International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-018-0136-8

REGULAR PAPER

Modeling behaviors and lifestyle with online and social data for predicting and analyzing sleep and exercise quality

Mehrdad Farajtabar[1] · Emre Kıcıman[2] · Girish Nathan[3] · Ryen W. White[2]

Received: 29 September 2017 / Accepted: 6 June 2018

Abstract

While recent data studies have focused on associations between sleep and exercise patterns as captured by digital fitness devices, it is known that sleep and exercise quality are affected by a much broader set of factors not captured by these devices, such as general lifestyle, eating, and stress. Here, we conduct a large-scale data study of exercise and sleep effects through an analysis of 8 months of exercise and sleep data for 20 k users, combined with search query logs, location information and aggregated social media data. We analyze factors correlated with better sleep and more effective exercise, and confirm these relationships through causal inference analysis. Further, we build linear models to predict individuals' sleep and exercise quality. This analysis demonstrates the potential benefits of combining online and social data sources with data from health trackers, and is a potentially rich computational benchmark for health studies. We discuss the implications of our work for individuals, health practitioners and health systems.

Keywords User modeling · Health tracker · Sleep and exercise quality · Online and social features · Prediction

1 Introduction

Improving sleep and exercise quality has many health benefits [1–3] and also leads to greater happiness and enhanced productivity [4–6]. Health practitioners consider sleep as an indicator of an individual's health; good quality sleep is a key part of a healthy lifestyle, benefiting the heart, mind, performance, and emotional balance [5], while poor sleep usually leads to daytime sleepiness, fatigue, and an impaired ability to learn and perform tasks [7]. Similarly, effective exercise helps individuals live longer and better, boosting mental wellness by relieving tension, anxiety, depression [8]; it also improves physical wellness by enhancing blood circulation, weight control and muscle strength [1]. Coupled with the increasing popularity of mobile and wearable devices (e.g, FitBit,[1] Jawbone,[2] Misfit,[3] and Sense[4]), the importance of exercise and sleep has led to the development of a number of non-clinical systems for monitoring and analyzing them.

Researchers have identified relationships between people's sleep and exercise quality and their demographic attributes and daily activities [9]. Extensive research has also been conducted to assess people's health from social media data, especially Twitter [10,11]. While it is generally accepted that a person's lifestyle (e.g., food habits) [12], stressful events (e.g., financial worries) [13], and events that disrupt a routine (e.g., travel, celebrations) can significantly impact sleep and exercise [14], there is a lack of research on incorporating such rich data sources into sleep and exercise analysis.

In this paper, we present a large-scale study of the sleep and exercise quality effects of a wide variety of individual behaviors. To expand the set of behaviors we can study,

✉ Mehrdad Farajtabar
mehrdad@gatech.edu

Emre Kıcıman
emrek@microsoft.com

Girish Nathan
ginathan@microsoft.com

Ryen W. White
ryenw@microsoft.com

[1] Georgia Tech, 266 Ferst Drive, Atlanta, GA, USA

[2] Microsoft Research, One Microsoft Way, Redmond, WA, USA

[3] Microsoft, One Microsoft Way, Redmond, WA, USA

[1] http://www.fitbit.com/.

[2] https://jawbone.com/up.

[3] http://misfit.com/.

[4] https://hello.is/.

Published online: 15 June 2018

we link health device data to non-traditional datasets, such as search query logs, location information, and aggregated social media data. While combining such varied data has been reported to be effective in other health domains (e.g., forecasting disease [15]), to our knowledge, such a large-scale study of diverse data sources has not been reported for sleep and exercise. Overall, our analysis encompasses approximately 20 k users' sleep and exercise data, containing 1.3 M sleep and 600 k exercise observations (cf. Table 2).

Following the practice of previous studies, we characterize sleep quality through the following measures: (i) time to fall asleep [16]; (ii) number of wakeups [17]; (iii) sleep efficiency [18] (the fraction of the total time spent in bed that the user is asleep). We quantify exercise and fitness quality based on (i) exercise intensity [19] (the rate of calorie burn during exercise); and (ii) resting heart rate [20] (the heart rate recorded during sleep, which is an indicator of heart health). We first present an exploratory characterization of our data. Using a regression analysis, we identify behaviors that are significantly correlated with sleep and exercise quality. We validate these associations using causal inference models [21] that reduce bias due to observed confounders. We further build linear regression models to predict sleep and exercise quality targets. In addition to examining the prediction accuracy on held-out data, we highlight specific features that are especially important for achieving good prediction performance, i.e., location and search features. Furthermore, we conduct experiments to analyze how the prediction accuracy can be improved and how features perform individually, which is helpful when data are limited. Finally, we predict sleep- and exercise-related variables and show that our models achieve low errors in prediction. The insights and analysis presented in Sects. 4.1 and 4.2 are most helpful for individuals and practitioners wishing to promote a healthy lifestyle. Results presented in Sect. 4.3 about prediction lead times and their effect on prediction quality (e.g., how early in the day low-quality sleep can be reliably predicted) may be most useful for wearable device manufacturers and health application developers.

2 Related work

In this section, we review prior research on monitoring, measuring, and predicting exercise and sleep and comment on their highlights and properties.

2.1 Exercise and sleep monitoring

Existing research related to sleep/exercise quality focuses primarily on monitoring and detection using signals from dedicated devices. Phone usage information has been utilized to find the onset of sleep [22], but not much about the quality of sleep. Researchers have estimated bedtime, wakeup time and sleep duration using phone sensor data and daily activities over 10 weeks, albeit limited to students only [23]. A mobile service was deployed that leverages built-in sensors (light sensor, accelerometer, microphone) on smartphones to detect sleep stages and sleep quality [24]. The bulk of existing research relies on special sensors or monitoring devices, reducing their general applicability. In a closely related study researchers built a practical system to monitor seven individuals' sleep quality using the smartphone microphone to detect events that are related to sleep quality, classify them by a decision-tree-based algorithm, and finally infer quantitative measures of sleep quality [25]. The number of subjects under study limits result generalizability. Leveraging the built-in sensors on smartphones, other researchers have integrated physical activities with the sleep environment, inherent temporal relations, and personal factors using statistical modeling for fine-grained sleep stage detection and to generate sleep quality reports [26]. The novelty is that it involved physical activity for sleep quality measurement. However, the sources of relevant information and features could have been much broader. To address the problem of monitoring exercise quality, Pernek and colleagues propose a network of wearable accelerometers and an off-the-shelf smartphone to measure exercise intensity [27]. They use a hierarchical algorithm, consisting of two layers of support vector machines to first recognize the type of exercise and then measure the exercise intensity. Unfortunately, wearing special sensors is not always feasible. In contrast, our work leverages common sources of information. Other researchers introduce a motion rehabilitation system for chronic patients based on smartphones [28], processing motion sensor data online to output real-time acoustic feedback regarding exercise quality. This study is limited to chronic patients requiring physical rehabilitation with the goal of maintaining the motivation to exercise in real-time.

2.2 Exercise and sleep analysis and prediction

While most existing research targets diseases such as insomnia and circadian rhythm sleep disorders, some prior work has tried to predict and analyze sleep and exercise quality in general settings. Bai and colleagues try to predict sleep quality using user data such as daily activity, living environment and social activity information [29]. They show that user context can predict sleep quality with a 78% accuracy. However, they used survey data, which can be challenging to obtain and featurize, and requires direct involvement from subjects. Other researchers use 1 month of phone sensor data and sleep diary entries from 27 participants to construct models that detect sleep and non-sleep states, daily sleep quality, and global sleep quality [30]. Sleep diaries are intrusive and burdensome for participants, which limits their scalability. Jayarjah and colleagues present a study on 400 undergraduate students over 15 months that quantifies the quality of sleep and tries to correlate this with aspects of their daily lives, especially individuals' usage of apps during the day and their physical environment that may impact sleep [31]. While this study has scale, undergraduate students have special habits and this again limits the generalizability of the results to the broader population. A set of causality analysis techniques is adopted in other research to find relationships between the environment and sleep quality via information collected using built-in sensors in off-the-shelf mobile and wearable devices [32]. Their model is adapted to their specific environment only and is not readily usable in other settings. Krishna and colleagues present an automated sleep quality monitor and sleep duration estimator for a user that combines features related to user surroundings and those related to user movements during sleep to generate personalized sleep quality measurements [33]. However, the movements are only useful for real-time detection and are not applicable for prediction.

Given the increasing availability of social data there have been notable efforts to use social media to study sleep [34–36]. In one article, the authors combine social media with data from a sleep-tracking app [34], conduct a large-scale study of sleep with more than 500 k users [35], and develop a novel way to predict individuals' sleep conditions by scrutinizing facial cues as sleep specialists would [37]. They find that higher social media activity levels are associated with lower sleep quality and duration. However, their method is not robust to noise within the data. Other research uses qualitative methods to study sleep patterns [36] or predicts sleep quality from sensor data using state-of-the-art deep learning techniques [38]. While very related, none of the aforementioned research has used a feature set as rich as that employed in this study. Furthermore, we study the factors affecting the prediction and support our findings on correlation with a causal analysis. All of these contribute to making our paper a comprehensive study of the subject and a reference for other related research.

3 Data, preprocessing, and methods

In this section, we explain data collection and the preprocessing steps to "sanitize" the data. This is done for all predictive features and the targets of interest. Data were collected between August 2015 and April 2016 and from users who agreed to link their Cortana data and Microsoft Health data (including their Microsoft Band device data) for use in generating additional insights or recommendations about their sleep or activity. Demographic variables are self-reported through the Microsoft Health app, which served as the companion application for the Band wearable device. While the user age and weight distributions closely track official estimates in the USA, we note that our sample is predominantly male. Furthermore, we acknowledge that the target of this study comprises those individuals who can afford to purchase a Microsoft Band device, which retailed for 150–200 USD. Further studies on different user populations or statistical methods for removing possible biases from study is left as future work.

3.1 Sleep and exercise measures

The top five rows of Table 1 denote five measures of sleep and exercise quality that are recorded by the Microsoft Band fitness device. In addition, our dataset includes age, weight and height, and daily activity features, such as step count and calories burned.

3.1.1 Sleep

Sleep data from wearable devices and smart phones provide objective measurements which have been preferred to subjective self-reports that may be significantly biased [39]. Among our sleep quality targets, the concept of "number of wakeups" and "time to fall asleep" is clear. "Sleep efficiency" is defined as the fraction of time spent in bed that the user is asleep. Sleep efficiencies under 85% are frequently reported in insomnia patients [18,40,41] and are used in both qualitative and quantitative sleep analysis. To detect sleep events, the Microsoft Band considers movement signals that utilize a three-axis accelerometer and gyrometer, and an optical heart rate sensor.

The Band employs internally vestigated proprietary algorithms for detection of sleep versus movement. The Band uses its hardware/software for sleep detection and computes sleep efficiency, number of wakeups. We use these quantities as our ground truth targets for training predictive models, as we detail later. Time in bed is either provided by manual input from the user (both before going to sleep and immediately after waking up) or automatically based on movement if the user does not provide manual input. The use of an event marker to denote bed timing is widely used in sleep research involving sleep diaries [42]. Following best practices [43,44], we filter out any sleep record with duration below 0.5 h and above 12 h of time in bed. Recent research [44] has verified that the Microsoft Band's measurements match published sleep estimates [43].

Table 1 Targets (dependent variables) and features (independent variables)

Target/feature	Health measure	Source	Range	Comment
Sleep quality	◇ Time to fall sleep	MS Band	Real	Recorded in minutes
	◇ Number of wakeups	MS Band	Integer	0, 1, 2,…
	◇ Efficiency	MS Band	Integer	Percentage of being actually sleep over being in bed (in 0–100)
Exercise and fitness	◇ Exercise intensity	MS Band	Real	Rate of burning calories while doing exercise in Cal/min (0–30)
	◇ Resting heart rate	MS Band	Integer	Heart rate recorded during sleep in bpm (typically in 40–80)
Basic	◇ Age	Profile	Integer	Recorded in years
	◇ Gender	Profile	Binary	Male, female
	◇ Weight	Profile	Integer	Typically in 80–250 (lb)
	◇ Height	Profile	Integer	Typically in 60–80 (in)
Daily	◇ Day of week	MS Band	Discrete	Fri, Sat, Sun, …, Thu
	◇ Hour of event	MS Band	Real	The hour (of sleep or exercise event) elapsed from last midnight
	◇ Calories burned	MS Band	Real	Average rate of burning calories in that day
	◇ Steps taken	MS Band	Integer	Total number of steps in that day (typically in 1000–20000)
	◇ Today's exercise	MS Band	Discrete	Bike, run, workout (just for analyzing sleep quality)
	◇ Prev. Night sleep 6 × 1 vector	MS Band	–	Time to fall asleep (mins), resting heart rate (beeps per mins) number of wakeups, sleep efficiency percentage, sleep duration time going to bed (for analyzing exercise quality)
Location	◇ 11 × 1 vector	Cortana	Mixed	Binary visit association to each of 10 categories: food, retail, health, entertainment, banking, education, sports, beauty, travel, service (in [0, 1]); offset time
Information	◇ 14 × 1 vector	Bing	Mixed	Binary association to 12 categories: food, exercise, health, celebration arts, police, religious, science, technology, business, positive, negative (each in [0, 1]); other/no category; offset time
Aggregated social	◇ 13 × 1 vector	Twitter	Mixed	Membership to 12 categories: Food, exercise, health, celebration arts, police, religious, science, technology, business, positive negative (each in [0, 1]); other/no category

3.1.2 Exercise

For exercise, users have an option to input the type of exercise they are engaging in—run, bike, or workout. The Band device tracks heart rate during exercise using its optical heart monitor. The device also uses GPS and movement sensors, the heart rate monitor, and the basic information to estimate calories burned. Following best practices, we excluded exercise events whose length (minutes) are below 5 or above 180.

3.1.3 Basic features

To remove outliers, we exclude users with age (in years) below 10 or above 100, weight (in lbs) below 50 or above 250,

Table 2 Summary statistics for sleep and exercise data

Task	# Users		# Events	
	Male	Female	Male	Female
Sleep	18,346 (93%)	1287 (7%)	1,204,558 (94%)	77,490 (6%)
Exer.	16,607 (93%)	1181 (7%)	596,345 (93%)	45,775 (7%)

and height (in inches) below 50 or above 80. All these features are commonly used to model sleep and exercise quality [45, 46].

3.1.4 Daily and activity features

Our list of daily features is as follows: (1) day of the week; (2) the hour when the event started; the next three express what the user has done in the day that the sleep event commenced: (3) steps taken; (4) calories burned, (5) type of exercise (if any). The last six are about the previous night and are used to predict exercise quality: last night's: (5) time to fall asleep; (6) resting heart rate; (7) sleep efficiency; (8) number of wake-ups; (9) sleep duration; and (10) bedtime.

After pruning the dataset, we obtain a final dataset of 20 K users and 1.3 M sleep and 600 K exercise records (see Table 2). We believe that this is a sufficiently large dataset for our analysis.

3.2 Characterizing behaviors and lifestyle

To connect information captured by health tracking devices to other behaviors and lifestyle factors, we aggregate data from various sources, specifically Cortana (personal digital assistant, offering location and search data) and Twitter. The combination of data from these disparate sources leads to a rich set of user features, including: (i) Locations visited during the day, such as restaurants, banks; and (ii) interests and intent inferred from both search queries and geo-aggregated tweets.

3.2.1 Location features

These features are collected by users' personal digital assistant (Cortana), where GPS and proprietary triangulation techniques are used to infer visits to points of interest, specifically businesses. The location features are split into various categories, summarized in Table 1 under the feature set "Locations." Since there are only a few categories, one-hot encoding is used to denote the presence or absence of a certain category in the user's visited location. For example, if a user visits a retail location that also has a food court, the relevant feature vector for such a location, by examining Table 1, is (1, 0, 1, 0, 0, 0, 0, 0, 0, 0). For each sleep or exercise event for a (user, day) tuple, we generate a location feature vector for that day by averaging over all such ten-dimensional vectors for that user on that day before the actual sleep or exercise event. In addition to the location visited, the time difference between when the location event began, and the actual sleep or exercise event also tends to be important and is recorded (referred to as "visit offset time" henceforth). Capturing this time for visited locations is useful for designing interventions and alerts related to the impact of visiting a particular location before a sleep or an exercise event. For example, with restaurants and food locations, it has been reported that food consumption close to bedtime is negatively associated with sleep quality [12]. We account for the temporal nature of location visits by expanding the location feature vector to 11 dimensions where the 11th coordinate is the average of all visit offset times of the user over all locations for that day before the sleep or exercise event. Averaging visit offset times is a useful indicator of how the individual's day relates to a sleep or an exercise event. For instance, we expect a negative correlation between low offset times between food visits and bedtime. To the best of our knowledge, the effect of visited location categories on sleep and exercise quality has not previously been studied at this scale.

The strong connection between locations and activities has been the subject of much research [47–50]. Researchers have observed that activities are often used instead of places when responding to a request for location [49], suggesting an interchangeable usage of activity and location. Other works search for a semantic representation for a location [48,50] that best represents it, for example associating "office" to "working". While associating a location with only one likely activity is clearly a simplification, associating a behavior or property to a class of location is common in activity recognition in location-based services [50–52].

3.2.2 Information features

Search logs (from Bing) offer insight into people's interests and intents, beliefs and thoughts [53], problems and health concerns [54]. Information features are extracted by membership proportion to preselected categories with a hand-crafted glossary summarized in Table 3. These features will be detailed shortly. Extracting textual features from a hand-crafted glossary is pragmatic and prudent since the development of automated tools (e.g., latent Dirichlet allocation) is not the focus of this paper. Furthermore, prior studies

Table 3 Text processing classes and subclasses

Class	Other subclasses	# Terms
Food	–	1629
Exercise	Fitness, hiking, sport, football	1046
Health	Wellness, dentistry, cancer	826
Celebration	Festival, wedding, christmas	980
Arts	Music, oscar, film, photo	853
Police	Army, law, detective, terrorist	1001
Religion	Faith, spirit, pope, passover	740
Science	Physics, chemistry, astronomy, biology	1136
Technology	Computer, internet, engineering	702
Business	Finance, economy, money, markets	1019
Positive	–	1385
Negative	–	3163

show that a manual textual feature extraction could result in superior performance when dealing with noisy texts [55], in our case search queries and tweets. Since there are only a few categories, one-hot encoding is used to denote the presence or absence of a certain category in user search queries. For each sleep or exercise event for a (user, day) tuple, we generate an information feature vector for that day by averaging over all such 12-dimension vectors for that user on that day before the actual sleep or exercise event. We account for the temporal nature of information features by expanding the location feature vector to 13 dimensions where the 13th coordinate is the average of all visit offset times of the user over all searches for that day before the sleep or exercise event takes place. Similarly, we add a 14th element to distinguish the case that the query does not fall into any of the above categories from the case that the user does not search at all. This is proverbially like utilizing the "Others" class in multi-class classification problems to account for patterns that do not match any of our pre-defined categories.

3.2.3 Social media features

Studies have shown that social data such as tweets are a rich source of information about people and their health trends [10]. They can also provide interesting insights into user behavior [56]. We do not have a way to identify a Microsoft Band user in the Twitter data and so do not use these data to directly build individual-level social features. Instead, we use location features to tag a user's event (sleep or exercise) with the appropriate textual features extracted from all tweets at that location, and use the English Twitter data to generate textual features for the locations users have ever visited.

Since Twitter posts and Bing queries are typically short and potentially noisy, following [55], we featurize this textual data by extracting features based on a hand-crafted glossary of topics. For terms related to food, we used the terms extracted in [57]. Positive and negative sentiments are also important in health analysis [58], and we use the glossary curated by [59] to extract sentiment from the contents. For the other categories, listed in "Aggregated Social" and "Information" portion of Table 1, we use contextual words for topical vocabulary games.[5] Each category captures words that appear in the associated context. For example, the business category has words commonly occurring in finance, economics, money, etc. Inspired by [60,61], the topics are hand-crafted based on subjects that might affect sleep and exercise such as fitness, travel, celebration. Table 3 summarizes the categories of our textual glossary. Note that there are 12 categories here. For each tweet/query, we count related terms from each category and normalize by query/tweet length and glossary size and average over all the query vectors of that day or tweets of that location. Similarly, for locations with no social features, we introduce a 13th feature. By using a glossary of topics or specified locations for extracting features, a natural question is whether we only detect things we know to look for *a priori*, like the effect of food consumption on exercise or effect of travel-related locations on sleep.

Here, we assume that the feature is extracted from the location, not the person. The hypothesis is that visiting a location with positive/negative emotion affects sleep and this is validated by the observed statistically significant correlation. We do not claim that we are considering the effect of the positive/negative sentiment of the user on their sleep/exercise quality.

We acknowledge that these features are noisy. For example, a visit to a location labeled as a food court is assumed to be associated with eating, but it could be due to using a restroom or going to a different store or office in the building. A related issue can occur for social features where the positive or negative feeling of an individual differs from the

[5] https://myvocabulary.com/.

Table 4 Normalizing constants for features and targets

Variable	Const	Variable	Const	Variable	Const	Variable	Const	Variable	Const
Age	11.5	Height	3.2	Weight	42	CalrsBrnd	459	SlpStrtHour	2.2
Steps taken	3914	Loc-Food	0.57	Loc-Retail	0.49	Loc-Health	0.26	Loc-Entertain	0.15
Loc-Banking	0.18	Loc-education	0.29	Loc-Sprots	0.19	Loc-Beauty	0.14	Loc-Travel	0.22
Loc-Service	0.29	Loc-Offset	4625	Soc-Food	0.05	Soc-Exercise	0.04	Soc-Health	0.03
Soc-Celebrate	0.05	Soc-Arts	0.03	Soc-Police	0.04	Soc-Religious	0.03	Soc-Science	0.02
Soc-Tech	0.03	Soc-Business	0.04	Soc-Positive	0.05	Soc-Negative	0.04	Soc-No-Twt	0.4
Inf-Food	0.12	Inf-Exercise	0.11	Inf-Health	0.09	Inf-Celebrate	0.11	Inf-Arts	0.08
Inf-Police	0.10	Inf-Religious	0.07	Inf-Science	0.08	Inf-Tech	0.12	Inf-Business	0.11
Inf-Positive	0.08	Inf-Negative	0.07	Inf-No-Bing	0.50	Inf-Offset	17,685	TimeToFallAsleep	791
NumOfWakeups	2.44	SleepEfficiency	6.05	RestHeartRate	3.65	ExerciseIntensity	2.89		

general public visiting a location. We believe that this uncertainty is reflected in the *p*-values and confidence intervals in regression. Furthermore, a small coefficient of a feature may be attributed to this noise. The more noisy the data, the less apparent the effects will be. Therefore, the results associated to these features (especially location and social features) are one-sided results, i.e., a statistically significant correlation (either positive or negative) shows a relation and is credible, however, lack of correlation may be due to noise. While we believe that the current features still lead to interpretable and significant results and are sufficient, we encourage follow-up work on de-noising, more feature engineering and curation to find yet more credible and meaningful results.

To be able to conduct cross-target and cross-feature analysis and comparisons, we standardize all feature and target values across the full dataset. They are normalized to have mean and standard deviation equal to zero and one, respectively. The normalizer coefficients can be found in Table 4. The regression coefficients can be multiplied by the values in this table to be transformed back to an interpretable scale in accordance with Table 1. We decided to transform the data for various reasons. First, recall that statistical significance tests assume that the modeling errors are uncorrelated and uniform, hence their variances do not vary with the effects being modeled. While least squares estimator (linear regression fitting) is unbiased in the presence of heteroscedasticity, it is inefficient because the true variance and covariance are underestimated. For example, in testing for differences between subpopulations (e.g., male and female in our case), standard tests assume that variances within groups are equal. Presence of heteroscedasticity entails the solution of ordinary least squares is not the "best linear unbiased estimator" and neither is its variance. Additionally, it provides us with a way to easily cross-compare the features and their contribution in a unified view given that we know their scaling and normalization factors. The same argument holds for the causal analysis even with higher priority as the confounder intervention in correlation is unobserved.

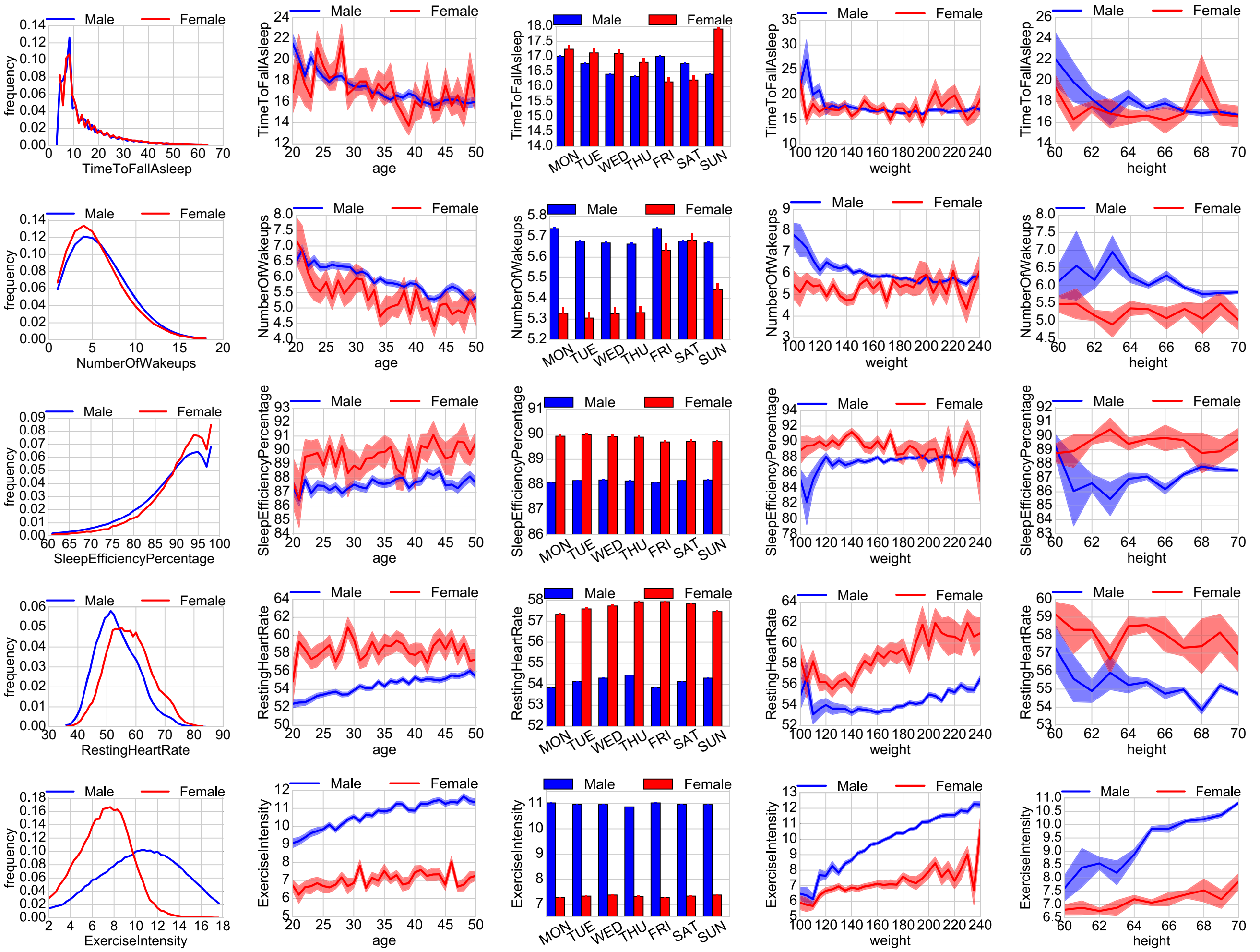
For the statistical inference, we used a simple linear regression model and fit the coefficients using ordinary least squares. Using *t*-test, a *p*-value is extracted to analyze the confidence interval for the inferred parameters. This provides us with a measure to express our certainty in the linear correlation observed between the targets and values. For the implementation details, we refer the reader to the standard scikit learn package in python for the regression analysis [62] that we leveraged in our research. We discuss findings and insights only on feature-target pairs that have a statistically significant correlation. In what follows, whenever we use the term "significance" it implies a statistical significance of $p \leq 0.05$. The hypothesis under test was if the feature has any linear relationship (either positive or negative) with the target variable or not. β is used to denote the regression coefficient corresponding to feature-target pair in the multiple regression fitted via ordinary least squares minimization.

4 Results

In this section, we first present the results of an exploratory analysis. We show how multiple regression results in some insights from data. Then, we apply a stratified propensity score analysis as a causal inference method to analyze this rich data set. Finally, we build predictive models and explore the ways that features can contribute to model performance.

4.1 Sleep and exercise analysis

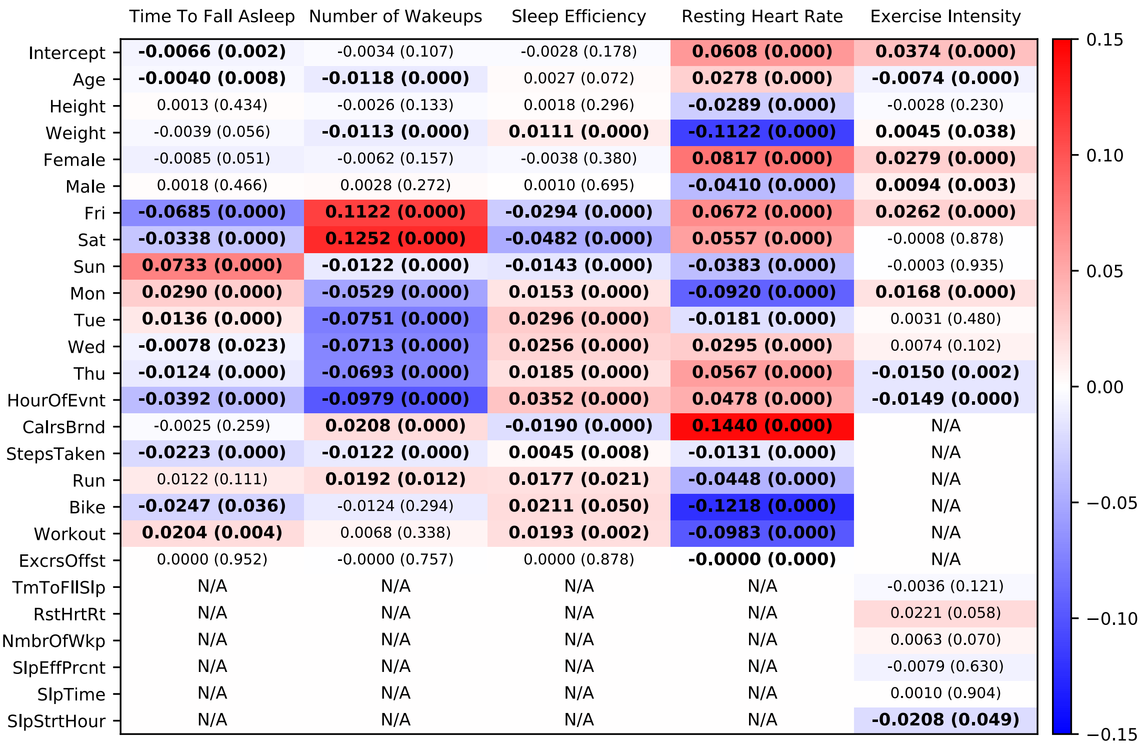
We focus our analysis on the relationship between people's activities throughout the day, including their location traces and web searches and the target variables capturing different aspects of sleep and exercise quality. Figure 1 shows the distribution of our target variables, as well as their cor-



(a) Distribution **(b)** Age **(c)** Day of week **(d)** Weight **(e)** Height

Fig. 1 Exploration of target variables with variation of basic information. Row 1: time to fall asleep; Row 2: number of wakeups; Row 3: sleep efficiency percentage; Row 4: resting heart rate; Row 5: exercise intensity. **a** Distribution, **b** age, **c** day of week, **d** weight, **e** height

relations with basic attributes of users and the day of the week. The influence of age and weight on sleep quality and exercise intensity is well-studied in the literature [45,46] and confirms our results. These figures show that the population sample of this study has similar properties as samples used in studies of the more traditional style (e.g., without social data). They also show the general trends and properties of these quantities at a glance and provide a quick way to get a sense of the data. As an example, let us consider Fig. 1c. Number of wakeups for females shows a significant increase during weekends compared to those for males. This alone may mean little while we study all aspect of this phenomena, but it can serve as an starting point for researchers to dive deeper into the underlying reasons. That additional research may lead to interesting and unknown facts about sleep habits of females and males on weekdays and weekends. Next, we will study correlation in a multiple linear regression models described in the previous section. The goal is to find the strength of the relation (coefficient) and a confidence score (p-value) to understand how features contribute to explaining the variability in the target variable. Figures 2 and 3 contain the full results of our regression analysis. Each entry in the matrix comes with the pair's coefficient followed with the p-value in parentheses. The statistically significant correlations are written in bold. Red and blue colors are used to indicate positive and negative correlations, respectively, while the strength of the color highlights the magnitude of the correlation. The matrix will serve as a reference for readers and researchers in this field. Analyzing all feature-target pairs is out of scope of the current study. The following contains a discussion on statistically significant correlations we found in our analysis and their connection to findings in the literature of sleep and exercise quality.



	Time To Fall Asleep	Number of Wakeups	Sleep Efficiency	Resting Heart Rate	Exercise Intensity
Intercept	**-0.0066 (0.002)**	-0.0034 (0.107)	-0.0028 (0.178)	**0.0608 (0.000)**	**0.0374 (0.000)**
Age	**-0.0040 (0.008)**	**-0.0118 (0.000)**	0.0027 (0.072)	**0.0278 (0.000)**	**-0.0074 (0.000)**
Height	0.0013 (0.434)	-0.0026 (0.133)	0.0018 (0.296)	**-0.0289 (0.000)**	-0.0028 (0.230)
Weight	-0.0039 (0.056)	**-0.0113 (0.000)**	**0.0111 (0.000)**	**-0.1122 (0.000)**	**0.0045 (0.038)**
Female	-0.0085 (0.051)	-0.0062 (0.157)	-0.0038 (0.380)	**0.0817 (0.000)**	**0.0279 (0.000)**
Male	0.0018 (0.466)	0.0028 (0.272)	0.0010 (0.695)	**-0.0410 (0.000)**	**0.0094 (0.003)**
Fri	**-0.0685 (0.000)**	**0.1122 (0.000)**	**-0.0294 (0.000)**	**0.0672 (0.000)**	**0.0262 (0.000)**
Sat	**-0.0338 (0.000)**	**0.1252 (0.000)**	**-0.0482 (0.000)**	**0.0557 (0.000)**	-0.0008 (0.878)
Sun	**0.0733 (0.000)**	**-0.0122 (0.000)**	**-0.0143 (0.000)**	**-0.0383 (0.000)**	-0.0003 (0.935)
Mon	**0.0290 (0.000)**	**-0.0529 (0.000)**	**0.0153 (0.000)**	**-0.0920 (0.000)**	**0.0168 (0.000)**
Tue	**0.0136 (0.000)**	**-0.0751 (0.000)**	**0.0296 (0.000)**	**-0.0181 (0.000)**	0.0031 (0.480)
Wed	**-0.0078 (0.023)**	**-0.0713 (0.000)**	**0.0256 (0.000)**	**0.0295 (0.000)**	0.0074 (0.102)
Thu	**-0.0124 (0.000)**	**-0.0693 (0.000)**	**0.0185 (0.000)**	**0.0567 (0.000)**	**-0.0150 (0.002)**
HourOfEvnt	**-0.0392 (0.000)**	**-0.0979 (0.000)**	**0.0352 (0.000)**	**0.0478 (0.000)**	**-0.0149 (0.000)**
ClrsBrnd	-0.0025 (0.259)	**0.0208 (0.000)**	**-0.0190 (0.000)**	**0.1440 (0.000)**	N/A
StepsTaken	**-0.0223 (0.000)**	**-0.0122 (0.000)**	**0.0045 (0.008)**	**-0.0131 (0.000)**	N/A
Run	0.0122 (0.111)	**0.0192 (0.012)**	**0.0177 (0.021)**	**-0.0448 (0.000)**	N/A
Bike	**-0.0247 (0.036)**	-0.0124 (0.294)	**0.0211 (0.050)**	**-0.1218 (0.000)**	N/A
Workout	**0.0204 (0.004)**	0.0068 (0.338)	**0.0193 (0.002)**	**-0.0983 (0.000)**	N/A
ExcrsOffst	0.0000 (0.952)	-0.0000 (0.757)	0.0000 (0.878)	**-0.0000 (0.000)**	N/A
TmToFllSlp	N/A	N/A	N/A	N/A	-0.0036 (0.121)
RstHrtRt	N/A	N/A	N/A	N/A	0.0221 (0.058)
NmbrOfWkp	N/A	N/A	N/A	N/A	0.0063 (0.070)
SlpEffPrcnt	N/A	N/A	N/A	N/A	-0.0079 (0.630)
SlpTime	N/A	N/A	N/A	N/A	0.0010 (0.904)
SlpStrtHour	N/A	N/A	N/A	N/A	**-0.0208 (0.049)**

Fig. 2 All coefficients and *p*-values: basic and daily features

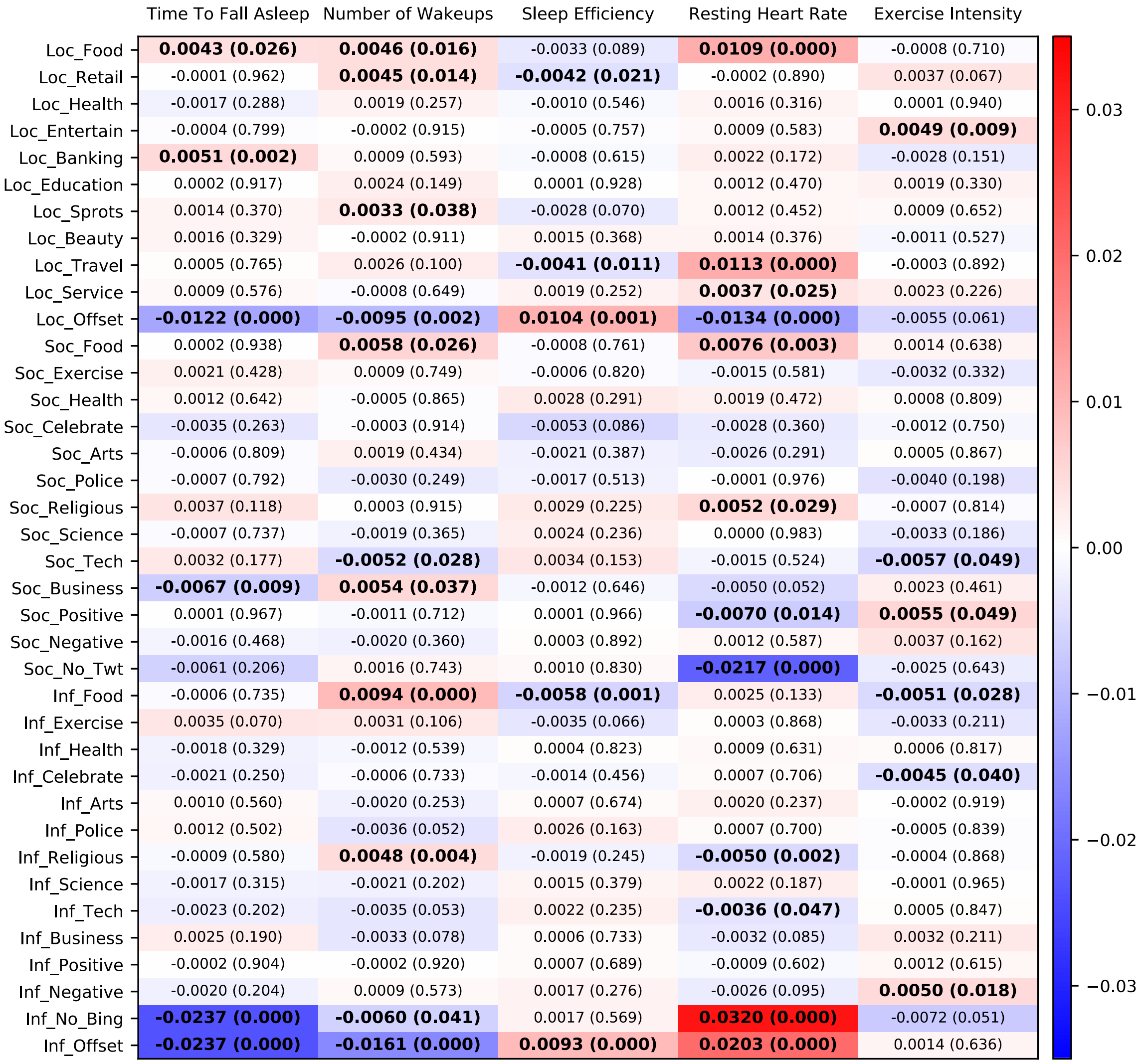
4.1.1 Time to fall asleep

Our analysis finds that exercise is usually correlated with a faster time to fall asleep that day, including steps ($\beta = -.022$) and bicycling ($\beta = -.024$), but not including gym workouts ($\beta = .02$). Note that coefficients are computed when all features are normalized to have mean 0 and variance 1. We find that people who visit food-related locations take longer to fall asleep ($\beta = .004$), indicating a relationship between restaurant food and sleep quality. More interestingly, perhaps, we find that those who visit banking-related locations take longer to fall asleep ($\beta = .005$), where bank visits may be acting as possible proxies for positive or negative life events causing stress (e.g., financial worries around life transitions). Burgard and colleagues [13] studied connections between common workplace experiences and sleep quality. Given the potential impact of home finances on sleep quality (similar to the effects of spousal-/partner-/child-related issues), they accounted for financial factors in their analysis and considered their interference with workplace issues affecting sleep.

Moreover, the closer a location visit is to bedtime, the longer it takes to fall asleep ($\beta = -.012$). Among the information features, we observe that users with no web searches before sleep usually take less time to fall asleep ($\beta = -0.023$), and the more time that elapses from the final search to bedtime, the quicker they fall asleep ($\beta = -0.23$). This is consistent with findings that the blue light emitted by devices disturbs sleep quality [63] and the importance of "winding down" before bedtime.

4.1.2 Number of wakeups

Our analysis finds that exercise has mixed effects on the number of wakeups. While taking more steps is correlated with fewer wakeups that day ($\beta = -.012$), running is correlated with more wakeups ($\beta = .019$) and, overall, burning more calories is correlated with more wakeups ($\beta = .020$). Interestingly, the medical and clinical literature has also reported mixed effects regarding the impact of exercise on sleep. While chronic exercise is believed to increase sleep quality [64], the acute and even sometimes chronic exercise has no consistent effect on sleep. The authors in [65] identified many factors contributing to the inconsistent effects of exercise on sleep. In their meta-analysis, they discussed moderator variables such as fitness of subjects, exercise heat load, duration and time of day, and subjects light exposure and sleep schedule. They concluded that studies that do not control for these factors will not give consistent results. Among people who visit food ($\beta = .0046$), retail ($\beta = .0045$), sports ($\beta = .0033$) locations, we observe more wakeups that day. This result is confirmed by previous work, which has shown that consuming food close to bedtime is negatively associated with sleep quality [12]. Furthermore, the



	Time To Fall Asleep	Number of Wakeups	Sleep Efficiency	Resting Heart Rate	Exercise Intensity
Loc_Food	**0.0043 (0.026)**	**0.0046 (0.016)**	-0.0033 (0.089)	**0.0109 (0.000)**	-0.0008 (0.710)
Loc_Retail	-0.0001 (0.962)	**0.0045 (0.014)**	**-0.0042 (0.021)**	-0.0002 (0.890)	0.0037 (0.067)
Loc_Health	-0.0017 (0.288)	0.0019 (0.257)	-0.0010 (0.546)	0.0016 (0.316)	0.0001 (0.940)
Loc_Entertain	-0.0004 (0.799)	-0.0002 (0.915)	-0.0005 (0.757)	0.0009 (0.583)	**0.0049 (0.009)**
Loc_Banking	**0.0051 (0.002)**	0.0009 (0.593)	-0.0008 (0.615)	0.0022 (0.172)	-0.0028 (0.151)
Loc_Education	0.0002 (0.917)	0.0024 (0.149)	0.0001 (0.928)	0.0012 (0.470)	0.0019 (0.330)
Loc_Sprots	0.0014 (0.370)	**0.0033 (0.038)**	-0.0028 (0.070)	0.0012 (0.452)	0.0009 (0.652)
Loc_Beauty	0.0016 (0.329)	-0.0002 (0.911)	0.0015 (0.368)	0.0014 (0.376)	-0.0011 (0.527)
Loc_Travel	0.0005 (0.765)	0.0026 (0.100)	**-0.0041 (0.011)**	**0.0113 (0.000)**	-0.0003 (0.892)
Loc_Service	0.0009 (0.576)	-0.0008 (0.649)	0.0019 (0.252)	**0.0037 (0.025)**	0.0023 (0.226)
Loc_Offset	**-0.0122 (0.000)**	**-0.0095 (0.002)**	**0.0104 (0.001)**	**-0.0134 (0.000)**	-0.0055 (0.061)
Soc_Food	0.0002 (0.938)	**0.0058 (0.026)**	-0.0008 (0.761)	**0.0076 (0.003)**	0.0014 (0.638)
Soc_Exercise	0.0021 (0.428)	0.0009 (0.749)	-0.0006 (0.820)	-0.0015 (0.581)	-0.0032 (0.332)
Soc_Health	0.0012 (0.642)	-0.0005 (0.865)	0.0028 (0.291)	0.0019 (0.472)	0.0008 (0.809)
Soc_Celebrate	-0.0035 (0.263)	-0.0003 (0.914)	-0.0053 (0.086)	-0.0028 (0.360)	-0.0012 (0.750)
Soc_Arts	-0.0006 (0.809)	0.0019 (0.434)	-0.0021 (0.387)	-0.0026 (0.291)	0.0005 (0.867)
Soc_Police	-0.0007 (0.792)	-0.0030 (0.249)	-0.0017 (0.513)	-0.0001 (0.976)	-0.0040 (0.198)
Soc_Religious	0.0037 (0.118)	0.0003 (0.915)	0.0029 (0.225)	**0.0052 (0.029)**	-0.0007 (0.814)
Soc_Science	-0.0007 (0.737)	-0.0019 (0.365)	0.0024 (0.236)	0.0000 (0.983)	-0.0033 (0.186)
Soc_Tech	0.0032 (0.177)	**-0.0052 (0.028)**	0.0034 (0.153)	-0.0015 (0.524)	**-0.0057 (0.049)**
Soc_Business	**-0.0067 (0.009)**	**0.0054 (0.037)**	-0.0012 (0.646)	-0.0050 (0.052)	0.0023 (0.461)
Soc_Positive	0.0001 (0.967)	-0.0011 (0.712)	0.0001 (0.966)	**-0.0070 (0.014)**	**0.0055 (0.049)**
Soc_Negative	-0.0016 (0.468)	-0.0020 (0.360)	0.0003 (0.892)	0.0012 (0.587)	0.0037 (0.162)
Soc_No_Twt	-0.0061 (0.206)	0.0016 (0.743)	0.0010 (0.830)	**-0.0217 (0.000)**	-0.0025 (0.643)
Inf_Food	-0.0006 (0.735)	**0.0094 (0.000)**	**-0.0058 (0.001)**	0.0025 (0.133)	**-0.0051 (0.028)**
Inf_Exercise	0.0035 (0.070)	0.0031 (0.106)	-0.0035 (0.066)	0.0003 (0.868)	-0.0033 (0.211)
Inf_Health	-0.0018 (0.329)	-0.0012 (0.539)	0.0004 (0.823)	0.0009 (0.631)	0.0006 (0.817)
Inf_Celebrate	-0.0021 (0.250)	-0.0006 (0.733)	-0.0014 (0.456)	0.0007 (0.706)	**-0.0045 (0.040)**
Inf_Arts	0.0010 (0.560)	-0.0020 (0.253)	0.0007 (0.674)	0.0020 (0.237)	-0.0002 (0.919)
Inf_Police	0.0012 (0.502)	-0.0036 (0.052)	0.0026 (0.163)	0.0007 (0.700)	-0.0005 (0.839)
Inf_Religious	-0.0009 (0.580)	**0.0048 (0.004)**	-0.0019 (0.245)	**-0.0050 (0.002)**	-0.0004 (0.868)
Inf_Science	-0.0017 (0.315)	-0.0021 (0.202)	0.0015 (0.379)	0.0022 (0.187)	-0.0001 (0.965)
Inf_Tech	-0.0023 (0.202)	-0.0035 (0.053)	0.0022 (0.235)	**-0.0036 (0.047)**	0.0005 (0.847)
Inf_Business	0.0025 (0.190)	-0.0033 (0.078)	0.0006 (0.733)	-0.0032 (0.085)	0.0032 (0.211)
Inf_Positive	-0.0002 (0.904)	-0.0002 (0.920)	0.0007 (0.689)	-0.0009 (0.602)	0.0012 (0.615)
Inf_Negative	-0.0020 (0.204)	0.0009 (0.573)	0.0017 (0.276)	-0.0026 (0.095)	**0.0050 (0.018)**
Inf_No_Bing	**-0.0237 (0.000)**	**-0.0060 (0.041)**	0.0017 (0.569)	**0.0320 (0.000)**	-0.0072 (0.051)
Inf_Offset	**-0.0237 (0.000)**	**-0.0161 (0.000)**	**0.0093 (0.000)**	**0.0203 (0.000)**	0.0014 (0.636)

Fig. 3 All coefficients and *p*-values: location, social, and information features

closer a location visit is to bedtime, the more wakeups a person experiences ($\beta = -.009$), congruent with our findings for time to fall asleep. Our aggregated social features reinforce our location visit features with a positive correlation between visiting locations with food and business terms ($\beta = .006$ and $\beta = .005$). In our informational feature set, we see again that food-related web searches are correlated with more wakeups ($\beta = .009$). We also see, however, that searches for religious and spiritual terms are correlated with more wakeups ($\beta = .005$). One possible explanation is that people are performing such searches more frequently when they are experiencing some stressful events, such as a death or sickness in the family; or that such searches may be correlated with religious activities such as waking up for daily prayers or church attendance. As before, not searching before bedtime, and increased time between the last web search and bedtime has a negative correlation with the number of wakeups ($\beta = -.006$ and $\beta = -.016$).

4.1.3 Sleep efficiency percentage

We find that exercise has a positive correlation with more efficient sleep, including taking more steps ($\beta = .005$), running ($\beta = .017$), biking ($\beta = .021$) and gym workouts ($\beta = .019$). Clinical studies (e.g., [64]) have shown that people who exercise will sleep more, which leads to an increase in the proportion of being asleep over the time in bed. We find that people who visit retail stores ($\beta = -.004$) and travel locations (e.g., airports) ($\beta = -.004$) have less efficient sleep. Researchers have found physiological evidence on sleep difficulty when traveling and visiting new locations [14]. The closer a location visit is to bedtime, the lower the sleep efficiency ($\beta = .01$). Food-related searches are correlated with poorer sleep efficiency ($\beta = -.006$), and offset between the last search and bedtime is correlated with more efficient sleep ($\beta = .009$).

4.1.4 Resting heart rate

In our data, there is a negative correlation between heart rate and exercise including running ($\beta = -.04$), biking ($\beta = -.12$), and gym workout ($\beta = -.09$), and steps taken ($\beta = -.01$), as reported in the literature [66]: those who take more steps, run, bike, or workout have better heart rates. We find positive correlation of resting heart rate with people who visit food ($\beta = .01$), travel ($\beta = .011$), and service oriented locations ($\beta = .003$), possibly due to restaurant food and disruptions associated with travel.

Among our aggregated social features, we find that people who visit locations with food-related terms have higher resting heart rates ($\beta = .007$), as do people who visit locations with religion-related terms ($\beta = .005$). Interestingly, people who visit locations with positive-sentiment terms are more likely to have a lower resting heart rate ($\beta = . - 007$). In our data, we also see that people who search for religion ($\beta = -.005$) and technology-related ($\beta = -.003$) terms are associated with lower resting heart rates. Understanding the underlying causes of this correlation will require further study.

4.1.5 Exercise intensity

The quality of the previous night's sleep has no statistically significant effect on exercise intensity, with the exception of the hour of sleep ($\beta = -0.02$): people with late bedtimes are less likely to have intensive exercise the following day. Tiredness has been shown to hinder serious exercise [67]. Location features contribute to exercise intensity significantly only via entertainment features with positive correlation ($\beta = .005$): those who visit entertainment and recreation centers are the ones who exercise more intensively. For social features, we observe positive and significant correlation of positive sentiments with exercise intensity ($\beta = .005$): those who visit a location with generally positive sentiment terms usually have a more intensive exercise afterward. This is in accordance with [68], which found a significant positive correlation between positive emotions and amount of physical activity. Among our informational features, we observe that the amount of food ($\beta = -.005$) and celebration ($\beta = -.004$) searches are negatively correlated with exercise intensity.

Overall, our findings indicate that web search traces and location traces do capture information about individual behaviors that are relevant to sleep quality and exercise intensity. While it is generally accepted that a person's lifestyle (food habits, for example) and disruptive events (travel, financial worries, celebrations) can have a significant effect on that person's health, today's automated fitness/sleep devices do not capture and take into account such information. Our analysis demonstrates that web search and location traces provide a possible avenue for capturing such information, providing better insights for individuals, health practitioners, and device manufacturers.

4.2 Causal inference analysis

The regression coefficients presented in Sect. 4.1 capture correlational relationships between our features and target variables. To more strongly establish possible causal relationships, we use a stratified propensity score analysis, one of a family of conditioning inference methods in the potential outcomes framework [69,70]. While we do not believe we can achieve the ideal identification of causal relationships, these methods are effective at reducing biases due to observed confounding factors.

Ideally, to determine the effect of some action, we would be able to observe and compare two potential outcomes for an individual i: an outcome $Y_i^{T=1}$, representing the outcome after a person takes a target action T, and another $Y_i^{T=0}$, representing the outcome after the same person in an identical context does not take the action. The causal effect of T is then $Y_i^{T=1} - Y_i^{T=0}$. Of course, it is impossible to observe both $Y_i^{T=1}$ and $Y_i^{T=0}$. Thus, the problem of causal inference is a problem of missing data, and causal inference techniques attempt to address it by estimating the missing counterfactual outcomes based on the outcomes of other, similar individuals. The challenge is to estimate these missing outcomes correctly despite the potentially confounding presence of covariates that influence both treatment likelihood and outcomes in observational (non-experimental) data.

Stratified propensity score analysis attempts to accomplish this through post-hoc identification of comparable subgroups within the observational data. Conceptually, the idea is to find pairs (generalizing to groups) of individuals in the observational data who are statistically similar to one another but where one has received a treatment and the other has not. Individuals with similar propensity scores are grouped into strata and, in aggregate, these individuals are likely to have similar covariates, allowing us to isolate and estimate the effects of the treatment itself [21].

The results of our re-analysis of the key relationships from Sect. 4.1 over 4 months of data are shown in Table 5. The set of covariates (features) are the same as in Sect. 4.1. We see, for example, that exercise reduced the time to fall asleep for individuals. The relative treatment effect (RTE) of 0.92 means that for that group of individuals, exercising led to a time to fall asleep that was 92% of the time taken by those in the control group; therefore, the treatment had an impact on this outcome variable. Other rows in this table should be interpreted similarly.

We note that we show, for brevity, the RTE for a few pairs of targets and treatments; however, our analysis confirms the correlations shown by the regression coefficients for all relevant target-feature combinations studied.

Table 5 Causal inference analysis results

Outcome	Treatment (T)	# Users	RTE	Outcome ($T = 1$)	Outcome ($T = 0$)
Time to fall asleep	Exercise	21,023	0.92	738.3 s	802.5 s
Number of wakeups	Exercise	21,023	0.94	5.2	5.5
Exercise intensity	Good sleep	22,323	1.01	15.6 cal/m	14.95 cal/m
Time to fall asleep	Web search	19,768	1.07	729.8 s	682.1 s
Time to fall asleep	Banking location	19,987	1.05	741.8 s	706.5 s
Exercise intensity	Food location	19,801	0.99	15.45 cal/m	15.61 cal/m

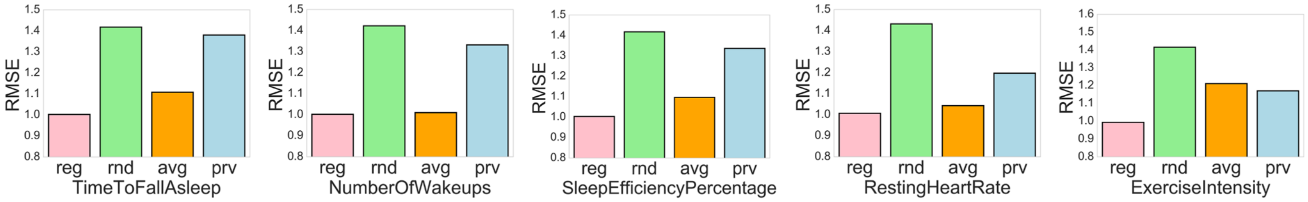


Fig. 4 Predictive performance of linear regression model (reg) compared to three baselines; random (rnd), the previous target value of the user (prv), the average of all past target values of the user (avg)

4.3 Sleep and exercise prediction

We were interested in whether the signals described thus far in the paper could be used to develop predictive models for sleep and exercise, for applications including early intervention. We trained linear regression models to predict sleep or exercise quality targets and investigate factors that affect the accuracy of our predictions. All the following experiments are run 10 times and the root mean squared error (RMSE) is used to report the accuracy of the prediction. Three baselines are used to compare the predictive performance of the linear regression model. Given target values y_t (where t is a discrete variable that indexes time) and corresponding predictions y_t^{pred}, for baseline `rnd` we have: $y_t^{\text{pred}} \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are the target's mean and variance, respectively. Baseline `prv` $y_t^{\text{pred}} = y_{t-1}$ is basically the most recent value, and finally we set $y_t^{\text{pred}} = \overline{y_{<t}}$ for baseline `avg`, where the overline denotes an average over all past target values. To train the linear regression model on 500 and 300 K randomly selected sleep and exercise events, respectively, based on a variety of features as predictors and test it on the remaining records which have been held out.

4.3.1 Model performance

Figure 4 compares the performance of the regression technique (`reg`) to the baselines `rnd`, `prv`, and `avg`. The regression model `reg` has the best performance with `avg` being the second best; moreover, `avg` is always better than `prv` except when the target is exercise intensity. It seems that exercise intensity is more temporally local than other targets: The previous day's exercise intensity is a better predictor than the average value of past exercises of the user. This is especially true when users gradually improve their exercise. As expected, `rnd` performs worst. Translating the numbers back to the unnormalized measures, the RMSE of our linear regression for time to fall asleep is around 13 min, for number of wakeups the RMSE is 2.4, for sleep efficiency percentage is almost 6, for resting heart rate it is 3.6 beats per second, and finally for exercise intensity RMSE is around 2.8 calories per minute. Linear regression is only used as a demonstration and investigating more advanced regression methods such as support vector and decision-tree regression remain as future work.

4.3.2 Predictive features

Figure 5 demonstrates the predictive performance of the linear regression model by incorporating different feature sets. Not surprisingly, *all* consistently performs the best; daily feature *act* is the second best. The considerable gap between these two and the rest hints at the fact that most daily features (day of week, hour of event, steps taken, etc.) are good predictors of our target variables. In the event that data collection is limited and not all features are available, good model performance can be achieved by priority daily and location features. This result could be of direct interest to companies investing on health trackers to reduce the cost of the wearable devices by incorporating the necessary sensors delivering the required accuracy at a minimal cost.

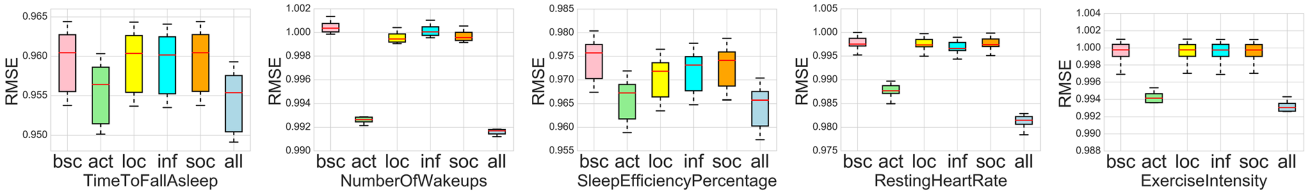


Fig. 5 Predictive performance of regression model by incorporating different feature sets: basic features (bsc) only, daily only (act), location only (loc), information feature only (inf), social feature only (soc), and all 5 feature sets together (all)

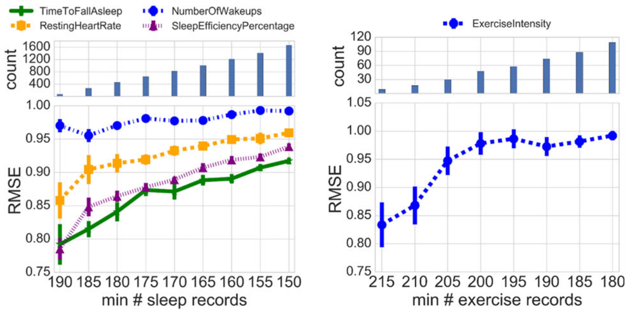


Fig. 6 Performance of linear model

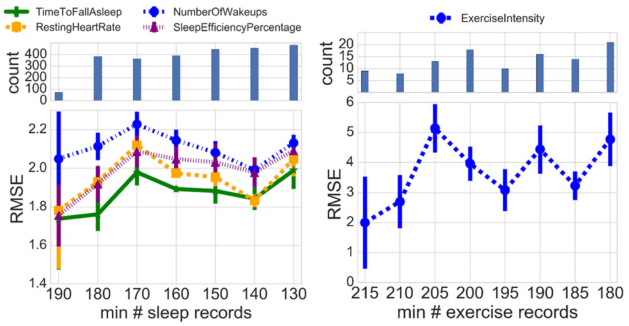


Fig. 7 Performance of personalized linear model when we filter out users with fewer (than a threshold) sleep/exercise events plotted with respect to varying threshold

4.3.3 Individual-data scale and prediction quality

One might expect that if we had more event records per user then we might better predict sleep or exercise quality. To test this, we first select a threshold for the number of sleep or exercise events, and exclude users with fewer events. Then we trained on a randomly chosen 80% of these data and test on the remaining 20%. We repeat this procedure 10 times and take the average.

Figure 6 demonstrates how the performance improves as the threshold increases. Time to fall asleep, efficiency percentage, and exercise intensity show improvement for prediction. However, we find that the number of wakeups and resting heart rate are not easily predictable even with increased number of data. This is an actionable insight for health application developers. Users and health practitioners should be careful about using the prediction results on these two measures relying on the abundance of historical data.

4.3.4 Personalized prediction models

To investigate personalized predictive models, we again put a lower bound on the number of sleep or exercise events, and filter users and data with fewer records. A per-person predictive model is built based on randomly selected 80% training data and is evaluated on the remaining 20% of the data, which was held out for testing. This procedure is repeated 10 times and the mean and standard deviation are plotted with respect to the threshold (see Fig. 7). We find that performance degrades with the application of personalized models. The explanation is quite simple: a limited number of events is insufficient to build a reliable predictive model.

The combination of these two experiments suggests that an ideal model for predicting sleep/exercise would be hierarchical; the top layer contains population parameters. Then, for each user, we build a personalized model. This is an interesting direction for future work. The implication for our three potential target audiences is that prediction of sleep quality and exercise intensity requires leveraging input from population behavior.

4.3.5 Early prediction

Practically speaking, one might be curious how the model works when we predict based on information available a few hours before sleep/exercise. If we only predict the event quality right before the event is happening (when we observe the complete data), then it might be too late to intervene and improve the quality. If we are able to predict a bad sleep in advance, e.g., a couple of hours before sleep, a system could intervene and recommend preventative action. Recall that finding the best intervention is not the purpose of this paper and is another interesting line of future work. A second compelling reason for determining the performance of models that have access only to “stale” information is that in most practical production systems, there is a time-lag between the actual event and when it becomes available for use in production systems.

With the above question and the two motivations in mind, we design the following experiment: We pick a time which the event is going to be predicted that soon and call it *predic-*

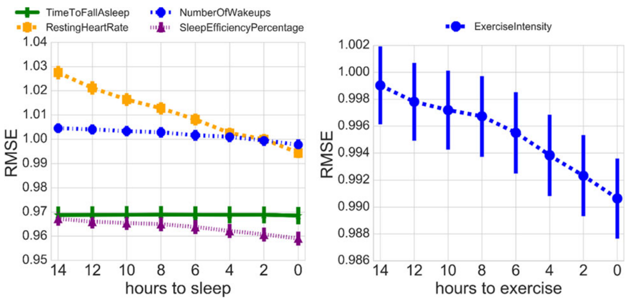


Fig. 8 Effect of prediction lead time on performance

tion lead time. We train the model based on randomly selected 500 and 300 K events for sleep and exercise, respectively. Then for the remaining test events, we use the information available only up to that threshold before the event and evaluate the performance. This procedure is run 10 times with random training/testing division and the results are reported in Fig. 8 by varying the lead time from 14 h until the moment of the event.

We observe that the performance of resting heart rate and exercise intensity is dependent on the lead time perhaps because they are local in time, e.g., the occurrence of a simple event, such as receiving bad news, consuming an alchoholic drink, or late-night exercise, might increase it easily. The performance of other tasks is affected to a lesser extent. Predictions made a couple of hours before are still reliable and could provide helpful information. However, for early prediction of resting heart rate and exercise intensity necessary precautions need to be devised considering the accuracy required by practitioners or the device.

5 Conclusion

In this study, we have demonstrated how digital traces of individual behaviors and lifestyles—captured through web search logs, digital assistants, and social media that are not conventionally associated with health studies—can be linked with fitness device data. The result is a new source of quantitative insights into the links between individual behaviors, habits and stress factors and individuals' sleep and exercise quality. In this section, we discuss the implications of our work for individuals, health practitioners, and health tracking systems. We also discuss the limitations of the current study and finally suggest some directions for future work.

6 Limitations

We acknowledge several limitations. The characteristics of the study participants, e.g., those who can afford to own and consistently wear a fitness band limit the generalizability of our findings. Furthermore, as discussed in the data section, most of our users are male and the data are collected during an 8-month interval which excludes the summer. We acknowledge that this limitation might bias the generalizablity of the results. Nonetheless, our main claim, demonstrating the advantages of data triangulation with online and social data sources, remains valid. Besides the detailed analysis of the causal inference section, the cross-sectional nature and the regression analysis of the study precludes conclusions about the causal relationship between sleep and exercise quality to the contextual information. Furthermore, we used linear regression to obtain early insights into data, which is a simple method unable to capture complex, nonlinear relationships that may exist.

Also, it is worth discussing the way we treat outliers and prune the dataset. The outliers removed were not simply extreme points but rather were clearly wrong input values or conceptually not sleep or exercise. For example, a sleep under 0.5 h is not regarded as a sleep event and considering it biases the outcome for short sleep durations (e.g., < 5 h). As another example, there are records in the data of people reporting an age of 200, which are clearly erroneous inputs. Assuming this to be the result of an inadvertent typing error, including such an example in our training data simply adds noise to the age variable. We refer the reader to the variance of the y-values for low and high x-values in Fig. 1. Such variances make training data noisy, and result in poor predictions from our models. Filtering these erroneous examples is common practice in the literature for sleep analysis (c.f. [34]).

In our analysis, we relied on the sleep detection of the Band device, which utilizes proprietary algorithms to detect sleep. These algorithms are not publicly available, so they, or validation studies thereof, cannot be cited in this paper. We acknowledge that this limits the generalizablity and the impact of the work.

7 Discussion

The potential audience of the current work is three groups: individuals, health practitioners and clinicians, and developers of health tracking systems and wearable device designers.

Individuals can get preliminary insight into how their sleep and exercise quality are related to basic features. There are many such insights. For example, it may help them to understand how biking or dining out is correlated with their sleep habits. Interest in celebrations, festivals, carnivals, etc. inferred from search activity suggests that people might exercise less efficiently. The scale of our study, both in terms of the number of users and the fmarrnumber of sleep/exercise events, allows health practitioners to perform reliable analysis resulting in credible insights. The predictive analysis makes it possible for the practitioners not only to forecast patients' sleep and exercise quality, but also combined with causal analysis it allows them to design effective, timely interventions to help their patients.

The third group which is possibly impacted by the results of our paper is the group of manufacturers and developers of health tracking devices and systems. Combining data to analyze and explore target variables show how the quality measures are distributed and how different features contribute to them, e.g., if an application provides a report on exercise intensity of a married couple, it should consider the differences that arise due to gender that emerge clearly from our data analysis. The report should consider these important factors, e.g., by normalizing individual measures based on statistics of the population. The other important area that applications can enter is trying to predict the quality of sleep/exercise. The data triangulation equips them with many interesting features with interpretable relationships to the target variable. The application can alert the user or their care team to take necessary action. Coupled with our causal inference, the application can even suggest preventive actions that lead to better overall outcomes. Additionally, our framework helps developers determine how to increase prediction accuracy and how early they can predict low-quality exercise or sleep and potentially intervene. Furthermore, we showed that for some quality measures, the abundance of historical data per user is not all that helpful, but for others, the accuracy of prediction can be significantly improved by collecting more data from users. The accuracy of regression models derived here serves as a benchmark for building future learning models with desired accuracies. Conversely, our analysis indicates that daily and location features are most relevant for prediction and should be considered if sensor usage is limited. Finally, our findings call for manufacturers to incorporate information and location traces in their health trackers and analyzers to better serve individuals' health needs. As people's interactions with mobile devices and online services continue to capture a more comprehensive view of their environment and actions, we believe that there is a significant opportunity to link such digital traces to health and fitness data and draw deep insights on the health influence of behaviors, habits, and stressors. Immediate future work includes more granular modeling of individual behaviors and environmental factors.

Compliance with ethical standards

Conflict of interest This work is funded 100% through the authors employment (a full-time internship for MF and full-time employment for the other authors) at Microsoft.

Research involving Human Participants All user identifying information was anonymized. We did not examine search queries with personally-identifiable information or other sensitive information. All data access and analysis performed for this research was done in accordance with the published end-user license agreement, which was worded as follows: "By connecting to Microsoft Health you agree to allow Microsoft to share your data between Cortana and Microsoft Health, to provide valuable personal insights and recommendations to help you reach your fitness and wellness goals." Visits to businesses were logged by Cortana to offer local services and is agreed to by users. Twitter data were not connected to specific users, but rather was based on publicly available tweets and were aggregated across many users who visited the business location. Our work was conducted offline, on data collected to support existing business operations, and did not influence the user experience. All data were anonymized and deidentified prior to analyses. Each user was represented by an anonymous identifier. We filtered search queries to only those matching a whitelist of keywords relevant to our study. The Ethics Advisory Committee at Microsoft Research considers these precautions sufficient for triggering the Common Rule, exempting this work from detailed ethics review.

Informed consent Our data were collected between August 2015 and April 2016 and from individuals who agreed to link their Cortana data and Microsoft Health data (including Band device data) for use in generating additional insights or recommendations about their sleep or activity.

References

1. Whelton, S., Chin, A., Xin, X., He, J.: Effect of aerobic exercise on blood pressure: a meta-analysis of randomized, controlled trials. Ann. Intern. Med. **136**, 493–503 (2002)
2. Petruzzello, S., Landers, D., Kubitz, A., Salazar, W.: A meta-analysis on the anxiety-reducing effects of acute and chronic exercise. Sports Med. **11**, 143–182 (1991)
3. Cappuccio, F.P., D'Elia, L., Strazzullo, P., Miller, M.A.: Sleep duration and all-cause mortality: a systematic review and meta-analysis of prospective studies. Sleep **33**, 585–592 (2010)
4. Reed, J., Ones, D.: The effect of acute aerobic exercise on positive activated affect: a meta-analysis. Psychol. Sport Exerc. **7**, 477–514 (2006)
5. Fortier, E., Beaulieu, S., Ivers, H., Morin, C.: Insomnia and daytime cognitive performance: a meta-analysis. Sleep Med. Rev. **16**, 83–94 (2012)
6. Rosekind, M., Gregory, K., Mallis, M., Brandt, S., Seal, B., Lerner, D.: The cost of poor sleep: workplace productivity loss and associated costs. J. Occup. Environ. Med. **52**, 91–98 (2010)
7. Pilcher, J., Huffcutt, A.: Effects of sleep deprivation on performance: a meta-analysis. Sleep **19**, 318–326 (1996)
8. Fox, K.R.: The influence of physical activity on mental well-being. Public Health Nutr. **2**(3a), 411–418 (1999)
9. Standage, M., Gillison, F., Ntoumanis, N., Treasure, D.: Predicting students physical activity and health-related well-being: a prospective cross-domain investigation of motivation across school physical education and exercise settings. J. Sport Exerc. Psychol. **34**, 37–60 (2012)
10. Fernández-Luque, L., Bau, T.: Health and social media: perfect storm of information. Healthc. Inform. Res. **21**(2), 67–73 (2015)
11. Culotta, A.: Estimating county health statistics with twitter. In: SIGCHI (2014)
12. Crispim, C., Zimberg, I., Diniz, R., Tufik, S., Mello, M.: Relationship between food intake and sleep pattern in healthy individuals. J. Clin. Sleep Med. **7**, 659 (2011)
13. Burgard, S., Ailshire, J.: Putting work to bed: stressful experiences on the job and sleep quality. J. Health Soc. Behav. **50**, 476–492 (2009)

14. Tamaki, M., Bang, J., Watanabe, T., Sasaki, Y.: Night watch in one brain hemisphere during sleep associated with the first-night effect in humans. Curr. Biol. **26**, 1190–1194 (2016)
15. Santillana, M., Nguyen, A., Dredze, M., Paul, M., Nsoesie, E., Brownstein, J.: Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput. Biol. **11**, e1004513 (2015)
16. Smith, M., Wegener, S.: Measures of sleep: the insomnia severity index, medical outcomes study (mos) sleep scale, pittsburgh sleep diary (psd), and pittsburgh sleep quality index (psqi). Arthritis Care Res. **49**, S184–S196 (2003)
17. Harvey, A.G., Stinson, K., Whitaker, K.L., Moskovitz, D., Virk, H.: The subjective meaning of sleep quality: a comparison of individuals with and without insomnia. Sleep **31**(3), 383 (2008)
18. Schutte, S., Broch, L., Buysse, D., Sateia, M.: Clinical guideline for the evaluation and management of chronic insomnia in adults. J. Clin. Sleep Med. **4**, 487 (2008)
19. American College of Sports Medicine et al.: ACSM's Guidelines for Exercise Testing and Prescription. Lippincott Williams & Wilkins (2013)
20. Waldeck, M.R., Lambert, M.I.: Heart rate during sleep: implications for monitoring training status. J. Sports Sci. Med. **2**(4), 133 (2003)
21. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
22. Chen, Z., Lin, M., Chen, F., Lane, N.D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., Campbell, A.T.: Unobtrusive sleep monitoring using smartphones. In: Pervasive Health (2013)
23. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: UBICOMP, pp. 3–14 (2014)
24. Gu, W., Yang, Z., Shangguan, L., Sun, W., Jin, K., Liu, Y.: Intelligent sleep stage mining service with smartphones. In: UBICOMP (2014)
25. Hao, T., Xing, G., Zhou, G.: iSleep: unobtrusive sleep quality monitoring using smartphones. In: SenSys (2013)
26. Gu, W., Shangguan, L., Yang, Z., Liu, Y.: Sleep hunter: towards fine grained sleep stage tracking with smartphones. IEEE Trans. Mob. Comput. **15**, 1514–1527 (2016)
27. Pernek, I., Kurillo, G., Stiglic, G., Bajcsy, R.: Recognizing the intensity of strength training exercises with wearable sensors. J. Biomed. Inf. **58**, 145–155 (2015)
28. Spina, G., Huang, G., Vaes, A., Spruit, M., Amft, O.: COPDTrainer: a smartphone-based motion rehabilitation training system with real-time acoustic feedback. In: UBICOMP, pp. 597–606 (2013)
29. Bai, Y., Xu, B., Ma, Y., Sun, G., Zhao, Y.: Will you have a good sleep tonight?: sleep quality prediction with mobile phone. In: BODYNETS (2012)
30. Min, J., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.: Toss'n'turn: smartphone as sleep and sleep quality detector. In: SIGCHI (2014)
31. Jayarajah, K., Radhakrishnan, M., Hoi, S., Misra, A.: Candy crushing your sleep. In: UBICOMP (2015)
32. Nguyen, A., Alqurashi, R., Halbower, A.C., Vu, T.: mSleep-Watcher: Why didn't i sleep well?. In: MCSE (2015)
33. Krishna, A., Mallick, M., Mitra, B.: Sleepsensei: an automated sleep quality monitor and sleep duration estimator. In: IoT of Health 2016 (2016)
34. Akbar, F., Weber, I.: # Sleep_as_android: feasibility of using sleep logs on twitter for sleep studies. In: ICHI (2016)
35. Wu, K., Ma, J., Zhumin, C., Ren, P.: Sleep quality evaluation of active microblog users. In: Asia-Pacific Web Conference (2015)
36. Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., Lawson, S: I can't get no sleep: discussing# insomnia on twitter. In: SIGCHI (2012)
37. Peng, X., Luo, J., Glenn, C., Zhan, J., Liu, Y.: Large-scale sleep condition analysis using selfies from social media. arXiv:1704.06853 (2017)
38. Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., Arora, T., Taheri, S.: Sleep quality prediction from wearable data using deep learning. JMIR Mhealth Uhealth **4**, e125 (2016)
39. Lauderdale, D.S., Knutson, K.L., Yan, L., Liu, K., Rathouz, P.J.: Self-reported and measured sleep duration: how similar are they? Epidemiology **19**, 838–845 (2008)
40. Natale, V., Léger, D., Bayon, V., Erbacci, A., Tonetti, L., Fabbri, M., Martoni, M.: The consensus sleep diary: quantitative criteria for primary insomnia diagnosis. Psychosom. Med. **77**(4), 413–418 (2015)
41. Lineberger, M.D., Carney, C.E., Edinger, J.D., Means, M.K.: Defining insomnia: quantitative criteria for insomnia severity and frequency. Sleep **29**(4), 479–485 (2006)
42. Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., Pollak, C.: The role of actigraphy in the study of sleep and circadian rhythms. Sleep **26**, 342–392 (2003)
43. Walch, O.J., Cochran, A., Forger, D.B.: A global quantification of normal sleep schedules using smartphone data. Sci. Adv. **2**, e1501705 (2016)
44. Althoff, T., Horvitz, E., White, R.W., Zeitzer. J.: Population-scale study of sleep and performance. In: WWW (2017)
45. Vargas, P., Flores, M., Robles, E.: Sleep quality and body mass index in college students: the role of sleep disturbances. J. Am. College Health **62**, 535–541 (2014)
46. Weeks, D., Borrousch, S., Bowen, A., Hepler, L., Sandau, A., Slevin, F.: The influence of age and gender of an exercise model on self-efficacy and quality of therapeutic exercise performance in the elderly. Physiother. Theory Pract. **21**, 137–146 (2005)
47. Dearman, D., Sohn, T., Truong, K.N.: Opportunities exist: continuous discovery of places to perform activities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2429–2438. ACM (2011)
48. Benetka, J.R., Balog, K., Nørvåg, K.: Anticipating information needs based on check-in activity. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 41–50. ACM (2017)
49. Iachello, G., Smith, I., Consolvo, S., Abowd, G.D., Hughes, J., Howard, J., Potter, F., Scott, J., Sohn, T., Hightower, J., et al.: Control, deception, and communication: evaluating the deployment of a location-enhanced messaging service. In: International Conference on Ubiquitous Computing, pp. 213–231. Springer (2005)
50. Yang, D., Zhang, D., Zheng, V.W., Yu, Z.: Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Trans. Syst. Man Cybern. Syst. **45**(1), 129–142 (2015)
51. Dearman, D., Truong, K.N.: Identifying the activities supported by locations with community-authored content. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, pp. 23–32. ACM (2010)
52. Hossain, N., Hu, T., Feizi, R., White, A.M., Luo, J., Kautz, H.: Inferring fine-grained details on user activities and home location from social media: detecting drinking-while-tweeting patterns in communities. arXiv:1603.03181 (2016)
53. White, R.: Beliefs and biases in web search. In: SIGIR (2013)
54. White, R.W., Horvitz, E.: Studies of the onset and persistence of medical concerns in search logs. In: SIGIR, pp. 265–274 (2012)
55. Stubbe, A., Ringlstetter, C., Schulz, K.U.: Genre as noise: noise in genre. Int. J. Doc. Anal. Recognit. (IJDAR) **10**, 199–209 (2007)

56. Kıcıman, E.: OMG, i have to tweet that! a study of factors that influence tweet rates. In: AAAI ICWSM (2012)
57. De Choudhury, M., Sharma, S., Kiciman, E.: Characterizing dietary choices, nutrition, and language in food deserts via social media. In: CSCW (2016)
58. Salathé, M., Vu, D., Khandelwal, S., Hunter, D.: The dynamics of health behavior sentiments on a large online social network. EPJ Data Sci. **2**, 4 (2013)
59. Liu, B.: Sentiment analysis and subjectivity. Handb. Nat. Lang. Process. **2**, 627–666 (2010)
60. Prier, K., Smith, M., Giraud, C., Hanson, C.: Identifying health-related topics on twitter. In: International Conference on Social Computing, Behavioral Modeling, Prediction (2011)
61. Ali, A., Magdy, W., Vogel, S.: A tool for monitoring and analyzing healthcare tweets. In: HSD Workshop, SIGIR. Citeseer (2013)
62. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. JMLR **12**, 2825–2830 (2011)
63. Czeisler, C.A.: Perspective: casting light on sleep deficiency. Nature **497**, S13 (2013)
64. Uchida, S., Shioda, K., Morita, Y., Kubota, C., Ganeko, M., Takeda, N.: Exercise effects on sleep physiology. Front. Neurol. **3**, 48 (2012)
65. Youngstedt, S., O'connor, P., Dishman, R.: The effects of acute exercise on sleep: a quantitative synthesis. Sleep **20**, 203–214 (1997)
66. Ashe, M.C., Khan, K.M.: Exercise prescription. J. Am. Acad. Orthop. Surg. **12**, 21–27 (2004)
67. Van Helder, T., Radomski, M.W.: Sleep deprivation and the effect on exercise performance. Sports Med. **7**, 235–247 (1989)
68. Lyubomirsky, S., King, L., Diener, E.: The benefits of frequent positive affect: Does happiness lead to success? Psychol. Bull. **131**, 803 (2005)
69. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)
70. Rubin, D.B.: Causal inference using potential outcomes. J. Am. Stat. Assoc. **100**, 322–331 (2011)