Feasibility Study of Monitoring Deterioration of Outpatients Using Multi-modal Data Collected by Wearables

DINGWEN LI, McKelvey School of Engineering, Washington University in St. Louis, USA
JAY VAIDYA, McKelvey School of Engineering, Washington University in St. Louis, USA
MICHAEL WANG, McKelvey School of Engineering, Washington University in St. Louis, USA
BEN BUSH, McKelvey School of Engineering, Washington University in St. Louis, USA
CHENYANG LU, McKelvey School of Engineering, Washington University in St. Louis, USA
MARIN KOLLEF, School of Medicine, Washington University in St. Louis, USA
THOMAS BAILEY, School of Medicine, Washington University in St. Louis, USA

In the paper, we explore the feasibility of monitoring outpatients using Fitbit Charge HR wristbands and the potential of machine learning models to predict clinical deterioration (readmissions and death) among outpatients discharged from the hospital. We developed and piloted a data collection system in a clinical study which involved 25 heart failure patients recently discharged. The results demonstrated the feasibility of continuously monitoring outpatients using wristbands. We observed high levels of patient compliance in wearing the wristbands regularly and satisfactory yield, latency and reliability of data collection from the wristbands to a cloud-based database. Finally, we explored a set of machine learning models to predict deterioration based on the Fitbit data. Through 5-fold cross validation, K nearest neighbor achieved the highest accuracy of 0.8667 for identifying patients at risk of deterioration using the data collected from the beginning of the monitoring. Machine learning models based on multimodal data (step, sleep and heart rate) significantly outperformed the traditional clinical approach based on LACE index. Moreover, our proposed Weighted Samples One Class SVM model with estimated confidence can reach high accuracy (0.9635) for predicting the deterioration using data collected within a sliding window, which indicates the potential for allowing timely intervention.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → *Health informatics*.

Additional Key Words and Phrases: medical data mining, ubiquitous computing, deterioration early warning, wearable tracker, heart failure

ACM Reference Format:
Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2019. Feasibility Study of Monitoring Deterioration of Outpatients Using Multi-modal Data Collected by Wearables. *ACM Trans. Comput. Healthcare* 1, 1, Article 1 (June 2019), 24 pages. https://doi.org/10.1145/3344256

Authors' addresses: Dingwen Li, McKelvey School of Engineering, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Jay Vaidya, McKelvey School of Engineering, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Michael Wang, McKelvey School of Engineering, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Ben Bush, McKelvey School of Engineering, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Chenyang Lu, McKelvey School of Engineering, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Marin Kollef, School of Medicine, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA; Thomas Bailey, School of Medicine, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO, 63130, USA.

2637-8051/2019/6-ART1 $15.00
https://doi.org/10.1145/3344256

1 INTRODUCTION

Hospital readmissions occur often and are difficult to predict. Many hospitals have dedicated resources to identify patients at risk for readmission, and to prevent such readmissions [26]. Heart failure is the most common principle hospital discharge diagnosis among Medicare beneficiaries and is among the most expensive conditions billed to Medicare [3]. Readmission rates following discharge for heart failure are high, with approximately 25% of patients being readmitted within 30 days. However, only about 35% of these patients are readmitted for heart failure [12], and efforts to prevent these readmissions, most of which have focused on the condition causing the index admission, have met with variable and incomplete success [47]. It is difficult to accurately diagnose heart failure in time [27]. In the medical field, clinicians use standardized LACE index[1] [37] (calculated based on length of hospital stay, acuity of admission, co-morbidity index and number of emergency department visit) to evaluate the risk of readmission or death after discharge. The problems of frequent readmission and hard-to-detect deterioration in patients with heart failure, along with the health care costs associated with these readmissions, suggest the need for continuously monitoring and early warning strategies that call attention to patients who are in need of intervention to prevent clinical deterioration and consequent need for emergency department visits or hospital readmission.

While wearable devices have become accessible for the general population, to date few *clinical* studies have assessed their potential role in patient care [4, 7, 39]. The increasing accessibility and improving quality of wearable devices are poised to incur an evolution in medical research by allowing clinicians to collect everyday health data in a cheap and friendly way. Studies showed Fitbit worked well in activity related monitoring and data collection [8]. Wearable devices can achieve high accuracy in steps measurement. An earlier study found that Fitbit Flex wristbands had an accuracy of 0.996 when measuring straight indoor walking [22]. Wearable devices also showed their accuracy in measuring everyday heart rate with errors ranging from 1.8% to 5.5% [34]. Due to their attractive designs, easy to use characteristics and accuracy, wearable fitness devices are promising devices for monitoring outpatients.

In this paper we explore the potential of wearables to monitor clinical deterioration among outpatients. Specifically, we aim to predict *clinical deterioration* defined as a composite outcome of either readmission or death among patients discharged from a hospital. We first developed a cloud-based database system to collect multi-modal data from outpatients using Fitbit Charge HR wristbands. The system can passively collect everyday health data including step counts, heart rates, and sleep duration and quality. We then conducted a clinical study to monitor 25 heart failure patients recently discharged from Barnes-Jewish Hospital, a major research hospital in St. Louis. We assessed the feasibility of collecting multi-modal data from outpatients using wearables by analyzing the data yield, reliability, and user compliance of our data collection system based on wristbands. Finally, we demonstrated the potential of applying machine learning to predict clinical deterioration among outpatients. Specifically, we explore two use cases of predicting clinical deterioration among outpatients: (1) *Deterioration Early Warning (DEW)* that continuously predicts upcoming deterioration using recent data collected in a sliding time window; (2) *Deterioration Risk Prediction (DRP)* that identifies patients at risks of deterioration using data collected from the beginning of monitoring. The prediction results could allow early or just-in-time intervention, such as contacting the patients to check up their health condition or delivering medical service if necessary.

[1]The LACE index is used to calculate the risk of readmission or death within 30 days. It includes four parameters: "L" represents length of stay, "A" represents the acuity of the admission, "C" represents co-morbidities, "E" represents the number of emergency department visits within the last 6 months.

While wearables have been studied in various contexts, the significance of this work lies in the exploration of wearable sensing and learning approaches in a clinical study with heart failure patients, a patient population that is particularly vulnerable to readmissions and in need of new monitoring capabilities in everyday settings. Specifically, the main contributions of our study are two-fold.

First, our experience in the clinical study demonstrated the feasibility of collecting multi-modal data from outpatients using wearables. In the study, our system collected more than 80% of per-minute step data from 92% of the patients. While the yield of heart rate data was lower, the median gap in heart rate data was only 4 minutes. The results suggest most of these patients were compliant with the study protocol and wore the wristband regularly. Furthermore, the cloud-based database collected 73% of the data within an hour, which potentially allowed timely intervention.

Second, we explored the potential to predict clinical deterioration among outpatients based on multi-modal data collected using wearables. For DEW, we propose *Weighted Samples One Class SVM (WOCSVM)* with confidence score on classification outcome, which extends One Class SVM by assigning weights to samples. The performance of WOCSVM (with accuracy of 0.9635) suggests the feasibility of predicting clinical deterioration ahead of time. The associated confidence score gives clinicians supportive details on assessing the predicting outcomes. For DRP, the evaluation of a set of predictive models show that the K nearest neighbor model achieved significantly higher accuracy (0.8667) than the traditional method based on LACE index (0.7826). Furthermore, combining multiple sensing modalities (step, heart rate and sleep) effectively improved predictive accuracy when compared to models based on a single modality, which suggests the significance of incorporating multiple modalities for training predictive models.

2 RELATED WORK

There has been significant interest in predicting clinical deterioration. Based on the sources of data used to perform the prediction, studies in this area can be categorized as prediction using clinical data collected within the hospital and that based on data collected by wearable devices.

Clinicians traditionally relied on scores derived from inpatient data [10, 13, 23, 24, 31, 37, 41] or patient mobility [16, 36]. In particular, LACE index has been widely adopted to predict readmission. The LACE index is manually calculated at the time of discharge to predict the risk of readmission of both medical and surgical patients after hospital discharge [41]. However, LACE index has variable performance for different use cases. Robinson and Hudali [31] showed that LACE index had fair discrimination in a study of 5,800 patients in Singapore [24] and poor discrimination in a study done on about 500 patients in UK with an average age of 85 years [10]. Wang [41] claimed that LACE index might not accurately predict 30-day readmission of congestive heart failure patients discharged from hospital. LACE-rt, as a real-time version of LACE index, was invented to predict readmission using the length of stay during the previous acute care admission [15]. However, LACE-rt underestimates the readmission rates by not taking early death into account [15].

As predicting hospital readmission via LACE index has variable performance, recent literature exploited learning-based approaches to train predictive models using clinical data. Logistic regression is a commonly used machine learning model to predict hospital readmission [19, 24, 35]. Low and Sushmita demonstrated that logistic regression outperformed LACE index in predicting 30-day readmission, while several other studies [24, 35, 38, 45] showed that random forest was also an accurate model in predicting hospital readmission. These learning-based approaches relied on medical records and vital signs collected while patients were *in* the hospital [20, 33, 40, 40, 48]. To date, there is scarce literature on predicting readmission using outpatient data [7].

As wearables can passively collect outpatient data in a continuous fashion, recent studies sought to employ wearables to predict readmission. Abdulmajeed et al. studied the feasibility of predicting

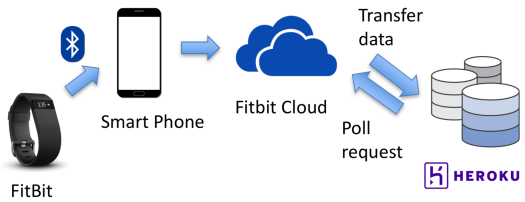

Fig. 1. System Overview

the readmission of heart failure patients via wearables [1]. Bae at al. [5] demonstrated the feasibility of predicting readmission of postsurgical cancer via sedentary behavior data collected by Fitbit. However, both of the studies utilized features from a *single* modality, i.e., the number of steps taken. As wearables incorporate more sensing modalities, it is important to explore the potential of *multimodal* data to improve model accuracy. Our system was built upon previous studies, but combined features derived from heart rate and sleep data in addition to step data. Our comparative evaluations demonstrated the advantage of exploiting multimodal data for predicting clinical deterioration. Moreover, while previous studies using wearables focused on predictive models only, we further provided in-depth analysis of patient compliance and the performance of data collection from heart failure patients after hospital discharge. Two clinical deterioration prediction scenarios and various machine learning models were investigated in the study. By demonstrating the feasibility of continuous data collection and deterioration prediction from outpatients using wearables, the findings from our clinical study have broad implications on future clinical studies that use wearables to monitor outpatients.

3 SYSTEM DESIGN

We have developed HIPAA-compliant[2] software to collect multimodal data from Fitbit wristbands, which then stores the data in a cloud-based database. The system architecture is shown in Figure 1. The wristband collects multimodal data including per-minute heart rate and step count, as well as the duration and quality of each sleep episode. The wristband synchronizes with the Fitbit App on an Internet connected device, in our case a smartphone, through Bluetooth. The smartphone then sends the collected data to the Fitbit cloud where all user data is stored indefinitely. If the wristband is unable to synchronize with a smartphone and push the data to the Fitbit cloud, it will store data locally on the device for up to 7 days. The data communication from the wristband to the Fitbit cloud is managed by software provided by Fitbit. Once data has been stored in the Fitbit cloud, our own cloud-based Heroku server retrieves the data of our participants through Fitbit's Web API and stores it in our PostgreSQL database. If the data indicates that there may be issues with patient compliance, then our server automatically generates an Excel file containing the anonymized patient's data and sends an alert through email to the nurses who are in charge of the study. The nurses review the patient's data and contact the patient if necessary. In the future, when the system is deployed for clinical applications involving a large cohort of patients, it may not be practical for a nurse to contact non-compliant patients. To scale up the system, we may automate the reminders by implementing a software service that monitors compliance and sends reminders to non-compliant patients via email or text messages.

[2]HIPAA (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information.

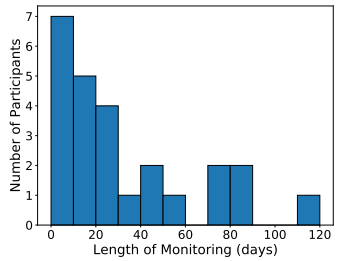

Fig. 2. Histogram of length of monitoring

The end-to-end latency of data collection from the wristband to our database is influenced by the frequency at which data flows through the system. When the Fitbit App is setup on the patient's smartphone, the setting to synchronize the Fitbit Charge HR with the smartphone every 15 minutes is enabled. Once the smartphone has received the data from the Fitbit, it immediately attempts to send this information to the Fitbit cloud over the Internet. From the Fitbit cloud, our application requests heart rate and step data as well as accessory data (battery life at last sync and time of last sync) periodically for all patients.

4 PERFORMANCE OF DATA COLLECTION

4.1 Study Protocol

We enrolled 25 participants from Barnes-Jewish Hospital in the clinical study that was IRB-approved by the Washington University Human Research Protection Office. The participants had to be heart failure patients with informed consent. In terms of device diversity, 8 participants had their own iPhone, 14 participants used Android (10 of them had their own devices and 4 of them were offered the devices by the study), and remaining 3 participants had smartphones with unknown platforms. All the participants used the same type of Fitbit Charge HR that was provided by study. During the study, participants were instructed to wear Fitbit Charge HR all the time and charge the device once per day. The monitoring started before the time of discharging from the hospital and can last for 60 days.

The longitudinal monitoring suggests, 16 participants wore the device less than 30 days, 4 participants wore the device between 30 and 60 days, and 5 participants wore the device longer than 60 days. The results are calculated by counting the number of days when the Cloud receives data, as reported in Figure 2. In order to assess the feasibility of using wearable to monitor the health conditions of outpatients, it is important to evaluate the patients' compliance, the data yield, as well as the reliability and latency of the data collection process.

4.2 Reliability

4.2.1 Yield. The *yield* is defined as the fraction of the expected samples that are successfully collected and stored in our database. The yield is calculated separately for heart rate, step and sleep data. The heart rate and step count is one per minute. For example, for a duration of 1 hour we are expecting 60 samples of heart rate and step count, respectively. Sleep data is collected once per day. Thus the sleep yield is defined as the fraction of days in which we collected sleep data.

Figure 3 shows the yield of heart rate, step and sleep data, respectively, for each participant. The average yield of step is much higher than that of heart rate. The median yield of step count is 0.999, and 92% of participants have step yield higher than 0.8, as shown in Figure 3. In comparison, the median yield of heart rate is 0.889, and 60% participants have heart rate yield higher than 0.8. There can be three potential causes for low yield: (1) user compliance (not wearing the device)

would lead to low yield for both heart rate and step; (2) wearing the device improperly can result in large discrepancy between step and heart rate yield. For instance, if the user wears Fitbit loosely on the wrist, we get unstable heart rate. However, we may still obtain stable steps, since the step is recorded as long as the device is on; (3) the last potential cause is the loss of connectivity. The Fitbit cloud may permanently lose some data due to the lack of Bluetooth connectivity to a mobile phone for a sustained period of time. If the end-to-end latency is longer than the period when Fitbit stores the unsynchronized data locally, we will permanently lose these data since they will be overwritten by the later coming data. In Section 4.3, we will show that this is not the cause of low yield for our system.

We observe that the Fitbit's step measurement usually has a higher yield than heart rate measurement. For example, Figure 3 shows that participant 6, 13, 16, 18 have much higher step yield than their heart rate yield. On the other hand, only participant 10 has a heart rate yield which is significantly larger than its step yield. While we could not directly measure participants' compliance in wearing the devices continuously during the study, we can estimate their compliance by cross-examining the yield of heart rate and step measurements. Intuitively, as long as a device managed to collect step or heart rate data, it is reasonable to assume that the user was wearing the device.

Therefore we can treat the maximum among heart rate yield and step yield as a proxy of the user compliance. Likewise, the maximum yield that is over 0.8 can be a proxy of high user compliance. As shown in Figure 3, 92% of participants are highly compliant with the study protocol. However, for those participants, such as 14 and 16, which have both low heart rate yield and step yield, the results imply the low user compliance.

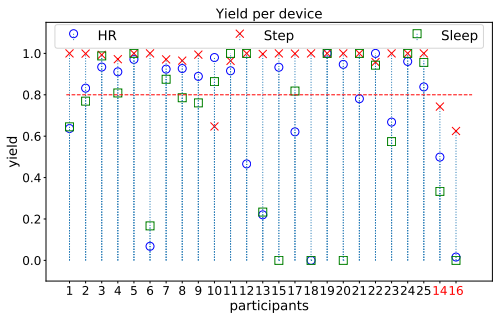

Fig. 3. The yield of each participant. The two participants to the right are those with compliant issue.

In addition, Figure 3, indicates that the sleep measurement is not as reliable as step and heart rate. Fitbit Charge HR detects user's sleep behavior via a combination of movement and heart rate patterns. An earlier study [11] suggested that sleep stages cannot be returned in three cases: the heart rate cannot be clearly detected throughout the night, sleep duration is less than three hours, or battery runs out of power during the sleeping period.

The yield of sensing modality varies a lot among participants, which raises the question of whether the compliance is correlated with the deterioration outcome. We use Analysis of Variance (ANOVA) to test the hypothesis as to whether the mean yield of the non-deteriorated group is significantly different from that of the deteriorated group. As shown in Table 1, the F statistics and p values are calculated for the three sensing modalities. F statistics reflect the ratio of between-group variability and within-group variability, while p values indicate whether the means of the groups

are significantly different. We choose η^2 as the effect size to reflect the proportion of total variance attributes to the variance of group means. The results of the ANOVA test indicate that there is no significant difference of yield between the non-deteriorated and deteriorated groups. For all the sensing modalities, all F statistics are low with extremely high p values ($p \gg 0.05$), which suggests the user compliance does not correlate with the deterioration outcome.

Table 1. ANOVA test of yields between the non-deteriorated and deteriorated group with between-group degree of freedom $df_1 = 1$ and within-group degree of freedom $df_2 = 23$.

Sensing modality	F	p	η^2
Heart rate	0.0979	0.7571	0.0651
Step	0.1021	0.7523	0.0018
Sleep	0.0414	0.8406	0.0044

4.2.2 Time to Failure and Time to Recovery. In this section, we will use two metrics to access system's reliability. *Time-to-failure* is defined as the time interval during which a component operates continuously without a failure and *time-to-recovery* is defined as the time interval from the occurrence of a failure until the component recovers [9]. Time-to-failure measures how frequently our system fails. On the other hand, time-to-recovery measures how quickly our system is able to recover from failure. In our system, a failure is the case when an expected data sample is missing. We analyze the system reliability in terms of heart rate and step data, since heart rate and step are continuously collected by Fitbit every minute. Step has an average of 0.019 failures per participant per day, which is significantly fewer than that of heart rate, 2.975 failures per participant per day. Thus, step sensing is less likely to fail than heart rate sensing. The median time-to-failure of heart rate sensing is 55 minutes, and the median time-to-failure of step sensing is 119 hours (about 5 days). The above observation explain the case where the heart rate yields of most participants are lower than step yields. However, when looking at time-to-recovery, heart rate sensing recovers from failure more quickly than step sensing. The median time-to-recovery of heart rate sensing is 4 minutes, and the median time-to-recovery of step sensing is 198 minutes (3.3 hours). Heart rate sensing usually takes less time to recovery, which can be observed when comparing Figure 4 and Figure 5. Usually, we care about the tail of time-to-recovery, since it can tell us whether our system can handle the extreme case of long duration failure. Figure 4 and Figure 5 plot the cumulative distribution function (CDF) of time-to-recovery for all devices. The 95% percentile time-to-recovery of heart rate is 8.4 hours and 137 hours (5.7 days) for step. Although the extremely large time-to-recovery occurs rarely for steps, it could still be disconcerting for the certain clinical application, since the clinicians may miss the appropriate time to provide just-in-time intervention if the system failure last for a long period.

4.3 Latency

We now assess the feasibility of using wearables for time-sensitive interventions by measuring the end-to-end latency of data collection in our system. In our analysis, the end-to-end system latency is defined as the time duration from when the data is collected by the Fitbit to when the data is stored in our database. Since the Fitbit synchronizes all types of sensing data simultaneously, we analyze latency based on step data because of its highest yield. We plot the CDF of the end-to-end latency, which can be seen in Figure 6. The median end-to-end system latency is 8.55 minutes and 99th percentile is 22.5 hours. As Fitbit can locally store data up to 7 days, the system latency is not an issue of causing data loss (low yield). Furthermore, 73% of data transmission finishes within an hour. While an hourly latency is not necessary in this particular study, the latency results suggest

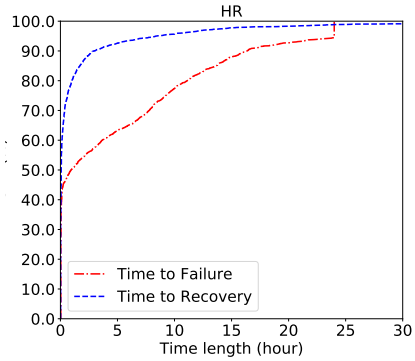

Fig. 4. CDF of time-to-failure and time-to-recovery of heart rate sensing

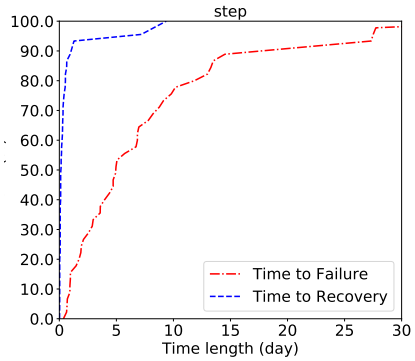

Fig. 5. CDF of time-to-failure and time-to-recovery of step sensing

the feasibility of using wearables for other applications that require timely alerts and intervention. Our data collection system is feasible of applying to hourly real-time monitoring and just-in-time intervention. The latency can be further reduced if the system is implemented natively in the Fitbit Cloud.

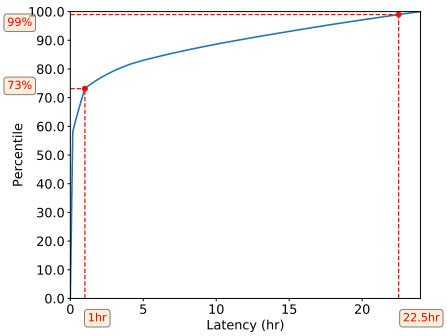

Fig. 6. CDF of system latency

5 PREDICTING CLINICAL DETERIORATION

In this section, we explore the feasibility of predicting clinical deterioration among outpatients using data collected with wearables. As alluded to in the introduction, we are interested in predicting clinical deterioration in two use cases: (1) Deterioration Early Warning (DEW) that predicts upcoming deterioration using data collected in a recent sliding time window, and (2) Deterioration Risk Prediction (DRP) that identifies participants at high risks of deterioration using data collected from the beginning of monitoring. In the rest of this section, we first describe the data processing and machine learning approaches, followed by an evaluation of predictive algorithms for DEW and DRP, respectively.

5.1 Data Preprocessing and Feature Extraction

The Fitbit API has its own data preprocessing protocols, thus we are only allowed to obtain the processed data provided by Fitbit Cloud. Time series data from Fitbit is preprocessed as summary

given for a certain granularity. In the study, the time granularity of heart rate, step count and sleep time series status is chosen as one minute. Sleep summary data is generated by Fitbit API for each sleep duration. Since patients are supposed to wear Fitbit all the time during the study period, there is inevitably some noisy data unrelated to the features we want to extract. Step count is continuously measured by Fitbit through the entire day, even when patients go to bed. As our purpose of measuring step count is to reflect patient's daily activity level when they awake, a time filter is applied to extract step count for the time period when patient is "awake". Sleep time is defined as the time when total step count is under 10 within 30 minutes after 7pm. Awake time is defined as the time when the first step is taken after 7am [5]. Thus, we only consider steps made within the duration between Awake time and Sleep time. To handle missing values, we apply carry forward imputation to the case where the interval of consecutively missing is less than a day. If the interval of consecutively missing is larger than a day, we discard that day's data for DEW problem, however we use the mean values of the entire historical data to fill the missing values on that day for DRP problem.

5.1.1 Statistical Features. The features used in the model training can be categorized as statistical features generated from time series data and semantic features derived from Fitbit API. Fitbit Charge HR gives time series data indicating the status of heart rate and step count and sleep at the minute granularity. As suggested by clinicians, we apply sliding window-based time series extraction techniques to obtain first- and second- order features as well as performing Detrended Fluctuation Analysis on the data. The purpose of generating statistical features is to identify patterns in time series data, which are informative for model training. The first-order statistical features used in our project are mean, maximum, minimum, skewness and kurtosis. The commonly used second-order time series features in medical data mining are co-occurrence features [25, 40]. Those features are shown to outperform other second-order features in the case of one-dimensional time series data [25]. Therefore, we generate the second-order features, such as energy, entropy, correlation, inertia and local homogeneity.

5.1.2 Detrended Fluctuation Analysis. In stochastic processes, chaos theory and time series analysis, detrended fluctuation analysis (DFA) is a method for determining the statistical self-affinity of a signal [21, 29]. Self-affinity is an important factor when analyzing time series which is supposed to have a regular patten. In our case, we apply DFA to heart rate and sleep time series, which evaluates long-range correlation of noisy time series data [21]. DFA can be used for analyzing non-stationary time series with slowly varying trend, such as heart rate [25].

DFA is the average fitting error over time series segments of different scale [25]. It first convert time series $\{x_i\}, 1 \leq i \leq N$ into an unbounded process $\{X_j\}, 1 \leq j \leq N$ by summation or integration:

$$X(j) = \sum_{i=1}^{j} [x(i) - \langle x \rangle], 1 \leq j \leq N \tag{1}$$

where $\langle x \rangle$ is the mean over the entire time series $\{x_i\}, 1 \leq i \leq N$. Then $\{X_j\}, 1 \leq j \leq N$ is divided into sub-series each of length n. A polynomial piecewise fit $\{Y_j\}, 1 \leq j \leq N$ is generated by minimizing local least square error within the sub-series. The fluctuation (F) of detrended time series is:

$$F(n) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (X(j) - Y(j))^2} \tag{2}$$

Finally, fluctuation measure is repeated over different length n.

We apply DFA to sleep status, an indication of sleep quality where high value means more fluctuation in sleep. Different time windows are generated to compensate the influence of unknown appropriate window size. We perform DFA on heart rate data, which will capture the heartbeat fluctuation along the day. Different time windows are also applied for extracting DFA of heart rate.

5.1.3 Semantic Features. Besides the original step count data which is directly given by Fitbit API, we derive new semantic features based on them. Sedentary bout is defined as a time duration where no steps are taken when patient is awake, as illustrated by Figure 7. Daily sedentary bout count refers to the total number of sedentary bouts per day. The Fitbit API also provides semantic features from sleep data. The useful features are Time in Bed, Minute to Fall Asleep, Minute Awake, Minute After Wakeup, Awake Count, Restless Count and Restless Duration.

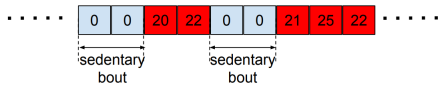

Fig. 7. Sedentary bout is the duration when there is 0 step count. Each block represents 1 minute granularity.

Finally we compute min, max, mean of those statistical and semantic features that have multiple measurements within a day. There are 51 features in total, representing the characteristics of activity, sleep and heart rate.

5.2 Weighted Samples One Class SVM with Estimated Confidence

In this section, we introduce our proposed approach for DEW. In most cases, the dataset of deterioration predicting is highly imbalanced, since clinical events (e.g., readmissions and death) occur rarely. In order to address the problem of extremely insufficient minority training examples, we proposed *Weighted Samples One Class SVM (WOCSVM).* A desirable feature of this approach is that it takes majority examples as input during training phase. Furthermore, our approach can provide a confidence score to support the clinical decision, which is important to clinical decision support systems.

One Class SVM (OC-SVM) is a semi-supervised machine learning method, which is designed for anomaly detection. OC-SVM is proposed assuming the anomaly examples are not available during the training phase. Therefore, it is trained with only normal examples. However, SVM classifiers are sensitive to outliers in the training examples [43] and need outlier suppression techniques in the training phase. Hence, we introduce the sample weights which are learned during the training phase to assign each training example a weight η_i to control its influence on the decision boundary. Therefore, the revised optimization objective of WOCSVM is:

$$\begin{aligned} \min_{\omega,\rho,\eta} \frac{\|\omega\|^2}{2} - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \eta_i \max(0, \rho - \omega^T \phi(x_i)) \\ \text{s.t. } e^T \eta \geq \beta n \end{aligned} \tag{3}$$

where ω is the normal vector to the decision boundary, ρ is the bias term, ν is the hyperparameter to control the upper bound of the fraction of training errors and the lower bound of the fraction of support vectors, β controls the maximum number of points allowed to be outliers. The hingle loss $\eta_i \max(0, \rho - \omega^T \phi(x_i))$ controls the influence of outlier on the optimization objective via forcing the corresponding sample weight η to be 0. The above optimization is not jointly convex in terms of

ω, ρ and η. The non-convex part $\frac{1}{\nu n}\sum_{i=1}^{n}\eta_i \max(0, \rho - \omega^T\phi(x_i))$ can be reformulated using concave duality [2, 44]. Then the above objective (3) is equivalent to:

$$\min_{\omega,\rho} R_{vex} + R_{cave}$$
$$R_{vex} = \frac{\|\omega\|^2}{2} - \rho, R_{cave} = \min_{\eta} \sum_{i=1}^{n} \eta_i \max(0, \rho - \omega^T\phi(x_i)). \tag{4}$$

Obviously, R_{vex} is a convex objective, and R_{cave} is a non-convex objective. We reformulated R_{cave} as:

$$R_{cave} = g(h(\omega))$$
$$\text{where } h(\omega) = \max(0, \rho - \omega^T\phi(x)), g(u) = \inf_{\eta\in\{0,1\}}[\eta^T u] \tag{5}$$

Since $h(\omega)$ is the point-wise infimum of a set of linear functions, $g(h(\omega))$ is concave on the domain $h(\omega) \in \Omega$ [46]. We can approximate R_{cave} by the multi-stage relaxation proposed by [44]. The basic idea of multi-stage relaxation is to first set η to be a vector of ones, indicating all the training examples have the same weights. Then the procedure iteratively minimizes the objective by alternatively fixing ω, ρ and η until convergence:

(1) Fix $\eta = \hat{\eta}$ and calculate $\hat{\omega}$ via solving

$$\hat{\omega} = \text{argmin}_{\omega} \frac{\|\omega\|^2}{2} - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\hat{\eta_i}\max(0, \rho - \omega^T\phi(x_i)) \tag{6}$$

(2) Fix $\omega = \hat{\omega}$ and calculate $\hat{h}(\hat{\omega}) = \max(0, \rho - \hat{\omega}^T\phi(x_i))$. In order to minimize the overall objective, η will be 1 for βn number of smallest h_i in $\hat{h}(\hat{\omega})$ based on the constrain $e^T\eta \geq \beta n$.

The WOCSVM begins with the standard OC-SVM by initializing the sample weights to be all ones. The multi-stage optimization framework could obtain a better solution than the standard OC-SVM after convergence [46]. In the experiment section, we will compare the WOCSVM with the standard one as well as other anomaly detection models to demonstrate its superior performance for solving the DEW problem.

Besides training the classification model for predicting deterioration outcomes, we create another probabilistic model that outputs the confidence of predicting each deterioration outcome. The confidence is estimated by a sigmoid function which takes the decision function of WOCSVM as the input variable:

$$p(X_i) = \frac{1}{1 + \exp(Af(X_i) + B)} \tag{7}$$

where $f(X_i)$ is the decision function evaluated on the data point X_i, A and B are the parameters of sigmoid function that needs to be tuned. The optimal parameters can be found by applying maximum likelihood estimation on the same dataset that is used for training the WOCSVM model [30]. In our case, the maximum likelihood estimation is equivalent to minimizing the cross entropy loss based on probability p:

$$\min -\sum_{i} y_i \log(p(X_i)) + (1 - y_i)\log(1 - p(X_i)) \tag{8}$$

where $p(X_i)$ is defined in (7), y_i is the corresponding label. After the optimal parameters A and B are chosen, we can calculate the estimated confidence associated with classification outcome. The estimated confidence score could be a supportive information for clinician to make decisions on what intervention to apply if deterioration is predicted.

5.3 Prediction Performance

We evaluate the performance of proposed models on the real data collected from 25 heart failure patients. The age of the participants ranges between 66 and 88 (mean: 71.6, median: 70, standard deviation: 5.3). Heart failure has been identified as a staggering clinical and public health problem [32], which is one of the most common reasons of hospitalization for the population aged 65 or over [17, 18]. This age group accounts for 71% of the total population with heart failure [18]. The cohort of our study is representative of the age group at high risk of heart failure. In our study, sixteen of the total participants are male and the remaining 9 are female, which is consistent with the observation from the general population that men have a higher rate of heart failure than women [28]. The participants report their ethnicity as 19 White and 6 African American. In comparison, the African Americans are more likely to have heart failure than other ethnic groups [28]. Future studies with larger cohorts are needed to confirm and generalize the findings in this preliminary study. The LACE index ranging from 3 to 15 (mean: 8.3, median: 8, standard deviation: 3.8). The statistics are summarized in Table 2. Five out of 25 patients became deteriorated (readmitted or deceased) within 30 days after being discharged from hospital. Two patients became deteriorated during the period between 30 days and 60 days.

Table 2. Patient Characteristics

Characteristic	n (range)
Age (year)	(66-88)
Mean	71.6
Median	70
Std	5.3
Ethnicity	
White	19
African American	6
Gender	
Male	16
Female	9
LACE index	(3-15)
Mean	8.3
Median	8
Std	3.8

Based on the clinical records, we perform survival analysis on deterioration to quantify the survival probability as time passes. Figure 8 demonstrates the survival probability generated via Kaplan-Meier estimator, which estimates the survival probability by taking into account both the deterioration events and the censored cases. The estimated survival probability indicates the survival probability drops significantly from day 2 to day 19 and from day 36 to day 49. The estimated survival probability after 60 days is 0.61. As observed from the survival curve, the patients are more likely to deteriorate at either the beginning of monitoring or the end of monitoring.

In this section, we will focus on the DEW and DRP problems which are commonly studied in the clinical research. The goal of DEW is to train a model to predict the deterioration event happening in the future based on the features collected in a recent sliding time window. The goal of DRP is to predict whether a patient will eventually deteriorate based on the features collected from the beginning of monitoring, namely identifying the patients with the risk of deterioration.

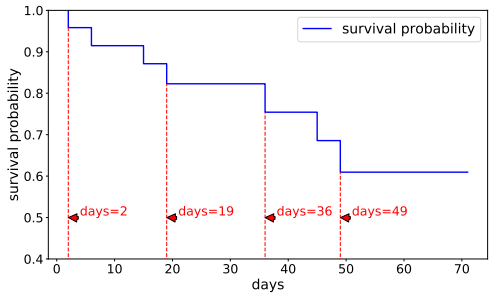

Fig. 8. Survival probability varies with the time passing. The event in our study case is the deterioration.

5.3.1 Deterioration Early Warning. There is a total of 438 days with valid data collected from the Fitbit devices, including 427 normal days and 11 days when deterioration happened. The deterioration day is defined as the first day of hospitalization or the day of death. We exclude the subsequent days in hospital that belong to the same hospitalization after we choose the first day as deterioration day. Also, we exclude the first day of hospitalization when there is a deterioration day selected within last 7 days. The 11 identified deterioration days are all from the 7 participants who were either readmitted or died during the monitoring period. The dataset in DEW consists of 438 samples; each one of them corresponds to one day. We divide the samples into a training set which contains only normal days and a testing set which contains both normal days and the days when deterioration happen. The testing set is designed in such a way to avoid favoring classifier which achieves high accuracy by only predicting the anomaly case for new unseen data.

We first explore the predictive performance of different combinations of time window size and the days ahead. Since there is a lack of empirical guidelines for how to choose the time window size and how the predictor performs for different numbers of days ahead, it is important to find the relation between time window size and the number of days ahead for prediction. We randomly split the normal days as 95% of them belonging to the training set and remaining 5% of them belonging to the testing set. The deterioration days are all counted into the testing set, since OC-SVM only use normal data for training. For each time of evaluation, we repeat the training and testing split for 100 times and average the results. In the experiment, we train a linear kernel WOCSVM with time window varying from 1 to 7 days for predicting deterioration ahead by 1 to 7 days. The hyperparameter β in WOCSVM is chosen as 0.1 via grid search and will be used throughout the experiments. The mean results are shown in Figure 9 with x-axis indicating the number of days ahead. For each number of days ahead, we compare the performance metrics by varying time window size, which are indicated by color bars. The accuracy and specificity are high when just using time window of 1 or 2 days. The combination of predicting 1 day ahead using 2 days data achieves the highest accuracy of 0.9635 and also the highest sensitivity of 1.0 with specificity of 0.9480 and precision of 0.8975. The results indicate the performance declines when predicting more days ahead, which is reasonable since the long-term future may not be related to the current status. The WOCSVM performs better when the time window size is small. The model is able to achieve over 0.9 overall accuracy when just using 1-3 day data. The results give us a practical guideline of using less than 3 days data to predict the deterioration in the near future. Figure 10 demonstrates the influence of ν on model's performance. ν is a hyperparameter of the soft-margin SVM, which controls the upper bound of training error and lower bound of support vectors. The curve serves as a guideline for how to tune the classifier to meet specific performance requirement. For instance, in order to avoid missing any deterioration, we choose $\nu = 0.09$ to achieve the highest sensitivity as well as high specificity.

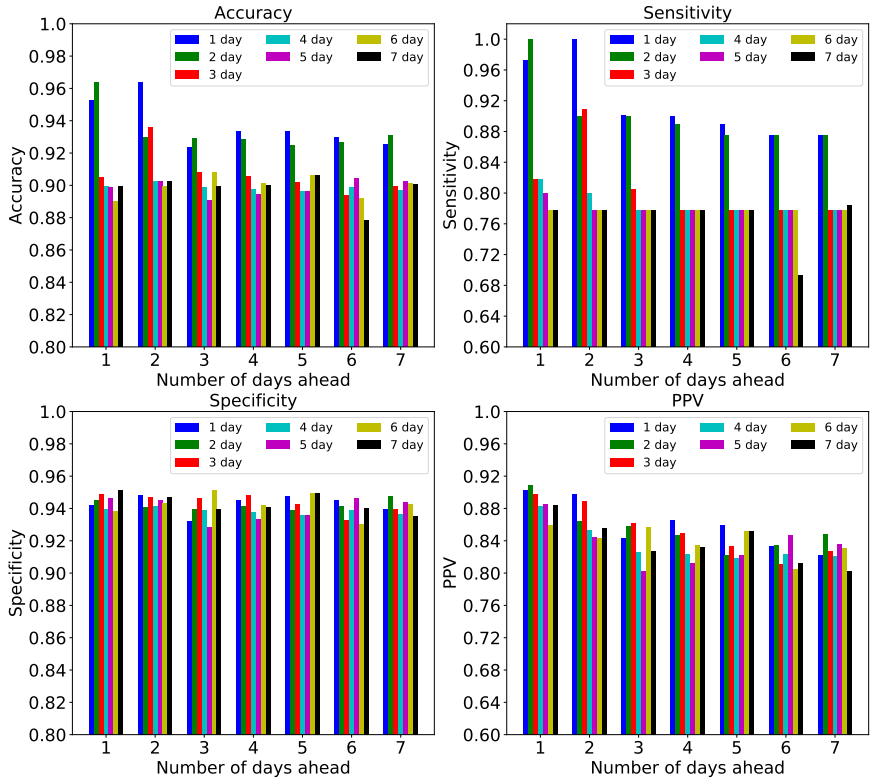

Fig. 9. Performance evaluation of WOCSVM with a varying size of time window and number of days ahead. The classifier is able to achieve good performance for predicting deterioration happening in the near future.

Then, we compare WOCSVM with other methods commonly used for anomaly detection, including the standard OCSVM, density-based methods such as local outlier factor (LOF) and clustering-based method such as K-means clustering. To reflect real-world clinical practice, we conduct leave-one-subject-out cross validation (LOSOCV) and leave-one-later-day-out cross validation (LOLDOCV) on the models, respectively. LOSOCV evaluates how a model performs on new patients and LOLDOCV evaluates the accuracy of models trained on earlier days to predict the deterioration events on later days. In these evaluations, the time window is set to 2 days, and we predict whether the deterioration event will happen after 1 day. In our study, the primary criteria for selecting a good model is to favor high accuracy, sensitivity and precision (PPV) while having high specificity (low false positive rate). We evaluate the model performance by comparing the accuracy, sensitivity and precision, while fixing the specificity around 0.95. We choose $\nu = 0.05$ for both OCSVM and WOCSVM in LOSOCV, and $\nu = 0.1$ in LOLDOCV. For those models that cannot achieve such high specificity, we choose the hyper-parameters that lead to the highest specificity. Table 3 summarizes the results for different models in the two cross-validation experiments. The LOF has low sensitivity and precision compared to K-means and OCSVM-based approaches, which implies that LOF is not suitable for our specific problem. The two OCSVM approaches outperform K-means in terms of specificity, precision and accuracy. In both cross-validation experiments, the proposed WOCSVM outperforms the standard OCSVM in sensitivity and precision while preserving the same high specificity. Aside from the performance, we also care about the features that have high influence on

the classification results. Those features can be treated as risk factors for later related clinical study. Figure 11 shows the top 10 feature importance, which is the absolute value of linear kernel SVM's coefficient. Sleep statistical features and heart rate statistical features have significant contribution to the classification results, especially these features that are generated from detrended fluctuation analysis (DFA). The observation implies that DFA is a valid feature extraction method for our study case and DEW task.

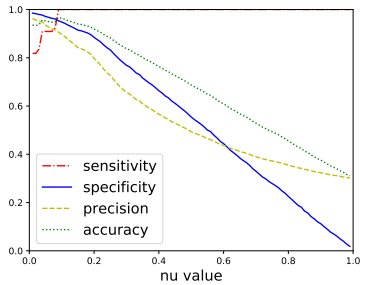

Fig. 10. Sensitivity, specificity, precision and accuracy vary along with ν. One example choice is $\nu = 0.09$ in order to make a trade-off between sensitivity and specificity.

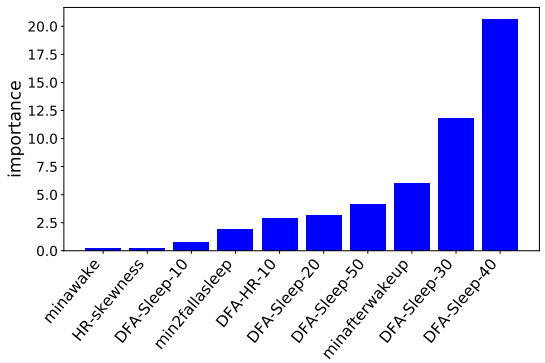

Fig. 11. Top 10 feature importance of WOCSVM. DFA shows its dominance on WOCSVM classification results.

Table 3. Performance evaluation of different anomaly detection methods with specificity fixed or close to 0.95

	Leave-one-subject-out				Leave-one-later-day-out			
Model	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision	Accuracy
K-means	**0.9091**	0.7500	0.6667	0.8065	0.9091	0.8500	0.5263	0.8592
LOF	0.0909	0.8500	0.1429	0.6863	0.7273	0.7833	0.3810	0.7746
OCSVM	0.8182	**0.9500**	0.8182	0.9216	0.9091	**0.9500**	0.7692	0.9437
WOCSVM	**0.9091**	**0.9500**	**0.9091**	**0.9355**	**1.000**	**0.9500**	**0.7857**	**0.9577**

These results demonstrate the potential of predicting clinical deterioration *early* to allow just-in-time intervention. We fit the sigmoid function to estimate the confidence score supporting

classification results, as shown in Figure 12. To obtain the estimated confidence for those points located on the other side of the separating hyperplane, we also incorporate positive examples. The fitted sigmoid curve precisely captures the trend of points that minimize the cross entropy loss of the confidence score. The curve is symmetric about the point that has 0.5 confidence score with zero distance to the separating hyperplane. The steepness of the fitted curve as well as the sparseness of the points near $x = 0$ indicates the trained model is highly confident about most of the classification results, especially for those which are far away from the separating hyperplane.

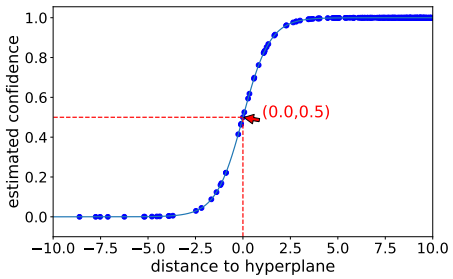

Fig. 12. The plot of sigmoid function for estimating confidence score. The fitted sigmoid is symmetric about (0.0,0.5).

5.3.2 Deterioration Risk Prediction. We aim to predict whether a certain participant will deteriorate within 60 days based on the features extracted from the data. Therefore, the dataset contains 25 samples with the series of Fitbit health data beginning from the time prior discharge. We choose 60-day prediction to contain all deteriorated samples since we have limited number of participants in this feasibility study. We note that the 60-day duration includes all deterioration within 30 days, which indicates that the methodology evaluated on 60-day duration can be generalized as predicting the deterioration risk within 30 day. First of all, we perform ANOVA test on all the features between non-deteriorated and deteriorated groups. The ANOVA test indicates whether there is a significant difference for the mean of features. It gives us an insight on whether the features of the two groups are separable, which is a preliminary evidence for the feasibility of learning a predictive model. As reported in Table 4, features such as the means of heart rate skewness ($F = 7.9125$, $p = 0.0099$), heart rate correlation ($F = 5.4789$, $p = 0.0283$), heart rate DFA 10 ($F = 5.3353$, $p = 0.0302$), restless duration ($F = 5.2912$, $p = 0.0308$) and time in bed ($F = 5.2663$, $p = 0.0312$) are significantly different between the two groups. ANOVA test gives us evidence that it is feasible to use predictive model for classifying patients with potential deterioration.

The State-of-Art models from previous work [20, 31, 42, 48], which demonstrated their superior performance in predicting readmission and clinical deterioration, are chosen as benchmarks. The models are logistic regression, SVM, random forest and neural network. In our case of limited sample size, we propose to use K nearest neighbor, which has good generalization ability on small dataset. The performance is evaluated by repeated 5-fold cross validation to tackle the model instability issue induced by the small dataset [6]. We repeatedly perform cross validation for 100 times and the results are averaged among the 100 repeated cross validation. In each iteration, we apply feature selection to the training set to remove the noisy features and improve the generalization accuracy for the testing set. The type of feature selection used in our experiment is the sequential forward feature selection with the accuracy criteria, which selects the best subset of features. We note that all models select the features derived from multiple modalities (step, HR and/or sleep), which suggests the potential benefits of exploiting multi-modal data for predicting clinical deterioration.

Table 4. ANOVA test of features between the non-deteriorated and deteriorated group with between-group degree of freedom $df_1 = 1$ and within-group degree of freedom $df_2 = 23$

Feature	F	p	η^2	Feature	F	p	η^2
HR corr	5.4789	0.0283	6.523e-6	restless count	4.2324	0.0512	2.347e-6
HR inertia	2.304	0.1427	0.1373	min awake	2.2073	0.1509	3.614e-5
HR skewness	7.9125	0.0099	5.149e-8	time in bed	5.2663	0.0312	0.4839
HR kurtosis	3.0178	0.0957	2.69e-6	awake count	4.0429	0.0562	0.0002
HR LH	4.0282	0.0566	2.87e-6	sleep skewness	0.9634	0.3366	3.152e-5
HR energy	2.2251	0.1494	5.035e-8	sleep kurtosis	0.2941	0.5928	1.547e-5
HR DFA 10	5.3353	0.0302	0.4257	sleep DFA 10	3.7881	0.0639	4.709e-3
HR DFA 20	4.8797	0.0374	1.057e-6	sleep DFA 20	0.822	0.374	0.0316
HR DFA 30	4.9808	0.0357	0.0417	sleep DFA 30	0.3489	0.5605	0.0034
HR DFA 40	4.9743	0.0358	4.603e-6	sleep DFA 40	0.5243	0.4763	8.115e-7
HR DFA 50	3.3127	0.0818	3.322e-7	sleep DFA 50	0.387	0.54	0.4423
HR DFA 60	3.3124	0.0818	2.234e-6	sleep DFA 60	0.2254	0.6395	0.0006
min after wakeup	4.0026	0.0574	2.095e-8	daily sed time	4.44	0.0462	0.0314
restless duration	5.2912	0.0308	6.11e-6	daily sed bout	4.8262	0.0384	0.0001
min to fall asleep	0.46	0.5044	3.353e-7	daily step	4.3625	0.048	0.3083
min asleep	4.3128	0.0492	1.903e-8	sed per bout	0.0056	0.9408	9.591e-7
efficiency	3.6833	0.0675	8.858r-9				

In particular, random forest, logistic regression and KNN each selected features from all three modalities. One finding regarding the feature selection is that KNN does not select sedentary per bout as part of model, which can be verified by the very little difference between the groups' mean ($F = 0.0056$ and $p = 0.9408$) as shown in Table 4. There are 6 features used by more than two models, which are DFA of sleep using 360-minute window, average minutes of being asleep, average daily steps, average awake counts and minimum minutes of being awake per sleep. Intuitively, we expect the selected features to follow different distributions among non-deteriorated group and deteriorated group. Therefore, we plot Figure 13, which are the distributions of the top 8 commonly selected features. As we expected, the distributions are different between the two groups. However, the distribution of DFA HR with 10-minute window is similar among the two groups.

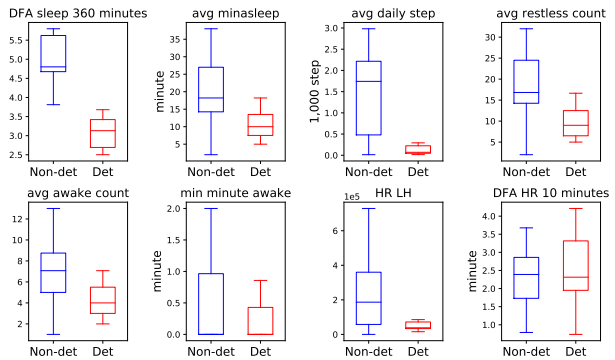

Fig. 13. Feature value distribution between non-deteriorated and deteriorated patient groups

In the following evaluations, the models are trained on the data collected in the consecutive 20 days starting from the time of discharge from hospital. We optimize the models by performing grid search to find the best hyperparameter setting. We compare the model performance in terms of

sensitivity, precision and accuracy by fixing the specificity at around 0.95, since our main goal is to correctly identify the risk patients with low false alarm rate. We also evaluate the area under the ROC curve (AUC-ROC) and the area under the precision-recall curve (AUC-PR). Since we have a skewed dataset, where the number of negative examples greatly exceeds the number of positive examples (18 vs. 7), AUC-PR is used to compare the number of false positives with that of true positives. Table 5 summarizes the evaluation results from all models with specificity fixed to around 0.95. We observe that KNN has the highest AUC-PR, sensitivity, specificity, precision and the second highest AUC-ROC. The results indicate that KNN can achieve the highest sensitivity compared with other models.

Fig. 14. Generalization performance with the inverse of number of nearest neighbors

The observation indicates that neural network has the worst performance in solving the DRP task. The reason may be that the neural network is overfitted with the small dataset. However, overfitting could also occur for the other models trained with the small dataset. A commonly used justification for overfitting is to compare training error with testing error. If the testing error is much higher than the training error, the model is very likely to be overfitted. On the other hand, if the testing error is very close to the training error, it suggests the model generalizes well. The results of performance evaluation shown in Table 5 indicate KNN is the best model for DRP. Hence, we focus on fine-tuning the parameters of KNN to make it generalize better. Figure 14 shows the influence of the number of nearest neighbor on the model's generalization ability. Moreover, as suggested by Occam's razor principle in machine learning, smaller model with the same accuracy is preferred over the complex model. Hence, we apply sequential forward feature selection to select the smallest feature subset when there are multiple subsets leading to same accuracy. The degree of overfitting is the smallest when $K = 2$, where the model can achieve the accuracy of 0.8667 for the unseen data. The KNN with 2 nearest neighbors is used throughout the rest of experiments, including the later sections where we evaluate the impacts of multiple data modalities and how early we can predict the deterioration.

To compare the predictive power of machine learning approach with the traditional approach implemented in the hospital where this clinical study was conducted, we evaluate the performance of using the LACE index to predict the clinical deterioration. LACE index is calculated by summing the assigned scores for the four parameters, where "L" stands for length of stay in the hospital, "A" stands for acuity of admission of patient in the hospital, "C" stands for co-morbidity, and "E" stands for number of emergency department visit. In our trial, the threshold of LACE index is 10 (as suggested by [15, 24, 41]), which means the patients are predicted as readmitted if their LACE index are larger than 10. The results reported in Table 5 indicate that KNN model outperforms the LACE index in terms of sensitivity, precision and overall accuracy, which demonstrates the

Table 5. Performance comparison of different models for predicting deterioration. The results shown in the table are based on fixing specificity at around 0.95. In general, KNN is a good model for predicting risk of deterioration.

Model	AUC-ROC	AUC-PR	Sensitivity	Specificity	Precision	Accuracy
RF	0.7434	0.5551	0.3077	0.9459	0.6667	0.780
SVM	0.5943	0.4016	0.1795	0.9459	0.5385	0.7467
LR	0.7829	0.6118	0.4872	0.9009	0.6333	0.7933
NN	0.4002	0.2048	0.077	0.9459	0.3333	0.720
KNN	0.7533	**0.6880**	**0.5385**	0.9820	**0.9130**	**0.8667**
LACE			0.7647	0.6250	0.5556	0.7826

Table 6. Performance metrics of models trained with different combination of features. The specificity is fixed at around 0.95. Better performance is achieved using features from multiple sensing modalities.

Model	AUC-ROC	AUC-PR	Sensitivity	Specificity	Precision	Accuracy
Step	0.6910	0.5839	0.3333	0.9550	0.7222	0.7933
HR	0.8064	0.6502	0.3590	0.9369	0.6667	0.7867
Sleep	0.7237	0.3439	0.0513	0.9099	0.1667	0.6867
Sleep, HR	0.8064	0.6502	0.3590	0.9369	0.6667	0.7867
Sleep, Step	0.6556	0.6304	0.4615	1.0	**1.0**	0.860
Step, HR	**0.8151**	**0.6991**	0.4872	0.9369	0.7308	0.820
All	0.7533	0.6880	**0.5385**	0.9820	0.9130	**0.8667**

feasibility of solving DRP of patient via the data passively collected by the Fitbit.

5.3.3 Impacts of Multiple Data Modalities. We now evaluate the contributions of different data modalities to prediction. The features used in our machine learning models are derived from three modalities collected by the Fitbit Charge HR: heart rate, step and sleep. In the following, we compare the performance of the KNN models when trained with different combinations of data modalities.

The results show that incorporating multi-modal data in machine learning models can improve the prediction accuracy compared with just using single-modal data. For instance, the best accuracy of KNN model trained with a combination of heart rate, sleep and step features is 0.8667, compared with 0.6867 when using features derived from sleep only. Table 6 summarizes the performance metrics for all models with specificity fixed at around 0.95. In general, using features from multiple sensing modalities produces better results, especially for the sensitivity, where using all features achieves 0.5385 compared with 0.0513 when only using sleep features. Moreover, we find that the model trained with sleep and step features can also yield good performance with an accuracy of 0.860 and a precision of 1.0. Overall, combining features derived from step, heart rate and sleep consistently improves the accuracy of the models.

5.3.4 How many days to use for training an accurate predictor? We investigate the number of days to be used for training an accurate model. In order to simulate the situations, we train and evaluate based on KNN models using 5-day, 10-day, 15-day and 20-day data from the beginning of the monitoring. The data belonging to any deterioration days is excluded to avoid "knowing" the deterioration in prior. As shown in Table 7, the specificity increases drastically as using more days when fixing the specificity near 0.95. In particular, 20-day data yields the highest overall accuracy 0.8677 as well as the highest specificity 0.5385. The results suggest that long-term monitoring data should be used to predict the risk of deterioration. From the results shown in Table 7, the model

trained with only 10-day data can already identify the patient at risk with an accuracy of 0.82 and precision of 1.0.

Table 7. Performance metrics for models trained with different length of day. The specificity is fixed at around 0.95.

Model	AUC-ROC	AUC-PR	Sensitivity	Specificity	Precision	Accuracy
5-day	0.4247	0.2114	0.0256	0.9009	0.0833	0.6733
10-day	0.7339	**0.6884**	0.3077	1.0	**1.0**	0.820
15-day	**0.7710**	0.5808	0.3333	0.9009	0.5417	0.7533
20-day	0.7533	0.6880	**0.5385**	0.9820	0.9130	**0.8667**

6 DISCUSSION

There has been limited research that reported a thorough approach from data collection, data analysis, to model training in a clinical study involving outpatients. In this section, we discuss the limitations of this study, summarize the implications of our results on future studies, and highlight recommendations for wearable-based clinical studies based on our experience and lessons learned.

6.1 Limitations

As a feasibility study, we have explored and assessed data collection protocols and machine learning methods to predict clinical deterioration based on multi-modal features collected from wearables. A limitation of our study is that the models are developed and evaluated based on a small dataset of congestive heart failure patients. Future studies involving larger cohorts of patients are needed to validate and generalize our methodology and models. Thus, the results in our study should be considered as preliminary. In addition, since we only validate the methods of creating predictive models for heart failure patients, the selected features and models may not generalize to other type of patients.

Another limitation of our work is that we detect the beginning of sleep based on steps within 30 minutes after 7pm. This approach may miss sleeps during daytime. A possible solution is to use activity recognition algorithms to detect all sleep periods and incorporate the additional sleep features into the predictive models.

6.2 Implications on Future Studies

As a preliminary study, this work establishes feasibility and explores methods for larger studies in the future. The findings from this study will provide guidance for future studies to predict clinical deterioration involving large cohorts of patients wearing fitness wristbands or smart watches.

- We establish the feasibility to continuously monitor outpatients using wearables with satisfactory data yield and patient compliance.
- We explore different machine learning approaches and develop predictive models suitable for deterioration early warning and deterioration risk prediction, respectively.
- We demonstrate the positive impacts of multi-modal sensor data on predictive performance, which points to the need to incorporate multi-modal sensing in wearable-based studies.
- We present data analysis and evaluation techniques to deal with challenges associated with wearable-based clinical prediction such as imbalanced dataset. For example, the compliance analysis, data preprocessing techniques and the idea of exploring optimal window size used in current study can be directly applied to future clinical studies or applications that predict clinical outcomes on wearable data.

6.3 Recommendations

Based on our experience and findings from the feasibility study, we have the following additional recommendations for researchers conducting similar studies.

- In this study, we rely on the nurses to review patients' data and contact them if the data indicates problems with compliance. If the system is deployed for clinical applications involving a large cohort of patients, it may not be practical for nurses to manually contact non-compliant patients. To scale up the study, we may automate the intervention by implementing a software service that automatically monitors compliance and sends reminders to non-compliant patients via email or text messages. In addition, since heart rates are identified as important features for predicting clinical deterioration, and heart rate yield is generally lower than step yield, we recommend implementing more granular compliance alerts for detecting low heart rate yield.
- In clinical practice, excessive false alarms generated by a clinical warning system can cause nurses to ignore important alarms [14]. It is therefore important to select the threshold of deterioration detection to achieve desired specificity based on the frequencies of false alarms that can be tolerated in specific clinical scenarios. For example, in our study, specificity is fixed around 0.95 to avoid overwhelming nurses with false alarms.
- Before developing machine learning models on a clinical dataset, it is prudent to first apply statistical analysis, e.g., ANOVA test, on the dataset to assess the feasibility of machine learning models on the dataset and outcomes of interest.
- A common challenge in clinical datasets is their imbalanced nature because clinical deterioration occurs infrequently. Learning methods therefore must be designed specifically to deal with imbalanced datasets. The Weighted Samples OC-SVM framework proposed in our study provides an example of effective methods to handle imbalanced dataset.

7 CONCLUSION

In this paper, we explore the feasibility and potential of using wearables to predict clinical deterioration among outpatients through a clinical study involving 25 congestive heart failure patients discharged from a hospital. Our primary contributions from the study are two-fold. First, our experience demonstrates the feasibility of collecting multi-modal data (step, sleep and heart rate) from outpatients using wristbands. The monitoring system achieves high data *yield*, as well as high levels of patient compliance in wearing the wristbands regularly. Second, we demonstrate the potential of using machine learning models to predict clinical deterioration among outpatients. The results show machine learning models can exploit multi-modal data to achieve high accuracy for identifying outpatients at risk of deterioration. This preliminary study establishes the feasibility and provides guidance for future studies to predict clinical deterioration involving large cohorts of patients wearing fitness wristbands or smart watches.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the Washington University Institute of Clinical and Translational Sciences grant UL1TR002345 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH. This work was also partially supported by the Fullgraf Foundation.

REFERENCES

[1] Raghad Abdulmajeed. 2016. *The Use of Continuous Monitoring of Heart Rate as a Prognosticator of Readmission in Heart Failure Patients.* Master's thesis. University of Toronto, Canada.

[2] Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. 2013. Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description (ODD '13)*. ACM, 8–15. https://doi.org/10.1145/2500853.2500857

[3] Roxanne M Andrews and Anne Elixhauser. 2007. The National Hospital Bill: Growth Trends and 2005 Update on the Most Expensive Conditions by Payer. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]* (2007). https://www.ncbi.nlm.nih.gov/books/NBK56314/

[4] Geoff Appelboom, Blake E Taylor, Eliza Bruce, Clare C Bassile, Corinna Malakidis, Annie Yang, Brett Youngerman, Randy D'Amico, Sam Bruce, Olivier Bruyère, Jean-Yves Reginster, Emmanuel PL Dumont, and E Sander Connolly Jr. 2015. Mobile Phone-Connected Wearable Motion Sensors to Assess Postoperative Mobilization. *JMIR mHealth uHealth* 3, 3 (28 Jul 2015), e78. https://doi.org/10.2196/mhealth.3785

[5] Sangwon Bae, Anind K. Dey, and Carissa A. Low. 2016. Using Passively Collected Sedentary Behavior to Predict Hospital Readmission. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, 616–621. https://doi.org/10.1145/2971648.2971750

[6] Claudia Beleites and Reiner Salzer. 2008. Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry* 390, 5 (01 Mar 2008), 1261–1271. https://doi.org/10.1007/s00216-007-1818-6

[7] Jason P Burnham, Chenyang Lu, Lauren H Yaeger, Thomas C Bailey, and Marin H Kollef. 2018. Using wearable technology to predict health outcomes: a literature review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1221–1227. https://doi.org/10.1093/jamia/ocy082

[8] Lisa Cadmus-Bertram, Bess H Marcus, Ruth E Patterson, Barbara A Parker, and Brittany L Morey. 2015. Use of the Fitbit to Measure Adherence to a Physical Activity Intervention Among Overweight or Obese, Postmenopausal Women: Self-Monitoring Trajectory During 16 Weeks. *JMIR mHealth uHealth* 3, 4 (19 Nov 2015), e96. https://doi.org/10.2196/mhealth.4229

[9] Octav Chipara, Chenyang Lu, Thomas C. Bailey, and Gruia-Catalin Roman. 2010. Reliable Clinical Monitoring Using Wireless Sensor Networks: Experiences in a Step-down Hospital Unit. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys '10)*. ACM, 14. https://doi.org/10.1145/1869983.1869999

[10] Paul E. Cotter, Vikas K. Bhalla, Stephen J. Wallis, and Richard W. S. Biram. 2012. Predicting readmissions: poor performance of the LACE index in an older UK population. *Age and Ageing* 41, 6 (05 2012), 784–789. https://doi.org/10.1093/ageing/afs073

[11] Massimiliano de Zambotti, Aimee Goldstone, Stephanie Claudatos, Ian M. Colrain, and Fiona C. Baker. 2017. A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiology International* 0, 0 (2017), 1–12. https://doi.org/10.1080/07420528.2017.1413578

[12] Kumar Dharmarajan, Angela F. Hsieh, Zhenqiu Lin, HÃĪctor Bueno, Joseph S. Ross, Leora I. Horwitz, JosÃĺ Augusto Barreto-Filho, Nancy Kim, Susannah M. Bernheim, Lisa G. Suter, Elizabeth E. Drye, and Harlan M. Krumholz. 2013. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA Internal Medicine* 309, 4 (2013), 355–363. https://doi.org/10.1001/jama.2012.216476

[13] Jacques D. Donzé, Mark V. Williams, Edmondo J. Robinson, Eyal Zimlichman, Drahomir Aujesky, Eduard E. Vasilevskis, Sunil Kripalani, Joshua P. Metlay, Tamara Wallington, Grant S. Fletcher, Andrew D. Auerbach, and Jeffrey L. Schnipper. 2016. International Validity of the "HOSPITAL" Score to Predict 30-day Potentially Avoidable Readmissions in Medical Patients. *JAMA internal medicine* 176, 4 (Apr 2016), 496–502. https://doi.org/10.1001/jamainternmed.2015.8462

[14] Barbara J. Drew, Patricia Harris, Jessica K. ZÃĺgre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloS one* 9, 10 (22 Oct 2014), 1–23. https://doi.org/10.1371/journal.pone.0110274

[15] Christo El Morr, Liane Ginsburg, Seungree Nam, and Susan Woollard. 2017. Assessing the Performance of a Modified LACE Index (LACE-rt) to Predict Unplanned Readmission After Discharge in a Community Teaching Hospital. *Interactive journal of medical research* 6, 1 (08 Mar 2017), e2–e2. https://doi.org/10.2196/ijmr.7183

[16] Steve R. Fisher, Yong-Fang Kuo, Gulshan Sharma, Mukaila A. Raji, Amit Kumar, James S. Goodwin, Glenn V. Ostir, and Kenneth J. Ottenbacher. 2013. Mobility After Hospital Discharge as a Marker for 30-Day Readmission. *The journals of gerontology. Series A, Biological sciences and medical sciences* 68, 7 (19 Jul 2013), 805–810. https://doi.org/10.1093/gerona/gls252

[17] Margaret Hall, Carol DeFrances, Sonja N Williams, Aleksandr Golosinskiy, and Alexander Schwartzman. 2010. National Hospital Discharge Survey: 2007 Summary. *National health statistics reports* 29 (10 2010), 1–20, 24. https://doi.org/10.3886/ICPSR28162.v1

[18] Margaret Hall, Shaleah Levant, and Carol DeFrances. 2012. Hospitalization for Congestive Heart Failure: United States, 2000-2010. *NCHS data brief* 108 (10 2012), 1–8.

[19] Danning He, Simon C. Mathews, Anthony N. Kalloo, and Susan Hutfless. 2014. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association* 21, 2 (2014), 272–279. https://doi.org/10.1136/amiajnl-2013-002151

[20] Arian Hosseinzadeh, Masoumeh T. Izadi, Aman Verma, Doina Precup, and David L. Buckeridge. 2013. Assessing the Predictability of Hospital Readmission Using Machine Learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13)*. AAAI Press, 1532–1538.

[21] Kun Hu, Plamen Ivanov, Zhi Chen, Pedro Carpena, and H Stanley. 2001. Effect of trends on detrended fluctuation analysis. *Physical review. E, Statistical, nonlinear, and soft matter physics* 64 (Jun 2001), 011114. Issue 1. https://doi.org/10.1103/PhysRevE.64.011114

[22] Kanitthika Kaewkannate and Soochan Kim. 2016. A comparison of wearable fitness devices. *BMC Public Health* 16, 1 (24 May 2016), 433. https://doi.org/10.1186/s12889-016-3059-0

[23] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. 2011. Risk prediction models for hospital readmission: A systematic review. *JAMA* 306, 15 (2011), 1688–1698. https://doi.org/10.1001/jama.2011.1515

[24] Lian Leng Low, Kheng Hock Lee, Marcus Eng Hock Ong, Sijia Wang, Shu Yun Tan, Julian Thumboo, and Nan Liu. 2015. Predicting 30-Day Readmissions: Performance of the LACE Index Compared with a Regression Model among General Medicine Patients in Singapore. *Biomed Res Int* 2015 (23 Nov 2015), 169870. https://doi.org/10.1155/2015/169870

[25] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2012. An Integrated Data Mining Approach to Real-time Clinical Monitoring and Deterioration Warning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, 1140–1148. https://doi.org/10.1145/2339530.2339709

[26] Eric Marks. 2013. Complexity science and the readmission dilemma: Comment on "potentially avoidable 30-day hospital readmissions in medical patients" and "association of self-reported hospital discharge handoffs with 30-day readmissions". *JAMA Internal Medicine* 173, 8 (2013), 629–631. https://doi.org/10.1001/jamainternmed.2013.4065

[27] Kenneth McDonald, Mark Ledwidge, John Cahill, Jean Kelly, Peter Quigley, Brian Maurer, Fiona Begley, Mary Ryder, Bronagh Travers, Lorna Timmons, and Teresa Burke. 2001. Elimination of early rehospitalization in a randomized, controlled trial of multidisciplinary care in a high-risk, elderly heart failure population: the potential contributions of specialist care, clinical stability and optimal angiotensin-converting enzyme inhibitor dose at discharge. *European Journal of Heart Failure* 3, 2 (2001), 209–215. https://doi.org/10.1016/S1388-9842(00)00134-3

[28] U.S. National Library of Medicine. 2019. Heart Failure. Retrieved June 3, 2019 from http://medlineplus.gov/heartfailure.html

[29] Chung-Kang Peng, Sergey Buldyrev, Shlomo Havlin, M Simons, H Stanley, and Ary Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E* 49 (Feb 1994), 1685–1689. Issue 2. https://doi.org/10.1103/PhysRevE.49.1685

[30] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 61–74.

[31] Robert Robinson and Tamer Hudali. 2017. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. *PeerJ* 5 (09 Jan 2017), e3137–e3137. https://doi.org/10.7717/peerj.3137

[32] Véronique L. Roger. 2013. Epidemiology of heart failure. *Circ Res* 113, 6 (30 Aug 2013), 646–659. https://doi.org/10.1161/CIRCRESAHA.113.300268

[33] Khader Shameer, Kipp W. Johnson, Alexandre Yahi, Riccardo Miotto, L. I. Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P. Sengupta, Sengupta Gelijns, Alan Moskovitz, Bruce Darrow, David L. David, Andrew Kasarskis, Nicholas P. Tatonetti, Sean Pinney, and Joel T. Dudley. 2016. *Predictive Modeling of Hospital Readmission Rates Using Electronic Medical Record-wide Machine Learning: a Case-study Using Mount Sinai Heart Failure Cohort*. Vol. 22. 276–287. https://doi.org/10.1142/9789813207813_0027

[34] Anna Shcherbina, C. Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W. Christle, Trevor Hastie, Matthew T. Wheeler, and Euan A. Ashley. 2017. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine* 7, 2 (2017). https://doi.org/10.3390/jpm7020003

[35] Shanu Sushmita, Garima Khulbe, Aftab Hasan, Stacey Newman, Padmashree Ravindra, Senjuti Basu Roy, Martine De Cock, and Ankur Teredesai. 2016. Predicting 30-Day Risk and Cost of "All-Cause" Hospital Readmissions. In *AAAI Workshop: Expanding the Boundaries of Health Informatics Using AI*.

[36] Tetsuya Takahashi, Megumi Kumamaru, Sue Jenkins, Masakazu Saitoh, Tomoyuki Morisawa, and Hikaru Matsuda. 2015. In-patient step count predicts re-hospitalization after cardiac surgery. *Journal of Cardiology* 66, 4 (2015), 286 – 291. https://doi.org/10.1016/j.jjcc.2015.01.006

[37] Carl van Walraven, Irfan A. Dhalla, Chaim Bell, Edward Etchells, Ian G. Stiell, Kelly Zarnke, Peter C. Austin, and Alan J. Forster. 2010. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal* 182, 6 (2010), 551–557. https://doi.org/10.1503/cmaj.091117

[38] Michael A. Vedomske, Donald E. Brown, and James H. Harrison. 2013. Random Forests on Ubiquitous Data for Heart Failure 30-Day Readmissions Prediction. In *2013 12th International Conference on Machine Learning and Applications*, Vol. 2. 415–421. https://doi.org/10.1109/ICMLA.2013.158

[39] Martijn Vooijs, Laurence L Alpay, Jiska B Snoeck-Stroband, Thijs Beerthuizen, Petra C Siemonsma, Jannie J Abbink, Jacob K Sont, and Ton A Rövekamp. 2014. Validity and Usability of Low-Cost Accelerometers for Internet-Based Self-Monitoring of Physical Activity in Patients With Chronic Obstructive Pulmonary Disease. *Interact J Med Res* 3, 4 (27 Oct 2014), e14. https://doi.org/10.2196/ijmr.3056

[40] Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer. 2017. Cost-sensitive Deep Learning for Early Readmission Prediction at A Major Hospital.

[41] Hao Wang, Richard D. Robinson, Carlos Johnson, Nestor R. Zenarosa, Rani D. Jayswal, Joshua Keithley, and Kathleen A. Delaney. 2014. Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders* 14, 1 (07 Aug 2014), 97. https://doi.org/10.1186/1471-2261-14-97

[42] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 110:1–110:24. https://doi.org/10.1145/3130976

[43] Linli Xu, Koby Crammer, and Dale Schuurmans. 2006. Robust Support Vector Machine Training via Convex Outlier Ablation. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1 (AAAI'06)*. AAAI Press, 536–542.

[44] Tong Zhang. 2009. Multi-stage Convex Relaxation for Learning with Sparse Regularization. In *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 1929–1936.

[45] Bichen Zheng, Jinghe Zhang, Sang Won Yoon, Sarah S. Lam, Mohammad Khasawneh, and Srikanth Poranki. 2015. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications* 42, 20 (2015), 7110 – 7120. https://doi.org/10.1016/j.eswa.2015.04.066

[46] Xi-chuan Zhou, Hai-bin Shen, and Jie-ping Ye. 2011. Integrating outlier filtering in large margin training. *Journal of Zhejiang University SCIENCE C* 12, 5 (04 May 2011), 362. https://doi.org/10.1631/jzus.C1000361

[47] Boback Ziaeian and Gregg C. Fonarow. 2015. The Prevention of Hospital Readmissions in Heart Failure. *Progress in Cardiovascular Diseases* 58, 4 (2015), 379–385. https://doi.org/10.1016/j.pcad.2015.09.004

[48] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, and Brian Muckian. 2013. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In *2013 IEEE International Conference on Big Data*. 64–71. https://doi.org/10.1109/BigData.2013.6691760