From Markers to Interventions - The Case of Just-in-Time Stress Intervention

Hillol Sarker, Karen Hovsepian, Inbal Nahum-Shani, Susan Murphy, Bonnie Spring, Emre Ertin, Mustafa al'Absi, Motohiro Nakajima, and Santosh Kumar

Abstract The use of sensor-based assessment of stress to trigger the delivery of just-in-time intervention has the potential to help people manage daily stress as it occurs in the person's natural environment. The challenge is to mine the continuous stream of sensor data and identify those few opportune moments for triggering an intervention — when there is sufficient confidence in the accuracy of the sensor-based stress markers, in order to limit interruptions to the daily lives. In this chapter, we describe the process of developing a real-time method to identify *stress* episodes, from a time series of stress markers, to inform the triggering of just-in-time stress-management interventions.

Hillol Sarker
University of Memphis, Memphis, TN, USA e-mail: hsarker@memphis.edu

Karen Hovsepian
Troy University, Troy, AL, USA e-mail: khovsepian@troy.edu

Inbal Nahum-Shani
University of Michigan, Ann Arbor, MI, USA e-mail: inbal@umich.edu

Susan Murphy
University of Michigan, Ann Arbor, MI, USA e-mail: samurphy@umich.edu

Bonnie Spring
Northwestern University, Chicago, Illinois, USA e-mail: bspring@northwestern.edu

Emre Ertin
The Ohio State University, Columbus, OH, USA e-mail: ertin.1@osu.edu

Mustafa al'Absi
University of Minnesota Medical School, Duluth, MN, USA e-mail: malabsi@d.umn.edu

Motohiro Nakajima
University of Minnesota Medical School, Duluth, MN, USA e-mail: mnakajim@d.umn.edu

Santosh Kumar
University of Memphis, Memphis, TN, USA e-mail: santosh.kumar@memphis.edu

1 Introduction

Data collected by wearable sensors can now be used to assess stress continuously in a person's natural environment [11]. Computational models convert data collected by wearable sensors into a continuous measure of stress by recognizing the physiological responses exhibited during stress [17, 19, 33]. These advancements have inspired new research to analyze and visualize the dense time series of stress measurements together with associated contexts (e.g., location, activity, driving, etc.) [36, 38]. The goal of these works is to inform the development of just-in-time stress interventions that can help individuals manage their daily *stress* in the natural environment.

Management of stress via providing just-in-time-intervention (JITI) at the most opportune moments can help in coping with stress. Managing stress in daily life can directly improve health and wellness. For example, it can help individuals deal with migraine and panic attacks. It can also help manage heart disease, diabetes, and addictive behaviors, such as smoking, drinking, illicit drug use, overeating, etc. [2, 7, 27, 37, 39, 43]. We use the case of smoking cessation to illustrate our proposed methods for designing just-in-time stress intervention.

Smoking cessation is an important health issue because smoking causes the largest number of deaths, accounting for one in every five death [10, 28, 28]. Smoking is very difficult to treat as most smokers trying to quit eventually lapse. Stress is one of the major triggers for smoking lapses [5, 9, 39], and it is usually elevated in early phases of smoking cessation, which is when most lapses occur [4, 9]. But, individuals who continue to be abstinent experience a gradual decrease in their stress level [8].

During abstinence, in addition to coping with nicotine withdrawal effects, participants have to deal with numerous other issues, especially if participating in a mHealth smoking cessation study. They are usually asked to wear sensors (in the form of a chest band and wrist bands) for measurement of stress and detection of smoking lapses. In addition, participants are asked to respond to frequent (about 10 per day) Ecological Momentary Assessments (EMAs) where they self-report their mental state and surrounding contexts, which are not readily available from sensors (e.g., experiencing craving). Therefore, just-in-time stress interventions (which can also be perceived as an interruption) should be limited to reduce the interruption burden on participants.

There are several other considerations in the design of an effective just-in-time stress intervention. First, when an intervention is triggered, we should have high confidence in sensor-derived stress assessments. Second, the timing of the intervention trigger should be selected to maximize efficacy. For example, providing an intervention when a user is found to be *stressed* may further increase their stress, whereas providing intervention during moments of low stress with high likelihood of stress in the near future may help them prepare to better tolerate a future stress event.

Third, stress assessments and the triggering of interventions occurs in real time on resource-constrained and battery-operated wearable sensors and smart phones.

Although there are major advancements in technology, battery life is still a major issue for continuous stress assessment in the natural environment. Therefore, the computational model for providing just-in-time stress intervention needs to be efficient computationally and in power consumption. Computational efficiency is also needed to ensure that the entire computation method keeps pace with the rapidly flowing stream of sensor data and does not fall behind. Otherwise, the computational process will introduce a lag between measurements and trigger generation that will grow larger with time. This chapter takes all of these constraints into account in designing a just-in-time stress intervention to help with stress management during smoking cessation.

Presented work analyzes the time series of stress measurements and identifies non-overlapping periods, classified as *stressed*, *unsure*, *not-stressed*, and *unknown*. The *unknown* class occurs when data is noisy, missing, or affected by confounders such as physical activity. The *unsure* class occurs when the physiological data cannot be classified into *stressed* or *non-stressed* with sufficient confidence. We use data collected in a lab stress study to train our models.

We applied our proposed model on data collected from a smoking cessation field study to discover the stress patterns among nicotine dependent participants in their natural environment. We found that experiencing stressful episodes increased the likelihood of additional *stress* episodes in the near future. Similarly, participants in a *not-stressed* state are likely remain in the same state. Furthermore, transitioning from *not-stressed* to *stressed* is less likely than transitioning from *not-stressed* to *unsure*, and then from *unsure* to *stressed*. Observations like these suggest that providing a stress intervention when a user experiences a stressful episode may help him/her better cope with future *stress* episodes.

2 Related Works

Continuous assessment of stress usually requires a continuous assessment of physiology. Significant advances have been made in assessing physiology continuously in the natural environment from wearable physiological sensors [11], electrodermal response [24], photoplethysmography from the fingertip [23], or near-infrared spectroscopy from the forehead [14]. The stress intervention method described in this chapter can be adapted to stress measurements obtained from any of the above methods.

The works focusing on assessing interruptibility or availability [12, 20, 21, 42] have a similar goal, i.e., of identifying appropriate moments from sensor data when the user can be interrupted to deliver a prompt for intervention, self-report, or a phone call. But, their goal is to decide when to defer or delay a trigger, and hence they can be used to decide whether to deliver a stress intervention after a trigger has been generated using the proposed method of this chapter. Also, their method of data analytics is not directly applicable to our problem because the goal in the interruptibility/availability work is mainly to assess the data for each moment (e.g.,

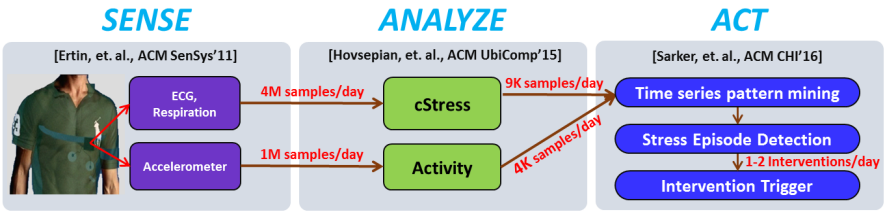

Fig. 1 Three stages of sensor-triggered intervention delivery process. First, sense using wearable sensor suite AutoSense [11] and a smart phone. Second, develop a computational model to analyze physiological data acquired from the first stage and assess stress [19]. Third, obtain stress time series, identify *stress* episodes, and act via triggering intervention at appropriate moments. This third stage is the main topic of this chapter.

minute) independently of the past to decide the current state of the user, whereas the goal here is to mine the time series to identify entire *stress* episodes.

The closest work related to the presented work is one of our recent works [36], where we developed a method to provide stress interventions using a *stress* episode detection method that addressed real life challenges such as physical activity confounds and missing data. However, this model involved very frequent stress assessments (every five seconds), which is not feasible to implement on a smartphone with limited computational capacity and battery life. Finally, the classification of *stress* episodes was not based on lab stress data, but left as a user-defined parameter that can be tuned on the basis of a global expected daily stress frequency. Stress occurrence in the field setting varies widely between individuals and between days for the same individual. Hence, the model has limited utility in real-life.

In contrast, the presented work uses data collected in a lab stress study for model development, where well-accepted stress tasks were performed. These protocol labels are used to learn the parameters of a *stress* episode detection model. Finally, the presented method is sensitive to the resource limitations of mobile phones, so it can be deployed in a real-life. In fact, the source code and the app version of our method is available for free use, as part of the MD2K software platform [1].

3 Overview of Sensors-to-Marker-to-Intervention

As shown in Figure 1, sensor-triggered mobile intervention has three main stages. First stage is the acquisition of data by sensing physiological parameters from wearable sensors in the user's free living condition. Sensor suites, such as AutoSense [11] can collect physiological signals (e.g., ECG, respiration, and accelerometer) at a high enough frequency (approximately five million samples per day) that suffices for continuous assessment of stress.

The second stage involves analysis and modeling of this high volume data obtained from the first stage. The outcome of this stage are personalized machine learning models that convert raw sensor data into bio-markers of health, behav-

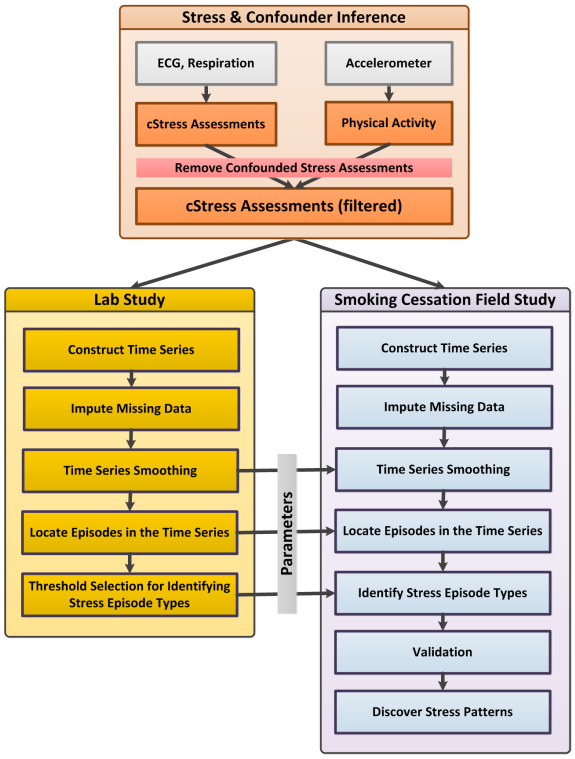

Fig. 2 Overview of the approach. First, we infer stress from ECG and respiration data, and confounder physical activity from accelerometer. Second, we remove physical activity confounded stress assessments. Third, we develop our *stress* episode identification model on lab study and apply the model on smoking cessation field study. Finally, we discover stress patterns from the smoking cessation field study.

ior, and environment (e.g., stress [19] and activity [34]). This stage reduces the data from 5 million per day to approximately ten thousand samples per day. Section 5 discusses the computational procedure for assessing stress and activity.

The third stage is tasked with identifying *stress* episodes from the stress marker time series obtained from the second stage. This stage reduces the data from ten thousand per day to usually 5 or less per day when an intervention should be delivered. This third stage is the main topic of this chapter.

Figure 2 shows an overview of the approach in this chapter. First, we infer stress from ECG and respiration data, and (confounding) physical activity from accelerometers. Second, we identify and filter out physical activity confounded stress assessments. Third, we develop our *stress* episode identification model on lab study data and apply the model on smoking cessation field study data. Finally, we present stress patterns observed in the smoking cessation field study data.

4 Data Description

Data collected in two user studies — a lab stress study and a smoking cessation field study — was used to train the stress inference model and design the just-in-time stress intervention. Each study was approved by the Institutional Review Board (IRB), and all participants provided written informed consent. This section provides an overview of the wearable sensor suite and a data description of lab stress study. The data description of smoking cessation field study is presented in Section 7.

4.1 Wearable Sensor Suite

The sensors worn by the all participants in both studies are part of a large suite of wearable biosensors, called AutoSense [11]. These unobtrusive sensors are worn mostly under the clothes, and include a two-lead electrocardiograph (ECG), 3-axis accelerometer, and respiration sensors, among others. Participants in the smoking cessation study also wore an inertial sensor on each wrist that includes a 3-axis accelerometer and a 3-axis gyroscope. Each sensor transmits the data continuously to a smartphone using a low-power wireless radio transmitter. The AutoSense chest band (with ECG, respiration, and accelerometer sensors) has its own 750 mAh battery that can last a week on a single charge. The phone, which collects GPS data continuously and keeps its wireless radio on for data reception, can last 13 hours on a single charge. The wrist sensor, using a 500 mAh battery, can last 3 days. The sampling rate is 128 Hz (downsampled to 64 Hz at the sensor) for the ECG sensor, 21.3 Hz for the respiration sensor, and 16 Hz for each axis of the accelerometer and gyroscope in both the chest band and wrist sensors.

Participants were given a smartphone to carry at all times. It receives and stores all sensor data. It is also used to fill out and store all the self-reports which capture instantaneous ground-truth assessments of stress and craving, as well as record various situational factors and events, such as physical activity levels, places visited, consumption of food and alcohol.

4.2 Lab Stress Study

We use ground-truth labeled data collected in a lab study that was reported in [19, 33]. The stress lab session lasts two hours including instrumentation (for 30 minutes), resting baseline (for 30 minutes), stress protocol (for 30 minutes), and post-stress rest (for 30 minutes) sessions.

Participants came to a lab where they wore the sensors for continuous data collection throughout the session. Participants were asked to sit in a comfortable chair and rest for 30 minutes during the initial baseline. The study includes three validated stress protocols, in the form of socio-evaluative, cognitive, and physical challenges.

During the socio-evaluative challenge, the participant was given a topic and asked to prepare (for 4 minutes) and deliver (for 8 minutes) a speech in front of a research staff. For a cognitive challenge (4 minutes), the participant was given a three digit number and asked to add three digits of that number, and then add the sum to the three digit number. Participants in the ***train*** study repeated this while seated and standing (counterbalanced). Participants in the ***test*** session completed only a single instance of this task while seated (because no significant effect of change in posture on stress response was observed in the ***train*** dataset). Finally, during the physical stressor, the participant was asked to leave his/her hand submerged in ice cold water, for 90 seconds. This was followed by a 30-minute rest period to allow the participants' physiology and mental state to return to baseline.

These tasks have been shown to reliably induce stress-related physiological changes [3]. Therefore, the lab protocol is used to label the data (i.e., gold standards) that are used to train and test the models. Time-stamping each distinct rest and stress period allows us to construct ground-truth labels for each minute of the lab-session, designating a minute as stressed, if the participant was undergoing a stress task during that minute, and not-stressed otherwise. These labels are subsequently used to train the *cStress* model and obtain continuous stress assessments.

5 Stress Inference from Physiological Data

The first step in stress intervention is the inference of stress from physiological sensor data in real time. In this section, we describe the procedure we used to infer physiological stress from wearable sensors. We adapt a recent model called *cStress* [19].

5.1 cStress Model for Stress Assessment

For the sake of completeness, we provide a brief summary of the *cStress* model that is presented in [19] and summarized in [36]. The *cStress* model uses electrocardiogram (ECG) and respiration data to infer stress. Acquiring these physiological signals in the field setting has several challenges. Wearable sensors sensing ECG and respiration signals, wirelessly transmits data to the smartphone. Data is timestamped when received by the phone. Data losses and software delays on the phone introduce variability in the time-stamping process. The granularity of stress is at the level of a minute while the errors in timestamps may be on the order of milliseconds. The main issue of time synchronization occurs due to data loss. A dynamic-programming based approach is used to correct the timestamps [19]. In addition, this time-stamp correction process identifies any losses in the sensor data stream. A small amount of missing data (1 packet) is imputed using cubic Hermite splines, which is known to be appropriate for interpolating physiological measurements [31]. Most packet losses involve only one packet, containing 5 samples (8%

of an ECG or respiration cycle). Imputation of 5 missing samples reduces the data loss rate from 10% to less than 1.5%.

ECG data processing contains three phases. First, identification of the acceptable portions of an ECG signal, which is considered acceptable if it retains characteristic morphologies of standard ECG, i.e., contains identifiable QRS complexes where R-peaks can be located. Otherwise, it is treated as unacceptable. Second, R-peaks are detected using Pan and Tompkins's algorithm [32]. The time difference between two successive R-peaks is R-R interval. Outlier R-R intervals (i.e., due to missing R-peaks) are removed from analysis. Third, the R-R intervals are normalized in order to develop a user-independent model. Respiration signal processing has similar phases, i.e., identifying and discarding unacceptable data, finding peaks and valleys, removing outliers, computing respiration features (i.e., inhalation duration), and normalizing the features.

As a next step in the stress assessment, a set of features is extracted from each non-overlapping minute's ECG and respiration sensor measurements. Based on this feature vector, the model determines whether that minute's sensor readings correspond to a physiological response to stressors. Among the many features used by the model are such ECG features as 80^{th} *percentile of R-R intervals* and *variance of R-R intervals*, and respiration features such as *mean IE ratio* and the *median of Stretch* [19]. This model was shown to classify stress and non-stress minutes collected in a lab stress protocol with 95% accuracy (F1 score of 0.78) on independent subject validation (different from the training set) [19]. In contrast to other stress inference works, such as [25, 26], which use only *Heart-Rate Variability (HRV)* features extracted from the ECG signal, the *cStress* model uses a richer feature set, containing other (non-HRV) ECG and respiration features. The authors of *cStress* paper show that adding these features significantly improves the performance of the model — F1 score jumps from 0.56 to 0.78.

Finally, the model was evaluated against self-reports collected in a week-long field study from an independent population of 23 participants and was found to have an F1 score of 0.71 [19]. In [36], the *cStress* model was evaluated with self report collected from another independent population of 38 participants who wore the sensors for 4 weeks and provided self-report of their stress level multiple times daily. In this validation, the F1 score was reported to be 0.72.

The *cStress* model provides a continuous measure of stress, scaled to be between 0 and 1, for every one minute of sensor data. These time-series of probability-like measures of stress is hereafter referred to as *stress likelihood*. To assess stress within intervals longer than a minute, we use a different measure, called *stress density*, from [36]. Stress density is defined as the area under the stress-likelihood time series divided by the length of the interval, which accounts for likely duration variation in contexts and activities (e.g. morning vs. afternoon, home vs. work).

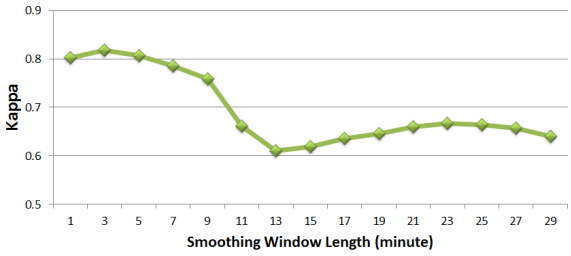

Fig. 3 Classification performances for different smoothing window length applied on stress likelihood time series in the lab study. We get the best performance with a kappa of 0.817 for a window length of 3 minutes.

5.2 Reducing The Impact of Physical Activity Confounds

Although physiology is influenced by several kinds of events in daily life, the main confounder for our sensor-based stress assessment is physical activity such as walking, which occurs frequently in our daily life. To isolate data affected by activity, we first detect physical activity from chest-worn 3-axis accelerometer data, using an existing model [34]. Although the stress assessment window is one minute, physical activity inference is available for every 10-second window. If the majority of 6 activity windows in a stress assessment minute window show presence of activity, the entire minute is excluded from stress assessment, i.e., considered missing.

Missing data due to sensor non-wear, sensor detachment, sensor loosening, sensor displacement [30, 34], or excluded due to the presence of physical activity confounds introduce discontinuity in the stress likelihood time series. In [36], missing data was imputed using via k-nearest neighbor method [13, 41, 44] where the imputation was based on other known contextual variables such as day of the week, time of day, previous stress levels, and the slope and intercept of previous time-series samples of the same user.

Such methods may be useful for offline analysis where we have access to an entire day's data, which is not the case during real-time computation on a smartphone. Therefore, we impute the missing stress assessments by simply carry forwarding the last known value. A *stress* episode containing majority of these imputed data is marked as *unknown* for intervention purposes. This may lead to some loss in accuracy, but makes it amenable to real-time efficient computation on a smartphone.

5.3 Time Series Smoothing

A basic fact of stress likelihood time series is that, because they are produced by a model that is imperfect, they undergo rapid fluctuations and may not be accurate for each minute. On the other hand, the number of stress interventions delivered per day should be limited (e.g., few times daily). It is also highly desirable to acquire high quality sensor outputs when triggering an intervention. Consequently, we first smooth the stress likelihood time series using a simple moving average. Then, in order to find the optimal window length, we compare the original labels (de-

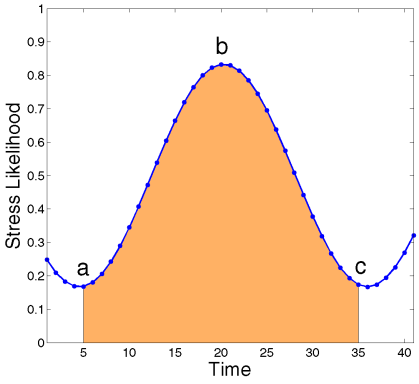

Fig. 4 A conceptual stress likelihood time series. We observe an increasing trend from '*a*' to '*b*' and a decreasing trend from '*b*' to '*c*'. An episode contains an increasing trend and immediately followed by a decreasing trend, marked as from '*a*' to '*c*'. For intervention (at '*c*') we compute the stress density from '*a*' to '*c*' and if stress density is above a specific cutoff we mark the episode as *stressed*.

rived from the lab stress protocol) with each one minute assessment in the smoothed *cStress*-based classification. Figure 3 shows classification performances for different smoothing window lengths. We get the best performance with a kappa of 0.817 for a smoothing window length of 3 minutes. We considered only odd-numbered window lengths to avoid introducing lag in the time series.

6 Determining the Timing of Intervention Delivery

Stress likelihood time series is a continuous time series of the outputs of *cStress* model for each minute. Just like any time series, the stress time series consists of peaks and valleys. The interval between two successive valleys is considered to be an episode. Figure 4 shows such a conceptual time series. In response to a stressor, stress likelihood starts increasing at '*a*'. At '*b*', stress likelihood starts decreasing down to point '*c*', where there is another upward trend. We define a *stress* episode as an increasing trend immediately followed by a decreasing trend. Based on this definition, we mark the entire period from '*a*' to '*c*' in the stress likelihood time series as a potential *stress* episode.

At the conclusion of an episode, we calculate the area under the stress likelihood time series of the concluded episode (at time '*c*'). The higher the area, more likely it is that the user had a stressful experience. However, duration of an episode is not constant. A short duration with a high area is more likely stressful in comparison with the same area for a longer duration. Hence, we divide the area by the duration of the episode and refer to it as stress density. A higher stress density indicates that the person has most likely experienced stress and the corresponding episode is a *stress* episode. On the other hand, a lower stress density in an episode indicates that the person is less likely to have experienced stress; hence we can mark the concluded episode as a *not-stressed* episode. If the concluded episode is identified as a *stress* episode, and the stress likelihood starts increasing again, as it does at '*c*', we can instantly provide an intervention (at '*c*'). An example of an appropriate intervention

can be the recommendation of a breathing exercise [22], allowing the person to be better prepared for subsequent stress occurrences.

In this chapter, we discuss the identification and delivery of an intervention at the conclusion of a *stress* episode (at '*c*'), which is also the beginning of an increasing trend for the next episode. As an alternate approach, we can consider the identification of the peak (at '*b*') and deliver an intervention when the person is highly likely to be experiencing stress. The approach proposed in this chapter can also be adapted to identify the *stress* episode when it is at peak ('*b*').

To generate triggers for stress intervention, we first need to locate and mark episodes in the stress likelihood time series. Next, we need to train a model to classify the episodes as *stressed* or *not-stressed*, which can then be used to decide the timing of stress interventions.

6.1 Locating Episodes in the Time Series

To provide an intervention, we first identify episodes in the rapidly varying stress likelihood time series. In addition, we need to identify increasing and decreasing trends in the time series. We now describe the computation of the starts and ends of all *stress* episodes. This approach is similar to the one proposed in [36], but the model parameters in [36] were based on field study data. In contrast, here we estimate the parameters based on a lab study where gold standard labels are known.

To find episodes and trends in our rapidly varying time-series data, we adapt the Moving Average Convergence Divergence (MACD) approach. This approach is commonly used in the stock market to inform buyers to purchase a stock when there is a positive trend in the time series and it is highly likely that the stock price will increase in near future. Similarly, it informs to sell the share when there is a negative trend in the time series. This MACD has recently been used to detect trends in physiological data [18, 36]. MACD estimates the trend based on short-term and long-term Exponential Moving Average (EMA). It provides one signal when the trend is going up and another signal when it is going down. When applied on the (simple moving average of the) stress likelihood time series, MACD can provide a signal when the stress likelihood is going up (positive trend) and another signal when the stress likelihood is going down (negative trend).

MACD is computed as follows:

$$\begin{aligned} M &= EMA(L; w_{slow}) - EMA(L; w_{fast}) \\ S &= EMA(M; w_{signal}), \end{aligned} \tag{1}$$

where L is the stress likelihood time-series, M is the so-called MACD line, and S is the so-called MACD Signal Line. As the formula shows, M is calculated by subtracting a fast-moving, short-term EMA line from a slow-moving, long-term EMA line. The intersection of M and S indicates a change in trend, and if $S - M > 0$ then the trend is positive, otherwise the trend is negative. Thus, MACD divides the stress-

likelihood time series into smaller variable length, increasing and decreasing stress trends in the time periods between intersections of M and S.

We tune the window length parameters, w_{slow}, w_{fast}, and w_{signal}, used in Equation (1) using the lab study data, seeking to maximize $gain/N$, where *gain* is defined as the total area under the stress likelihood time series curve during positive-trend intervals, whereby the start and end of each positive-trend interval are dictated by the MACD rule, mentioned above, and N is the number of positive-trend intervals. Dividing by N discourages window lengths that result in a very large number of short positive-trend intervals. To estimate parameters $<w_{slow}, w_{fast}, w_{signal}>$ we conduct a grid search with progressive zoom, with initial grids covering the range from 1 minute to 30 minutes for each parameter, with the goal to maximize $gain/N$. In our analysis, we found that the optimal window lengths are: w_{slow} = 19 minutes, w_{fast} = 7 minutes, and w_{signal} = 2 minutes, which maximize $gain/N$. In the lab time series using the specified parameters, we obtained 119 episodes across 21 participants.

6.2 Threshold Selection for Identifying Stress Episodes

Conclusion of an episode also marks the start of an increasing trend for the next episode. We need to assess whether the just concluded episode is a candidate *stress* episode worthy for an intervention.

However, there are missing data (imputed) in the episodes of the time series, which can be attributed to sensor detachment, equipment non-wear, lack of good quality data, or discarded data due to the presence of confounder physical activity. If more than 50% of the minutes in an episode are missing, we mark the entire episode as *unknown* and discard the episode from the threshold selection step. If a detected episode in the time series contains the majority of a lab stressor, we mark it as a *stress* episode.

In the lab, we have the precise timings of the start of lab stressors, allowing us to easily identify each *stress* episode. In the field, when we do not have such markings of stressors, we require a metric for assessing or marking an episode as *stressed* or *not-stressed*. We found that the aforementioned stress density is a great candidate for such a metric. A high stress density identifies a *stress* episode and low stress density identifies a *not-stress* episode. However, using a single stress density cutoff to make this binary decision can lead to misidentifying those 'gray-area' episodes having stress density near the decision cutoff. To address this issue, we assign all such gray-area episodes into class *unsure*. Thus, rather than picking one threshold, we pick two thresholds for these three episode classes.

In summary, an episode is classified as *not-stressed* if its stress density is below the first threshold (threshold 1), as *stressed* if its stress density is above the second threshold (threshold 2), and as *unsure* if its stress density is between the first and second thresholds. Using this approach allows us to identify *stressed* and not-*stressed* episodes with high confidence.

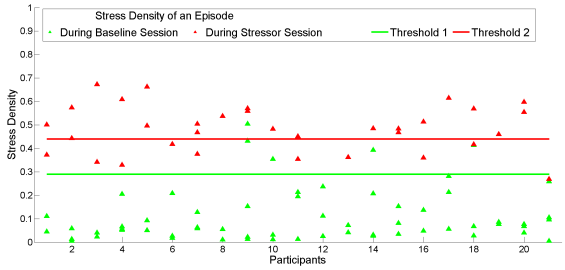

Fig. 5 Stress density of each session in the lab study. Discarding episodes with stress density between two thresholds (0.29 and 0.44) ensures both precision and recall of *stressed* and *not-stressed* class above 95% with episodes discarded due to being *unsure* is minimum.

Table 1 Computation of *stress* episodes classification performance metric — precision and recall from Figure 5

Precision of *stressed* =	Number of red triangles above threshold2 / Total triangles above threshold2
Recall of *stressed* =	Number of red triangles above threshold2 / Total red triangles above threshold2 or below threshold 1
Precision of *not-stressed* =	Number of green triangles below threshold1 / Total triangles below threshold1
Recall of *not-stressed* =	Number of green triangles below threshold1 / Total green triangles below threshold1 or above threshold2

Table 2 Confusion matrix of *stress* episode identification for thresholds 0.29 and 0.44, ensuring 95% precision and recall, where we excluded 13 *unsure* episodes and 24 *unknown* episodes.

		Classified by Model		
		Stress	Not stress	Total
Actual	Stress	23 (95.8%)	1 (4.2%)	24
	Not stress	1 (1.7%)	57 (98.3%)	58
	Total	24	58	82

Out of 119 episodes in the lab study, 24 are *unknown* due to missing data or poor quality data. Figure 5 shows the stress density for each of the remaining 96 episodes in the lab study. Labeling episodes with stress density between two thresholds (0.29 and 0.44) as *unsure* ensures both precision and recall for *stressed* and *not-stressed* class above 95% while keeping the *unsure* episode count as low as possible. Table 1 summarizes the calculation of precision and recall for *stressed* and *not-stressed* class. Table 2 presents the confusion matrix. Precision and recall for *stressed* class are 95.8% and 95.8%, respectively and for *not-stressed* class are 98.3% and 98.3%, respectively.

In case we want to ensure 90% precision and recall in identifying *stress* episodes, we can pick different thresholds — <0.29, 0.42>. For 85% precision and recall, the thresholds are <0.29, 0.29>; in this case there is no *unsure* class and the two threshold method simplifies to a binary decision with a single threshold. Table 3 summarizes these results.

7 Smoking Cessation Field Study

Stress is prevalent among nicotine-dependent individuals, especially during their abstinence. We applied our proposed model on smoking cessation field study data to observe the stress patterns of abstinent smokers during their first 3 post-quit days.

7.1 Data Description

Participants: We use data collected in a smoking cessation study that was reported in [35]. In this study, the participants were cigarette smokers who reported smoking 10 or more cigarettes per day for at least 2 years, and who reported high motivation to quit. To qualify, participants had to pass a screening session prior to being enrolled in the study. The screening includes assessment of current medical and mental health status and history of any major medical and psychiatric illness. Screening also includes assessment of smoking behavior, mood, and other behavioral health measures. Participants were excluded if they had ongoing major medical or psychiatric problems and if they had other comorbid psychiatric and substance use problems. Also, participants who did not follow a normal day/light diurnal cycle were excluded to control for variation in diurnal physiological activity and behaviors.

Protocol: Once enrolled, the participants picked a smoking quit date. Two weeks prior to their quit date, subjects wore the sensor suite for 24 hours in their natural environment. After completion of the 24 hour monitoring, which we call the pre-quit session, subjects come back to the lab for their second visit. Smoking cessation counseling is provided starting at this second visit to the lab. Then the subjects come back to the lab on the assigned quit date to attend a counseling session and to begin the 72 hours of monitoring in the field; this is referred to as the post-quit session. They come back to the lab each day to confirm smoking status by capturing an expired breath sample in a carbon monoxide (CO) monitor. During each day of monitoring (24 hour pre-quit and 72 hour post-quit), the participants wear the sensor suite during awake hours, and complete 12 Ecological Momentary Assessments (EMAs) [40] daily.

Data Collected: We collected data from 53 participants. The participants wore the sensor suite for a total of 2,706 hours with 1,350 hours of stress assessments after excluding intermittently missing data, and excluding all stress assessments confounded by physical activity. A total of 2,526 EMA prompts were delivered (11.9 per day) with a completion rate of 94.2%.

We apply the proposed model on this smoking cessation field study data to observe the stress patterns in the first 3 days after quitting. We compute the stress likelihood for each minute from ECG and respiration data, impute the missing data, apply simple moving average to smooth the time series, identify the *stress* episodes using the MACD based approach, and mark them as *stressed*, *unsure*, *not-stressed*, and *unknown* based on the stress density of each episode.

7.2 Validation of Stress Assessments in the Smoking Cessation Study

The *cStress* model was validated against lab study and independent field studies [19, 36] as described earlier. To validate the *cStress* assessments in this new data set, we followed the similar approach presented by Hovsepian et al. [19]. First, we check the consistency of self-reports as they are subject to bias and careless responding [36].

We use Cronbach's alpha [6] to assess the consistency of the self-reported responses. This metric is widely used in the field of psychometrics. Cronbach's alpha measures the internal consistency of items that are intended to measure the same psychological construct. An alpha score of 0.7 or higher is regarded as acceptable [6] in most studies. We compute the Cronbach's alpha using 5 affect items of self-report — "*Cheerful?*", "*Happy?*", "*Frustrated/Angry?*", "*Anxious/Tense?*", and "*Sad?*" (The two positive items, "*Cheerful?*" and "*Happy?*", were reverse-coded). The overall consistency score across all participant's self-reports is 0.76, suggesting an acceptable consistency (≥ 0.7).

We then compare the sensor-inferred stress markers (for each minute) with participant's self-reported EMA. We used F1 as a metric, which is a harmonic mean of precision and recall. Figure 6 summarizes the F1 scores across participants from this smoking cessation field study. They range from 0.36 to 1.0 with a median of 0.65. This is lower as compared to those reported in the two previously reported field studies, i.e., 0.71 in [19] and 0.72 in [36].

There are several potential reasons for a lower F1 score. First, the presented work validates stress assessments in a smoking cessation phase when participants may not fully available to provide accurate self-reports. We find some evidence of it in that the self-report consistency of this presented study is significantly lower as compared to [36] (0.76 vs. 0.84). In general, the median F1 score of 0.72 in [36] should be viewed against its self-report consistency of 0.84, while the median F1 score of 0.65 for the present study should be viewed against its self-report consistency of 0.76.

We compute Cronbach's alpha for the participants who have F1 score below median (see Figure 6). They have unacceptable self-report consistency scores with a median Cronbach's alpha of 0.58. Participants with above median F1 score have median Cronbach's alpha 0.68. Median F1 score for participants with acceptable Cronbach's alpha score ($\geq$0.7) is 0.68 while for participants with unacceptable Cronbach's alpha score ($<$0.7), F1 score is 0.63. In summary, in cases of poor agree-

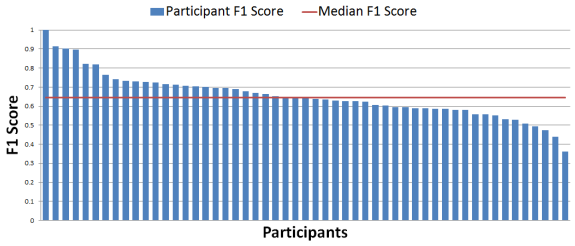

Fig. 6 F1 score between self-report and sensor assessment range from 0.36 to 1.00 with median 0.65.

ment between self-reports and *cStress* assessments, the consistency of self-reports are poor, which may prevent obtaining a good F1 score.

Second, in comparison to [19]) that excluded missing or physical activity confounded data from validation analysis, we use all the data (with imputation where necessary). Imputation was also done in [36], but using a heavy-weight and potentially more accurate method. In contrast, we use a simple and computationally efficient method for imputation to make it feasible to run in real time on the phone. This may have also introduced some loss in accuracy.

Finally, in comparison to [36], which used overlapping windows with a 5 second moving increment for smoothing the time series (resulting in computation of 12 stress values during a minute worth of data), we do not use any overlapping windows for computational efficiency and to avoid any lag between data and generation of stress trigger due to computational delays. This may have led to some additional loss in accuracy.

The above validation is for the minute-level output from the *cStress* model. To evaluate *stress* episodes rather than the minute-level outputs, we compare them against self-report response to the item "*Anxious/Tense?*." To remove participant's biases in self-report, we compute z-scores from the self-report. By using this z-score, we can directly compare one participant's response to another. Values of z-score above 0 indicates *stressed* while values of less than 0 indicates *not-stressed*. Out of the 2,526 prompted EMAs at random moments, 22 were triggered at moments when our model identified that the participant was *stressed*. We found a median z-score of 0.21 in such cases which indicates *stressed* from self-report. For the 673 EMAs triggered during when our model suggests *not-stressed*, we found a median z-score of -0.20 indicating *not-stressed* from self-report.

7.3 Stress Patterns Observed in the Smoking Cessation Study

We apply the approach proposed in Section 5 and Section 6 on smoking cessation field study data collected from 53 participants. We obtain *stressed*, *unsure*, *not-stressed*, and *unknown* episodes in the field using stress density as a metric.

As discussed in Section 6.2, to ensure 95% precision and recall for both *stressed* and *not-stressed* class we need to pick stress density threshold $<0.29, 0.44>$. As shown in Table 3, we find 28.3 *not-stressed*, 2.7 *unsure*, and 1.5 *stress* episodes per day on average. Figure 7 shows the episodes for one participant and on pre-quit day.

If we relax the constraint by considering above 90% precision and recall, we can pick stress density thresholds $<0.29, 0.42>$ for episode assessing. We observe 1.7 *stress* episodes per day as compare to 1.5 in case of 95%. In case we relax even further, for 85% precision and recall we get stress density thresholds $<0.29, 0.29>$ meaning there is only one threshold and no unsure class. We observe 4.2 *stress* episodes per day in such a case.

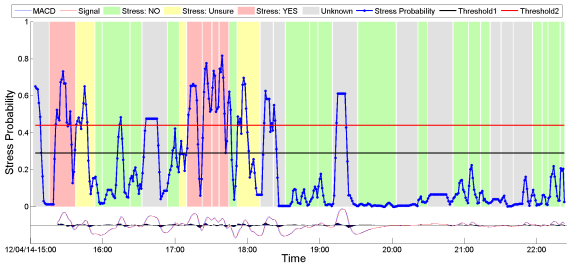

Fig. 7 Time series of stress likelihood of one participant on pre-quit day.

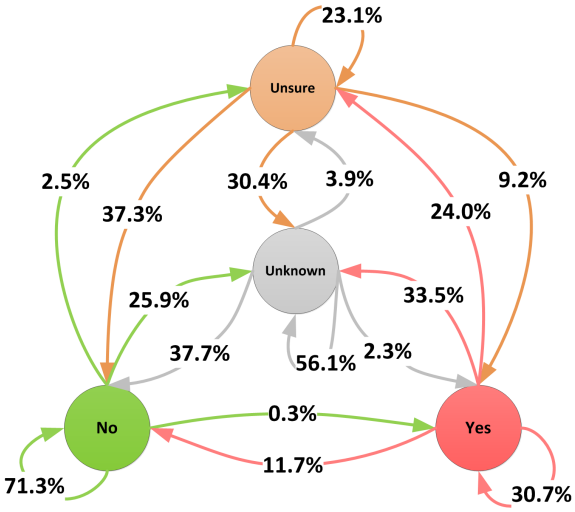

Fig. 8 State transition probabilities between different *stress* episode types, *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*.

7.4 Transitions Between Episodes of Different Classes

Stress episodes are classified as *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*. We analyze transition probabilities among these classes which can inform the intervention design and the modeling of the time-series data. Figure 8 shows the estimated transition probabilities between these types of episodes for the field study of 53 participants.

Table 3 *Stress* episodes classification statistics for ensuring different precision and recall (95%, 90%, and 85%).

		Precision and Recall		
		95%	90%	85%
Lab Study (Stress Density)	Threshold 1	0.29	0.29	0.29
	Threshold 2	0.44	0.42	0.29
Field Study (per day)	Not-stressed	28.3	28.3	28.3
	Unsure	2.7	2.5	0
	Stressed	1.5	1.7	4.2

Stress episodes more likely to be of similar kinds in successive episodes. From Figure 8, we observe transition probabilities for *no-no* (71.3%), *unsure-unsure* (23.1%), and *yes-yes* (30.7%). It was shown in our earlier work in [36] as well that there is a correlation between the durations of successive *stress* episodes. This can be explained by theory and evidence [15, 16, 29] suggesting a spiral process where current exposure to stressors attenuate the stress coping capability of the person. This can lead to subsequent reactivity to other stressors. For example, a person in a conflict with a colleague at work produces negative feelings and emotions that makes it difficult for the person to manage his or her workload during the day, making him/her more prone to making mistakes at work, which can lead to further stress.

If a person is *not-stressed* in the current episode it is likely that next episode in the time series is also going to be a *not-stressed* one with probability 71.3%. It is less likely to make a transition directly to *stressed* state (0.3%). The more likely transition is from *not-stressed* to *unsure* (2.5%), and then to *stressed* (9.2%).

Observations like these suggest that providing a stress intervention when the person experiences a *stressed* episode or an *unsure* episode followed by a *not-stressed* episode can help that person to cope with future stress occurrences. As an alternate application, we can also feed the previous minute's stress estimate into the computational model (such as *cStress*) for estimating stress in the current minute. Such recursive relationships may increase the accuracy of stress assessment.

8 Discussion, Limitation, and Future Work

There are several limitations in the presented work. First, in addition to physical activity, stress can be confounded by pharmacological factors such as caffeine, smoking, or drugs. Automated detection of such events can improve stress assessment accuracy.

Second, wearing of ECG and respiration sensors in a chest band is not very convenient and unlikely to scale widely. Collection of physiological data from other devices such as smartwatches may capture stress more conveniently. Also, assessment of stress from multiple sensors (e.g., PPG and galvanic skin response in smartwatches) can improve data yield. In case data is missing from one modality, one can use data from the other modality for stress assessment.

Third, this work demonstrates a mechanism for determining the timing for an intervention. It does not directly provide any efficacious intervention, which requires making choices on not only the timing of delivery, but also the right content, the adaptation mechanisms for personalizing it to the individual, the user's context, and the selection of the right modality for delivery (e.g., on the phone, on a smartwatch). Right now, it's not clear whether we should provide an intervention when somebody is going through a stressful experience and may not be receptive to receiving intervention. On the other hand, we may consider providing an intervention when somebody is *not-stressed* so that they can better tolerate future *stress* episodes. These issues can be investigated via conducting a micro-randomized trial.

Fourth, the presented work identifies a *stressed* or *not-stressed* episode at the conclusion of the episode. Intervention delivered at that time is aimed to prepare the person for future stress occurrences. As an alternative approach, we can also identify the timing for proactive intervention. When stress likelihood is in an increasing trend, a rapid rise in the stress likelihood time series may indicate that the episode may build up to become a *stress* episode. Machine learning models can be developed that can look into such time series patterns (e.g., slope, prior stress density, and skewness) and predict whether the episode is going to be a stressful one. As soon as the model is confident enough, a proactive stress intervention can be triggered.

Finally, we have presented the relationship between *stress* episodes among the nicotine dependent individuals who are going through abstinence. Detection of the first lapse during abstinence [35] made it feasible to investigate the relationship between *stress* episodes and smoking relapse via objective sensor based approach. Discovery of additional insights from such data can contribute to designing an efficacious smoking cessation intervention.

9 Conclusion

Identifying the appropriate timing of intervention is a critical component in a just-in-time stress intervention. Providing frequent interventions will increase user burden and hence it is critical to identify the opportune moments when there is sufficient confidence in sensor-based stress assessment. In this study, we presented such an approach to determine the timings of *stressed* and *not-stressed* episodes from sensor based measurements in the context of smoking cessation. While there are numerous ways to further improve the presented approach and the eventual intervention, the overall framework for data analysis may be applicable to several other biomarkers obtained from sensor data.

10 Acknowledgments

The authors acknowledge support by the National Science Foundation under award numbers CNS-1212901 and IIS-1231754 and by the National Institutes of Health under grants R01CA190329, R01MD010362, and R01DA035502 (by NIDA) through funds provided by the trans-NIH OppNet initiative, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative. We also thank Soujanya Chatterjee and Barbara Burch Kuhn from University of Memphis.

References

1. mCerebrum: An Open Source Software Suite for Mobile Sensor Data. https://md2k.org/software/ (2016)
2. Al'Absi, M.: Stress and addiction: Biological and psychological mechanisms. Academic Press (2011)
3. Al'Absi, M., Bongard, S., Buchanan, T., Pincomb, G.A., Licinio, J., Lovallo, W.R.: Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. Psychophysiology **34**(3), 266–275 (1997)
4. al'Absi, M., Hatsukami, D., Davis, G., Wittmers, L.: Prospective examination of effects of smoking abstinence on cortisol and withdrawal symptoms as predictors of early smoking relapse. Drug and Alcohol Dependence **73**(3), 267–278 (2004)
5. Baer, J.S., Lichtenstein, E.: Classification and prediction of smoking relapse episodes: an exploration of individual differences. Journal of consulting and clinical psychology **56**(1), 104 (1988)
6. Bland, J., Altman, D.: Statistics: notes cronbach's alpha. BMJ **314**(7080), 572–572 (1997)
7. Bunker, S.J., Colquhoun, D.M., Esler, M.D., Hickie, I.B., Hunt, D., Jelinek, V.M., Oldenburg, B.F., Peach, H.G., Ruth, D., Tennant, C.C., et al.: " stress" and coronary heart disease: psychosocial risk factors. The Medical Journal of Australia **178**(6), 272–276 (2003)
8. Cohen, S., Lichtenstein, E.: Perceived stress, quitting smoking, and smoking relapse. Health Psychology **9**(4), 466 (1990)
9. Cummings, K.M., Jaén, C.R., Giovino, G.: Circumstances surrounding relapse in a group of recent exsmokers. Preventive Medicine **14**(2), 195–202 (1985)
10. for Disease Control, C., (CDC, P., et al.: Smoking-attributable mortality, years of potential life lost, and productivity losses–united states, 2000-2004. MMWR. Morbidity and mortality weekly report **57**(45), 1226 (2008)
11. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., Shah, S.: Autosense: Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In: ACM SenSys, pp. 274–287 (2011)
12. Fogarty, J., Hudson, S., Lai, J.: Examining the robustness of sensor-based statistical models of human interruptibility. In: ACM CHI, pp. 207–214 (2004)
13. Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., Botstein, D.: Imputing missing data for gene expression arrays (1999)
14. Hirshfield, L.M., Solovey, E.T., Girouard, A., Kebinger, J., Jacob, R.J., Sassaroli, A., Fantini, S.: Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In: ACM CHI, pp. 2185–2194. ACM (2009)
15. Hobfoll, S.E.: Conservation of resources: A new attempt at conceptualizing stress. American psychologist **44**(3), 513 (1989)
16. Hobfoll, S.E., Vinokur, A.D., Pierce, P.F., Lewandowski-Romps, L.: The combined stress of family life, work, and war in air force men and women: A test of conservation of resources theory. International Journal of Stress Management **19**(3), 217 (2012)
17. Hong, J., Ramos, J., Dey, A.: Understanding physiological responses to stressors during physical activity. In: ACM UbiComp, pp. 270–279 (2012)
18. Hossain, S., Ali, A., Rahman, M., Ertin, E., Epstein, D., Kennedy, A., Preston, K., Umbricht, A., Chen, Y., Kumar, S.: Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In: ACM IPSN, pp. 71–82 (2014)
19. Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S.: cstress: towards a gold standard for continuous stress assessment in the mobile environment. In: ACM UbiComp, pp. 493–504 (2015)
20. Iqbal, S., Zheng, X., Bailey, B.: Task-evoked pupillary response to mental workload in human-computer interaction. In: ACM CHI Extended Abstracts, pp. 1477–1480 (2004)
21. Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., Bailey, B.P.: Towards an index of opportunity: understanding changes in mental workload during task execution. In: ACM CHI, pp. 311–320 (2005)

22. Konrad, A., Bellotti, V., Crenshaw, N., Tucker, S., Nelson, L., Du, H., Pirolli, P., Whittaker, S.: Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction. In: ACM CHI, pp. 3829–3838 (2015)
23. Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y., Kameyama, K.i.: Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In: ACM CHI, pp. 857–866 (2015)
24. Matthews, M., Snyder, J., Reynolds, L., Chien, J.T., Shih, A., Lee, J.W., Gay, G.: Real-time representation versus response elicitation in biosensor data. In: ACM CHI, pp. 605–608 (2015)
25. McEwen, B.: Stress, adaptation, and disease: Allostasis and allostatic load. Annals of the New York Academy of Sciences **840**(1), 33–44 (2006)
26. McEwen, B.: Physiology and neurobiology of stress and adaptation: Central role of the brain. Physiological Reviews **87**(3), 873–904 (2007)
27. McEwen, B.S.: Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. Annals of the New York Academy of Sciences **1032**(1), 1–7 (2004)
28. Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L.: Actual causes of death in the united states, 2000. Journal of the American Medical Association (JAMA) **291**(10), 1238–1245 (2004)
29. Nahum-Shani, I., Hekler, E., Spruijt-Metz, D.: Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychology
30. Ni, K., Ramanathan, N., Chehade, M., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M., Srivastava, M.: Sensor network data fault types. ACM TOSN **5**(3), 25 (2009)
31. Nielsen, P., Le Grice, I., Smaill, B., Hunter, P.: Mathematical model of geometry and fibrous structure of the heart. American Journal of Physiology-Heart and Circulatory Physiology **260**(4), H1365–H1378 (1991)
32. Pan, J., Tompkins, W.: A real-time qrs detection algorithm. IEEE Transactions on Biomedical Engineering **32**(3), 230–236 (1985)
33. Plarre, K., Raij, A., Hossain, S., Ali, A., Nakajima, M., Al'absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., et al.: Continuous inference of psychological stress from sensory measurements collected in the natural environment. In: IEEE/ACM IPSN, pp. 97–108 (2011)
34. Rahman, M., Bari, R., Ali, A., Sharmin, M., Raij, A., Hovsepian, K., Hossain, S., Ertin, E., Kennedy, A., Epstein, D., Preston, K., Jobes, M., Beck, G., Kedia, S., Ward, K., alAbsi, M., Kumar, S.: Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors. In: ACM BCB, pp. 479–488 (2014)
35. Saleheen, N., Ali, A.A., Hossain, S.M., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., Kumar, S.: puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In: ACM UbiComp, pp. 999–1010 (2015)
36. Sarker, H., Tyburski, M., Rahman, M., Hovsepian, K., Sharmin, M., Epstein, D.H., Preston, K.L., Furr-Holden, C.D., Milam, A., Nahum-Shani, I., al'Absi, M., Kumar, S.: Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In: ACM CHI (2016)
37. Sauro, K.M., Becker, W.J.: The stress and migraine interaction. Headache: The Journal of Head and Face Pain **49**(9), 1378–1386 (2009)
38. Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J.G., Vhaduri, S., Preston, K., Kumar, S.: Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In: ACM UbiComp, pp. 505–516 (2015)
39. Shiffman, S.: Relapse following smoking cessation: a situational analysis. Journal of consulting and clinical psychology **50**(1), 71 (1982)
40. Shiffman, S., Stone, A., Hufford, M.: Ecological momentary assessment. Annual Review of Clinical Psychology **4**, 1–32 (2008)
41. Speed, T.: Statistical analysis of gene expression microarray data. CRC Press (2004)
42. Tan, C.S.S., Schöning, J., Luyten, K., Coninx, K.: Investigating the effects of using biofeedback as visual stress indicator during video-mediated collaboration. In: ACM CHI, pp. 71–80 (2014)

43. Torres, S.J., Nowson, C.A.: Relationship between stress, eating behavior, and obesity. Nutrition **23**(11), 887–894 (2007)
44. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)