no. 1, pp. 39–45, 2014.

[21] N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to Combine Representations for Medical Records Search," *Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 833–836, 2013.

[22] J. H. Morra, Zhuowen Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Comparison of AdaBoost and Support Vector Machines for Detecting Alzheimer's Disease Through Automated Hippocampal Segmentation," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 30–43, 2010.

[23] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008.

[24] S. Shen *et al.*, "A Bayesian model for estimating multi-state disease progression," *Comput. Biol. Med.*, vol. 81, no. November 2016, pp. 111–120, 2017.

[25] B. Schuster-Böckler and A. Bateman, "An introduction to hidden Markov models.," *Curr. Protoc. Bioinformatics*, vol. Appendix 3, p. Appendix 3A, 2007.

[26] E. Birney Clamp, M., Durbin, R, "GeneWise and Genomewise.\r," *Genome Res.*, vol. 14, no. 4, pp. 988–995, 2004.

[27] L. A. Cadmus-bertram, B. H. Marcus, R. E. Patterson, B. A. Parker, and B. L. Morey, "Physical Activity Intervention for Women," *Am. J. Prev. Med.*, vol. 49, no. 3, pp. 414–418, 2015.

[28] J. B. Wang *et al.*, "Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial," pp. 18–23, 2014.

[29] S. Benedetto, C. Caldato, E. Bazzan, D. C. Greenwood, V. Pensabene, and P. Actis, "Assessment of the Fitbit Charge 2 for monitoring heart rate," *PLoS One*, vol. 13, no. 2, p. e0192691, 2018.

[30] M. A. Tully, C. McBride, L. Heron, and R. F. Hunter, "The validation of Fibit ZipTM physical activity monitor as a measure of free-living physical activity," *BMC Res Notes*, vol. 7, p. 952, 2014.

[31] K. M. Diaz *et al.*, "Fitbit®: An accurate and reliable device for wireless physical activity tracking," *Int. J. Cardiol.*, vol. 185, pp. 138–140, 2015.

[32] R. K. Reddy *et al.*, "Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise : Evaluation Study Corresponding Author :," vol. 6, 2018.

[33] E. Jo, K. Lewis, D. Directo, M. J. Kim, and B. A. Dolezal, "Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking," no. July, pp. 540–547, 2016.

[34] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: A guide for non-statisticians," *Int. J. Endocrinol. Metab.*, vol. 10, no. 2, pp. 486–489, 2012.

[35] R. D. Hays, J. B. Bjorner, D. A. Revicki, K. L. Spritzer, and D. Cella, "Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items," *Qual. Life Res.*, vol. 18, no. 7, pp. 873–880, 2009.

[36] B. M. R. Spiegel *et al.*, "Development of the NIH patient-reported outcomes measurement information system (PROMIS) gastrointestinal symptom scales," *Am. J. Gastroenterol.*, vol. 109, no. 11, pp. 1804–1814, 2014.

[37] J. O. Ogutu, H. P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proc.*, vol. 5, no. SUPPL. 3, pp. 3–7, 2011.

[38] R. J. Petrella, J. J. Koval, and D. A. Cunningham, "A Self-Paced Step Test to Predict Aerobic Fitness in Older Adults in the Primary Care Clinic," pp. 632–638, 2001.

[39] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuff-Less Blood Pressure Estimation Algorithms for Continuous Health-Care Monitoring.," *IEEE Trans. Biomed. Eng.*, vol. 9294, no. c, pp. 1–11, 2016.

[40] F. Fallo *et al.*, "Circadian blood pressure patterns and life stress," *Psychother. Psychosom.*, vol. 71, no. 6, pp. 350–356, 2002.

[41] A. Stolcke and S. Omohundro, "Hidden Markov Model Induction by Bayesian Model Merging," *Neural Inf. Process. Syst.*, vol. 5, no. Ml, pp. 11–18, 1993.

[42] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," pp. 1–18, 2015.

[43] A. Rajkomar *et al.*, "Scalable and accurate deep learning for electronic health records," *npj Digit. Med.*, no. January, pp. 1–10, 2018.