# Toss 'N' Turn: Smartphone as Sleep and Sleep Quality Detector

**Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, Jason I. Hong**
Human-Computer Interaction Institute, Carnegie Mellon University
{loomlike, adoryab, jwwiese}@cs.cmu.edu, shahriyar@cmu.edu, {johnz, jasonh}@cs.cmu.edu

## ABSTRACT

The rapid adoption of smartphones along with a growing habit for using these devices as alarm clocks presents an opportunity to use this device as a sleep detector. This adds value to UbiComp and personal informatics in terms of user context and new performance data to collect and visualize, and it benefits healthcare as sleep is correlated with many health issues. To assess this opportunity, we collected one month of phone sensor and sleep diary entries from 27 people who have a variety of sleep contexts. We used this data to construct models that detect sleep and wake states, daily sleep quality, and global sleep quality. Our system classifies sleep state with 93.06% accuracy, daily sleep quality with 83.97% accuracy, and overall sleep quality with 81.48% accuracy. Individual models performed better than generally trained models, where the individual models require 3 days of ground truth data and 3 weeks of ground truth data to perform well on detecting sleep and sleep quality, respectively. Finally, the features of noise and movement were useful to infer sleep quality.

## Author Keywords
Smartphone; sleep; machine learning; sensors

## ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## INTRODUCTION
The meteoric adoption of smartphones places a rich sensor platform in the pockets, purses, and backpacks of many people. Interestingly, many people choose to use their phone as an alarm clock, placing these sensors in proximity of the bed. A recent study by the Pew Internet and American Life project found that 44% of mobile phone owners (83% of teens) sleep with their phone on or near their bed, with many using their phone as an alarm clock [36]. This advance in technology and associated change in behavior offers the possibility of cheaply and effectively tracking people's daily sleep behaviors without the need for additional hardware or for a significant change in behavior.

From a broader perspective, monitoring a person's sleep patterns offers three opportunities for the HCI community. First, sleep can be considered important context information for UbiComp systems. Having these systems model if a person is asleep or awake could allow them to modify their behavior to act more appropriately. Second, there is a growing interest in HCI around personal informatics and quantified self, where people are increasingly engaged in tracking and visualizing their personal behaviors [15]. The ability to detect and log sleep and sleep quality can add to this growing area of interest. Third, sleep and sleep quality have a strong connection to healthcare. Chronic sleep problems have been associated with diabetes, heart disease, and depression. In addition, even a few nights of poor sleep can impact alertness, memory, mood, and cognitive function [2,30,34,45]. Better tools for monitoring sleep could help improve diagnoses as well as help people understand their own needs and trends.

The goal of our work is to investigate how well a commodity smartphone can sense and model sleep and sleep quality without requiring significant changes in people's behavior. More specifically, we built Toss 'N' Turn (TNT), an Android app that logs seven different sensor inputs (an accelerometer, microphone, ambient light sensor, screen proximity sensor, running process, battery state, and display screen state). We provided TNT to 27 participants and collected data from them for one month. Each day, participants entered a sleep diary to provide ground truth. Using this dataset, we built models that classify if a person was asleep or not asleep (within a 10 minute window) with 93.06% accuracy. We detect when a person went to bed (within ±35 minutes on average), when they woke up (within ±31 minutes on average), and their sleep duration (within ±49 minutes on average). We also classify their daily sleep quality into good or poor with 83.97% accuracy and their global sleep quality (whether they are good or poor sleeper) with 81.48% accuracy.

Our contributions include the Toss 'N' Turn app that collects phone sensor data and ground truth data needed for sleep inference, a month long data collection study, an analysis of the collected data to identify the features that best predict sleep and sleep quality, models that detect sleep and sleep quality, and insights on how a commodity smartphone based sleep detector might benefit UbiComp, personal informatics, and healthcare.

## RELATED WORK

Sleep is a natural periodic state of rest where the human senses and motor activity are relatively suspended. *Sleep quality* can be defined in several ways, for example, having enough sleep or not [33], daytime sleepiness [24], and the subjective feeling about sleep along with objective factors such as the number of sleep disturbances and use of sleep medications [10]. Poor quality sleep is linked to many negative health outcomes, including diabetes, obesity, cardiovascular health, and depression [5,13,17,21,29,44].

Sleep is not necessarily a regular, singular, static activity, and thus there exist many challenges in monitoring sleep. People sleep during the daytime (sometimes napping) as well as night. They sleep together, with romantic partners, pets, and/or children. Different factors impact sleep quality including sleep partners, environments (e.g., an uncomfortable mattress or pillow, noise, light, too high or low temperature level), stress and anxiety, and consuming food and drinks; many of which are difficult to detect [7].

Below, we summarize the state of the art in sleep data collection and detection, organizing related work into three sections: tools for capturing sleep data, sleep monitoring apps, and sleep research within HCI.

### Sleep Measurement

The gold standard of assessing sleep is polysomnography, a system that combines an all-night electroencephalogram with measures of muscle tone and eye movements. Polysomnography requires lots of special equipment and trained technicians to setup the equipment, such as attaching the electroencephalogram sensors to a patient. It is impractical outside of a sleep lab.

Actigraphy provides a simpler approach with just one metric, movement captured by an accelerometer. Actigraphy is surprisingly accurate. Today watch-sized actigraphs are used in clinical settings to sense basic sleep patterns, such as hours slept, sleep efficiency, and number of wakings [4]. Consumer-oriented, wearable sleep sensors include the wrist worn FitBit [19], Jawbone [23], and WakeMate [46]. These devices cost approximately $100 USD, and people must remember to put them on before going to bed.

Data describing a person's sleep patterns are most commonly captured through self-reports via paper-based surveys and diaries. Examples include the Sleep Timing Questionnaire [33] and the Epworth Sleepiness Scale [24]. We used the Pittsburgh Sleep Quality Index (PSQI), which assesses sleep quality and disturbances over a one-month period [10]. PSQI integrates a wide variety of factors associated with sleep quality, including subjective quality ratings, sleep time, efficiency (time spent trying to fall asleep), frequency, and severity of sleep-related problems.

One challenge with surveys is that they require people to accurately recall details of past sleep behavior, and this can prove difficult [32]. A complementary approach involves keeping a sleep diary. While tedious to collect, a diary-based approach has proven to be reliable [8]. Espie et al. concluded in their seminal work that daily self-report may be a valid index of sleep disturbance where the correlations between the self-report and sleep assessment device (SAD) measures were consistently high [18]. Sleep diaries have been found to be reliable for bedtime and wake-time estimates via actigraphy [47] and ambulatory electroencephalographic monitoring [30]. Diaries have also been used to test sleep-detecting technologies including actigraphy [9]. With these respects, we used a sleep diary (designed based on the PSQI) for ground-truthing daily sleep data in addition to the PSQI that we used to measure the global quality from a month of sleeps.

### Sleep Monitoring Apps

Recently, a number of smartphone apps have emerged to help people manually track sleep, e.g., Tylenol PM Sleep Tracker [43], YawnLog [48], Sleep Journal [40], and Wellness Diary [31]. The output includes descriptive longitudinal displays, statistical analyses, and visualizations of sleep time, duration and subjective quality. Some apps try to automate sleep logging by tracking movement in bed with smartphone accelerometers ([39,41,42]). These systems require people to keep their phones on the bed while sleeping [15]. Researchers have also investigated using smartphones to detect sleep. Krejcar et al. presented a system that tracks sleep phases based on sound features [26]; Chen et al. monitored phone usage patterns to detect sleep duration [14]; and Bai et al. used the previous day's mobility, activity, and social interaction data to predict sleep quality with 78% of accuracy [6].

### Sleep Research in HCI

There has been increasing work in HCI and mobile computing focused on sleep [20,28]. Choe et al. identified challenges and design opportunities for using computing to support healthy sleep behaviors, as well as a design framework for mapping the design space of sleep technologies [15]. Aliakseyeu et al. examined how external factors of sleep like the location and family settings can affect sleep to highlight potential design opportunities [1]. Kay et al. utilized UbiComp sensors in the bedroom that allow people to reflect on the unconscious experience of sleep through sophisticated visualizations [25].

The most recent studies have focused on smartphones as a cheap and easy-to-use sleep technology. Researchers investigated people's interaction with the smartphones before and during bedtime. Bauer et al. presented Shuteye, a mobile app that promotes good sleep behaviors [7]. Lawson et al. identified the need for objective, reliable, and scalable methods of measuring and recording sleep, and designed an interactive mobile app based on a stimulus-response paradigm to record periods of sleep and wakefulness [27]. Shirazi et al. tested an idea for capturing sleep duration and sharing sleep information via social networks and found that monitoring can increase induce healthier sleep habits [38].

The closest past work to ours is by Chen et al. [14]. They measured sleep duration using smartphone sensors with data from eight solo-sleeping subjects collected over one-week. Their system detected sleep duration within ±42 minutes in average. We saw similar accuracy for sleep duration (within ±49 minutes in average) over more complex sleep contexts including people sleeping with partners, pets, and babies. In addition, we detect bedtime, waketime, daily sleep quality, and global sleep quality, all factors closely associated with health, as well as of interest to UbiComp and personal informatics.

## OVERVIEW OF TOSS 'N' TURN (TNT)

We designed and implemented TNT as an Android app that captures sensor data for use in inferring sleep and sleep quality. The app collects sensor data, asks people to fill out a daily sleep diary, and uploads the data every day to a server (see Fig. 1).

### Sensing and Data Uploading

We designed our system to collect various sensor data that might be relevant to sleep and sleep quality including sound amplitude (via the microphone), light (via the ambient light sensor), and movement (via the accelerometer) as listed in Table 1. While some recent smartphones come equipped with more than one microphone sensitive enough to capture even very small sounds; we used only the main microphone in the bottom of the phone and captured only maximum sound amplitude for simplicity and privacy reasons. Light intensity may be less reliable as people keep phones in pockets and bags. Therefore, we also collected screen proximity sensor values. Device states, such as screen on/off, processes (apps running on the phone), and the battery-charging state are also potentially informative in detecting sleep. For example, screen on (using the phone) is a good signal that a person is probably not asleep, but the screen is also sometimes automatically turned on for incoming calls or text messages, and by notification alarms from apps. Thus, other data, such as motion, should be used with the screen state to detect people's actual use of device. People often charge their phone before going to bed, and they often use the phone as an alarm clock. Both provide clues about bedtime and waketime. Fig. 2 shows examples of smartphone sensor data along with visualization of a participant's sleep.
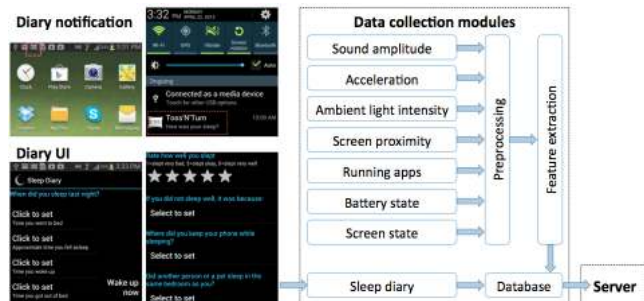


**Figure 1. Overview of Toss 'N' Turn. The app shows a notification for a daily sleep diary, collects sensor data, and uploads the data every day to a server.**

| Sensed value (frequency) | Data collection cycle | | | |
|---|---|---|---|---|
| | Night | Day | Btry<30% | Btry<15% |
| Sound amplitude (1hz) | Cont. | Every other minute | | |
| Acceleration (5hz) | Cont. | Every other minute | | |
| Light intensity and screen proximity (1/5hz) | Cont. | Every other minute | | |
| List of running apps (1/10hz) | Collect when screen is turned on | | | Stop |
| Battery states | Collect when battery level is changed or the power cable is plugged in/out | | | |
| Screen states | Collect when screen is turned on/off | | | |
| Sleep diary | Notification to enter ground truth every morning | | | |

**Table 1. Types of sensor data and frequency captured by Toss n' Turn (Cont. = collect continuously)**
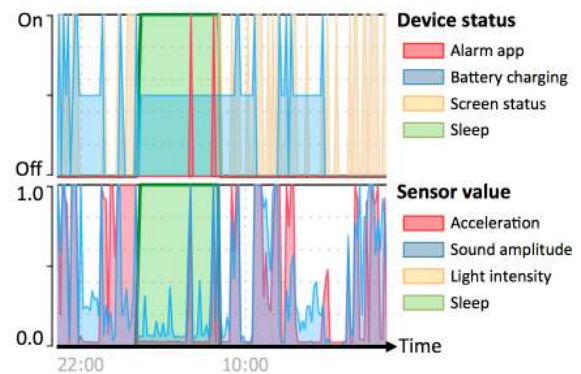


**Figure 2. Phone status values (the upper graph), sensor values (the lower graph), and sleep (the range in both upper and lower graph with the green shaded background)**

One challenge in continuous data collection with smartphones is battery life. We used two strategies to minimize power use. First, TNT changes the frequency of data collection based on the time of interest (in our case, we assume most people sleep from 10PM to 10AM). Second, it reduces duty cycles when the phone battery level is less than 30% and stops when below 15% (see Table 1).

TNT stores sensed data in a database residing on the protected storage of the phone. It creates a new database each day and uploads the previous database to the server. This strategy reduces the risk of data loss and complications that can come when attempting to upload large files.

### Sleep Diary

TNT collects information about people's last night sleep by using a diary designed based on the PSQI (see Fig. 1). Items in the diary include sleep time (when they went to bed, fell asleep, and woke up), subjective sleep quality (five scale rating, 1 = poor to 5 = very good), sleep disturbances (e.g. could not fall asleep within 30 minutes, woke up in the middle of night, etc.), and daytime sleepiness. The diary also asked about basic sleep environment, such as phone's location during sleep (on bed, near the bed, in the bedroom, not in bedroom) and co-sleeping (e.g., alone, with partner

or pet, with kids), for data analysis purpose. TNT reminds people to enter the sleep diary every morning, by using the Android notification service.

## DATA COLLECTION

To assess the performance of TNT to detect sleep and sleep quality, we conducted a data collection study. We recruited participants living in United States, over age 18, who use an Android phone (version 4.0+) as their primary phone. We recruited participants across a range of sleep contexts and who both had and did not have trouble sleeping.

### Procedure

All participants input demographic information, sleep environment details (e.g. have a pet or baby age < 2 in the bedroom), and sleep habits using our survey website. Next, participants installed TNT from the Google Play Store and used it for a month. We encouraged them to keep their phone turned on and to keep it in their bedroom while sleeping. We never instructed them as to where to place their phone in the bedroom. Each morning, a popup notification asked them to enter a daily sleep diary, described above.

On the last day of the study, we asked participants to fill out the PSQI survey online. This asked for a self-reported rating of sleep quality for the duration of our month long data collection. The PSQI consists of 19 self-rated questions and five questions rated by a bed partner or roommate [10]. We only used the self-rated questions in our online survey, which includes:

1. Subjective sleep quality
2. Sleep latency (amount of time it takes to fall asleep)
3. Sleep duration
4. Habitual efficiency (hours slept / hours spent in bed)
5. Sleep disturbances
6. Use of sleeping medications
7. Daytime dysfunction

Each item was rated on a 0-3 scale. These seven component scores are then summed to yield a global score, which has a range of 0-21; higher scores indicate worse sleep quality.

We used the sleep diary and PSQI data as ground truth in building models about sleep based on smartphone sensor data. We selected PSQI because 1) it provides a reliable and standardized measure of sleep quality as well as habitual bedtimes and waketimes, 2) it also collects a variety of sleep disturbances that might affect sleep quality, and 3) it has been used to discriminate between "good" and "poor" sleepers [10,11], which we can use this as a ground-truth for detecting global sleep quality.

Participants were paid $2 for each diary entry for a maximum of $72, which includes a completion incentive.

### Participants

Thirty participants signed up for the study and 27 (19 female) completed it. Two withdrew and one was eliminated because of sensor failure on the phone. Participants ranged in age from 20 to 59 with an average
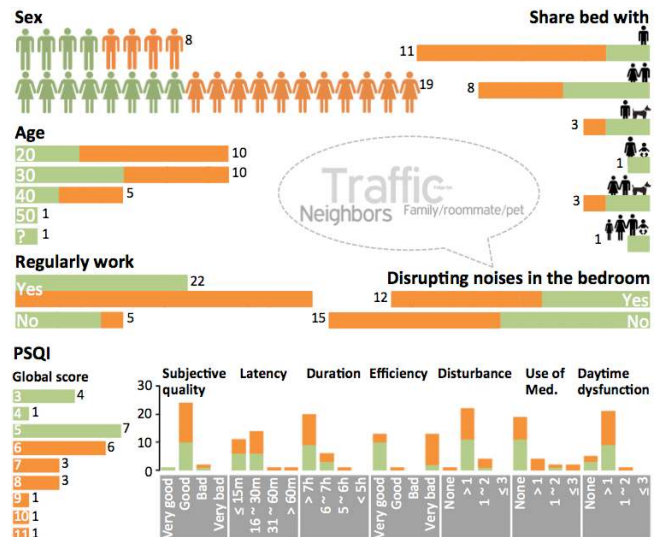


**Figure 3. Demographic information and the PSQI scores of our participants (Green and orange denote good and poor sleepers, respectively, where good and poor are distinguished by PSQI global score > 5). Center texts present the frequency of self-reported categories about disrupting noises.**

age of 34 (Std. = 9.8). Approximately 80% of participants reported working or going to school during daylight hours and sleeping at night. Participants reported several different numbers of other sleepers in their bedroom (14 alone; 3 with a pet; 8 with a partner; 3 with a partner and a pet; 1 with a partner and a baby, and we had 1 with a partner, baby, and a small child).

Fig. 3 shows demographic information and PSQI scores of our participants. We used PSQI global score > 5 to distinguish good and poor sleepers as presented in the previous literature (> 5 indicates that a subject is having severe difficulties in at least two PSQI components, or moderate difficulties in more than three components [10]). In our dataset, 66% had poor sleep where 3 reported that they have sleep abnormality (two participants talk in their sleep and one has insomnia).

## METHOD

In this section, we describe how we processed our data, extracted features, and built our models to detect sleep and infer the sleep quality.

### Data Processing

During our one-month study, participants submitted 795 diary entries. Some skipped days and some entered more than one diary for a single day. We hand-cleaned the data, fixing obvious bugs in the diaries. For example, participants were occasionally confused by the 24-hour time entry.

As a time unit for data analysis and feature extraction, we divided the time-series sensor data into a series of non-overlapped 10-minute windows. We used 10-minute windows as our basic unit of analysis for two reasons: 1) a median sleep latency (a transition time from wake to sleep) of participants was 10.9 minutes in average, and 2) 10-

minutes was the level of granularity used by most participants to report bedtime and waketime. We labeled the windows in-between reported bedtime and waketime from the diaries as *sleep* and the other windows as *not-sleep*. We did not use the sensor data of a day when there was no label information (a participant skipped entry of the sleep diary for that day). We did not use a diary when there was no sensor data (a participant turned their phone off). The resulting dataset included 90,097 windows (60,197 not-sleep and 29,900 sleep) and 1439 segments (711 not-sleep and 728 sleep).

We used data preprocessing techniques including outlier elimination and smoothing to reduce noise in the raw sensor values. We found that each phone seems to be different in terms of sensor ranges; therefore, we normalized the sensor values for each participant.

**Feature Extraction**
To infer different sleep contexts, we developed a number of features (see Table 2). To detect sleep (bedtime, waketime, and sleep duration), we used sensor level information, such as if the room is dark and quiet. We extracted statistical features including minimum (Min.), the $1^{st}$ quartile (Q1), median (Med.), the $3^{rd}$ quartile (Q3), maximum (Max.),

| Category | Modality | Feature variables |
|---|---|---|
| Sleep detection (32 features were extracted for each window) | Noise level | {Min., Q1, Med., Q3, Max., Avg., Std.} of sound amplitudes |
| | Movement | {Min., Q1, Med., Q3, Max., Avg., Std.} of the changes of acceleration |
| | Light intensity | {Min., Q1, Med., Q3, Max., Avg., Std.} of light intensities |
| | | {Min., Q1, Med., Q3, Max., Avg., Std.} of screen proximities |
| | Device state & usage pattern | Duration of screen-on time |
| | | The battery state {charging, not-charging, plugging-in/out} and alarm app usage {use, not-use} |
| | Regular sleep time | Timestamp of window |
| Daily sleep quality (122 features were extracted for each sleep segment) | Sleep duration | Bedtime, waketime, and sleep duration (extracted by using a sleep detection model) |
| | Sleep latency, habitual efficiency, & disturbances | {#peaks, Avg. width of peaks, Avg. height of peaks, interval of peaks, position of peaks} of sensor values, {Min., Q1, Med., Q3, Max., Avg., Std.} of sensor values, {Med., Std.} of each feature for the window detection, and yesterday's sleep quality (use the previously inferred result) |
| Global sleep quality (198 features were extracted for each participant) | Sleep regularity | {Med., Avg., Std.} of bedtimes, waketimes, sleep durations and qualities for a month of sleeps (use the previously detected and inferred results) |
| | Regular sleep latency, efficiency, & disturbances | {Med., Avg., Std.} of each feature for the daily sleep quality inference |

**Table 2. Feature sets for sleep detection, daily sleep quality inference, and global sleep quality inference**

average (Avg.), and standard deviation (Std.) of the sensor values for each 10-minute window. We also used screen on/off, battery states, and the timestamp of the window as additional sleep detection features.

We defined sleep quality features based on four factors of PSQI. These include sleep duration, latency, efficiency, and disturbances (we excluded three factors, subjective sleep quality rating, sleep medication, and daytime dysfunction, as these cannot easily be sense with a smartphone). We used the sleep detection results (bedtime, waketime, and sleep duration) as the inputs for the sleep quality inference models. We also added Med., Std., and peak of sensor values during sleep as features. Peaks at the inferred bedtime might indicate that the participant did not fall asleep completely (related to high sleep latency and low efficiency). On the other hand, peaks in the middle of inferred sleep could indicate sleep disruptions, e.g. waking up in the middle of night.

Finally, to infer global sleep quality (i.e. distinguishing if people are generally good or poor sleepers), we extracted participant's sleep regularity by calculating {Med., Avg., Std.} of sleep time. We also used {Med., Avg., Std.} of the sensor values for the sleep segments regarding sleep latency, efficiency, and disturbances.

In total, we extracted 32 features for sleep detection, 122 features for daily sleep quality inference, and 198 features for global sleep quality inference.

**Classification Models**
To detect sleep, we designed an algorithm based on a series of sleep/not-sleep classification as follows. First, the algorithm observes the sensor logs for 10 minute windows and classifies the window as a sleep or not-sleep state. It then uses a low-pass filter on the series of classified windows, eliminating possible sleep state detection errors such as temporal noises and/or disrupted states between very quiet and stationary situations. We used 30 minutes for the filter width to distinguish between a sleep disturbance and just noise. In other words, the algorithm has to look up the previous and next 30 minutes to make an accurate decision about if a participant is in a sleep or not-sleep state at that time. The filter width was chosen based on the rationale that more than 30 minutes could be regarded as not-sleep or a sleep disturbance, since one of the sleep disturbances described in the PSQI is "Cannot get to sleep within 30 minutes." Finally, it detects bedtime as the start point of the series of sleep-state windows, waketime as the end point of the windows, and sleep duration as the time difference between the bedtime and waketime.

To infer daily sleep quality, we formulated it as a two-class (good sleep and poor sleep) classification problem based on the PSQI's global score calculation function. PSQI considers all different factors regarding sleep quality to estimate global quality of sleep that can be used to distinguish good and poor sleepers. As we mentioned earlier, we designed our sleep diary based on the PSQI,

therefore we used the same measurement to get the ground truth of daily sleep quality from the sleep diary.

As our final task, we classified each participant as good or poor sleeper (global sleep quality from PSQI) based on the aggregated sleep data for a month, in which the results can be used to detect the changes of regularity of sleep quality.

We used two machine-learning algorithms, decision tree (C4.5) and Bayesian network (BN), along with a correlation-based feature selection method (FS) in which the features correlated to the target classes are used to build the models [22]. We also tested some features separately by using a 1R classifier that builds a set of rules based on only one feature. As baseline accuracies for our problems, we presented the result of a Random classifier that predicts a test sample as the most common class label in the training set. To handle the unbalanced class sizes in our problem, we used random resampling in the training set when we built models.

## RESULTS
### Sleep Detection
We tested general models (trained with data from other participants) based on leave-one-user-out cross validation and evaluated the individual models (trained on each participant's previous data) based on leave-one-day-out cross validation (a day = *6PM to next day's 6PM*). Table 3 shows the comparison result between two approaches with averaged accuracies (%) in classifying 10-minute windows as sleep or not-sleep and sleep detection results (precision, recall, and *F*-score). Note that 1R_Time uses only timestamp of window, 1R_Scrn uses screen on/off, 1R_Btry uses battery state, 1R_Light uses a feature from the light sensor, 1R_Acc uses an accelerometer (movement) feature, and 1R_Mic uses a microphone (noise level) feature.

The best performance was achieved by the individual models with BN FS (BN w/ feature selection) that shows an average of 94.52% accuracy for the sleep and not-sleep

| | General model | | | | Individual model | | | |
|---|---|---|---|---|---|---|---|---|
| | **%** | **Pre.** | **Rec.** | **F** | **%** | **Pre.** | **Rec.** | **F** |
| **Random** | 66.81 | 0.00 | 0.00 | 0.00 | 42.29 | 0.29 | 0.82 | 0.42 |
| **1R_Time** | 87.53 | 0.77 | 0.91 | 0.83 | 90.81 | 0.82 | 0.93 | 0.87 |
| **1R_Scrn** | 77.91 | 0.63 | **0.95** | 0.76 | 78.83 | 0.67 | **0.97** | 0.78 |
| **1R_Btry** | 80.73 | 0.75 | 0.67 | 0.71 | 80.96 | 0.73 | 0.74 | 0.72 |
| **1R_Light** | 72.79 | 0.59 | 0.71 | 0.64 | 80.61 | 0.70 | 0.84 | 0.75 |
| **1R_Acc** | 85.04 | 0.79 | 0.79 | 0.79 | 89.11 | 0.84 | 0.84 | 0.84 |
| **1R_Mic** | 89.45 | 0.84 | 0.86 | 0.85 | 89.33 | 0.86 | 0.82 | 0.84 |
| **C4.5** | 91.30 | **0.88** | 0.86 | 0.87 | 93.76 | **0.89** | 0.93 | 0.91 |
| **C4.5 FS** | 92.36 | 0.87 | 0.92 | 0.89 | 94.00 | **0.89** | 0.95 | 0.91 |
| **BN** | 90.19 | 0.84 | 0.89 | 0.87 | 93.73 | **0.89** | 0.93 | 0.91 |
| **BN FS** | **93.06** | 0.87 | 0.94 | **0.90** | **94.52** | **0.89** | 0.95 | **0.92** |

**Table 3. Averaged accuracies in classifying 10-minute windows as sleep/not-sleep and sleep detection results (precision, recall, and F-score). Bold shows the best performance for each column.**
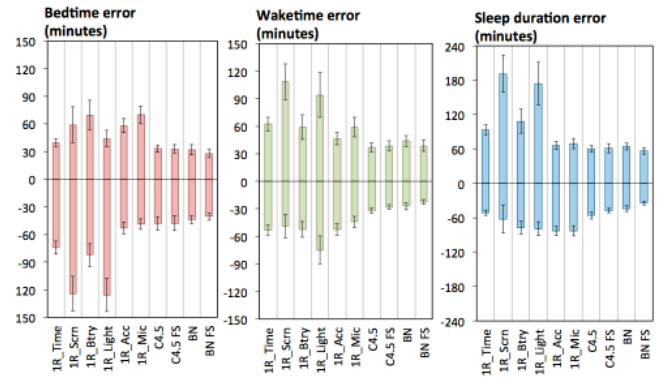


**Figure 4. Averaged errors (minutes) and the standard errors in detecting bedtime, waketime, and duration by using the individual-model approach**

classification (*F* = 0.92 in detecting the sleep windows). The model produces ±35 minutes, ±31 minutes, and ±49 minutes errors in detecting the bedtime, waketime, and sleep duration, respectively (Fig. 4 shows plus and minus time errors separately). The general models show comparative performance, for example, BN FS yielded 93.06% accuracy (*F* = 0.9) for detecting sleep states, which is similar to the individual approach result (see Table 3). The global models, however, were less accurate in detecting bedtime and waketime, where the global BN FS made predictions within ±44 minutes, ±42 minutes, and ±64 minutes of the ground truth data for bedtime, waketime, and duration, respectively.

The top five selected features were time, battery state, Min. of movement, Std. of sound amplitude, and Q3 of sound amplitude. 1R algorithm results in Table 3 and Fig. 4 show how accurate the information from a single sensor can detect sleep. When the model used the regular time of sleep information (1R_Time), it yielded average errors of 1 to 1.5 hours in detecting sleep time. Accelerometer (1R_Acc) and microphone (1R_Mic) features produced similar accuracies to 1R_Time, which showed the models' flexibility at detecting unusual sleeping hours. On the other hand, the screen-on states (1R_Scrn), battery states (1R_Btry), and ambient light sensor (1R_Light) features performed poorly. The 1R_Scrn shows high recall rate (0.95 and 0.97 for global and individual model, respectively) since people cannot use phone (turn the screen on) during sleeping, while it showed very low precision rate (0.63 and 0.67) denoting that not using the phone does not mean people are sleeping. In the same vein, 1R_Scrn outputs biased errors toward the minus direction in detecting bedtime (i.e. it infers a person went to sleep earlier than the actual bedtime) and overestimated the waketime and sleep duration. Interestingly, 1R_Light also showed similar results to 1R_Scrn, especially for bedtime detection. It is because the ambient light sensor was largely affected by the phone's screen light at nighttime.

If a sleep monitor system employs an individual model, a participant has to manually supply a certain amount of
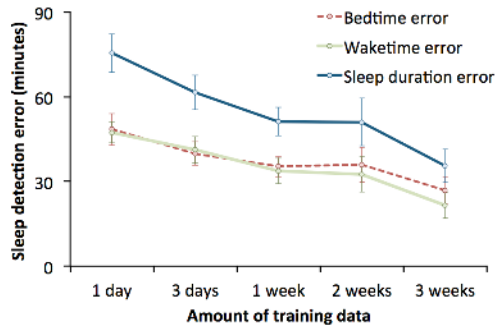
**Figure 5. Averaged errors (minutes) and the standard errors in detecting bedtime, waketime, and duration over different size of training data (model = individual BN FS)**

ground truth to train the system. We investigated how much training data a participant would have to enter to expect reasonably good prediction performance. We tested our best model, individual BN FS, over different amounts of training data where we picked the training data randomly from the past of the test date. As shown in Fig. 5, three days to one week of manual tagging would be required to train the system to achieve more than one-hour level of accuracy for sleep detection.

**Daily Sleep Quality Inference**
To classify each participant's daily sleep quality as good or poor, we first detected sleep by using the detection algorithm presented in the previous section (for training, we used ground truth sleep time), extracted features from the detected sleep segment, and then input the features to our sleep quality inference models. Here, to detect sleep quality, we used our best-performing model (BN FS built for individual participants). We could not use the sleep data not detected by our algorithm (1.62% of total sleep sessions). In addition, we did not include the participants who had only one or less days of poor sleep after the automatic sleep detection, since the trivial classifier of just saying all the sleep is good quality would perform more than 90% of accuracy in these cases. As a result, total 362
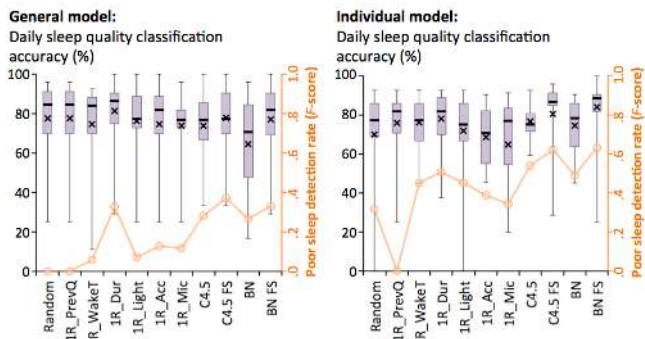


**Figure 6. Quartiles (x marks = average) of the accuracy (%) in classifying daily sleep quality. The detection rate (*F*-score) of poor sleeps is presented as a line plot where higher values denote better performance. A general model was evaluated by using a leave-one-user-out cross validation, while an individual model was tested by using a leave-one-day-out cross validation.**
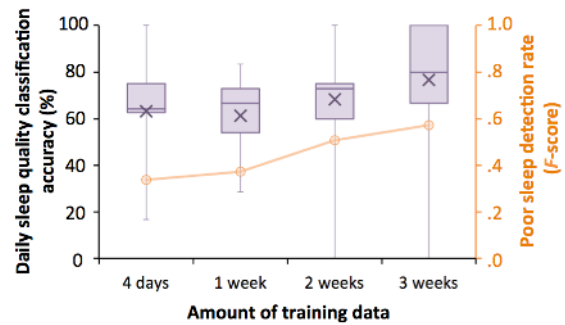


**Figure 7. Quartiles of the accuracy (%) in classifying daily sleep quality over different amount of training data. Poor sleep detection rate (*F*-score) is presented as a line plot (model = BN FS built for individual paticipants).**

sleep sessions from 17 participants were used in this specific experiment. Fig. 6 shows the classification accuracy and poor sleep quality detection rate by using the generally trained models and individual models (we plotted the accuracy as quartile graphs here because of the large variation for each participants result). BN FS built for individuals achieved the best performance on average with 83.97% accuracy in classifying good and poor quality for each sleep (Min. = 25, Q1 = 81.82, Med. = 88.46, Q3 = 90.48, Max. = 100%) and $F = 0.63$ in detecting the sleeps of poor quality. Please note the class distribution of our dataset is highly imbalanced where even the Random model could achieve around 70% of classification accuracy by predicting all the sleeps as the good class. However, it fails to detect poor sleep quality, and the detection of poor sleep is more important.

Using a feature selection phase increased the performance much more than in the sleep detection task. It was because the feature dimension was too high for the small number of samples in the quality inference task. Features that were most frequently selected while classifying sleep quality include the detected sleep time (bedtime, waketime, and sleep duration), Std. of movement, and the previously inferred sleep quality (previous day's quality). Interestingly, when we use only the previous day's quality feature, the model (1R_PrevQ) failed to infer daily sleep quality ($F = 0.0$). When we used only the waketime or duration feature with the individual model, we could detect around half of the poor days with poor sleep (1R_WakeT $F = 0.45$ and 1R_Dur $F = 0.51$).

As shown in Fig. 7, the individual model requires 3 weeks of ground truth data in order to produce results comparable to our best result (inter-quartile range of the accuracy = 66.67 to 100%).

**Global Sleep Quality Classification**
For the good and poor sleeper classification problem (note, poor sleeper: PSQI global score > 5), we evaluated models based on leave-one-user-out cross validation (i.e. built general models). We used the individual models to detect sleep and infer daily quality, then input their results into the global sleep quality classifiers as features along with all the
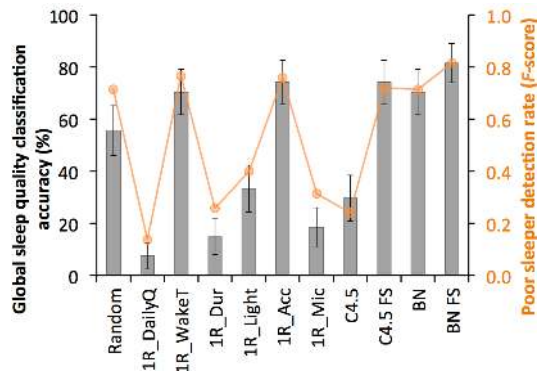
**Figure 8. The averaged accuracy (%) and the standard error in classifying participants into good and poor sleeper (global sleep quality classification). The detection rate (*F*-score) of poor sleepers is presented as a line plot.**

other sensor features. Since we were not classifying daily quality but were detecting good and poor sleepers, we included all participant data (27 participants) in contrast to the daily classification task (17 participants) where we rejected the subjects who had one or no days with poor sleep quality. As shown in Fig. 8, BN FS produced the best performance with an average 81.48% of accuracy for classifying good and poor sleepers and $F = 0.81$ in detecting poor sleepers. Features closely correlated with sleep disturbances, such as Avg. width of peaks in sound noises, Std. of movement, and Max. of light intensity, were selected for global sleep quality detection.

**DISCUSSION**
In this work, we studied several feasibility aspects related to using commodity smartphones to detect sleep and sleep quality. Given the availability, popularity, and capabilities of smartphones, our objective was to assess if smartphones might be a reliable/adequate tool to automate collection of sleep behavior.

**Detection of Sleep Time and Duration**
Our analysis in estimating of sleep time and duration showed that we are able to detect bedtime, waketime, and sleep duration with ±35 minutes, ±31 minutes, and ±49 minutes errors on average compared to the ground truth time reported by participants. These ranges seem relatively larger than commercial actigraphs like Jawbone and Zeo devices that have error rates lower than 10 minutes [14]. However, given the flexibility of our data collection and not restricting participants' options in placing their phone in certain proximity, our achieved accuracy is significant. Besides, many health recommendations related to sleep duration assume hour-level accuracy [3] and according to sleep experts, precise sleep measurements are not necessarily needed to have a meaningful picture of sleep behaviors and trends [15]. Hence, our system yielded reasonable results in monitoring sleep status.

Another unique aspect of our study was the ability to detect participant's sleep state in more complex situations such as

being with a bed partner, baby, and/or a pet. The average error in sleep duration in our study was ±49 minutes, which is the same as reported by Chen et al. in their mobile sleep monitor [14] where only sleeping alone situation was considered.

One limitation of our system is that people might not remember their exact sleep time when they enter the diary in the morning, which could result in some evaluation errors (even the diaries have proven to be reliable as we described in Related Work). For example, in certain situations it was shown that people underestimated the amount of time they slept and overestimated sleep onset latency [12]. One way to mitigate this kind of error would be to measure the bias in the reported sleep time with commercial actigraphs and then repeat the study. Since the intention of our work is to see if a smartphone is "good enough" as a sleep monitor, using other equipment is out of scope of this paper.

Our approach, based on detection of 10-minute windows classified as sleep or not sleep has several advantages compared to the direct detection of bedtime and waketime from the series of sensor values. For example, our approach can provide instant detection of participants' sleep or not-sleep state for a UbiComp system, without tracking people falling asleep or waking up time. Our approach also provides additional information about the kinds of abnormalities (such as waking up in the middle of night, sleep talking, severe snoring) by detecting disruptions as awake-like states in-between sleep states. While our current design only detects bedtime, wake time, and sleep duration, a low-pass filter could be used to resolve misclassified states.

**Classification of Sleep Quality**
In inferring daily sleep quality, we achieved an average accuracy of 83.97%. Our data collection strategy (including co-sleep options and flexible placement of the phone) makes inferencing more challenging. For example, the phone captures movements if participants place their phone on the bed, which results in more accurate detection of sleep disruptions. The false positives in the number of detected disruptions among solo sleepers might be less than for co-sleepers.

One way of acquiring more accurate data about the sleep environment and people's sleep status for the previous night is to ask people to monitor their recorded sleep and label disruptions. This option will give the people opportunities to capture and remember significant sleep events [25] as well as help train models to infer sleep quality.

Sleep quality inference proved to be more difficult than sleep detection. This may be partially caused by the challenge of collecting accurate ground truth data. Sleep quality can be an ambiguous, making it difficult to report. Our query "How was your sleep last night? Rate it on a one to five scale score" does not capture the full extent of a sleep session. Consider the following scenarios. A person
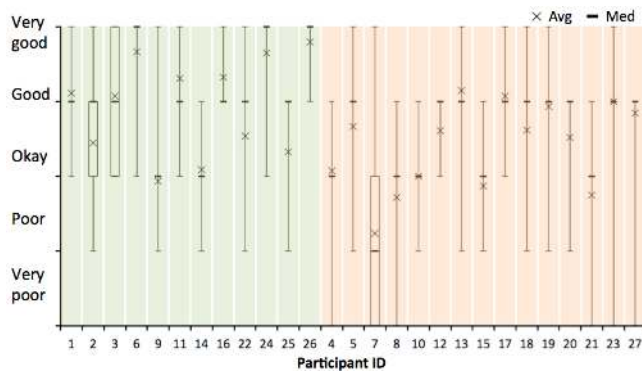
**Figure 9. Quartile distribution of daily subjective sleep quality (self-reported score rated by 1 "very poor" to 5 "very good") for each participant. Green and orange denote good and poor sleepers, respectively, where good and poor are distinguished by PSQI global score > 5 [10].**

might sleep only a few hours, rising early to catch a flight. The quality can be good but the quantity is not enough; they do not feel rested. A person might have lots of trouble falling asleep, but once asleep they sleep deeply and feel rested when they rise 12-hours after going to bed. Finally, a person might be awakened by a pet in the middle of the night. The sleep was disturbed, but they might or might not feel rested. For example, it was shown that people tend to overestimate sleep quality with noise disturbances like music [35]. In all of these situations the single sleep quality question is ambiguous. Better questions could ask people to rate if people feel rested, if they feel they got enough sleep, and to rate the quality of sleep while asleep.

Fig. 9 shows that many participants had a small interquartile range for daily subjective sleep quality where the median quality was "Good." Nevertheless, subjective sleep quality is very important. For example, recurrence of major depression is often preceded by a drop in subjective sleep quality, but only moderately associated with objective sleep quality [16].

One study limitation is that there is only one month of data. Because of this, we lacked enough examples of poor quality sleep as these happen with less frequency than good sleep.

Finally, we could classify good and poor sleepers with 81.48% of accuracy based on their one month of sleeps. The results from this model could be used to assess global sleep quality as well as to detect changes of regular quality. For example, one of our participants showed a decrease in sleep quality reported: "*I generally sleep well, but haven't been sleeping or feeling well for the past ten days. I have a lot of congestion and bronchitis which makes it hard to sleep.*" We expect to use our model over longer period of time to detect those health-related changes.

## CONCLUSION
The goal of our work was to investigate how well a commodity smartphone can sense and model sleep and sleep quality without requiring the purchase of any new hardware or a significant change in people's behavior. We also wanted to produce a set of insights to help guide the development of future sleep technologies, such as the amount of ground truth data needed for reasonably good detection. Our one-month field trial with 27 participants showed that we can detect an individual's sleep with average errors of 35, 31, and 49 minutes for bedtime, waketime, and sleep duration. We can also infer an individual's daily and global sleep quality with average accuracies of 83.97% and 81.48%, respectively. In our future work, we plan to use our dataset to test the feasibility of identifying types of sleep disturbances and to study recurring patterns in sleep behaviors that might predict upcoming problems. Finally, we plan to compare the performance of our mobile application with other commercially used wearable devices such as actigraphs.

## REFERENCES
1. Aliakseyeu, D., Du, J., Zwartkruis-Pelgrim, E., & Subramanian, S. Exploring interaction strategies in the context of sleep. In Proc. INTERACT '11, (2011) 19-36.
2. Altena, E., van der Werf, Y., Strijers, R.L.M., & van Someren, E.J.W., Sleep loss affects vigilance: Effects of chronic insomnia and sleep therapy. Journal of Sleep Research 17 (3), (2008) 335-343.
3. Alvarez, G.G. & Ayas, N.T. The impact of daily sleep duration on health: A review of the literature, Progress in Cardiovascular Nursing 19(2), (2004) 56-59.
4. Ancoli-Israel, S. et al. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 26(3), (2003) 342-392.
5. Ayas, N.T. et al. A Prospective Study of Sleep Duration and Coronary Heart Disease in Women. Arch. Intern. Med., 163(11), (2003) 205-209.
6. Bai, Y., Xu, B., Ma, Y., Sun, G., & Zhao, Y. Will you have a good sleep tonight?: Sleep quality prediction with mobile phone. In Proc. Int. Conf. BodyNets '12, (2012) 124-130.
7. Bauer, J.S. et al. ShutEye: Encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In Proc. CHI '12, (2012) 1401-1410.
8. Bootzin, R.R. & Engle-Friedman M. The assessment of insomnia. Behavioral Assessment 3, (1981) 107-126.
9. Borazio, M. & Van Laerhoven, K. Predicting sleeping behaviors in long-term studies with wrist-worn sensor data. Ambient Intelligence, LNCS 7040, (2011) 151-156.
10. Buysse, D.J., Reynolds, C.F., Monk, T.H., Berman, S.R., & Kupfer, D.J. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. Psychiatry Res. 28(2), (1989) 193-213.

11. Buysse, D.J. et al. Quantification of subjective sleep quality in healthy elderly men and women using the PSQI, Sleep 14(4), (1991) 331-338.

12. Carskadon, M.A. et al. Self-reports versus sleep laboratory findings in 122 drug-free subjects with complaints of chronic insomnia. Am J Psychiatry 133(12), (1976) 1382-1388.

13. Chandola, T., Ferrie, J.E., Perski, A., Akbaraly, T., & Marmot, M.G. The effect of short sleep duration on coronary heart disease risk is greatest among those with sleep disturbance. Sleep 33 (6) (2010) 739–744.

14. Chen, Z. et al. Unobtrusive sleep monitoring using smartphones, Proc. Int. Conf. PervasiveHealth '13, (2013) 145-152.

15. Choe, E.K., Consolvo, S., Watson, N.F., & Kientz, J.A. Opportunities for Computing Technologies to Support Healthy Sleep Behaviors. Proc. CHI '11, (2011) 3053-3062.

16. De, S. Subjective assessment of quality of sleep in chronic obstructive pulmonary disease patient and its relationship with associated depression. Lung India. 29(4) (2012) 332-335.

17. Drake, C.L., Roehrs, T., & Roth, T. Insomnia causes, consequences, and therapeutics: An overview. Depress Anxiety 18, (2003) 163–176.

18. Espie, C.A. et al. Use of the sleep assessment device to validate insomniacs' self-report of sleep pattern. J. Psychopathology and Behavioral Assessment 11, (1989) 71-79.

19. FitBit, http://www.fitbit.com

20. Fogg, B.J. Persuasive Technology: Using Computers to Change What We Think and Do. Science & Technology Books (2002).

21. Gottlieb, D.J. et al. Association of sleep time with diabetes mellitus and impaired glucose tolerance. Arch. Intern. Med. 165(8), (2005) 863–867.

22. Hall, M.A. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998).

23. Jawbone, https://jawbone.com

24. Johns, M.W. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. Sleep 14(6), (1991) 540-545.

25. Kay, M. et al. Lullaby: A capture & access system for understanding the sleep environment. In Proc. UbiComp '12, (2012) 226-234.

26. Krejcar, O., Jirka, J., & Janckulik, D. Use of mobile phones as intelligent sensors for sound input analysis and sleep state detection. J. Sensors 11(6), (2011) 6037-6055.

27. Lawson, S. et al. Validating a mobile phone application for the everyday, unobtrusive, objective measurement of sleep. In Proc. CHI '13, (2013) 2497-2506.

28. Li, I., Dey, A., Forlizzi, J., Höök, K. & Medynskiy, Y. Personal informatics and HCI: Design, theory, and social implications. In. Ext. Abstracts CHI '11, (2011) 2417-2420.

29. Livingston, G., Blizard, B. & Mann, A. Does sleep disturbance predict depression in elderly people? A Study in inner London. The British Journal of General Practice 43(376), (1993) 445-448.

30. Maquet, P. The role of sleep in learning and memory. Science 294(5544), (2001) 1048-1051.

31. Mattila, E. Mobile diary for wellness management--Results on usage and usability in two user studies. IEEE Trans. Inf. Technol. Biomed. 12(4), (2008) 501-512.

32. Monk, T.H., Buysse, D.J., Rose, L.R., Hall, J.A., & Kupfer, D.J. The sleep of healthy people--A diary study. Chronobiol Int. 17, (2000) 49-60.

33. Monk, T.H. et al. Measuring sleep habits without using a diary: the sleep timing questionnaire, Sleep 26(2), (2003) 208-212.

34. Moturu, S.T., Khayal, I., Aharony, N., Pan, W., & Pentland, A. Using social sensing to understand the links between sleep, mood, and sociability. In Proc. IEEE Int. Conf. SocialCom '11, (2011) 9-11.

35. Muze, A., Environmental noise, sleep and health. Sleep Medicine Reviews 11(2), (2007) 135-142.

36. Pew Internet and American Life Project. http://pewinternet.org

37. Rogers, A.E., Caruso, C.C., & Aldrich, M.S. Reliability of sleep diaries for assessment of sleep/wake patterns. Nurs. Res. 42(6), (1993) 368-372.

38. Shirazi, A.S. et al. Already up? Using mobile phones to track & share sleep behavior. Int. J. Human-Computer Studies 71(9), (2013) 878-888.

39. Sleep as Android. https://sites.google.com/site/sleepasandroid/

40. Sleep Journal App. http://sleeptrackerapp.blogspot.com

41. SleePBot. http://mysleepbot.com

42. SleepCycle. http://www.sleepcycle.com

43. Tylenol PM Sleep Tracker. http://www.tylenol.com/pain-sleeplessness/sleeptracker

44. Vgontzas, A.N. et al. Insomnia with objective short sleep duration is associated with type 2 diabetes: a population-based study. Diabetes Care 32(11), (2009) 1980-1985.

45. Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. Sleep inspires insight. Nature 427(6972), (2004) 352-355.

46. WakeMate. http://www.wakemate.com

47. Wilson, K.G. et al. Daily diary and ambulatory activity monitoring of sleep in patients with insomnia associated with chronic musculoskeletal pain. Pain 75, (1998) 75-84.

48. YawnLog. http://beta.yawnlog.com