Wearable Device-Independent Next Day Activity and Next Night Sleep Prediction for Rehabilitation Populations

Allison Fellger, Gina Sprint, *Member, IEEE*, Doug Weeks, Elena Crooks, and Diane J. Cook, *Fellow, IEEE*

Abstract Wearable sensor-based devices are increasingly applied in free living and clinical settings to collect fine-grained, objective data about activity and sleep behavior. The manufacturers of these devices provide proprietary software that labels the sensor data at specified time intervals with activity and sleep information. If the device wearer is an unhealthy individual, such as a patient undergoing inpatient rehabilitation, these labels and their values can vary greatly from manufacturer to manufacturer. Consequently, generating outcome predictions based on data collected from patients attending inpatient rehabilitation wearing different wearable sensor devices can be challenging, which hampers usefulness of these data for patient care decisions. In this paper, we present a data-driven approach to combining datasets collected from different device manufacturers. With the ability to combine datasets, we merge data from three different device manufacturers to form a larger dataset of time series data collected from 44 patients receiving inpatient therapy services. In order to gain insights into the recovery process, we use this dataset to build models that predict a patient's next day physical activity duration and next night sleep duration. Using our data-driven approach and the combined dataset, we obtained a normalized root mean square error prediction of 9.11% for daytime physical activity and 11.18% for nighttime sleep duration. Our sleep result is comparable to the accuracy we achieved using the manufacturer's sleep labels (12.26%). Our device-independent predictions are suitable for both point-of-care and remote monitoring applications to provide information to clinicians for customizing therapy services and potentially decreasing recovery time.

***Index Terms*—Actigraphy, activity and sleep prediction, inpatient rehabilitation, machine learning, wearable sensors**

I. Introduction

When an individual experiences an injury or illness that requires inpatient rehabilitation, the individual's physical activity and sleeping patterns are often affected. Common reasons for undergoing inpatient rehabilitation include recovering from a traumatic brain injury (TBI), a stroke, cardiac disorders, lower extremity fractures, and various orthopedic surgeries. Specifically for patients with TBI, research has found that more than 66% of patients experience sleep disorders [1], while that number is as high as 78% for individuals recovering from a stroke [2]. Individuals recovering from TBI or stroke are often admitted to an inpatient rehabilitation facility to receive therapy services. Unfortunately, prescribed therapy may not be equally effective for each patient due to fatigue during the day and sleep disorders at night. Together, fatigue and sleep disorders can negatively impact the rest-activity circadian rhythm cycle that may slow the recovery from an injury or illness, even affecting quality of life [3]. Therefore, in inpatient rehabilitation, objective physical activity and sleep data can offer insights for clinicians to help customize therapy sessions with the goal of shortening the recovery process.

Objective data can be collected using wearable sensor-based devices that collect fine-grained physical activity and sleep data. Sensor-based physical activity and sleep measurements offer several benefits over human observation by therapists and by the patients themselves. Data collected from sensors remove the inaccuracy that is common amongst measurements that are self-reported by patients. For both healthy and unhealthy individuals, it is difficult to objectively self-characterize activity and sleep. We often tend to either overestimate or underestimate our activity, with correlations to direct measurement varying from -0.71 to 0.96 [4]. Secondly, data collected from sensors is not subject to variability due to inter-rater reliability. Sensor-based devices continuously track 24-hour physical activity and sleep in the same format and under the same conditions, allowing consistent data collection. Also, the technology has advanced enough to require minimal effort on the part of the clinician. This is primarily due to

Paper submitted for review on March 15, 2020. This work was supported in part by a Computing Research Association-Widening Participation (CRA-WP) CREU grant and by the National Institutes of Health (NIH) grant number R25AG0146114.

A. Fellger and G. Sprint are with Gonzaga University, Spokane, WA 99258 USA. (e-mails: afellger@zagmail.gonzaga.edu, sprint@gonzaga.edu).

D. Weeks is with St. Luke's Rehabilitation Institute, Spokane, WA, 99202 USA (e-mail: dlweeks@st-lukes.org).

E. Crooks is with Eastern Washington University, Spokane, WA 99202. (e-mail: ecrooks@ewu.edu).

D. Cook is with Washington State University, Pullman, WA 99164 USA. (e-mail: djcook@wsu.edu).

shorter device setup times and longer battery lives, permitting 24-hour recordings without the subjectivity that is frequently introduced by human observation.

Despite these advantages of sensor-based devices, their efficacy, accuracy, and applicability in inpatient rehabilitation settings remain areas of significant research. The majority of studies have focused on evaluating sensor-devices as well as their associated activity and sleep algorithms for healthy populations. It is difficult to generalize such results to unhealthy populations, such as those undergoing inpatient rehabilitation. Typically when these devices and their algorithms are applied to individuals with mobility impairments or sleep disorders, the results are highly variable [5] and can produce inexplicable results [6]. Additional challenges that arise from analyzing data collected from clinical settings include how to combine datasets from different sensor devices and how to use the combined data to help clinicians provide therapy services. There are several wearable sensor device manufacturers, and each one produces slightly different measurements of physical activity and sleep. While research has studied the validity of various manufacturers and their devices, the discrepancies across devices makes it difficult for sleep researchers and clinicians to combine datasets and interpret the results [7].

To help alleviate this challenge and advance clinical activity and sleep research, we utilized research-grade Actigraph devices and consumer-grade pedometer devices in an inpatient rehabilitation facility to collect data from patients during their recovery process. Specifically, we utilized Ambulatory Monitoring Inc (AMI) MotionLogger devices, Philips Actiwatch Spectrum Plus devices, and Fitbit Charge devices with heart rate measurement capability. Using a data-driven approach, we combined data from these devices to implement a machine learning-based approach to measure and predict a patient's future physical activity and sleep duration. Our results provide accurate predictions of activity levels for the forthcoming day and sleep duration for the forthcoming night. Our approach to manufacturer-independent physical activity and sleep prediction support point-of-care and remote patient monitoring that can help meet the needs of precision medicine by individualizing healthcare services [8].

II. Related Work

Wearable sensor-based devices, like Actigraphs and Fitbits, are wrist-worn devices that are less obtrusive and less expensive alternatives to gold-standard methods. For sleep analysis, the commonly-used gold-standard technique for wearable sensor evaluation is polysomnography [9]. For physical activity, the gold-standard techniques include direct observation and motion capture systems [10]. At a minimum, wrist-worn devices typically contain tri-axial accelerometers that measure the acceleration of the wearer's wrist over the duration of a short time interval, such as a second. Manufacturers of these devices process the acceleration time series data to determine a more clinically-relevant measure of physical activity than the original raw acceleration values, namely activity counts in the case of Actigraphs and step counts in the case of pedometers. When raw acceleration signals are combined with other sensor signals, such as heart rate or ambient light, algorithms can accurately label time intervals as "sleep" or "wake." These labels are used by researchers and clinicians to determine if the wearer is sleeping or awake.

Manufacturers and researchers have investigated the error between gold-standard measurements and the output of manufacturer processing algorithms; however, this research has primarily used healthy subjects for evaluation [11]–[14]. While these algorithms perform well on healthy populations, the algorithms can have higher error for unhealthy subjects, such as those recovering from an injury or illness like stroke or TBI, who exhibit highly irregular sleep and activity patterns [15]. To address this, recent research studying both healthy and unhealthy populations has focused on evaluating wearable sensor-based devices specifically for counting steps [10], detecting sleep periods [5], [16], and measuring sleep characteristics [17]–[20], such as total sleep time, sleep efficiency, number of awakenings, sleep onset latency, and wake after sleep onset.

Since we deployed three devices from different manufacturers for continuous data collection, we will summarize the research investigating the accuracy of these specific devices for healthy and unhealthy populations. The three devices include MotionLogger, Actiwatch, and Fitbit. Beginning with physical activity measurements, for healthy adults in free-living conditions, a wrist-worn Fitbit has been shown to not differ significantly from a waist-worn ActiGraph GT3X for counting steps taken per minute over a 24-hour period [14]. For patients in a cardiac rehabilitation population, a wrist-worn Fitbit has been shown to correlate well with step count estimates from an Actigraph (r=0.95); however, the Fitbit tended to over-count steps [21]. When compared to direct observation, a wrist-worn Fitbit was reported to underestimate step count by 16% during a self-paced walking test performed by older adults with impaired ambulation [22]. In another study of subjects with multiple sclerosis performing a 2-min walk test, the Fitbit step count correlation (r=0.69) was lower than the Actigraph correlation (r=0.76) [23]. For the MotionLogger device, no clear difference between the accelerometer counts measured by the device and indirect calorimetry were detected during level walking [24].

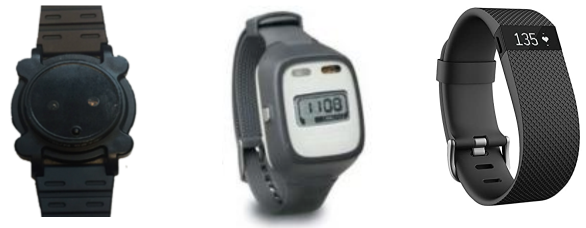
For sleep measurements, strong correlations have been reported in healthy adults between estimates of total sleep time between a Philips Actiwatch and polysomnography (r=0.94), as well from a wrist-worn Fitbit and polysomnography (r=0.97) [25]. In the same study, sleep efficiency measures from the Actiwatch and Fitbit did not differ from sleep efficiency measured by polysomnography. Therefore, total sleep time and sleep efficiency appear to be monitored by Actiwatch and Fitbit with reasonable accuracy. On the other hand, the MotionLogger device was found to underestimate total sleep time by almost 24 minutes and overestimate wake time by 25 minutes in healthy children and adolescents [26]. The same study found that the Philips Actiwatch did not demonstrate significant differences for total sleep time when controlling for age and sleep disordered breathing.

These studies have found that though wearable devices do not demonstrate perfect measurements of activity and sleep for unhealthy populations, they do produce reasonable estimates. The next research step is to utilize these activity and sleep estimates to determine if they can help customize therapy for individual patients. One way to provide additional insights for customization is for an automated machine learning system to make predictions about an individual's future physical activity and sleep performance. Machine learning models generally benefit from being trained with large datasets. To acquire as much data as possible for human activity learning, several studies have investigated device-orientation independent methods for data collection [27], [28], fusing data from multiple sensors [29], and transfer learning approaches [30], [31].

Research using machine learning models for activity and sleep applications have primarily focused on classifying types of physical activity [32], [33] and various sleep characteristics for healthy populations [11], [34], [35]. The work that is most similar to this paper is that of Sathyanarayana and colleagues [11]. Sathyanarayana and colleagues collected Actigraph data from 92 healthy adolescents wearing ActiGraph GT3X+ devices for one week. Machine learning models trained with the collected daytime physical activity data were used to classify good and poor sleep efficiency with an area under the receiver operating curve of 0.9449. In our recent work, we expanded this research to investigate the applicability of sleep prediction for individuals with sleep disorders [36]. For this study, we deployed AMI MotionLogger devices in an inpatient rehabilitation setting. We continuously collected activity and sleep data from 17 subjects with identified sleep problems due to recovering from a stroke or TBI. Using this data, we constructed machine learning regression models to predict a patient's future night sleep duration. Our regression approach achieved a 14.40% normalized root mean square error predicting next night sleep minutes.

III. Methods and Procedures

In this paper, we validate our approaches to data fusion and activity/sleep prediction for a pool of 44 subjects. Data were collected from subjects receiving inpatient therapy services for a variety of ailments, including stroke, TBI, cardiac disorders, pulmonary disorders, and lower extremity fractures. These subjects wore one of three different wearable-sensor devices: a MotionLogger, an Actiwatch, or a Fitbit. Because of these different devices, we apply a data-driven approach to support normalizing and combining the data from different manufacturers. We employ this combination of minute-by-minute activity and sleep data to make predictions about future nighttime sleep total inactive minutes (TIM) and total sleep time (TST) as measured by the device manufacturers. In addition to predicting a patient's next night sleep duration, we also predict a patient's next day total active minutes (TAM) to gain insight about daytime behavior.



a) Ambulatory Monitoring Inc MotionLogger (AMI dataset)

b) Philips Actiwatch Spectrum Plus (Philips dataset)

c) Fitbit Charge with heart rate (Fitbit dataset)

Fig 1. The three devices used for data collection.

A. Data Collection

For data collection and analysis purposes, we define a 24-hour day as a period beginning at 06:00:00 and ending at 05:59:00 the following calendar day. Using known controlled lighting times at the inpatient facility, we determined the daytime (DT) period to coincide with when the lights were typically on in patients' rooms, which was from 06:00:00 to 20:59:00. The nighttime (NT) period corresponded to the time period when lights were off in patients' rooms, which was from 21:00:00 to 05:59:00. For ease of explanation, we denote successive NT and DT periods using the >> symbol. For example, the notation NT >> DT >> NT describes a nighttime, then daytime, then the following nighttime sequence which represents 9 + 15 + 9 = 33 continuous hours. We identify a period in a sequence of successive DT and NT periods using a subscript, such as DT_1 >> NT_1 >> DT_2, where DT_1 and NT_1 are sampled from the same 24-hour period and DT_2 is from the next period.

We deployed and collected data from three different wearable sensor-based devices. These three devices represent three datasets we collectively analyze in this paper to evaluate our prediction approach. The three datasets and their devices, with sample sizes, are as follows:

1. The "AMI" dataset: Ambulatory Monitoring Inc Basic MotionLogger Actigraph devices (N=17)
2. The "Philips" dataset: Philips Actiwatch Spectrum Plus Actigraph devices (N=19)
3. The "Fitbit" dataset: Fitbit Charge with Heart Rate pedometer devices (N=8)

For reference, Figure 1 includes images of these devices. In total, we collected continuous data from 44 patients undergoing inpatient rehabilitation. Patients admitted to the hospital following an injury or illness, such as stroke or TBI, were recruited to participate in these studies if their therapist stated they were experiencing irregular sleeping patterns. The data collection protocols for all three datasets were approved by our local institutional review board and all patients provided written informed consent to participate. Each subject continuously wore one of these three devices during both the daytime and nighttime periods until they were discharged from the rehabilitation facility.

For the AMI dataset, we analyzed data from 17 subjects (age 64.11 ± 17.05 years; 11 females and 6 males), for which the data collection periods ranged from 9 days to 30 days [36]. For the Philips dataset, we analyzed data from 22 subjects (age 63.96 ± 17.93 years; 5 females and 17 males),

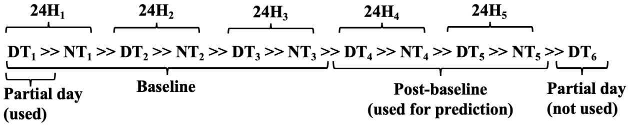

24H₁
24H₂
24H₃
24H₄
24H₅
DT₁ >> NT₁ >> DT₂ >> NT₂ >> DT₃ >> NT₃ >> DT₄ >> NT₄ >> DT₅ >> NT₅ >> DT₆
Partial day (used)
Baseline
Post-baseline (used for prediction)
Partial day (not used)

Fig. 2. Example extracting daytime (DT) and nighttime (NT) periods into baseline and post-baseline for a subject with five 24-hour (24H) days of data.

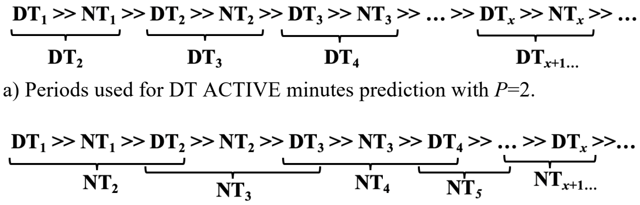
for which the data collection periods ranged from 4 days to 30 days. For the AMI and Philips datasets, the device manufacturers provided Actigraph-style activity counts and binary "sleep" or "wake" labels for each minute of data collection. We used these activity counts to represent physical activity.

For the Fitbit dataset, we originally collected data from 15 subjects who participated in the study during the duration of their inpatient rehabilitation stay [6], [37]. For several of these 15 subjects, there were entire nights with missing sleep data, likely due to patients taking the device off and/or the Fitbit sleep algorithms not properly detecting the wearer's abnormal sleeping patterns. Therefore, in this study we only used the data collected from eight participants for whom high-integrity sleep data were available every night of data collection (age 66.25 ± 12.89 years; 6 females and 2 males), for which the data collection periods ranged from 5 days to 17 days. Instead of recording activity counts, Fitbit labels each minute with a number of "steps" taken. We used steps as a similar measure to the aforementioned activity counts to estimate a subject's physical activity. For labeling sleep, Fitbit provides four levels of sleep: 0 (no sleep) or 1, 2, 3, (increasing levels of deeper sleep). To align this data with that of the AMI and Philips datasets, we reclassified these four sleep levels into binary sleep/wake labels where a value of 0 was re-labeled as "wake" and a value of 1, 2, or 3 was re-labeled as "sleep.". In summary, across all three datasets there was a total of 596 days of data collected in this study.

B. *Data Preprocessing*

The three datasets consisted of minute-by-minute physical activity and sleep/wake time series data. We preprocessed these time series to prepare the data for consistent analyses across the different wearable device manufacturers. We framed each subject's time series data to start on the first day with at least 400 consecutive minutes of recorded activity and to end on the day with at least 800 consecutive minutes of no recorded activity. We then normalized each subject's activity counts (AMI and Philips datasets) or steps (Fitbit dataset) to be between 0 and 1.

Because the three datasets were each sampled from different devices, we computed our own normalized labels for each individual subject. For each subject, we provided an "active" or "inactive" label for each minute in the time series data. These labels represent platform-independent labels that offer an alternative to individual device manufacturer's activity and sleep labels. To assign our individualized minute labels, we held out the first three days of data collection for each subject to serve as a baseline period. Using a subject's own data as a multiple-day baseline allowed us to account for extreme variability across subjects' data. We divided the baseline data into DT and NT periods, for a total of three baseline DT periods and three baseline NT periods. Figure 2 provides an example of how the baseline period was extracted from an example subject with five days of data collection. From these baseline periods, we extracted baseline activity means (BAM), namely, the DT baseline activity mean DT_{BAM} and the NT activity mean NT_{BAM}. We decided to use the mean of baseline activity because it was highly correlated with manufacturer sleep and wake labels (see Section IV for results). Using the DT_{BAM} for each subject, we labeled the remaining post-baseline DT minutes for the subject as active if its activity value was greater than DT_{BAM}, or inactive if it was less or equal to DT_{BAM}. We repeated this process for the NT periods, using NT_{BAM}. Our BAM labeling algorithm provided subject-specific and device-independent labels for daytime activity and nighttime activity. For nighttime activity, we hypothesized that the inactive labels were indicative of sleep and we evaluated this hypothesis by comparing the inactive labels to the manufacturer-provided sleep/wake labels.


DT₁ >> NT₁ >> DT₂ >> NT₂ >> DT₃ >> NT₃ >> … >> DTₓ >> NTₓ >> …
DT₂
DT₃
DT₄
DTₓ₊₁…
DT₁ >> NT₁ >> DT₂ >> NT₂ >> DT₃ >> NT₃ >> DT₄ >> … >> DTₓ >>…
NT₂
NT₃
NT₄
NT₅
NTₓ₊₁…

a) Periods used for DT ACTIVE minutes prediction with *P*=2.

b) Periods used for NT sleep duration prediction with *P*=3.

Fig. 3. Example of daytime (DT) and nighttime (NT) periods used for DT ACTIVE minutes prediction (a) and used for sleep duration prediction (b) with example P values.

C. *Feature Extraction*

From the time series data for each subject, we extracted relevant physical activity and sleep quality features from the DT and NT periods separately. To determine physical activity during DT periods, we used both the manufacturer activity counts and our aforementioned BAM labels. For each 24-hour day, we counted the number of BAM-labeled DT active minutes, as well as the number of transitions from active to inactive. We also computed a daytime activity ratio, which is the daytime sum of the manufacturer-measured activity divided by the corresponding 24-hour sum.

We computed NT features using both the minute-by-minute manufacturer's sleep/wake time series and our BAM-labeled inactive minutes. We extracted nighttime TST, number of sleep transitions, sleep onset latency (number of minutes from the start of nighttime before sleep), longest sleep bout length, and wake after sleep onset [17]. In addition to daytime and nighttime features, we included the number of days since each subject's injury or illness as a feature. To summarize, Table 1 lists the labels and features that were used for predicting DT TAM, NT TIM, and NT TST, respectively.

We extracted the aforementioned features from a sequence of *P* number of DT and NT periods to predict TAM for the following DT period, TIM for the following

TABLE I
PREDICTED VALUES AND THEIR LABELS USED FOR FEATURE EXTRACTION

Predicted Value	DT Feature Labels	NT Feature Labels
DT TAM	BAM active mins	BAM inactive mins
NT TIM	BAM active mins	BAM inactive mins
NT TST	BAM active mins	Manufacturer sleep/wake mins

BAM=baseline activity means, TAM=total active minutes, TIM=total inactive minutes, TST=total sleep time labeled by the manufacturer.

TABLE II
CORRELATIONS BETWEEN NIGHTTIME MANUFACTURER SLEEP/WAKE VALUES AND BAM INACTIVE VALUES

Dataset	Activity Metric	Threshold	Correlation
AMI	Activity count	Mean	0.78
	Activity count	70th	0.72
	Activity count	80th	0.68
Philips	Activity count	Mean	0.59
	Activity count	70th	0.56
	Activity count	80th	0.55
Fitbit	Steps	All	0.53

All correlations were significant at $p<0.01$. Correlations for all Fitbit dataset thresholds tested were the same value.

TABLE III
BEST DAYTIME TOTAL ACTIVE MINUTES (TAM) PREDICTION RESULTS

Subgroup	P	K	MAE	RMSE	NRMSE	r
Gender	4	125	55.73	73.92	9.11%	0.89
All	4	150	55.45	73.96	9.12%	0.89
All	4	250	56.04	74.54	9.19%	0.88
Gender	4	50	57.18	74.65	9.20%	0.88
Gender	4	75	57.04	74.67	9.21%	0.88

MAE=mean absolute error, RMSE=root mean squared error, NRMSE=normalized RMSE, and r=Pearson correlation coefficient. All correlation results were statistically significant at $p<0.0001$.

TABLE IV
BEST NIGHTTIME TOTAL INACTIVE MINUTES (TIM) PREDICTION RESULTS

Subgroup	P	K	MAE	RMSE	NRMSE	r
Dataset	4	20	43.61	60.37	11.18%	0.84
Dataset	4	15	43.70	60.85	11.27%	0.83
Dataset	4	10	43.22	61.37	11.37%	0.83
Dataset	5	10	44.45	62.16	11.51%	0.83
All	4	300	46.17	63.63	11.78%	0.82

MAE=mean absolute error, RMSE=root mean squared error, NRMSE=normalized RMSE, and r=Pearson correlation coefficient. All correlation results were statistically significant at $p<0.0001$.

TABLE V
BEST NIGHTTIME TOTAL SLEEP TIME (TST) PREDICTION RESULTS

Subgroup	P	K	MAE	RMSE	NRMSE	r
All	5	200	50.56	66.07	12.26%	0.85
All	4	200	50.37	66.22	12.29%	0.84
All	5	150	50.52	66.41	12.32%	0.84
All	4	150	51.12	66.53	12.34%	0.84
All	4	300	51.06	66.86	12.40%	0.84

MAE=mean absolute error, RMSE=root mean squared error, NRMSE=normalized RMSE, and r=Pearson correlation coefficient. All correlation results were statistically significant at $p<0.0001$.

NT period, or TST for the following NT period. For example, if P=1 and we are predicting TST, then we use the manufacturer's sleep/wake features from P-sequence DT_1 to predict NT_1, DT_2 to predict NT_2, and so forth. If P=3, then we use the P-sequence DT_1 >> NT_1 >> DT_2 to predict NT_2, DT_2 >> NT_2 >> DT_3 to predict NT_3, and so forth (see Figure 3 for a diagram showing both DT and NT predictions with different P values). Because we did not include the three-day baseline for prediction and there was a subject in the Fitbit dataset with only five days of data collection, the maximum P value was P=2 for DT TAM predictions and P=3 for NT TST predictions. Additionally, excluding the three-day baseline reduced the total combined dataset size to 464 days.

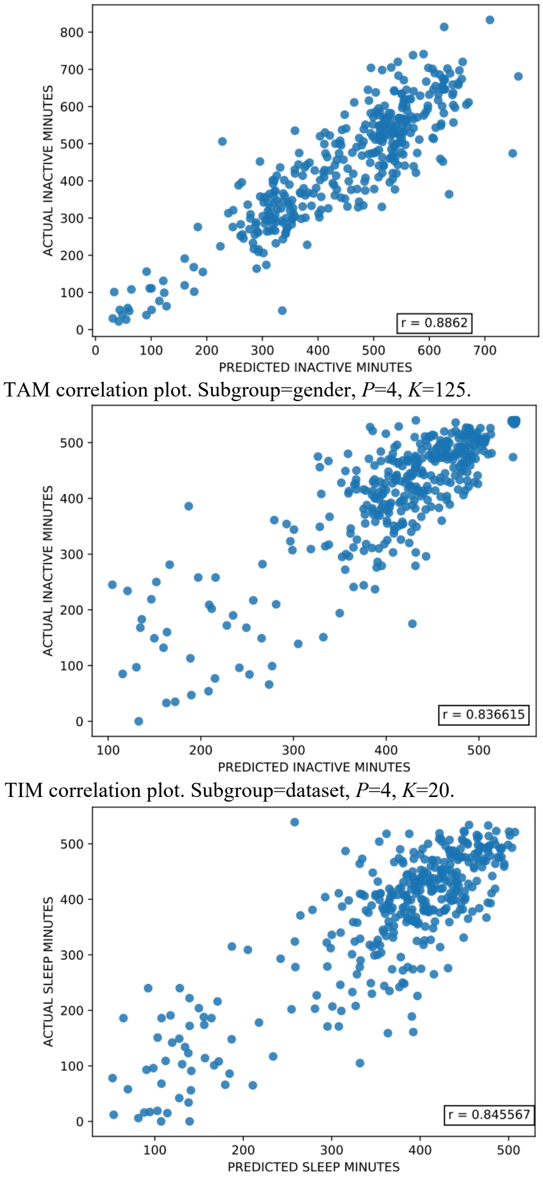
D. Prediction Models

To predict TAM, TIM, and TST based on the extracted features, we utilized a 100-tree random forest with bagging regression algorithm with leave-one-out-cross-validation. For leave-one-out-cross-validation, each of the participant periods was held out as a test sample while the remaining P-sequences were used for training. Our initial nighttime predictions started on the first night following the three-day baseline. We excluded P-sequences if they were futuristic data from the same participant as the test participant. For example, when P=1 there were 464 total NT periods, from which we held out one nighttime period, NT_x for leave-one-out-cross-validation. We then excluded all periods > x from training that were collected from the same participant as NT_x. To improve our random forest prediction accuracy, we utilized a K-nearest neighbors (KNN) algorithm to select "similar" P-sequences from the training set to form a smaller, more specialized training set. Two P-sequences were considered similar if they were in the same subgroup and had a small Euclidean distance computed between their feature vectors. We investigated alternatives to comparing based on feature vectors, including using dynamic time warping to compare two P-sequences. Experiments revealed the best results were achievable using feature vectors. The subgroup parameter restricted which feature vectors were considered similar to the held-out feature vector. Subgroups we explored included dataset, gender, and no subgroupings (using all P-sequences). With the K parameter we aimed to train on a minimal set of P-sequences that historically were similar to the current P-sequence for which we were making a prediction.

IV. RESULTS

To evaluate our BAM data-driven approach to normalizing data collected from different devices, we correlated the original manufacturer-provided sleep/wake time series with our BAM-labeled time series. We experimented with various threshold values for determining active/inactive, including the baseline mean and percentiles in increments of ten. Table II shows the top three correlations for the AMI and Philips datasets. For the Fitbit dataset, the correlations for mean and all percentiles tested were the same, r=0.53. This is because Fitbit typically measures zero steps when the wearer is laying down and doesn't exhibit wrist motion similar to an arm swing during a step. This means that during NT periods, there were almost exclusively zero values, causing any minute with more than zero steps to be labeled as active.

To standardize features for input to machine learning models, we subtracted the mean and scaled to unit variance. Next, we trained and tested the prediction models using leave-one-out-cross-validation. We evaluated the random

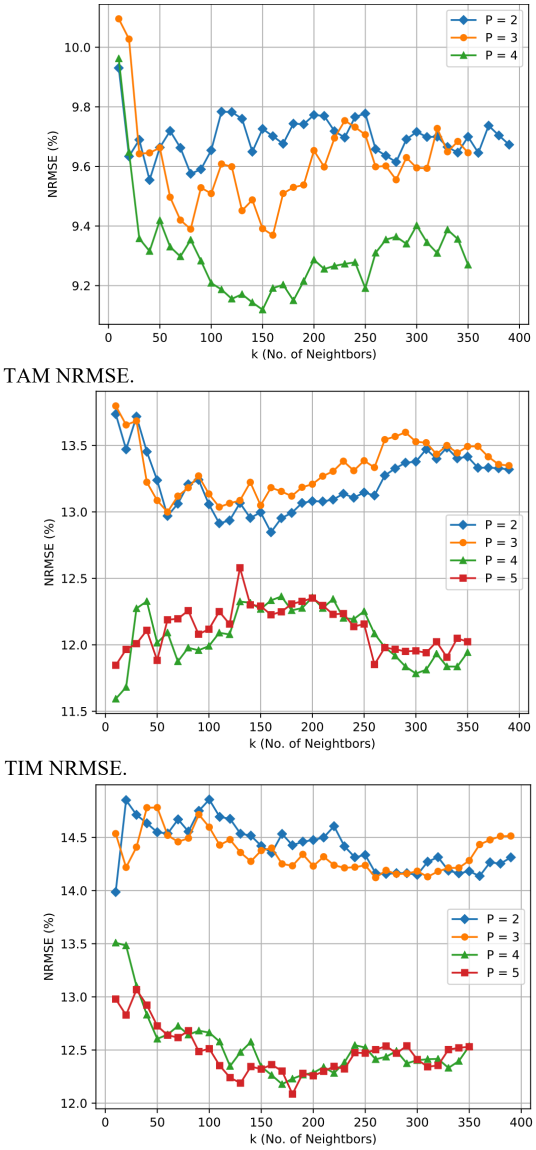


a) TAM correlation plot. Subgroup=gender, *P*=4, *K*=125.

b) TIM correlation plot. Subgroup=dataset, *P*=4, *K*=20.

c) TST correlation plot. Subgroup=all, *P*=5, *K*=200.

Fig. 4. Correlation plots showing top regression results for total active minutes (TAM), total inactive minutes (TIM), and total sleep time (TST) actual versus predicted values.



a) TAM NRMSE.

b) TIM NRMSE.

c) TST NRMSE.

Fig. 5. Normalized root mean square error (NRMSE) as a function of *K* (number of similar *P* sequences used for training) for different values of *P*. Plots for total active minutes (TAM), total inactive minutes (TIM), and total sleep time (TST) are shown.

forest regression results using mean absolute error (MAE), root mean squared error (RMSE), normalized RMSE (NRMSE), and Pearson correlation coefficients (*r*). We computed NRMSE as RMSE divided by the difference between the maximum and minimum actual ground truth values:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{prediction}_i - \mathrm{actual}_i)^2}{n}}}{\mathrm{actual}_{\max} - \mathrm{actual}_{\min}} \qquad (1)$$

Tables III-V show five random forest results and their parameter configurations, in ascending NRMSE order, for DT TAM, NT TIM, and NT TST, respectively. To provide context for interpreting the prediction results, DT TAM, demonstrated a mean and standard deviation of 431.20 ± 162.45 minutes (coefficient of variation equal to 37.67%) across the participant group, while the NT TIM and TST had a mean and standard deviation of 414.22 ± 108.61 minutes (coefficient of variation equal to 26.22%) and 370.27 ± 124.20 minutes (coefficient of variation equal to 33.54%), respectively. To more thoroughly investigate the prediction results, we provide scatter plots of predicted total minutes versus actual total minutes for the best TAM (Figure 4a), TIM (Figure 4b), and TST (Figure 4c) results from Tables III-V. For each of the figures, the correlation for the data is shown as an annotation in the bottom right corner of the plot.

We ran several experiments to explore the effects of different value combinations of parameters, *P* (the number of periods preceding the NT period) and *K* (the number of similar *P*-sequences used by *K*NN to determine the training set). Figure 5 shows NRMSE values for alternative values of *P* as a function of alternative values of *K*. *P*=1 exhibited relatively large error compared to the other *P* values so we exclude it from the plots of Figure 5 to clearly illustrate the patterns of the lower *P* value NRMSE results.

V. Discussion

In this paper, we propose the BAM data-driven approach for predicting daytime total active minutes and nighttime total inactive minutes for patients undergoing inpatient rehabilitation. Our approach allows data collected from different wearable device manufacturers to be combined, compared, and used for prediction. Of the different mean and percentiles from the three-day baseline period we explored as active minute thresholds, the baseline means exhibit the highest correlation to the manufacturer's sleep and wake labels (see Table II; AMI dataset *r*=0.78; Philips dataset *r*=0.59). Since the correlations were < 1.0, we anticipated the TIM prediction results using the BAM labels would be less accurate than the TST prediction results that used the manufacturer's sleep/wake labels. The random forest results indicate that the top TAM and TST results are comparable (see Tables IV and V; TIM NRMSE 11.18%; TAM NRMSE 12.26%). This suggests that our BAM approach could be used not only for prediction, but also for comparisons between subject's data collected from different devices. When training the random forest regressors on each dataset individually, the results are not as strong as the combined dataset. For the AMI dataset, the best NRMSE result for TIM is 14.11% and for TST is 14.98%. For the Philips dataset, these numbers are 11.45% for TIM and 13.12% for TST. For the Fitbit dataset, these numbers are 18.05% for TIM and 21.23% for TST. A likely reason that the Fitbit-based prediction is weaker than the other datasets is the small sample size of the Fitbit dataset (N=8) and low detection of nighttime active minutes. This observation further supports the idea of pooling data together to form larger datasets, which would provide more similar *P*-sequences detected with *K*NN and used for training.

For the TAM results, there are no manufacturer "active" or "inactive" labels with which to compare our BAM approach; however, we do see lower prediction error for TAM (see Table III; 9.11% NRMSE and *r*=0.89). Breaking this result down by dataset reveals that, like the NT period predictions, the prediction results vary by device and are strongest when pooled together. The AMI dataset's best TAM NRMSE is 13.96%, while that number is 10.48% for the Philips dataset, and 11.48% for the Fitbit dataset. These NT and DT prediction results further suggest the importance of combining datasets to increase training set size and consequently prediction accuracy. To deploy models to a clinical setting, we would want the lowest error possible.

To more thoroughly explore the 9-12% error rates, we include correlation plots in Figure 4. The prediction results align with our intuition. The random forest regressors are more accurate when the actual minutes being predicted are closer to the mean, which is the case for the NT periods. In Figures 4b and 4c, there is a cluster of high actual total inactive minutes and sleep minutes. With a larger, more diverse dataset, we anticipate custom models could be built for the outlier subjects with lower active and inactive minutes to improve accuracy.

Next we explore the effects of the *P* and *K* parameters on the TAM, TIM, and TST prediction accuracies. From investigating the plots in Figure 5, there are large incremental benefits to increasing *P* from 1 to 2, from 2 to 3, and from 3 to 4. The NRMSE curves for *P*=4 and *P*=5 are fairly similar, suggesting that accurate predictions can be made with as few as four post-baseline training periods. This constitutes a 48-hour period for a TAM prediction and a 63-hour period for a TIM prediction. We plan to investigate techniques to shorten this overall time from when a device is first worn by a patient to when an accurate next period prediction can be made. A few approaches include trimming periods from the baseline and constructing more individualized models for specific injuries or illnesses, such as stroke or TBI. The tradeoffs for the *K* parameter are not as clear as for the *P* parameter. *K* appears to be more sensitive to which value is being predicted. For TAM, *K*=150 seems to capture the majority of the prediction improvement. For TIM, *K*=50 and for TST, *K*=150 are reasonable minimum values for achieving accurate results. Since larger values of *K* do not greatly improve prediction accuracy, as datasets get larger, computational overhead from exploring optimal *K* values is not expected to increase. However, the search space for determining the *K* nearest neighbors will grow. Individualized models for different patient subgroups could also help limit the number of required *K*NN comparisons.

VI. Conclusions

We investigated applying device-independent physical activity and sleep labels determined from a baseline period to allow data collected from multiple wearable devices to be combined into larger datasets. Larger datasets can be used to train machine learning models to predict a patient's next day physical activity and next night sleep duration with greater accuracy. We demonstrated such prediction models trained with data collected from 44 inpatient rehabilitation subjects can achieve NRMSE values near 9% for daytime physical activity prediction and near 11% for nighttime sleep duration prediction. These results were an expansion over own prior work with data from a single sensor device [36]. For future work, we plan to continue growing our sample size to provide additional historical data sequences for *K*NN to select from. We anticipate this will help prediction accuracy for outlier *P*-sequences sampled from subjects with highly irregular activity and sleep behavior. We also plan to apply deep learning models [11] in an effort to further reduce the nighttime prediction error. Our eventual goal is to deploy models that are accurate enough for clinicians to use to help customize individual patient therapy programs. If such a system can make accurate predictions in near real time, clinicians could use this additional information about a patient's next day physical activity and next night sleep requirements to adapt forthcoming therapeutic activities and potentially shorten the recovery process.

ACKNOWLEDGMENT

We wish to thank Alexa Andrews, Jordana Dahmen, and Alyssa La Fleur for their help with data collection. We also wish to thank our therapist collaborators at St. Luke's Rehabilitation Institute for providing analysis and interpretation insights.

REFERENCES

[1] J. L. Mathias and P. K. Alvaro, "Prevalence of sleep disturbances, disorders, and problems following traumatic brain injury: a meta-analysis," *Sleep Med.*, vol. 13, no. 7, pp. 898–905, Aug. 2012.

[2] Z. Pasic, D. Smajlovic, Z. Dostovic, B. Kojic, and S. Selmanovic, "Incidence and Types of Sleep Disorders in patients with Stroke," *Med. Arch.*, vol. 65, no. 4, p. 225, 2011.

[3] P. F. Innominato *et al.*, "Circadian Rhythm in Rest and Activity: A Biological Correlate of Quality of Life and a Predictor of Survival in Patients with Metastatic Colorectal Cancer," *Cancer Res.*, vol. 69, no. 11, pp. 4700–4707, Jun. 2009.

[4] S. A. Prince, K. B. Adamo, M. E. Hamel, J. Hardt, S. C. Gorber, and M. Tremblay, "A Comparison of Direct Versus Self-Report Measures for Assessing Physical Activity in Adults: A Systematic Review," *Int. J. Behav. Nutr. Phys. Act.*, vol. 5, p. 56, 2008.

[5] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Med. Rev.*, vol. 15, no. 4, pp. 259–267, Aug. 2011.

[6] G. Sprint *et al.*, "Analyzing Sensor-Based Time Series Data to Track Changes in Physical Activity during Inpatient Rehabilitation," *Sensors*, vol. 17, no. 10, p. 2219, Sep. 2017.

[7] A. M. Berger, K. K. Wielgus, S. Young-McCaughan, P. Fischer, L. Farr, and K. A. Lee, "Methodological Challenges When Using Actigraphy in Research," *J. Pain Symptom Manage.*, vol. 36, no. 2, pp. 191–199, Aug. 2008.

[8] M. E. G. Bigelow *et al.*, "Point-of-Care Technologies for the Advancement of Precision Medicine in Heart, Lung, Blood, and Sleep Disorders," *IEEE J. Transl. Eng. Health Med.*, vol. 4, pp. 1–10, 2016.

[9] L. de Souza, A. A. Benedito-Silva, M. L. N. Pires, D. Poyares, S. Tufik, and H. M. Calil, "Further validation of actigraphy for sleep studies," *Sleep*, vol. 26, no. 1, pp. 81–85, Feb. 2003.

[10] D. Treacy, L. Hassett, K. Schurr, S. Chagpar, S. S. Paul, and C. Sherrington, "Validity of Different Activity Monitors to Count Steps in an Inpatient Rehabilitation Setting," *Phys. Ther.*, vol. 97, no. 5, pp. 581–588, May 2017.

[11] A. Sathyanarayana *et al.*, "Sleep Quality Prediction From Wearable Data Using Deep Learning," *JMIR MHealth UHealth*, vol. 4, no. 4, p. e125, Nov. 2016.

[12] M. M. Ohayon, M. A. Carskadon, C. Guilleminault, and M. V. Vitiello, "Meta-Analysis of Quantitative Sleep Parameters From Childhood to Old Age in Healthy Individuals: Developing Normative Sleep Values Across the Human Lifespan," *Sleep*, vol. 27, no. 7, pp. 1255–1273, Oct. 2004.

[13] P. Alinia, C. Cain, R. Fallahzadeh, A. Shahrokni, D. Cook, and H. Ghasemzadeh, "How Accurate Is Your Activity Tracker? A Comparative Study of Step Counts in Low-Intensity Physical Activities," *JMIR MHealth UHealth*, vol. 5, no. 8, p. e106, 2017.

[14] G. M. Dominick, K. N. Winfree, R. T. Pohlig, and M. A. Papas, "Physical Activity Assessment Between Consumer- and Research-Grade Accelerometers: A Comparative Study in Free-Living Conditions," *JMIR MHealth UHealth*, vol. 4, no. 3, p. e110, Sep. 2016.

[15] S. Roomkham, D. Lovell, J. Cheung, and D. Perrin, "Promises and Challenges in the Use of Consumer-Grade Devices for Sleep Monitoring," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 53–67, 2018.

[16] M. J. Makley *et al.*, "Objective measures of sleep and wakefulness in patients with moderate to severe brain injury on an inpatient rehabilitation unit. Pearls and pitfalls of actigraph monitoring," *NeuroRehabilitation*, vol. 43, no. 3, pp. 277–285, 2018.

[17] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, A. G. Ravelo-García, and T. Penzel, "A Review of Approaches for Sleep Quality Analysis," *IEEE Access*, vol. 7, pp. 24527–24546, 2019.

[18] H.-Y. Chiu, P.-Y. Chen, N.-H. Chen, L.-P. Chuang, and P.-S. Tsai, "Trajectories of sleep changes during the acute phase of traumatic brain injury: A 7-day actigraphy study," *J. Formos. Med. Assoc.*, vol. 112, no. 9, pp. 545–553, Sep. 2013.

[19] D. Leger *et al.*, "Using actigraphy to assess sleep and wake rhythms of narcolepsy type 1 patients: a comparison with primary insomniacs and healthy controls," *Sleep Med.*, vol. 52, pp. 88–91, Aug. 2018.

[20] R. Sainath, K. Pruthvisai, B. Akhil, and S. Palaniswamy, "Sleep Pattern Monitoring and Analysis to Improve the Health and Quality of Life of People," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 1900–1905.

[21] M. Alharbi, A. Bauman, L. Neubeck, and R. Gallagher, "Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population," *Eur. J. Prev. Cardiol.*, vol. 23, no. 14, pp. 1476–1485, Sep. 2016.

[22] T. A. Floegel, A. Florez-Pregonero, E. B. Hekler, and M. P. Buman, "Validation of Consumer-Based Hip and Wrist Activity Monitors in Older Adults With Varied Ambulatory Abilities," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 72, no. 2, pp. 229–236, Feb. 2017.

[23] V. J. Block *et al.*, "Continuous daily assessment of multiple sclerosis disability using remote step count monitoring," *J. Neurol.*, vol. 264, no. 2, pp. 316–326, Feb. 2017.

[24] K. R. Westerterp, "Physical activity assessment with accelerometers," *Int. J. Obes.*, vol. 23, no. 3, pp. S45–S49, Apr. 1999.

[25] J. Mantua, N. Gravel, and R. M. C. Spencer, "Reliability of Sleep Measures from Four Personal Health Monitoring Devices Compared to Research-Based Actigraphy and Polysomnography," *Sensors*, vol. 16, no. 5, 05 2016.

[26] L. J. Meltzer, C. M. Walsh, J. Traylor, and A. M. L. Westin, "Direct Comparison of Two New Actigraphs and Polysomnography in Children and Adolescents," *Sleep*, vol. 35, no. 1, pp. 159–166, Jan. 2012.

[27] S. Thiemjarus, "A Device-Orientation Independent Method for Activity Recognition," in *2010 International Conference on Body Sensor Networks*, 2010, pp. 19–23.

[28] J. Wu and R. Jafari, "Orientation Independent Activity/Gesture Recognition Using Wearable Motion Sensors," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1427–1437, Apr. 2019.

[29] M. Skubic, R. D. Guevara, and M. Rantz, "Automated Health Alerts Using In-Home Sensor Data for Embedded Health Assessment," *IEEE J. Transl. Eng. Health Med.*, vol. 3, pp. 1–11, 2015.

[30] R. Saeedi and A. H. Gebremedhin, "A Signal-Level Transfer Learning Framework for Autonomous Reconfiguration of Wearable Systems," *IEEE Trans. Mob. Comput.*, vol. 19, no. 3, pp. 513–527, Mar. 2020.

[31] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: a survey," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 537–556, 2013.

[32] K. Lee and M.-P. Kwan, "Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results," *Comput. Environ. Urban Syst.*, vol. 67, pp. 124–131, Jan. 2018.

[33] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiol. Meas.*, vol. 35, no. 11, pp. 2191–2203, Oct. 2014.

[34] W. Hidayat, T. D. Tambunan, and R. Budiawan, "Empowering Wearable Sensor Generated Data to Predict Changes in Individual's Sleep Quality," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, 2018, pp. 447–452.

[35] M. Farajtabar, E. Kıcıman, G. Nathan, and R. W. White, "Modeling behaviors and lifestyle with online and social data for predicting and analyzing sleep and exercise quality," *Int. J. Data Sci. Anal.*, Jun. 2018.

[36] A. Fellger, G. Sprint, A. Andrews, D. Weeks, and E. Crooks, "Nighttime Sleep Duration Prediction for Inpatient Rehabilitation Using Similar Actigraphy Sequences," in *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*, 2019, pp. 41–44.

[37] J. Conci, G. Sprint, D. Cook, and D. Weeks, "Utilizing Consumer-grade Wearable Sensors for Unobtrusive Rehabilitation Outcome Prediction," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4.