11-23-2016

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

Hillol Sarker

FROM MARKERS TO INTERVENTIONS – THE CASE OF JUST-IN-TIME STRESS INTERVENTION

by

Hillol Sarker

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Computer Science

The University of Memphis

December 2016

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Santosh Kumar for mentoring me to grow as a research scientist. His consistent guidance, encouragement, and inspiration helped me to pursue my research. I would also like to thank my committee members, Dr. Vasile Rus, Dr. Scott Fleming, and Dr. Duen Horng (Polo) Chau, for their constructive comments and suggestions, and serving as my committee members.

I would like to thank Dr. Emre Ertin from The Ohio State University, Dr. Moushumi Sharmin from Western Washington University, Dr. Karen Hovsepian from Troy University, Dr. Matthew Tyburski, Dr. David Epstein, and Dr. Kenzie Preston from The the NIH National Institute on Drug Abuse, Dr. Susan Murphy and Dr. Inbal Nahum-Shani from the University of Michigan, Dr. Mustafa al'Absi from the University of Minnesota Medical School, Dr. C. Debra Furr-Holden and Dr. Adam Milam from Johns Hopkins University Bloomberg School of Public Health, Dr. Ida Sim from the University of California, San Francisco, Dr. Bonnie Spring from Northwestern University, and Dr. Amin Ahsan Ali, Dr. Md. Mahbubur Rahman, Syed Monowar Hossain, Rummana Bari, Nazir Saleheen, Soujanya Chatterjee, and Barbara Burch Kuhn from the University of Memphis for their contributions. I would also like to thank Jeremy Luno and Lucas Salazar regarding their help in conducting the high burden user study in the natural environment which is an integral part for my thesis.

A special thanks to my family including my supportive parents, my beloved wife Gitaly Das, and my wonderful kid, Amio Sarker, for their sacrifices during the journey and being a constant source of inspiration.

This work was supported by the National Science Foundation under award numbers CNS-1212901 and IIS-1231754 and by the National Institutes of Health under grants 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.

Finally, I would like to thank the University of Memphis for providing me with an excellent research platform as well as the constant support throughout my academic career.

ABSTRACT

Sarker, Hillol. Ph.D. The University of Memphis. December, 2016. From Markers to Interventions – The Case of Just-in-Time Stress Intervention. Major Professor: Dr. Santosh Kumar.

Wearable wireless sensors for health monitoring are enabling the design and delivery of just-in-time interventions (JITI). Critical to the success of JITI is to time its delivery so that the user is available to be engaged. This dissertation takes a first step in modeling users' availability by analyzing 2,064 hours of physiological sensor data and 2,717 self-reports collected from 30 participants in a week-long field study. Delay in responding to a prompt is used to objectively measure availability. Presented work compute 99 features and identify 30 as most discriminating to train a machine learning model for predicting availability. Findings suggest that location, affect, activity type, stress, time, and day of the week, play significant roles in predicting availability. Users are least available at work and during driving, and most available when walking outside. Proposed model finally achieves an accuracy of 74.7% in 10-fold cross-validation and 77.9% with leave-one-subject-out.

Management of daily stress can be greatly improved by delivering sensor-triggered just-in-time interventions (JITIs) on mobile devices. In addition to assessing the availability of a person, the success of such JITIs critically depends on being able to mine the time series of noisy sensor data to find the most opportune moments. This dissertation proposes a time series pattern mining method to detect stress episodes in a time series of discontinuous and rapidly varying stress data. This model is applied to two separate human subject studies on physiological, GPS, and activity data collected from 91 (38+53) users in their natural environment to discover patterns of stress in real life. Findings suggest that the duration and the type of a prior stress episode predict the duration and the type of the next stress episode. Stress in mornings and evenings is lower than

during the day. The work then analyzes the relationship between stress and objectively rated disorder in the surrounding neighborhood and suggests a model to identify the proactive or reactive timing for JITI.

TABLE OF CONTENTS

Content	Page
List of Tables	**x**
List of Figures	**xi**
1 Introduction	**1**
1.1 Background and Motivation	1
1.2 Problem Statement	3
1.3 Summary Results	5
1.3.1 Availability for JITI	5
1.3.2 Trigger for JITI	6
1.4 Key Contributions	10
1.5 Organization	12
2 Sensor-Triggered mHealth Interventions	**14**
2.1 Vision of Sensor Triggered JITI	14
2.2 Overview of the Approach	15
2.2.1 SENSE	15
2.2.1.1 AutoSense Sensor Suite	16
2.2.1.2 Mobile Phone	17
2.2.2 ANALYZE	17
2.2.2.1 cStress Model for Stress Assessment	17
2.2.2.2 Activity Inference from Accelerometer	19
2.2.2.3 Inference of Semantic Location	21
2.2.2.4 Driving Detection from GPS	23
2.2.3 ACT	24
2.2.3.1 Content of Intervention	24
2.2.3.2 Modality of Intervention	26
2.2.3.3 Timing of Intervention	27
3 Trigger Generation for Sensor-Triggered Just-In-Time Intervention — The Case of Stress Intervention	**30**
4 Related Works	**31**
4.1 Availability for Intervention	31
4.2 Trigger Generation for Intervention	32
5 Determining Availability for Intervention	**34**
5.1 Introduction	34
5.2 Related Works	37
5.3 Study Design	40
5.4 Sensor Inference	43
5.5 Metric for Measuring Availability	43
5.5.1 Response Delay	44

5.5.2 Completion Time 44
5.5.3 Labeling of Available and Unavailable States 47
5.6 Findings 48
5.7 Predicting Availability 52
5.7.1 Feature Computation 52
5.7.2 Feature Selection 56
5.7.3 Model 57
5.8 Limitations and Future Work 58
5.9 Chapter Summary 60

6 Identifying Stress Episodes Based on Field Stress Data 61
6.1 Introduction 61
6.2 Related Works 63
6.3 Data Description 65
6.3.1 Devices and Sensor Measurements 65
6.3.2 Field Study Procedure 65
6.4 Stress Inference from Physiological Data 67
6.4.1 Stress Likelihood & Stress Density 67
6.4.2 Need for Personalization 68
6.5 Reducing the Impact of Confounding Factors 68
6.5.1 Data Based Approach 70
6.5.2 Model Based Approach 71
6.6 Missing Data Imputation 73
6.7 Field Validation of Stress Assessment 74
6.8 Locating Stressful Episodes 76
6.9 Parameterization of Episode Identification 79
6.9.1 Approach 1: Based on Field Study 79
6.9.2 Approach 2: Based on Lab Study 82
6.10 Chapter Summary 82

7 Identifying Stress Episodes Based on Lab Stress Data 84
7.1 Introduction 84
7.2 Overview 86
7.3 Data Description 86
7.3.1 Wearable Sensor Suite 88
7.3.2 Lab Stress Study 88
7.4 Stress Inference from Physiological Data 89
7.4.1 Reducing the Impact of Physical Activity Confounds 90
7.4.2 Time Series Smoothing 91
7.5 Determining the Timing of Intervention Delivery 92
7.5.1 Locating Episodes in the Time Series 93
7.5.2 Threshold Selection for Identifying Stress Episodes 94
7.6 Smoking Cessation Field Study 96
7.6.1 Data Description 96

7.6.2 Validation of Stress Assessments in the Smoking Cessation Study 98
7.6.3 Stress Patterns Observed in the Smoking Cessation Study 101
7.7 Chapter Summary 101

8 Applications of Our Model 103
8.1 Application 1: Triggering of the Self-Report 103
8.2 Application 2: Patterns of Stress 103
8.2.1 Role of Prior Stress 103
8.2.2 State Transition Probability 104
8.2.3 Temporal Effect on Stress 106
8.2.4 Effect of Activity on Stress 107
8.2.5 Environmental Effect on Stress 107
8.3 Application 3: Intervention 110
8.3.1 Reactive Stress Intervention 110
8.3.2 Proactive Stress Intervention 112
8.3.3 Phone Implementation 115
8.4 Other Applications 116
8.5 Chapter Summary 116

9 Conclusion and Future Directions 117
9.1 Summary and Key Contribution 117
9.2 Future Directions 119
9.2.1 Availability Assessment 119
9.2.2 Just-in-Time Stress Intervention 120

References 123

LIST OF TABLES

Table		Page
1	Confusion Matrix for the Semantic Labeling model [104]. Restaurant is sometime confused with store. Precision and recall related performance metrics are available in Table 2.	20
2	Accuracy using semantic labeler model [104]. TP = true positive rate, FP = false positive rate, P = precision, R = recall, F = F-Measure, and AUC = area under the curve.	21
3	For 6 values of N (30 seconds, 1 min, 2 min, 3 min, 4 min, and 5 min), the above four derivative features are computed for stress and activity, producing 24 features for each.	52
4	Selected 30 features ranked (R) according to information gain. Detailed location offers the highest information gain.	55
5	Confusion Matrix for predicting availability using SVM model on RBF kernel built on **Representative Dataset**. Overall Accuracy is 74.7% against a base accuracy of 50%, with a kappa of 0.494.	57
6	Confusion Matrix for predicting availability using SVM model on RBF kernel built on **Democratic Dataset**. Overall Accuracy is 69.2% against a base accuracy of 50%, with a kappa of 0.383.	58
7	In total there are 9,087 stress episodes with an expected count per day of 9.2. A duration threshold of 13.5 minutes labels 498 significant stress episodes, with an expected daily count of 0.5.	81
8	Computation of *stress* episodes classification performance metric — precision and recall from Figure 20	97
9	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.	97
10	*Stress* episodes classification statistics for ensuring different precision and recall (95%, 90%, and 85%).	102
11	Performance of the model for predicting Significant Stress Episodes for duration thresholds of 13.5, 7.3, and 2.4 minutes.	115

LIST OF FIGURES

Figure		Page
1	Three stages of sensor-triggered intervention delivery process. First, sense using wearable sensor suite AutoSense [59] and a smart phone. Second, develop a computational model to analyze physiological data acquired from the first stage and assess stress [89]. 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 dissertation.	16
2	Using cut-off point 0.21384 we observe that subjects were physically active for around 20% of their total wearing time.	20
3	A sample GPS trace for one day from a participant. The red line shows the path commuted by the participant. The pinned locations are the location at the time of EMA prompt.	22
4	Two level semantic labeling of GPS clusters.	23
5	Delay distribution is fitted with a Gamma distribution with shape parameter κ=1.2669 and scale parameter θ=35.5021. We use the cut-off of $p = 0.95$ that occurs at 124.1 seconds, as the grace period. A response delay beyond this grace period is marked as *unavailable*.	44
6	We plot Cronbach's alpha value for various thresholds of z-score of completion times. We observe that the alpha is always acceptable (i.e., ≥ 0.7). This holds even when we consider various subsets of items that require recollection or multiple choice selection.	46
7	Probability of participants being available across different contexts. Here morning is defined as before 9 AM and evening as after 5 PM. Arrived at a location means arrival within 30 minute, while leaving means 30 minute prior to leaving. Red line is drawn for $p(A) = 0.5$.	49
8	Mean response delay across different contexts. Morning, evening, arrival, and leaving are defined as in Figure 7. Red line represents the overall mean of 49.5s ($\pm$116.0s).	50
9	Time of day when participants wear the wearable sensors and contributed in the data collection campaign. X-axis indicates time of day, Y-axis indicates number of person days. The number over each histogram indicates average minutes of data collected on that hour of day.	66

10 (a) Stress density for each participant. We observe wide between person variation here. (b) Day wise stress for the participant with highest stress density. We observe wide between day-to-day variation for this (and other) participants. 69

11 R-R interval decreases (and heart rate increases) due to physical activity and recovers exponentially after the conclusion of activity. 70

12 Exponential recovery parameter τ is learned for each participant. Black curves show 99% exponential recovery (Equation 6.1). In this case, before the heart rate fully recovers from the first episode, another activity episode occurs. Hence, baseline heart rate is carried forward. 72

13 F1 score between self-report and sensor assessment range from 0.130 to 0.917 with median 0.717. Bottom 5 have unacceptable self-report consistency score with median cronbach's alpha score 0.335 while overall consistency score is 0.843. 75

14 Timing of just-in-time stress intervention for momentary and significant stress episode. Starting of a rectangular region indicates timing of potential proactive interventions as generated by MACD. 76

15 The likelihood of stress follow beta distribution with shape parameter $\alpha = 0.222$ and $\beta = 1.027$. The threshold for significant stress is 0.782 (p=0.95). 79

16 A momentary stress episode with high likelihood of stress (95th percentile) (see Figure 15) and a duration of more than duration threshold is marked as a significant stress episode. Duration threshold of 7.3 minute leads to an average of one significant stressful episode per day (in 10+ hours of sensor wearing time). 80

17 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. 87

18 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. 91

19 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*. Similarly for an intervention at 'b' we compute stress density upto 'b'. 92

20 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. 96

21 F1 score between self-report and sensor assessment range from 0.36 to 1.00 with median 0.65. 98

22 Time series of stress likelihood of one participant on pre-quit day. 102

23 Next stress duration as a function of current stress duration. A healthy correlation of 0.4243 is observed here. 104

24 State transition probabilities between different *stress* episode types, *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*. 105

25 Role of time and activity level on stress density. Here, morning is before 8 AM, day time is 8 AM to 7 PM, and night is after 7 PM. Horizontal (red) line represents the overall stress density. 106

26 Effect on stress density across different location contexts detected with $\kappa > 0.7$. Noisy environment is highly associated with stress. 108

27 The likelihood of stress for one participant overlaid on the disorder map. Disorder here is the aggregated posterior probability value for top 10 NIfETy variables (see Figure 26) with $\kappa > 0.70$. 110

28 Timing for proactive or reactive intervention. 111

29 Trade-off analysis for triggering frequency of stress intervention. The x-axis represents model-proposed triggering frequency of stress intervention per day and two y-axes represent precision and recall for predicting SSEs. 114

PUBLICATIONS

1. **Sarker, H.**, Hovsepian, K., Nahum-Shani, I., Murphy, S., Spring, B., Ertin, E., al'Absi, M., Nakajima, M., and Kumar, S. From markers to interventions-the case of just-in-time stress intervention. In Mobile Health: Sensors, Analytic Methods, and Applications, Rehg, J., Murphy, S. A., and Kumar, S., Eds. Springer, **(2017)**.

2. **Sarker, H.**, Tyburski, M., Rahman, M., Hovsepian, K., Sharmin, M., Epstein, D., Preston, K., Debra Furr-Holden, C., Milam, A., Nahum-Shani, I., al'Absi, M., and Kumar, S. Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data. In Proceedings of **ACM CHI (2016)**, 4489-4501. (13 pages) (Acceptance Rate = 23.4%)

3. Chatterjee, S., Hovsepian, K., **Sarker, H.**, Saleheen, N., al'Absi, M., Atluri, G., Ertin, E., Lam, C., Lemieux, A., Nakajima, M., Spring, B. mCrave: continuous estimation of craving during smoking cessation. In Proceedings of **ACM UbiComp (2016)**, 863-874. (12 pages) (Acceptance Rate = 24%)

4. Saleheen, N., Ali, A. A., Hossain, S. M., **Sarker, H.**, Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., and Kumar, S. puffMarker: A Multi-sensor Approach for Pinpointing the Timing of First Lapse in Smoking Cessation. In Proceedings of **ACM UbiComp (2015)**, 999-1010. (12 pages) (Acceptance Rate = 22%)

5. **Sarker, H.**, Sharmin, M., Ali, A., Rahman, M., Bari, R., Hossain, M., and Kumar, S. Assessing the Availability of Users to Engage in Just-in-Time Intervention in the Natural Environment. In Proceedings of **ACM UbiComp (2014)**, 909-920. (12 pages) (Acceptance Rate = 20.7%)

SUMMARY OF DOCTORAL WORK

I have joined University of Memphis as a Ph.D. student in Fall 2012. My doctoral research on mobile health (mHealth) is an interdisciplinary work on Data Mining, Machine Learning, Statistical Data Analysis, Wearable Computing, Human Computer Interaction (HCI), and Health Informatics. As a part of the large research initiative NIH Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K) my aim is to improve the health and well-being of a person by continuously monitoring his or her health status, recognize behavior, diagnose medical condition, and provide just-in-time-intervention (JITI) in the person's natural environment.

When I joined our research lab there was a newly funded R01 project from the National Institutes of Health (NIH). The aim was to identify the sensor-inferred risk factors for smoking lapses. A user study was about to begin at the University of Minnesota Medical school where we were about to launch a wrist watch containing a 3-axis accelerometer and a 3-axis gyroscopes. I immediately began working on the mobile phone software platform that our group had developed for the collection of high-frequency mobile sensor data from wearable sensors using ANT radio. I took the responsibility of integrating the wrist sensors on to our mobile phone software platform so that we can store wrist sensor data in the phone SD card and later analyze them. I conducted some preliminary analysis to show that smoking gestures have distinct signature in inertial sensor data collected from wrist-worn sensors (e.g., smartwatches). In contrast to some recent work that used 9-axis (3-axis accelerometer, 3-axis gyroscopes, and 3-axis magnetometer) to capture smoking gesture, I have shown that it is possible with 6-axis (3-axis accelerometer and 3-axis gyroscopes) which led to battery efficient smoking gesture recognition. I have also shown that smoking hand-to-mouth gesture has distinct signature in compare to other closely related gestures (e.g., eating). The software was used in the 75-person smoking cessation field study where participants wear a AutoSense containing a chest sensor collecting ECG, respiration, and accelerometer data and wrist sensor collecting accelerometer and gyroscope data. My work laid the foundation for the computational model for detecting smoking puffs from wrist movements and respiration patterns. This work was published and presented in ACM UbiComp'15 which later has been covered in various media outlets due to its ability to pinpoint the precise timing of first smoking lapses.

Next, I began analyzing the high-frequency mobile sensor data collected in various user studies in free living condition. I took several statistics courses that later helped me in analyzing high-frequency mobile sensor data. For example, I have shown in my Master's Thesis that there are sensor detectable contexts in our daily life in which user's self-report reliability may not be consistent and should be excluded from analysis.

My Master's work led me to a field of interruptibility where over a decade work exists. But I was more interested about extending these works for assessing availability of user's in their free living condition for just-in-time intervention and that too from sensor inference. Majority of the prior works on interruptibility use self-report for assessing interruptibility which is subjective by nature. In contrary, I have used the concept of an objective metric (response delay) for assessing availability. The use of the term availability in compare to interruptibility captures the notion that the user must be available physically, cognitively, and socially to be engaged in a just-in-time intervention. I have shown that people are most likely available when they are walking outside and least likely available while walking at work. This work was published at ACM UbiComp'14 and has been cited well in both computing and in behavioral science field.

This work gave me a direction towards my dissertation. I developed an overall framework for sensor-triggered just-in-time intervention. My availability work provides a model that assesses the opportune moment about when to deliver an intervention so that the user is available for significant user engagement. Another part is when to generate a trigger for an intervention via investigating the time-series of markers.

Next I started working on determining the timing of sensor-triggered intervention. As a specific case I started investigating sensor-triggered just-in-time stress intervention. I developed a time-series pattern-mining method to analyze the time series of minute-level stress markers obtained from physiological sensors. There are several design challenges for an effective just-in-time stress intervention. First, intervention should only be triggered when we have high confidence in sensor-interred stress assessments. Because, triggering too many stress interventions may interrupt a person in his or her daily life. So, we only need to provide intervention at the most opportune moments. Doing so is especially challenging because sensor-inferred stress measurements from physiological parameters are by their very nature rapidly varying and include intermittent missing data. Second, intervention should only be triggered to maximize its efficacy. Third, stress inference and the triggering of intervention occur in real time on resource-constrained and battery-operated wearable sensors and computing devices (e.g., smart phone and smart watch). Hence, the method should be computationally efficient.

My work addressed each of these challenges. First, I developed a method to deal with confounder, such as, physical activity which occurs frequently in our daily life. Second, I applied the cStress model, imputed the missing data, and validated the output of cStress model against self-reported stress. I found that for a small subset of participants for whom the agreement is poor between self-report and the sensor-inferred stress assessments, those participant’s self-report consistency is also questionable. Third, I trained a stock prediction model that is Moving Average Convergence Divergence (MACD) to locate the increasing and decreasing trend in the time series and identified the episodes in the time series. Then, I classified those episodes as stressed, unsure, not-stressed, and unknown. Fourth, I applied the model on a study data from 38 participants and 4-week long. I found that active day is more stressful in compare to morning or evening. I then analyzed the relationship between stress and objectively rated disorder in the surrounding neighborhood and develop a model to predict stressful episodes. This work appeared as an ACM CHI'16 paper.

This work generated widespread interest in the behavioral science community. Later, I collaborated with the behavioral scientists to adapt my stress intervention method on smokers who are going through abstinence. Stress in prevalent among this population. Management of stress is critical for the success of the abstinence. I adapted my method for this population where each participant wore the sensor suite during their abstinence period. I collaborated with Dr. Bonnie Spring from the Northwestern Medical School who is widely known for her behavioral interventions work and Dr. Susan Murphy, a National Academy member, who is widely recognized for her micro-randomized trial design. My methods are now being used to evaluate the first sensor-triggered stress intervention study in smoking cessation population at Northwestern Medical school.

Finally, I showed the feasibility of detecting brushing and flossing behavior from wrist-worn inertial sensors. This work led to a successful R01 grant application led by Dr. Vivek Shetty from the UCLA Medical School.

Chapter 1

Introduction

1.1 Background and Motivation

Mobile technology has a potential to provide unprecedented visibility into the health status of users in their natural environment [107]. Sensors embedded in smart phones (e.g., GPS, microphone), and wireless sensors worn on the body (e.g., electrocardiography (ECG), accelerometers) can continuously monitor an individual's health, behavior, and the surrounding environment. Machine learning algorithms have been developed to obtain measures of behavior and exposure to the environment such as activity from accelerometers, geo-exposure from GPS, stress from physiology, and social context from microphone. These automated measures of behavioral and environmental contexts enable the design of just-in-time interventions (JITI) to support maintenance of healthy behaviors.

The effectiveness of an intervention depends on the timing, content, and modality of the delivered intervention. Considerable amount of research have focused on the design and use of scheduled and context-sensitive interventions targeted to support maintaining healthy life-style [68, 97]. Scheduled interventions can trigger recurring motivational messages, instructions from caregivers, or medication reminders to maintain healthy behavior. However, such scheduled interventions lack the knowledge about receiver's context and therefore can become an additional source of distress [143]. In order to address this issue context-sensitive interventions are proposed [97]. Context-sensitive interventions extend this area of research by utilizing individual's contextual information such as location and activity. However, such context-sensitive interventions do not assess or consider the cognitive, physiological, or physical 'availability' of the individual to adapt to the content or modality of the triggered intervention. Research on scheduled and context-sensitive intervention thus can benefit from the knowledge of an individual's availability to engage in an intervention.

We use smoking cessation to illustrate the potential of JITI and the importance of timing the delivery of JITI. Smoking is responsible for most deaths in the US, accounting for one in five deaths [128]. Although a majority of daily smokers want to quit, according to Center for Disease Control, less than 10% actually succeed to quit. The highest lapse rate among newly abstinent smokers is in the first week which is observed to be over 50% of the total lapses [13]. Smoking lapse is impulsive and the first lapse usually leads to full relapse [166]. Hence, it is critical to help abstinent smokers break their urge when and where it occurs (within first few days of quitting). There has been some technological breakthroughs to address this problem. First, advent of wearable sensors now enables us to detect the precipitants who are vulnerable to lapse (e.g., due to stress [89, 145]). Sensors such as smart eyeglasses can also detect smoking cues (e.g., cigarette filters scattered around, other people smoking close-by) that can cause smoking urges. Second, the mobile phones connected with these sensors then can be used trigger a JITI to break the urge. However, such technology will succeed only if the user is available to be engaged with the intervention when the JITI is delivered. Otherwise, we may lose the precious opportunity to prevent the potent first lapse.

In addition to inferring the availability of users, successful JITI depends on correctly assessing the most appropriate timing of intervention via investigating the time series of the sensor inferred markers. We use just-in-time stress intervention to illustrate this case. Chronic stress induce vulnerability towards addiction [168] in both developing and lapse phase. 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. [10, 36, 125, 159, 164, 177]. Despite high prevalence of stress, 33% of Americans never discuss about managing stress with their health provider [1]. Recent advances in wearable sensors and computational modeling have made it feasible to obtain continuous assessment of stress in the natural environment [86, 89, 145]. Given the widespread adverse health consequences of stress (both in the short term and in the long term) [43, 115, 121, 124, 156], these advances hold tremendous promise to improve public health and well-being. But delivering a sensor-triggered stress intervention (e.g., breathing or relaxation exercises) is feasible only if there exists a method to detect significant stress episodes in real time that can be used to trigger the intervention at most opportune moments. To trigger a stress intervention, we need to locate significant stress episodes in the sensor data stream. One of the goal of this dissertation is to establish the foundation on which a just-in-time stress intervention can be developed.

1.2 Problem Statement

JITI is aimed at improving the user's health and require appropriate engagement of the user. Asking users to rate their availability in-the-moment can become an additional source of disruption. With the help of contextual information derived from sensor measurement, we need to develop a model to assess the availability of individuals for just-in-time-intervention in their natural environment.

To trigger a stress intervention, we need to locate significant stress episodes in the sensor data stream. This introduces several challenges.

- Stress measurements obtained from sensors usually have to be inferred from physiological data, which by their very nature rapidly varying, similar to real-time tracking of stock prices. Unlike stock-price data, the time series of stress is discontinuous due to factors such as sensor detachment and wireless losses [136, 151].

- Sensor measurements are frequently confounded by physical activity (23% of the time [151]), that needs to be filtered out for an accurate assessment of stress.

- The decision to trigger the intervention must be made quickly so that the intervention can be effective. Hence, simple methods that can be efficiently implemented on mobile devices are needed.

- Too-frequent prompts of an intervention can lead to alarm fatigue [96] and render the system useless. Ideally, the intervention policy should be personalized to the tolerance level of the individual and the frequency of intervention (e.g., once per day) desired by the user. Given a piecewise continuous time series of stress assessments we want to identify the precise timing for just-in-time stress intervention personalized to the individual preference of frequency.

- When an intervention is triggered, we should have high confidence in sensor-derived stress assessments.

- 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.

1.3 Summary Results

1.3.1 Availability for JITI

In this dissertation, we develop a model to predict availability in the natural environment. The proposed model is derived from data collected from a week-long mobile health study with 30 participants. The goal of this study was to investigate the relationship among stress, smoking, alcohol consumption, and their mediators (e.g., location, conversation) by measuring these via wearable sensors, rather than via self-reports. During the study, participants wore a wireless physiological sensor suite that collected ECG, respiration, and accelerometry, and carried a smart phone that included GPS and accelerometers. Participants were prompted by a smartphone to complete Ecological Momentary Assessment (EMA) self-reports consisting of 42 items, multiple times daily. Answering these 42-items required a level of engagement expected in JITI. Each EMA was associated with micro-incentive to encourage compliance [132].

To address the biases in human estimates of availability [20], we use delay in responding to EMA as an objective metric to measure the availability of a participant. When an EMA is prompted, a participant can act in five ways, (s)he can i) answer the EMA without delay, ii) answer the EMA within a grace period (around 2 minutes), iii) answer after the grace period iv) request to delay the EMA, and v) ignore and not answer the EMA at all. Acts iii), iv) and v) indicate unavailability in answering the EMA, while Act i) indicates immediate availability., while the first act shed light on situations where the participant was available to attend to the EMA. However, when users respond within 2 minutes, we are unsure about their availability states. Utilizing delay as a metric of availability, we seek to identify affective states of participants when they will be cognitively, physically, and physiologically available to engage in a JITI.

To predict availability, we use GPS traces to identify participants' location and driving state, infer their physical activity states from on-body accelerometers, and stress from ECG and RIP sensor data. In addition, we use time of day, day of the week, and self-reported affect, activity type, and conversation status. We compute a total of 99 features. We identify 30 most discriminating features and train a machine learning model to predict the availability of a user. We find that several features derived from sensors such as location, activity type, time, and day of the week, play significant roles in predicting availability. In particular, features derived from stress (inferred from physiological sensors) play a significant role in predicting availability. We find that the machine learning model can predict availability with 74.7% accuracy (against a base accuracy of 50%). This compares favorably against existing works on predicting interruptibility, where the prediction accuracy was reported to be 79.5% against a base accuracy of 70.1% in the office environment [65], and an accuracy of 77.85% against a base accuracy of 77.08% in the natural environment [147]. We find that users are usually available when walking outside of their home or work, or even if just outside of their home or work location. But, they are usually not available when driving or at work. We also find that participants are more available when they are happy or energetic versus when they are stressed.

1.3.2 Trigger for JITI

Although physiology is affected by several kinds of events in daily life, the main confounder for stress assessment is physical activity. To isolate data affected by activity, we first detect physical activity from chest-worn 3-axis accelerometer data, using an existing model [151]. Second, we estimate the time it takes for physiology to recover from the effect of a just concluded activity episode. Both data are then excluded. Physiological readings generally return to baseline within 2 minutes after the conclusion of physical activity (unless the activity is especially intense) [60]. However, the majority of activity episodes in our daily life are of short durations. Although our participants were physically active 22.7% of their sensor-wearing time, 95% of their activities lasted less than 2.1 minutes. Discarding 2 minutes of data after each activity episode would result in excluding 35.0% of additional data (for a total of 57.7% of all data). We, therefore, proposed a more systematic person- and situation-specific method to **estimate recovery time**. According to [69, 87], heart-rate after an arousal (e.g., activity) recovers exponentially. We have learned the exponential recovery rate (τ) for each participant to estimate the recovery duration once physical activity is over. Using this parameter, in addition to the entire physical activity interval, the estimated recovery interval that follows is excluded from analysis, i.e., considered missing for the purpose of stress inferencing. With this approach, only 7.4% of data (as opposed to 35%) are excluded due to recovery from physical activity, in addition to 22.7% that are directly affected by physical activity (for a total of 30.1% of all data). Computation of the recovery rate in the natural environment could also serve as an indicator of cardiovascular fitness, similar to the 6-minute walk tests [34, 152] done in clinics.

Standard methods for finding trends in time-series data [17, 35] require continuous data streams. To apply these methods, we needed a method to impute the missing data. Missing data in time series of stress assessments can be due to unavailability of data or due to presence of confounder such as physical activity. Before imputation, we need to rule out the possibility that the data are Missing Not At Random (MNAR) [53]. We found that missing data in stress assessments are not MNAR. Missing data could be considered Missing At Random (MAR) [39, 53] because stress can be explained by other known contextual variables [61, 70, 150] such as day of the week, time of day, previous stress levels, and the slope and intercept of previous time-series samples. We use these variables to impute the missing data using the K-Nearest Neighbor method proposed in [77, 169, 179].

Although cStress model was validated in both lab and field settings [89], before using it on two separate dataset obtained from polydrug users and smoking abstinent users, we validate it against their field self-reports using F1 score as metric. The participant F1 scores range from 0.130 to 0.917 with a median of 0.717. Although the F1 scores are acceptable for majority of the participants, there are 5 participants having low F1 score seem to suggest poor agreement between self-reported stress and the model output. We, therefore, analyze the consistency of their self-reports, because they may be subject to consistent bias or careless responding. We use Cronbach's alpha [27] to assess the consistency of the self-reported responses. The overall consistency score across all participant's self-reports is 0.843. We compute Cronbach's alpha for the 5 participants from Figure 13 who show poor F1 score. They have unacceptable self-report consistency scores with a median Cronbach's alpha of 0.335. Furthermore, the participant with the smallest F1 score (0.13) answered "3" on item "Nervous/Stressed?" in 173 out of 177 self-reports, suggesting a bias toward neutral self-assessment. These observations also demonstrate the value of an objective sensor-based model of stress.

We applied the cStress model [89], imputed the missing data, and validated the output of cStress (together with its imputation) against self-reported stress. Next, we trained a stock prediction method called Moving Average Convergence Divergence (MACD) [17] to locate the time of an increase in stress in rapidly varying continuous time-series data. We estimated the probability distribution of the likelihood of stress assessments and the probability distribution of stress durations (in the smoothed time series). We found that the likelihood of stress follow beta distribution with shape parameter $\alpha = 0.222$ and $\beta = 1.027$ and stress durations follow the LogNormal distribution with parameters $\mu = 2.064$ and $\sigma = 0.871$. To personalize the algorithm for each individual, the threshold on stress likelihood can correspond to tolerance level, and the duration can be selected to meet the daily intervention frequency preference. If, in a candidate window, the likelihood of stress crosses the high likelihood threshold and remains elevated for a threshold duration, then this window represents significant stress episode (SSE).

When we apply the above model to our dataset, we find that the duration of a stress episode predicts the duration of the next stress episode ($r = 0.42$). This correlation can be explained by theory and evidence [84, 85, 133] suggesting a spiral process where current exposure to stressors can lead to subsequent reactivity to other stressors by attenuating the state coping capability of the person. We find that stress in the morning and evening are lower than during the day (0.105 and 0.133 vs. 0.186). Our participants are more likely to be stressed after an activity episode (0.186 vs. 0.117). We assessed relationships between stress and the neighborhood environment with independently obtained data from the Neighborhood Inventory for Environmental Typology (NIfETy) [71]. We found that *noisy* locations; the presence of *graffiti*, *cigarette butts*, *trash in street*, and *bars* are associated with high stress likelihood. In contrast, locations where the NIfETy raters had seen *male adults involved in positive interaction* and *youth playing* are associated with lower stress than average.

We investigated the feasibility of predicting whether a rapid rise in stress would lead to a significant stress episode (SSE) from spatio-temporal context and the users' prior history. Proposed model is able to predict SSEs with a duration of 13.5 minutes with accuracy of 94.8% and $\kappa = 0.444$.

In addition, 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.

1.4 Key Contributions

In summary, this dissertation makes the following key contributions for just-in-time-intervention:

- Proposed a novel objective approach to determine user's availability to engage in a task which requires significant user involvement (as compared to [65, 90, 147]). Proposed a model that performs with 74.7% accuracy (over 50% base accuracy) and 0.494 kappa to predict availability in the natural environment using data collected from a real-life field study with wearable sensors. To the best of our knowledge this is the first study related to interruptibility which uses micro-incentives [132] to obtain a stronger indicator of unavailability.

- Taken first steps towards the development of JITI and develop time-series-pattern mining methods to detect significant stress episodes in discontinuous ambulatory data. Presented model can suggest the timing for just-in-time stress intervention in a real-time fashion on resource-constrained and battery-operated wearable sensors and smart phones.

- Validation of sensor inferred stress is the field setting is challenging due to lack of gold standard truth. Self-reported stress is commonly used for this validation [89]. Presented work show that lack of agreement between self-reported stress and sensor inferred stress is subject to inconsistent self-report which we observed for a small subset of the population.

- Sensor detachment or wireless signal loss causes missing data in the time series. Physical activity is a confounder for stress assessments which occurs frequently in our daily life. In addition to data missing due to variety of factors, discarding physical activity related confounded stress assessments; further increases missing data in the time series. This causes discontinuity in the time-series. There are several time-series pattern mining methods which require a continuous time series to analyze and predict trends. In order to obtain a continuous time series of stress assessments we need to impute the missing data. But we can't do imputation if missing data is Missing Not At Random (MNAR). We found that missing stress assessments are not MNAR.

- Heart rate increases during physical activity. At the end of activity period heart rate recovers exponentially. Proposed model-based approach in Chapter 6 is able to compute the recovery rate of a person in their daily life without active user engagement. Computation of the recovery rate in the natural environment could serve as an indicator of cardiovascular fitness, similar to the 6-minute walk tests [34, 152] done in clinics. This represents interesting future work opportunities.

- We analyzed the relationship between successive *stress* episodes. *Stress* episodes more likely to be of similar kinds in successive episodes. We found that occurrence of a *stress* episode increases the likelihood of future *stress* episodes. If a person is *not-stressed* in the current episode it is highly likely that next episode in the time series is also going to be a *not-stressed*. Similarly, if the person experiences a stress episode it is likely that the next episode is also going to be a stress episode. This can be explained by theory and evidence [84, 85, 133] suggesting a spiral process where current exposure to stressors can lead to subsequent reactivity to other stressors by attenuating the state coping capability of the person. For example, stressors such as facing financial troubles may decrease the person's stress coping capacity. This may lead the person to respond with subsequent stress to an event or an environment that would, in other circumstances, be easy to deal with, such as being in a noisy environment. Observations like these suggest that providing a stress intervention when the person experiences a *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.

1.5 Organization

Chapter 2, Sensor-Triggered mHealth Interventions, presents the vision and overall architecture of sensor triggered JITI. There are three major stages. First, sense the physiological signals via wearable sensors and mobile sensors in the person's free living condition. Second, analyze these physiological signals and obtain markers (e.g., stress). Third, investigate the time series of markers and act via providing just-in-time intervention.

Chapter 3, Trigger Generation for Sensor-Triggered Just-In-Time Intervention, discuss the two major components of this trigger generation. First, assess the availability of the user to be engaged. Second, investigate the time series of markers for generating an intervention trigger.

Chapter 4, Related Works, presents the works related to the timing of just-in-time-intervention. First set of works focuses on assessing the availability of individuals for JITI. Second set of works focuses on the timing to trigger just-in-time stress intervention.

Chapter 5, Determining Availability for Intervention, informs the necessity of successful user engagement for an effective JITI. This chapter proposes a novel objective approach to determine user's availability to engage in an intervention which requires significant user involvement.

Chapter 6, Identifying Stress Episodes Based on Field Stress Data, proposes a model to identify the stress episodes based on real-life stress-likelihood time series. It contains methods to deal with confounding physical activity and discontinuities in the time-series data, and identify significant stress episodes (SSEs) in the stress-likelihood time series.

Chapter 7, Identifying Stress Episodes Based on Lab Stress Data, proposes an alternate model to identify the stress episodes based on parameters computed from a lab based stress-likelihood time series where ground truth stress markers are available.

Chapter 8, Applications of Our Model, discusses about the applications of identifying stress episodes. First, trigger a self-report prompt to understand the causality of someone being stressed. Second, observe and identify the patterns of stress that will help intervention designers to devise appropriate intervention. Third, provide a proactive or a reactive intervention. Fourth, generalize the proposed method to other interventions, such as, via investigating the time series of craving for cigarette, food, or drug.

Chapter 9, Conclusion and Future Directions, concludes the dissertation and discusses about future research directions that is set by this dissertation.

Chapter 2

Sensor-Triggered mHealth Interventions

Mobile health (mHealth) aims to improve the health and well-being of a person by continuously monitoring his or her health status, recognize behavior, diagnose medical condition, and provide just-in-time-intervention (JITI) in the person's natural environment.

2.1 Vision of Sensor Triggered JITI

Mark Weiser, the scientist who introduced the concept of ubiquitous computing, believed that technologies should be designed to disappear [180] into the background and serve users by anticipating their needs. Ubiquitous computing can work in the background to assess users' physiological states, predict when users need help managing stress, and deliver just-in-time interventions (JITIs). Such prediction and prevention could improve health and quality of life for the entire society, given the ubiquity of stress in human society and its wide-ranging adverse impact on physical, psychological, behavioral, and social health.

Since the time smartphone was introduced, people throughout the world embraced this new technology. Smartphone sales surpassed feature phone globally in the year 2013 [3]. Pew data from 2015 indicate that 92% US households have a cell phone and 64% US adults now owns a smartphone [8].

In addition to mobile calling features available in a feature phone, smartphone consist of wide range of sensors. Motion sensors such as accelerometer and gyroscope can assess person's activity context. GPS, WiFi, barometer, and Bluetooth data can provide us person's location and social context [104]. Microphone capturing surrounding conversation can inform social interaction. Light sensors and temperature sensors can assess whether a person is indoor versus outdoor. Heart rate monitor or phone camera can provide us ECG data which eventually can be used to assess physiological stress [75]. These contextual information provides us unprecedented visibility into person's physical, physiological, behavioral, and social context in user's natural environment.

Wearable computing makes this mobile health technology pervasive by intertwining into our natural life. Smart watch, smart glass, sensors embedded in clothing, or other wearable on-body sensors sensors provide us rich information about person's context in daily life. App stores such as Android, iOS, and Amazon Appstore provides us a platform to write and deploy software. Software can collect phone sensor data, along with wirelessly collected physiological data from wearable sensors while person is doing their daily activity. Researchers can apply data mining and machine learning method on these collected data and infer adverse health behavior aiming towards just-in-time-intervention. smartphone application, such as, 6-min walk test (6MWT) can assess cardiovascular fitness of an individual [57, 152]. Internet connectivity of smartphone and wearables enables us to send these information to caregivers in real time. Solution of such a kind is leading towards detection of disease at a early developing stage in a proactive manner.

2.2 Overview of the Approach

Sensor-triggered mobile intervention has three main stages (see Figure 1). First, sense the physiological signals via wearable sensors and mobile sensors in the person's free living condition. Second, analyze these physiological signals and obtain markers (e.g., stress). Third, investigate the time series of markers and act via providing just-in-time intervention.

2.2.1 SENSE

First stage is the acquisition of data by sensing physiological parameters from wearable sensors in the user's free living condition.

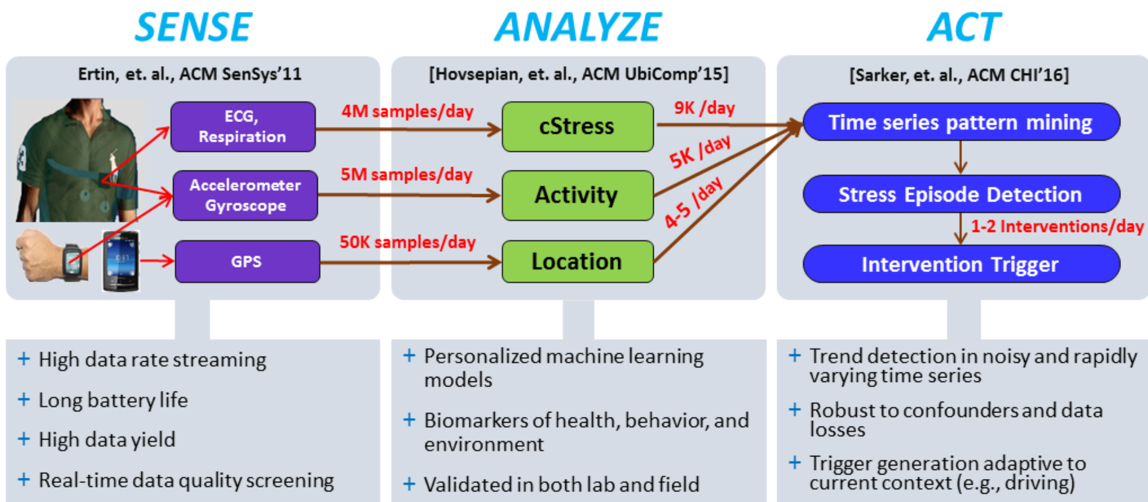


Figure 1: Three stages of sensor-triggered intervention delivery process. First, sense using wearable sensor suite AutoSense [59] and a smart phone. Second, develop a computational model to analyze physiological data acquired from the first stage and assess stress [89]. Third, obtain stress time series, identify *stress* episodes, and act via triggering intervention at appropriate moments. This third stage is the main topic of this dissertation.

2.2.1.1 AutoSense Sensor Suite

Sensor suites, such as, AutoSense [59] can sense physiological signals. During the study, participants wear this wireless suite of physiological sensors under their clothes. The sensor suite consisted of an unobtrusive, flexible band worn around the chest. It provided respiration data by measuring the expansion and contraction of the chest via inductive plethysmography (RIP) and included a two-lead electrocardiograph (ECG), and a 3-axis accelerometer. The measurements were transmitted wirelessly using ANT radio [5] to an Android smartphone. The sampling rates for the sensors were 128 Hz for ECG, 64 Hz for respiration, and 32 Hz for each accelerometer axis. They were downsampled at the sensor before wireless transmission at the rate of 28 packets/second, where each packet has 5 samples. There are approximately five million samples per day. This high enough frequency of physiological sensor data suffices for continuous assessment of stress.

2.2.1.2 Mobile Phone

Participants carrying a smart phone have four roles. First, it robustly and reliably receive and store data wirelessly transmitted by the sensor suite. Second, it store data from GPS and accelerometers sensors in the phone. These measurements are synchronized to the measurements received from wearable sensors. Third, participants use the phone to complete system-initiated self-reports in the field (e.g., drinking and smoking events). Fourth, at random moments smart phone prompt participants to complete Ecological Momentary Assessment (EMA) self-report items. These items consist of affect items, single choice items, multiple choice items, or recall based items. Most of these EMA items are not readily detectable from sensor data (e.g., happy?), while some are used for the validation of sensor inference (e.g., stressed?). There is also provision to associated micro-incentive with each EMA to encourage compliance [132].

2.2.2 ANALYZE

The second stage involves analysis and modeling of this high volume physilogical sensor 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, behavior, and environment (e.g., stress [89], activity [151], and location [104]). Validation of these markers is also challenging, as in most of the cases there is a lack of gold standard while users are in their free living condition.

2.2.2.1 cStress Model for Stress Assessment

The *cStress* [89] 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 [89]. 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 [137]. 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 [140]. 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* [89]. 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) [89]. In contrast to other stress inference works, such as [122, 123], 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 [89]. In [158], 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.

2.2.2.2 Activity Inference from Accelerometer

To infer whether a subject is in motion or not, we use a simple threshold based activity detector using the 3-axis on-body accelerometer (placed on chest). Phone accelerometer data was not used because the phone may not be on the person and thus may miss some physical activity. We adapt the physical movement detection approach in [19, 141]. As the placement of the accelerometer and the participant population is different from that presented in prior works, we used an existing approach proposed in [151] to infer activity. Training data was used to determine an appropriate threshold for detecting activity. There was labeled data under walking and running (354.16 minutes), and stationary

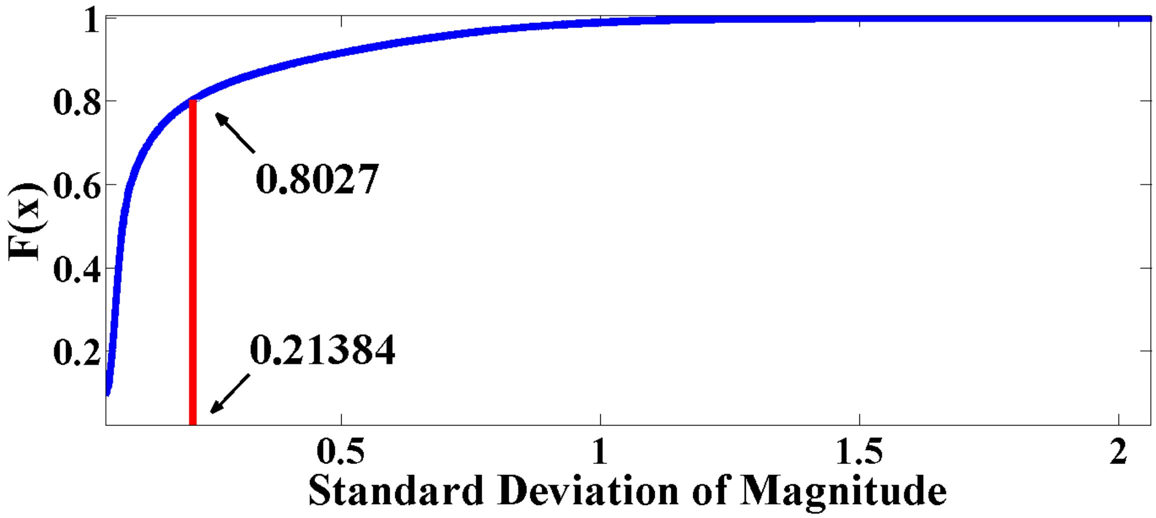


Figure 2: Using cut-off point 0.21384 we observe that subjects were physically active for around 20% of their total wearing time.

Table 1: Confusion Matrix for the Semantic Labeling model [104]. Restaurant is sometime confused with store. Precision and recall related performance metrics are available in Table 2.

	Classified as				
	Home	Work	Store	Restaurant	Other
Home	617	11	10	0	4
Work	12	708	1	0	1
Store	8	7	203	6	9
Restaurant	4	1	43	27	3
Other	62	14	40	1	96

(1426.50 minutes) states from seven pilot participants who wore the same sensor suite. We filtered the raw signal, removed the drift, and extracted the standard deviation of magnitude, which is independent of the orientation of the accelerometers and recommended in literature [19, 141]. The distinguishing threshold for our accelerometer to be 0.21384, which is able to distinguish stationary from non-stationary states with an accuracy of 97% in 10-fold cross-validation. Figure 2 shows that subjects were physically active for around 20% of their total wearing time.

Table 2: Accuracy using semantic labeler model [104]. TP = true positive rate, FP = false positive rate, P = precision, R = recall, F = F-Measure, and AUC = area under the curve.

	TP	FP	P	R	F	AUC
Home	0.95	0.08	0.86	0.95	0.90	0.98
Work	0.97	0.03	0.96	0.97	0.97	0.99
Store	0.79	0.06	0.67	0.79	0.73	0.96
Restaurant	0.37	0.01	0.60	0.37	0.46	0.92
Other	0.43	0.02	0.77	0.43	0.55	0.89
	0.86	0.05	0.85	0.86	0.85	0.97

2.2.2.3 Inference of Semantic Location

Locations of interest and their semantic labels are determined from GPS traces that were collected on the phone. Figure 3 shows a typical GPS trace of a participant for one day. Places of interest for a participant were places where the participant spent a significant amount of time. We first apply a clustering algorithm to the GPS data using the method proposed in [130]. Distance threshold of 100 meters and temporal threshold of 5 minutes are used to find the spatio-temporal clusters throughout the day for each participant. These clusters represent the locations of interest. Next, we assign semantic labels to these locations using Semantic Context labeler from [104].

Label assignment is based on demographic, temporal and business features. Demographic features include the age and gender of the participant, which are obtained from recruitment forms. The temporal features include *the arrival time, visit midpoint time, departure time, season, holiday,* and *the duration of stay* at that location. These features were computed from the GPS traces and clusters. Lastly, the business features include the count of different types of business entities such as *Arts/Entertainment, Food/Dining, Government/community, Education,* etc. within different distance thresholds from the current location (see [104] for details). To compute the business features, we

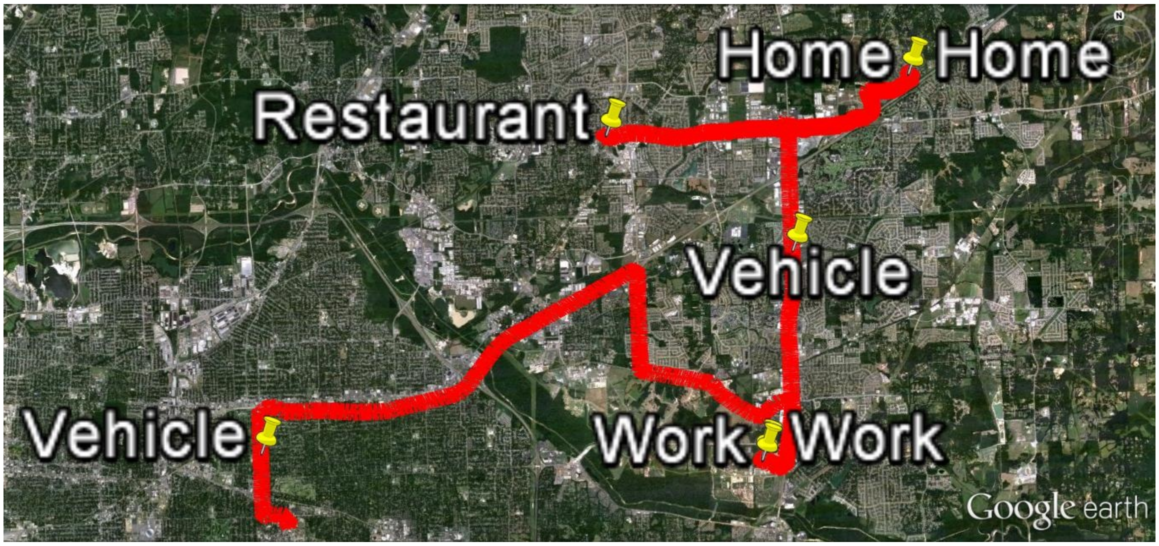


Figure 3: A sample GPS trace for one day from a participant. The red line shows the path commuted by the participant. The pinned locations are the location at the time of EMA prompt.

used Google Places API. For this model, we obtain an accuracy of 85.8% (and $\kappa = 0.80$). Table 1 presents the confusion matrix for this semantic context labeler model where F-measure is 0.85 and area under the curve is 0.97. We observe that *Home, Work,* and *Store* are detected quite well. But, *Restaurant* is confused with *Store*, because a *Store* and a *Restaurant* can be co-located. We correct the labels (if necessary) by plotting the GPS traces in Google earth and by visually inspecting it. These location labels were considered as ground truth. But, in some cases we could not reliably distinguish between a store and a restaurant (due to inherent GPS inaccuracy). We discard these data points by marking them unknown.

We also obtain a detailed level of semantic labeling. For *Home*, detailed label can be *Indoor Home, Dormitory,* and *Backyard*. Figure 4 shows a detailed breakdown of the labels. Our labeling concept of these details evolved over time [106] (e.g., by adding new levels). Hence, we made multiple iterations to obtain consistent labels.

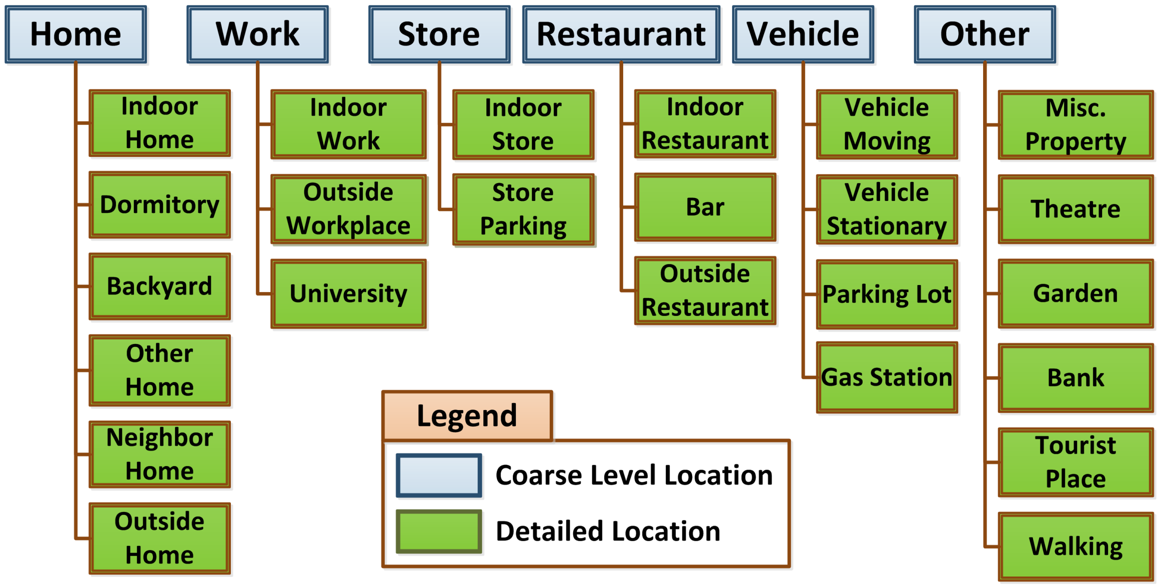


Figure 4: Two level semantic labeling of GPS clusters.

2.2.2.4 Driving Detection from GPS

Driving is detected from GPS-derived speed and by applying a threshold for maximum gait speed of 2.533 meters/sec [29]. A driving session is composed of driving segments separated by stops, e.g., due to a traffic light being red. Stops usually are of short duration unless there is a congestion. The end of a driving session is defined as a stop (speed=0) for more than 2 minutes. Otherwise, two driving segments sandwiched by a less than 2 minute stop is considered to be part of the same driving session. In case of loss of GPS signal for more than 30 seconds we also end the driving session at the timestamp when we received the last GPS sample. In order to determine whether participant is driving or just riding a vehicle we use the EMA question *"If you commuted since the last interview, what type?"*, where possible responses are *"Driving"*, *"Biking"*, *"Walking"*, *"Riding as a Passenger"*, *"Riding Public Transportation"*, and *"Did not commute"*. Finally, if an EMA prompting time is between start and end of a driving session, and the self-report response mentions *"Driving"*, we mark that EMA to occur during driving.

This stage reduces the data from 5 million per day to approximately 10 thousand samples per day.

2.2.3 ACT

The third stage is just-in-time intervention via investigating the markers obtained from the second stage. The success of just in time intervention depends on the timing of intervention that is “when” to provide an intervention?, the content of intervention that is “what” to provide in that intervention? And the modality that is “how” to provide an intervention. To identify the appropriate content and modality of intervention we first need to identify the timing for intervention. This proposed work address this timing part of the JITI that enables behavioral scientist to find appropriate content and modality.

2.2.3.1 Content of Intervention

In a smoking cessation program caregiver can provide short message (e.g., SMS, leaflet) to the quitters to guide them in the abstinence phase. Research show that sending materials tailored to each individuals increases efficacy of the program [135, 172]. Because, those who receive the tailored material perceive it as being written especially for them, and read it thoroughly in compare to those who received non-tailored material. Text message tailored to each individual contains information, such as, risks of smoking, monetary costs of smoking, social norms of smoking, outcome expectancies, and motivation to increase the impact of the intervention and reduce smoking [78, 135].

In an automated coaching system for stress reduction system sets goal for individuals (e.g., exercise, mediation, and accessibility). Users may fail to achieve daily stress reduction goals in case goals are too easy or too difficult to complete. An effective theoretically grounded system can set adaptive goals based on the individual’s prior performances [102] and increase compliance [102].

Interventions, such as, social networking, playing games, guided acupressure, and guided breathing are effective for stress reduction [143].

Self-reflection or biofeedback is another category of content for intervention. Biofeedback has been used clinically in relaxation skills training aiding us to help reduce stress related symptoms [32]. Relax2Win is a biofeedback game [163] where player control their character to go faster by being more relaxed (EDA level). [175] investigates the effects of using biofeedback as visual stress indicator during video-mediated collaboration between instructor and worker. Instructors and workers using the biofeedback as compared to using interfaces with facial view, reported lower mental workload and stress. In [162], authors developed a visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. AffectAura [120] is an emotional prosthetic that allows users to reflect on their emotional states over time. System logs physiological state using audio, visual, sensors, and user activities and aims to support reflection via visualization. Visualization is replaced by a wearable butterfly in [113] that helps users reflect on their stress level and regulate it. A self-reflective visualization of Blood pressure (BP) can enable users associate among stress, food, and daily routines [98].

An animated conversational agent on a wallmounted display can act as a virtual exercise advisor [26]. In [46, 178] a group of users share their step counts with each other via mobile phone. In case of daily goal is met system provides rewards, such as, a symbol next to the user's step count. *UbiFit Garden* [47] encourage individuals for physical activity. System uses on-body sensing for real-time assessment of physical activity. Smart phone glanceable display shows a wallpaper of a garden with butterflies. System sets a weekly goal of physical activity. Small butterflies in the garden indicate recent goal attainments, while large butterfly indicates this week's goal is met. The absence of flowers means no activity in current week. In a similar study Fish‘n’Steps [109] sets personal goal of activity. User’s foot step count is linked to the growth, and activity of an animated virtual fish in a virtual fish tank containing fishes of other users.

2.2.3.2 Modality of Intervention

Mobile phone SMS based intervention aim to motivate participants about negative consequence of adverse health practice (e.g., smoking) and helps individuals guiding in the cessation phase [78, 135, 172]. In a systematic review [54] about the usage of mobile devices for mobile health care, authors listed five smoking cessation studies. In each of the cases participant group receiving mobile phone based (mostly text messaging) intervention are more likely achieve abstinence comparing with those under controlled condition.

Intervention systems [26, 47, 109] use computer or mobile phone based virtual assistant as a modality of intervention to encourage individuals for physical. In comparison to a virtual assistant or text based intervention, a socially intelligent robotic personal assistant is more effective when an intervention is provided [111]. Socially intelligent robots can show empathy and compassion which led to better compliance [100]. In addition, robotic interaction is more natural and trusted in compare to a virtual assistant or text based interface. Because many people are habituated in getting information from a computer or mobile screen rather than receiving it in spoken text from robot with an ability to express emotions.

In [38] authors presented a JITI system to prevent emotional food intake. Another example is [149] that proposed a system where earpieces (to monitor chewing and swallowing), augmented-reality glasses (for capturing food consumed), and a physiological sensor (for heart rate) are connected to a mobile-phone application that processes the data and gives feedback to the user.

MoodLight [117] system uses ambient light as a modality of stress intervention. The ambient light is interactive and provides visual representation of momentary experiences of arousal (e.g., stress) by changing its color from blue to red. Electro-Dermal Activity (EDA) sensors collect biometric data about the current arousal level of an individual and change the color of the light accordingly to show internal state of individuals and increase awareness. Authors were motivated by a belief that cultivating self-awareness is one of the most important potential contributions of affective computing to the problem of stress management.

Lee et al. [108] developed a patina engraving system which engraves patina-like patterns on an wrist wearable fashionable activity tracker according to user's activity logs. Patina motivates participants to increase physical activity for engraving aesthetic patinas and triggers spontaneous social interactions.

Smart glass based system can provide intervention by facilitating visually impaired persons to engage in a social conversation [16]. System provides feedback to the visually impaired person about the facial expression and emotional state of the person he or she is interacting with.

2.2.3.3 Timing of Intervention

Successful delivery of the content of an intervention using a modality is critically depends on the appropriate timing of intervention. As a first step towards identifying the timing of intervention, we need to detect adverse health practices of individuals in their natural environment.

Inertial sensors embedded in the smart watch can track wearer's wrist movements and has capability to provide us health related information. We can detect whether a person is smoking [142, 155], eating [176], doing physical activity [7], or driving [99, 110] tracking these wrist movements. In addition to inertial sensors smart watch also contains Photoplethysmographic (PPG) sensors. PPG signals are obtained by pulse oximeter. It illuminated wearer's skin via a light-emitting diode (LED) and measures the change in light absorption in the skin. The periodicity of this PPG signals indicates cardiac activity. We can estimate heart-rate (HR) from this signal [187] leading towards the detection of physiological stress [89].

Visual exposure to specific cues induce craving for adverse health practice. Persons going through a smoking abstinent phase will feel strong urge to smoke viewing a tobacco advertisement or an outlet. Similarly, gazing at a lucrative advertisement of palatable food in front of a restaurant will most likely induce craving for food among people who are going through dietary restriction (e.g., obese). An outward camera in the smart glass can record the video of the surroundings and assess such visual exposures [62]. Incorporation of an inward camera can detect where a person is gazing at in the field of view. System can detect exposure to cue (e.g., smoking, drug, and food) and can provide intervention in real time. In addition, an inward camera in the smart glass can track the eye movement [118, 119, 186] and can assess health condition, such as, Parkinsons [24, 31, 170], Alzheimers [129], Autism [160], and others. Visually impaired person can get non-verbal cue during a social interaction (e.g., smile and yawn) [16] enabling them better engaged in a social conversation.

Wearables such as Zephyr BioHarness [2] and AutoSense [59] can collect ECG, respiration, and skin temperature in wearer's natural environment. Systems have been developed to detect smoking [14, 155], drug intake [88, 134], mental workload [44], and physiological stress [89, 145] from collected physiological data. Iqbal et al. [92, 93] proposed a measure of mental workload for interactive tasks using of task-evoked pupillary response. The magnitude of the pupillary dilation appears to be a function of processing load, or the mental effort required to perform the cognitive task. Functional Near Infrared Spectroscopy (fNIRs) measures blood flow in different part of the brain and can be used as a metric of mental workload [82]. Chest worn camera can take picture of food and assess the calorie intake of the person [174]. Sensors embedded in the clothing can collect physiological signal (e.g., ECG, respiration, and posture) in an unobtrusive way and reduce burden of wearing additional wearables [73, 131, 141, 183].

A home reminder system for medication and healthcare was reported in [97]. Smart home and wearable sensors were used to identify a person's contextual information for triggering an intervention. A similar study was conducted using smart home sensors to remind patients about their medications in [79], which considered availability of the patient when triggering a prompt, e.g., the system did not trigger a reminder when the patient was not at home, was in bed, or was involved in a phone conversation.

A majority of these works so far are about detecting markers related adverse health conditions from ubiquitous sensors. Presented work in this dissertation is a bridge between this detection part and the delivery of the intervention.

Chapter 3

Trigger Generation for Sensor-Triggered Just-In-Time Intervention — The Case of Stress Intervention

As a next step towards sensor-triggered just-in-time intervention we need to identify the timing for intervention. In this dissertation, we use stress intervention as a running example because of its well-known adverse effect on health. For example, depression, heart disease, diabetes, and addictions [10, 11, 43, 55, 56, 80, 124, 125, 153].

There are two parts in timing. First, identify when to deliver based on whether the person is available for a task like intervention which requires significant user involvement. Chapter 5 discusses about this availability part. Second, identify the timing about when to generate a trigger for an intervention via investigating the patterns in the time series. This is a noisy and rapidly varying time series which can be attributed three major factors. First, participants wearing the sensors may face problems, such as, intermittent loosening, improper attachment, jerks, wireless packet loss, etc. Second, machine learning models that obtain markers from the time series of physiological data are rarely perfect and may contain inaccuracies. Third, physiology can be confounded by several events in our daily life. The next major challenge is missing data due to wireless signal loss or due to presence of confounder. In addition, stress assessments and the triggering of interventions occurs in real time on resource-constrained and battery-operated wearable sensors and smart phones. Therefore, the computational model for providing just-in-time stress intervention needs to be efficient computationally and in power consumption. Chapters 6 and 7 propose methods about addressing each of these challenges of this triggering part. This stage reduces the data (markers) from ten thousand per day to usually 5 or less (interventions) per day.

Chapter 4

Related Works

In this chapter, we will discuss two broad categories of literature that is related to the timing for JITI. First, we will discuss close related works that assess the availability of an individual for Just-In-Time-Intervention (JITI). Second, we will discuss related works about generating the trigger for intervention.

4.1 Availability for Intervention

Research on interruption is closely related to the availability of individuals for intervention. A majority of research in this area primarily focuses on the impact of interruption in workplace. Information available from sensors installed in the workplace (e.g., door magnetic sensor, keyboard/mouse logger, microphone) can be used to develop a machine learning model to assess person's interruptibility level at workplace [65, 90]. Research on interruption in the natural environment has primarily focused on determining the receptivity of a user in receiving a phone call [83]. Interruptibility has been assessed from Ecological Momentary Assessment (EMA) [65, 83, 90] which can be biased due to number of factors, such as subjective biases, urgency of complete, lack of motivation, and lack of attention [28, 165, 184]. To address such challenges, objective measure such as the state of the ring tone of person's phone was used to assess his/her interruptibility level [63, 147]. We argue that the state of the phone ringer is a broad measure of the interruptibility of a user and it does not assess the availability of the user to engage in JITI.

Presented work complements prior works in several ways. First, a majority of the related works either recruit volunteer with no compensation [63, 147] or a fixed compensation [83, 91, 115, 126] for participation. Research shows that introduction of micro-incentives in scientific studies helps achieve better compliance with the protocol [132]. Our work uses micro-incentive to enhance participant's motivation and compliance. In addition, a person unavailable even though he or she may sacrifice the micro-incentive, provides us a stronger measure of unavailability. Second, we use wearable sensors in addition to smart phone sensors. Third, we report higher accuracy of 74.7% and a kappa of 0.494 compared to prior works [147]. Finally, to the best of our knowledge, this is the first work to directly inform the timing of delivering EMA prompts in scientific studies that use micro-incentives.

We will discuss the details of the related works related to availability for intervention in Chapter 5 section 5.2.

4.2 Trigger Generation for Intervention

The first step towards finding the timing for just-in-time stress intervention is detection of stress in a continuous manner. Recent advances of wearable sensors and mobile sensors enables us to get physiological parameters (e.g., ECG) to detect stress [59, 75, 81, 173]. Stress detection can be done from a variety of physiological parameters including ECG and respiration [89, 145], electrodermal response [117], photoplethysmography from fingertip [112], or near-infrared spectroscopy from forehead [82]. Our method of generating trigger for JITI can be applied to stress measurements obtained from any of the above methods.

Self-reflective visualization of person's current stress level along with other associated contextual information helps users to manage theirs stress [113, 120]. MoodLight [117] finds episodes of arousal from electro-dermal activity (EDA) in the lab environment. System regulates the color of a desk lamp to reflect the user's current stress level. When users reduce their stress level, the light color changes from red to blue. In [95], the authors present a method to predict the time series of heart-rate variability (HRV) using a first-order Hidden Markov Model. The algorithm was tested in a simulated patient environment using a beta distribution ($\alpha = 0.1$ and $\beta = 1$). In contrast to these works, our model addresses real-life challenges of discontinuity and rapid variability in the time series of stress assessments.

We will discuss the details of the related works related to the timing for just-in-time intervention in Chapter 6 section 6.2.

Chapter 5

Determining Availability for Intervention

5.1 Introduction

The success of Just-In-Time-Intervention (JITI) depends on successful user engagement. Intervention will succeed only when the recipient user is available physically, cognitively, and socially to attend to the intervention. Assessing availability is even more critical in a smoking cessation program. Smoking is responsible for every one in five death in US [128]. Although the majority of the daily smokers want to quit, 90% of them eventually relapse according to the Center for Disease Control. The first week of the abstinence is the most critical period for the newly abstinent smokers when over 50% of them lapse [13]. Smoking lapse is impulsive and the first lapse usually leads to full relapse [166]. Hence, it is critical to help abstinent smokers break their urge when and where it occurs (within first few days of quitting). Although wearable sensors now provide us an ability to detect the potential precipitants (e.g., stress [145] or smoking cues detected via smart eyeglasses) and trigger a JITI to break the urge, it will succeed only if the user is available to be engaged when the JITI is delivered. Otherwise, we may lose the precious opportunity to prevent the potent first lapse. Hence, timing a JITI is critical.

Considerable research have been conducted in a closely related topic of interruptibility [64, 96]. These works largely aim to detect interruptibility of a user at workplace by analyzing the user's computer activity (e.g., key stokes), workplace status via audio and/or video capture of the workplace, phone status, and physical activity status via wearable sensors. Research on interruptibility provides insights about tasks or social contexts where a person is more interruptible, however, lessons from these studies cannot adequately guide the design of JITIs. This is because, unlike the case of interruption that may disrupt concentration of a task, JITI is aimed at improving the user's health and require appropriate engagement of the user. Further, these works asked users to rate their availability in-the-moment. Such reports are subjective, can become an additional source of disruption, and do not assess user's capability to engage in a JITI. It is desirable to investigate whether availability of an individual can be inferred objectively from data collected using lightweight wearable sensors. We also investigate whether context of the user such as stress (inferred from physiological sensors) and activity level (inferred from accelerometers) can be utilized to infer user's availability. Moreover, there is a subtle difference between interruption and intervention. Interruption is for the the benefit of interrupter (e.g., marketing call) while Intervention is for the benefit of receiver (e.g., smoking cessation).

In this chapter, we developed a model to predict availability in the natural environment. The proposed model is derived from data collected from a week-long mobile health study with 30 participants. The goal of this study was to investigate the relationship among stress, smoking, alcohol consumption, and their mediators (e.g., location, conversation) by measuring these via wearable sensors, rather than via self-reports. During the study, participants wore a wireless physiological sensor suite that collected ECG, respiration, and accelerometry, and carried a smart phone that included GPS and accelerometers. Participants were prompted by a smartphone to complete Ecological Momentary Assessment (EMA) self-reports consisting of 42 items, multiple times daily. Answering these 42-items required a level of engagement expected in JITI. Each EMA was associated with micro-incentive to encourage compliance [132].

To address the biases in human estimates of availability [20], we used delay in responding to EMA as an objective metric to measure the availability of a participant. To predict availability, we used GPS traces to identify participants' location and driving state, infer their physical activity states from on-body accelerometers, and stress from ECG and RIP sensor data. In addition, we use time of day, day of the week, and self-reported affect, activity type, and conversation status. We computed a total of 99 features.

We identified 30 most discriminating features and train a machine learning model to predict the availability of a user. We found that several features derived from sensors such as location, activity type, time, and day of the week, play significant roles in predicting availability. In particular, features derived from stress (inferred from physiological sensors) play a significant role in predicting availability. We found that the machine learning model can predict availability with 74.7% accuracy (against a base accuracy of 50%). This compares favorably against existing works on predicting interruptibility, where the prediction accuracy was reported to be 79.5% against a base accuracy of 70.1% in the office environment [65], and an accuracy of 77.85% against a base accuracy of 77.08% in the natural environment [147]. We found that users are usually available when walking outside of their home or work, or even if just outside of their home or work location. But, they are usually not available when driving or at work. We also find that participants are more available when they are happy or energetic versus when they are stressed.

In summary, presented work makes the following contributions: 1.) we propose a novel objective approach to determine user's availability to engage in a task which requires significant user involvement (as compared to [65, 90, 147]), 2.) we propose a model with 74.7% accuracy (over 50% base accuracy) and 0.494 kappa to predict availability in the natural environment using data collected from a real-life field study with wearable sensors, and 3.) to the best of our knowledge this is the first study related to interruptibility which uses micro-incentives [132] to obtain a stronger indicator of unavailability.

We note that EMAs are widely used in scientific studies on addictive behavior [127, 161, 165], pain [171], and mental health [15, 114, 139, 181]. While EMAs have obvious benefits, prompting EMAs at inopportune moments can be very disruptive for the recipients' current task [154] or social situation [21, 154]. The work presented here can directly inform the appropriate timing for delivering EMA prompts.

5.2 Related Works

In an era of mobile computing and ubiquitous sensors, we have unprecedented visibility into user's contexts (e.g., physical, psychological, location, activity) and this awareness can be used to guide the design of interventions.

A home reminder system for medication and healthcare was reported in [97]. Smart home and wearable sensors were used to identify a person's contextual information for triggering an intervention. A similar study was conducted using smart home sensors to remind patients about their medications in [79], which considered availability of the patient when triggering a prompt, e.g., the system did not trigger a reminder when the patient was not at home, was in bed, or was involved in a phone conversation. A context sensitive mobile intervention for people suffering from depression was developed in [37]. Data from phone sensors such as GPS and ambient light, and self-reported mood were used to infer contextual information of the patient and predictions were made about future mental health related state to trigger an appropriate intervention. A system to assist diabetes patients was reported in [148] to keep track of their glucose level, caloric food intake, and insulin dosage by logging user contexts (e.g., location from GSM cell tower, activity) and used these logged data to learn trends and provide tailored advice to the user. This thread of research highlights the tremendous capabilities and utility of mobile sensor inferred context-sensitive interventions. However, research in this area focuses primarily on determining the time of triggering the intervention. A timely intervention may still not be effective if the receiver is not available physiologically or cognitively to engage in that intervention. Thus, assessing the cognitive, physical, and social availability of a user in the natural environment will extend and complement research in this area.

Research on interruption is closely related to availability of an individual. A vast majority of research in this area focused on understanding the impact of interruption in workplaces. A feasibility study for detecting interruptibility in work environment used features extracted from video capture (a simulated sensor) [90]. Subjective probe of interruptibility in Likert scale was converted to binary labels of interruptible and highly non-interruptible. A machine learning model was able to classify these states with an accuracy of *78.1% (base=68.0%)*. An extension of this research used sensors (e.g., door magnetic sensor, keyboard/mouse logger, microphone) installed in the office [65], which improved the accuracy to *79.5% (base=70.1%)*.

These studies provide insights on interruptibility in carefully instrumented controlled environment (i.e., office), but may not capture the user's receptivity outside of these environments. For instance, a smoking urge may occur outside of office setting, where most of the above used sensors (e.g., video, keystrokes, etc.) may not be available. In addition, the approach of probing users at regular intervals to gauge their interruptibility may not indicate their true availability due to subjective biases as pointed out in [63].

Research on interruption in the natural environment has primarily focused on determining the receptivity of a user to receive a phone call. In [83], a one-day study was conducted with 25 users who wore accelerometers and responded to prompted EMA's on whether they are currently receptive to receiving phone calls. Using accelerometers to detect transition, it is shown that people are more receptive during postural transition (e.g., between sitting, standing, and walking).

The first work to use an objective metric was [63] that conducted a week-long study with 5 users. It collected the moments when users changed their ring tones themselves and also in response to a prompt generated every 2 hours. By using phone sensors (e.g., GPS, microphone, accelerometer, proximity) to infer phone posture, voice activity, time, and location, and training a person-specific model, it was able to predict the ringer state with an average accuracy of 96.1%. The accuracy dropped to 81% if no active queries were used. We note that predicting the state of ringer is a broad measure of the interruptibility of a user to receive calls and it does not indicate the user's availability to engage in a JITI.

The closest to our work is a recent work [147] that conducted a large-scale study (with 79 users) to predict user's availability to rate their mood on 2 items when prompted by their smartphones. The prompt occurred every 3 hours, if the phone was not muted. The notification is considered missed if not answered in 1 minute. The users can also actively reject a notification. A model is developed based on phone sensor data (location provider, position accuracy, speed, roll, pitch, proximity, time, and light level) to predict availability. It reports an accuracy of 77.85% (base=77.08%, kappa=0.17), which is only marginally better than chance.

The work presented here complements and improves upon the work reported in [147] in several ways. First, [147] recruited volunteers without any compensation. Other works in the area of interruptibility also either used no compensation [63, 147] or a fixed compensation [83, 91, 115, 126] for participation. Micro-incentives are now being used in scientific studies to achieve better compliance with protocols [132]. Ours is the first work to use micro-incentive to enhance participant's motivation. In [147], participants answered only 23% of the prompts (1508 out of 6581), whereas in our study participants responded to 88% of the prompts (2394 out of 2717) within the same 1 minute cutoff used in [147].

This is despite the fact that our EMA's are more frequent (upto 20 per day) and require a deeper involvement (to complete 42 item questionnaires), which may be the case with JITI that require frequent and deeper engagement. Therefore, our work complements all existing works by providing a stronger measure of unavailability, not considered before. Second, we use wearable sensors in addition to a subset of smartphone sensors used in [147]. Third, we report a significantly higher accuracy of 74.7% (over 50% base accuracy) and a kappa of 0.494 compared to [147]. Finally, to the best of our knowledge, this is the first work to directly inform the timing of delivering EMA prompts in scientific studies that use micro-incentives.

5.3 Study Design

In this dissertation, we analyzed data collected in a scientific user study that aimed to investigate relationship among stress, smoking, alcohol use, and their mediators (e.g., location, conversation, activity) in the natural environment when they are all measured via wearable sensors, rather than via traditional self-reports. The study was approved by the Institutional Review Board (IRB), and all participants provided written informed consents. In this section, we discussed participant demographics, study setup, and data collection procedure.

Participants: Students from a large university (approximately 23,000 students) in the United States were recruited for the study. Thirty participants (15 male, 15 female) with a mean age of 24.25 years (range 18-37) were selected who self-reported to be “daily smokers” and “social drinkers”.

Wearable Sensor Suite: Participants wore a AutoSense wireless physiological sensor suite underneath their clothes. The wearable sensor suite consisted of two-lead electrocardiograph (ECG), 3-axis accelerometer, and respiration sensors. A description of this AutoSense sensor suite is available in Chapter 2 and Section 2.2.1.1.

Mobile Phone: Participants carried a smart phone that had four roles. First, it robustly and reliably received and stored data wirelessly transmitted by the sensor suite. Second, it stored data from GPS and accelerometers sensors in the phone. These measurements were synchronized to the measurements received from wearable sensors. Third, participants used the phone to complete system-initiated self-reports in the field. Fourth, participants self-reported the beginning of drinking and smoking episodes by pressing a button.

Self-report Measures: The mobile phone initiated field questionnaires based on a composite time and event based scheduling algorithm. Our time-based prompt was uniformly distributed to provide an unbiased experience to participants throughout the day. However, using only time-based prompts may not facilitate EMA collection about interesting events such as smoking or drinking. To capture these, a prompt was also generated around a random subset of self-reported smoking and drinking events.

For availability modeling, we only used random EMAs that are similar to sensor-triggered JITI in unanticipated appearance. The 42-item EMA asked participants to rate their subjective stress level on a 6-point scale. In addition, the EMA requested contextual data on events of interest (stress, smoking, and drinking episodes). For example, in case of a stress, users were asked about social interactions, for smoking episodes they were asked about presence of other smokers, and for drinking, they were asked about the number of drinks consumed. EMAs pose burden on the users [96] and we adopted several measures to reduce this burden. First, the smart phone software was programmed to deliver no more than 20 questionnaire prompts in a day. Second, two subsequent EMA prompts were at least 18 minutes apart. Third, the anticipated completion time of the EMA was designed to range between 1 and 3 minutes. As selection of different answers leads to different paths, we reported a time range considering the maximum and the minimum possible path length. Fourth, participants had the option of delaying an EMA for up to 10 minutes. If the participant did not respond to the prompt at the second opportunity, the prompt would disappear. Fifth, participants were encouraged to specify time periods in advance (every day before beginning the study procedure) when they did not wish to receive prompts (e.g., during exams).

Participant Training: A training session was conducted to instruct participants on the proper use of the field study devices. Participants were instructed on the proper procedures to remove the sensors before going to bed and put them back on correctly the next morning. In addition, participants received an overview of the smart phone software's user interface, including the EMA questionnaires and the self-report interface. Once the study coordinator felt that the participant understood the technology, the participant left the lab and went about their normal life for seven days. For all seven days, the participant was asked to wear the sensors during working hours, complete EMA questionnaires when prompted, and self-report smoking and drinking episodes.

Incentives: We used micro-incentives to encourage compliance with EMA's [132]. Completing a self-report questionnaire was worth $1, if the sensors were worn for 60% of the time since last EMA. An additional $0.25 bonus was awarded if the questionnaire was completed within five minutes. A maximum of 20 requests for self-reports occurred each day. Thus, the participant could earn up to $25 per day ($1.25 x 20 self-report requests), adding up to $175 over seven days of field study ($25 x 7). Since wearing physiological sensors and answering 42-items questionnaire upto 20 times daily are highly burdensome, level of compensation was derived from the prevailing wage in similar behavioral science studies [132] that involve wearable sensors. Most interruptibility studies provided fixed incentive to participants for completing the study [83, 91, 115, 126], while some studies were purely voluntary [147]. We believe that micro-incentive associated with each EMA helps obtain a stronger measure of unavailability.

Data Collected: Average number of EMA prompts delivered per day was 13.33, well below the upper limit of 20 per day. This EMA frequency is consistent with prior work [65]. EMA compliance rate was 94%. An average of 9.83 hours per day of good quality sensor data was collected from physiological sensors across all participants.

5.4 Sensor Inference

We adapt existing algorithms to infer context from physiological and mobile sensors. GPS data is used to infer semantic location and driving. To infer whether a subject is doing physical activity, we use chest worn accelerometer data. Measurements from the ECG and respiration sensors are used to infer physiological stress. Chapter 2 and Section 2.2.2 describe these computational procedures.

5.5 Metric for Measuring Availability

We define *availability* as a state of an individual in which (s)he is capable of *engaging* in an incoming, unplanned activity. For example, consider a software engineer who has just quit smoking, is working on a project, when (s)he receives a JITI on the mobile phone, perhaps triggered by an acute stress detection from sensors. In response, (s)he could – 1) stop ongoing work and engage in the intervention (e.g., do a biofeedback exercise to relax), 2) continue working on the project for a short time (pre-specified threshold) and then stop the work to engage in JITI, 3) continue working on the project but attend to JITI later, or 4) completely ignore the JITI. In our proposed definition, for cases 1 and 2 the software engineer will be considered as available while for cases 3 and 4 (s)he will be considered unavailable. We first consider delay in starting to answer a randomly prompted EMA as a metric for measuring availability.

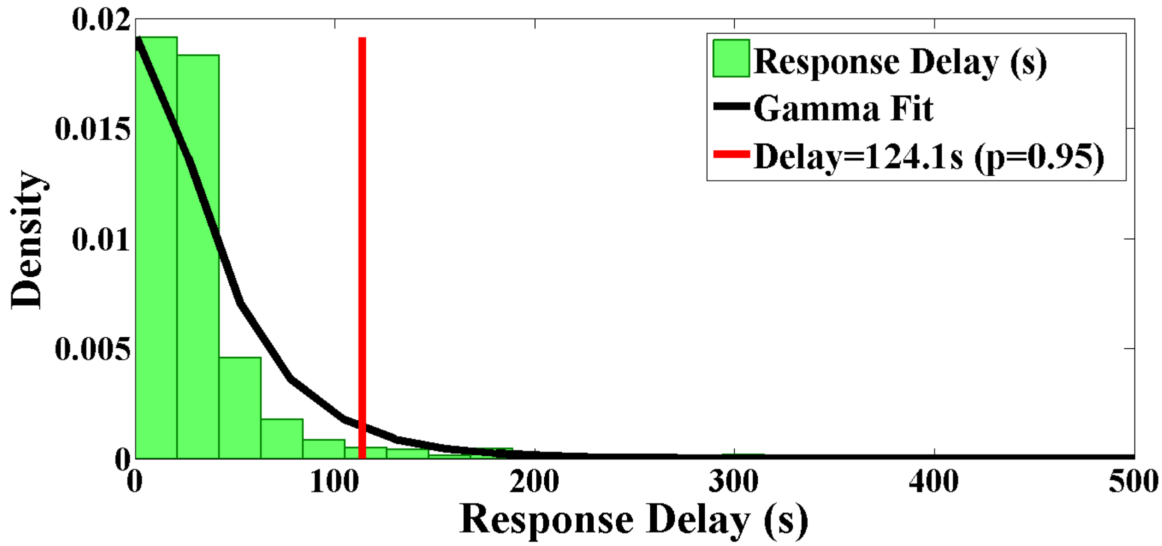


Figure 5: Delay distribution is fitted with a Gamma distribution with shape parameter κ=1.2669 and scale parameter θ=35.5021. We use the cutoff of $p = 0.95$ that occurs at 124.1 seconds, as the grace period. A response delay beyond this grace period is marked as *unavailable*.

5.5.1 Response Delay

Response delay for an EMA is the duration between the prompt time and the time of completion of the first item in the EMA. Figure 5 shows the probability distribution of response delay across all participants. Delay distribution fits a Gamma distribution with shape parameter $\kappa = 1.2669$ and scale parameter $\theta = 35.5021$. We use the $p = 0.95$ cutoff (which occurs at 124 seconds) as the grace period to obtain a good separation between the *available* and *unavailable* states.

Since each EMA is associated with a micro-incentive, it is plausible that some participants may be financially sensitive and fill out each EMA in a timely fashion, even when not fully available. In such cases, they may complete some EMA's quickly without sufficient care. We, therefore, consider completion time as another metric to complement response delay.

5.5.2 Completion Time

Completion Time for an EMA is the ratio of total completion time to the number of items answered. However, time to answer the first item includes the time to take the phone out. Therefore, we compute completion time, starting from the second item. Finally, there is between person difference in completion time due to participant's cognitive processing, typing variations, and affinity to micro-incentive. To remove these biases, we compute the z-score of completion time for each participant and then use this z-score in further analysis.

To investigate if there is a threshold such that a completion time of lower than this threshold indicates urgency and lack of care in answering an EMA, we measure the consistency of response to the EMA. For this purpose, we use a measure of consistency that is used widely in psychometrics. It is called Cronbach's alpha [27]. For a given set of items in an EMA (with numerical responses) that measure the same psychological construct, Cronbach's alpha is given by

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum s_i^2}{s_T^2}),$$

where k is the number of items in the set, s_i^2 is the variance in response to the i^{th} item, and s_T^2 is the variance of the total scores formed by summing up the responses to all the items in the set. We observe that if all the items in the set have equal variance and thus were perfectly correlated, we obtain $\alpha = 1$. On the other hand, if all the items in the set are independent, $\alpha = 0$. An $\alpha \geq 0.7$ is regarded as acceptable [27].

In our 42-item EMA, there are several affect items that measure the same psychological construct. These items are *Cheerful?*, *Happy?*, *Energetic?*, *Frustrated/Angry?*, *Nervous/Stressed?*, and *Sad?*, where participants respond on a Likert scale of 1–6. To compute alpha, items that assess positive affect (*Cheerful*, *Happy*, and *Energetic*) are retained as scored and items that assess negative affect (*Frustrated/Angry*, *Nervous/Stressed*, and *Sad*) are reverse coded

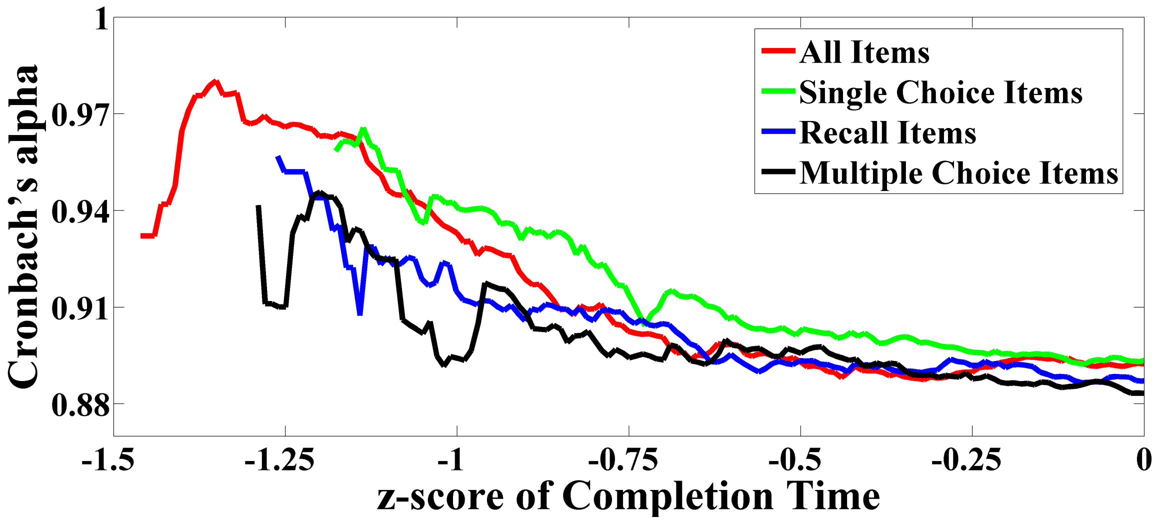


Figure 6: We plot Cronbach's alpha value for various thresholds of z-score of completion times. We observe that the alpha is always acceptable (i.e., ≥ 0.7). This holds even when we consider various subsets of items that require recollection or multiple choice selection.

(e.g., 1 becomes 6). To test whether these six items indeed measure the same psychological construct, we computed the overall alpha score for all responses from all participants. The overall $\alpha = 0.88$ indicates a good agreement [72].

We next computed the Cronbach's alpha score for various thresholds of (z-scores of) completion times. We observed that the 42-item EMA questionnaire contains different types of item. First, there are single choice items which participant can answer right away. Second, there are multiple choice items which requires going through various possible answers, which may take more time. Third, there are recall based items where participants need to remember about past actions. An example of such an item is *"How long ago was your last conversation?"*. Such items may require longer to complete. We consider subsets of EMA items in each of the above three categories and compute their corresponding z-scores. Figure 6 plots the alpha values for various thresholds on completion times for four cases — when the completion time for all items is considered and when the completion times for each of the above three subset of

EMA items is considered. Since our goal is to find a lower threshold such that completing EMA items quicker than this threshold may indicate lack of availability, we only plot completion times lower than average (i.e., z-score of 0). We observe that in each case, $\alpha \geq 0.7$, which implies that completing EMA items quickly does not indicate inconsistent response. Hence, completion time is not a robust estimator of unavailability and we retain only the response delay as our metric of unavailability.

5.5.3 Labeling of Available and Unavailable States

When an EMA prompt occurs, the phone beeps for 4 minutes. If the participant begins answering or presses the delay button, this sound goes away. There are 4 possible outcomes for each such prompt — i) *Missing:* Participant neither answers the EMA nor presses delay, i.e., just ignores it, ii) *Delayed:* Participant explicitly delays the EMA, and plans to answer it when (s)he becomes available, iii) *After Grace:* Participant answers after a grace period, which is defined in Figure 5, iv) *Before Grace:* Participant answers within the grace period. We mark the first three scenarios as *Unavailable*.

To identify available EMAs, we use two different approaches. In the first, we take n quickest answered EMA's from each participant, where n is the number of EMA prompts when this participant was found to be unavailable. We mark each such EMA as *available*. We call this a **Representative** dataset, because it gives more weight to those participants' data, who sometimes forego micro-incentives by missing or delaying EMA's when they are not available. This may be similar to the situation in a class where several students may have a question, but only a few speak up, thus helping others who may be shy. This dataset gives less weight to data from those participants who are always prompt in answering EMA's, due to their sincerity, scientific devotion to the study, or affinity to micro-incentives. This dataset thus recognizes and respects wide between person variability inherent in people.

Counting missed, delayed, or delayed above grace period (124.1s), we label 170 EMA's as triggered when participants were unavailable. Number of instances when a participant was unavailable ranges from 0 to 15. By marking n quickest answered EMA from each participant as available, where n is the number of EMA prompts for which that particular participant was unavailable, we obtain a total of 340 EMA's for training data. This dataset provides a robust separation of delay between the *available* and *unavailable* class (with a mean of $141.4s \pm 51.7s$ and a minimum separation of 107.7s). This kind of wide separation helps us mitigate the effect of delay in taking out the phone to answer an EMA.

Due to the definition of *Representative dataset*, 3 participants are completely ignored due to always being compliant, responding within grace period, and never delaying an EMA. Hence, we construct a **Democratic** dataset, where we consider equal number of EMA's from each participant. To obtain a similar size of training data as in the *Representative dataset*, we use 6 quickest EMA from each participant as *available* and 6 slowest (including delayed or missed) as *unavailable*. We thus obtain 12 samples from each participant, making for a total of 360 samples. The delay separation between *available* and *unavailable* class in this dataset has a similar mean of 169.8s, but a higher standard deviation of 193.8s, and a smaller minimum separation of 5.2s.

5.6 Findings

Before presenting our model for predicting availability, we conducted a preliminary analysis of various factors in this section to understand their role in predicting availability. We investigated various contextual factors (e.g., location, time, etc.), temporal factors (e.g., weekend vs. weekdays, time of transition, etc.),

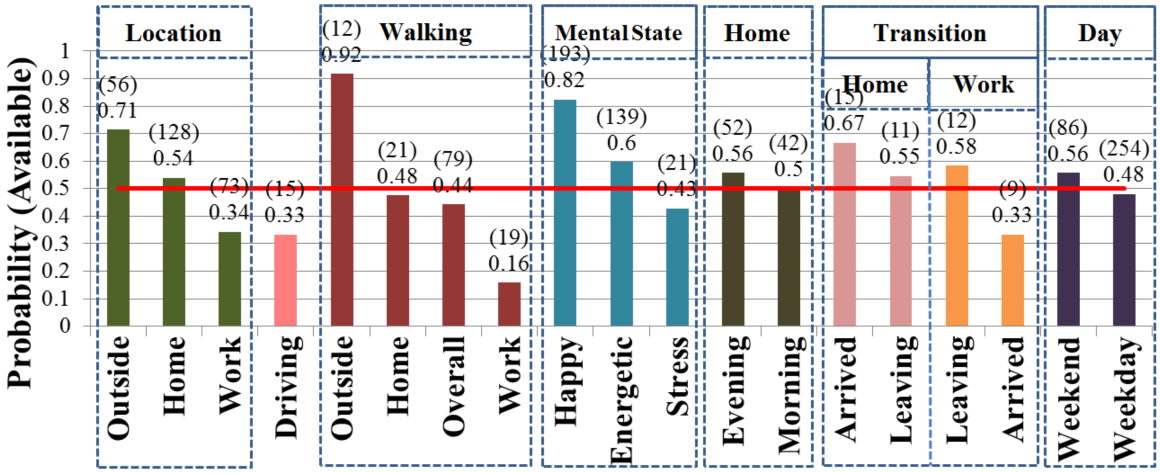


Figure 7: Probability of participants being available across different contexts. Here morning is defined as before 9 AM and evening as after 5 PM. Arrived at a location means arrival within 30 minute, while leaving means 30 minute prior to leaving. Red line is drawn for $p(A) = 0.5$.

mental state (e.g., happy, stressed, etc.), and activity state (e.g., walking, driving, etc.).

Figures 7 and 8 present the probability of participants being available and the mean response delay across different contexts (e.g., location, activity, mental state, and time) respectively. In these figures, outside refers to outside of home, work, store, restaurant, and vehicle. We observe in Figure 8 that the response delay has high variance (range 23.2-137.5) across different contexts, which can be attributed to the Gamma distribution of response delay (see Figure 5).

Location: From Figure 7, we observe that participants are more likely to be available ($p(A) = 0.71$[1]) when they are outside and they are most likely to be unavailable at work ($p(A) = 0.34$). When participants are outside, their response delay is also lower than any other location (mean=41.0s; $p = 0.074$ on Wilcoxon rank-sum) (see Figure 8). This lower response delay may be because when participants are outside, they are unlikely to be engaged in a time-sensitive activity (e.g., deadline, driving a vehicle) and thus can attend to the incoming

[1]We use $p(A)$ to denote the probability of being available.

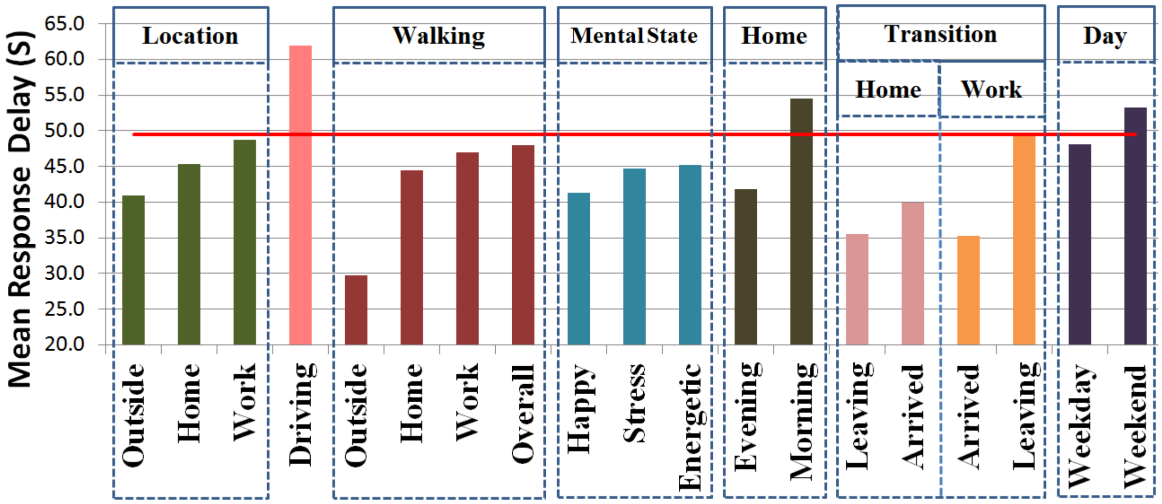


Figure 8: Mean response delay across different contexts. Morning, evening, arrival, and leaving are defined as in Figure 7. Red line represents the overall mean of 49.5s ($\pm$116.0s).

prompt relatively quickly. As expected, during driving participants are usually unavailable ($p(A) = 0.33$) and the delay in response during driving is significantly higher than other times ($p = 0.019$ on Wilcoxon rank-sum).

Walking: In contrast to [83], which found posture change as an indication of being interruptible, we find that in daily life, walking by itself does not indicate availability ($p(A) = 0.44$). Interestingly though, walking outside indicates a highly available state ($p(A) = 0.92$), while walking at work indicates a highly unavailable state ($p(A) = 0.16$). We observe a mean response delay of 29.7s when participants are walking outside, which is not significantly lower than stationary ($p = 0.318$ on Wilcoxon rank-sum), but significantly lower ($p = 0.008$ on Wilcoxon rank-sum) when compared with other locations (e.g., home, work, etc.).

Mental State: When participants are in a happier state, they are more likely to be available ($p(A) = 0.82$) and we observe a lower response delay (mean=41.3s; $p = 0.008$ on Wilcoxon rank-sum). Similarly, when participants are feeling energetic, they are more available ($p(A) = 0.6$). But, unlike happy, in the energetic state the delay (45.2s) decrease is not significant ($p = 0.144$ on

Wilcoxon rank-sum). On the other hand, participants being stressed reduces the probability of being available ($p(A) = 0.43$). A good news for JITI that may be triggered upon detection of stress is that participants are not found to be highly unavailable when stressed as is the case at work. Such JITI, therefore, may still be attended to by users. Investigation of the receptivity of stress-triggered JITI may represent an interesting future research opportunity.

Home: Since being at home indicates only a marginally available state ($p(A) = 0.54$), we investigate whether time of day makes a difference. We find that availability at home is lower during morning ($p(A) = 0.5$), and higher in the evening ($p(A) = 0.56$), but not by much. However, response delay in the morning (54.5s) is higher than that in the evening (41.8s) ($p = 0.052$ on Wilcoxon rank-sum). This indicates that participants are more pressed for time in the morning.

Transition: To further investigate the effect of location on availability, we analyze the availability of participants when they are about to leave a place (within 30 minutes of departure) or have just arrived at a place (30 minute since arrival). We find that the availability of participants when leaving home ($p(A) = 0.55$) is similar to when in home generally. But, their availability is higher when they have just arrived home ($p(A) = 0.67$). The scenario is reversed at work. The availability at work upon arrival ($p(A) = 0.33$) is similar to the overall availability at work. But, their availability is higher when about to leave work ($p(A) = 0.58$).

Day: Finally, we analyze the effect of weekday vs. weekend. We find that participants are more likely to be available on weekend ($p(A) = 0.56$) than on weekdays ($p(A) = 0.48$). Interestingly, the response delay on weekends is higher than that during weekdays ($p = 0.061$ on Wilcoxon rank-sum).

Although one could investigate several combinations of factors, we next develop a model that uses several features derived from these factors to predict availability of participants.

Table 3: For 6 values of N (30 seconds, 1 min, 2 min, 3 min, 4 min, and 5 min), the above four derivative features are computed for stress and activity, producing 24 features for each.

All-N	Event occurred in **every** past window within N second of corresponding sensor prior to random EMA prompt
Any-N	Event occurred in **any** past window within N second of corresponding sensor prior to random EMA prompt
Duration-N	Duration of occurrence of event within past N second prior to EMA prompt
Change-N	Number of change where event occurred in one window followed by non-event within N second prior to EMA prompt

5.7 Predicting Availability

In this section, we develop a model to predict availability. We first discuss the features we compute, feature selection methods to find the most discriminating features, and then the machine learning model to predict availability. We conclude this section by reporting the evaluation of our availability model.

5.7.1 Feature Computation

To predict availability, we compute a variety of features. Majority of them come from sensors, but we also obtain several from self-reported EMA responses because the sensor models for their detection is not mature enough today to detect them with reasonable confidence. We expect that these features will also become reliably detectable from sensors in near future. In total, we compute 99 features.

Clock: Time and Day (6 features): We compute several time related features. We include *"day of the week"* since there may be a day-specific influence, *"elapsed hour"* in a day to identify work vs. non-work hours, and *"Time since last EMA"* to capture the cumulative fatigue caused by frequent EMA prompts. We also include binary features such as *"Working Hour?",* which is defined to be between 9 AM and 5 PM, *"Weekend?",* and *"Holiday?"*.

Sensor: Stress (4+24 features): As discussed earlier, we infer stress level for each 30 second window. Since sometimes EMA prompts itself may cause stress, we used binary stress levels in the 30 second windows prior to the generation of an EMA prompt. From these windows, 24 derived features are computed (see Table 3), similar to that in [64]. We note that if the participant is physically active during a 30-second window, we mark the stress feature as undefined for this window (due to stress being confounded by physical activity). Hence, stress level in each of the 30 second windows for derived features may not be available. Consequently, we compute four other features. The first two of these come from the first window preceding the prompt where stress inference is available (i.e., unaffected by physical activity). Binary stress state and probability of being stressed are used as features from this first window. The remaining two features are the number of windows where the participant is stressed over the prior 3 (and 5) windows preceding the prompt, for which stress inference is available. These windows must occur within 5 minutes prior to the prompt.

Sensor: Location, Place, Commute Status (7 features): We compute several location related features. This includes coarse level location such as *Home, Work, Store, Restaurant, Vehicle,* and *Other* and detailed location such as *inside home, dormitory, backyard*, etc. We also include *"Previous Location"* and *"time spent in current location"* because it is likely that after immediate arrival at home from work or from other locations people are less likely to be available. We include a binary feature for *"driving"* because driving requires uninterrupted attention and distraction during driving can result in injury, loss of life, and/or property. It is also illegal in several parts of the world to engage a driver in a secondary task such as texting. Since EMA prompts are generated randomly (as per the norm in behavioral science [167]), some EMA prompts did occur during driving. Participants were instructed to find a safe place to park the car in such cases before answering. A binary feature *outdoor* is also included since we observe participants being more available when they are outdoors and walking.

Sensor: Physical Activity (3+24 features): Since physical activity can also indicate availability, we use physical activity data from the chest accelerometer sensor as a binary feature, and intensity of activity as a numeric feature. EMA questionnaire contains items such as *"Describe physical movement"* with possible answers *"Limited (writing)", "Light (walking)", "Moderate (jogging)",* and *"Heavy (running)"*. We include features such as *writing* as a categorical feature because *writing* state may affect availability and we are unable to infer it activities from our sensors with reasonable confidence today.

EMA: Activity Context (13 features): EMA questionnaire asked participants to describe their ongoing activity using the following items: *"How would you describe your current activity?"*, with possible responses as *"Leisure/Recreation"* or *"Work/Task/Errand"*, a multiple choice item *"What's going on?"* with possible responses as *Meeting, E-mail, Reading, Phone Call, Writing, Sports, Video Game, Surfing the Internet, Watching Movies/TV/Video, Listening to Music,* and *Other*. Each possible response is used as a binary feature. We also use binary response to the *"Taken alcohol?"* item.

EMA: Social Interaction (6 features): Research on interruption has revealed that situations involving social engagement are considered less interruptible [23, 66, 76]. To model availability, we used participants' responses for the social interaction related EMA queries that includes *"In social interaction?", "Talking?", "If talking, on the phone?", "If talking, with whom?", "If not talking, how long ago was your last conversation?",* and *"Who was it with?"*.

EMA: Mental State (9 features): We also include emotional state due to

Table 4: Selected 30 features ranked (R) according to information gain. Detailed location offers the highest information gain.

R	Feature	R	Feature
1	Detailed Location	16	Stress probability
2	Coarse Location	17	Stress count in 5 previous window
3	Weekday	18	StressChange-300
4	Outdoor?	19	StressChange-240
5	Sleepy?	20	ActivityAll-120
6	Happy?	21	StressAny-180
7	Energetic?	22	StressChange-180
8	Commute Mode?	23	StressAny-240
9	Recreation?	24	StressAny-60
10	Activity type	25	StressChange-30
11	Weekend?	26	StressDuration-30
12	Talking on phone?	27	ActivityAll-180
13	Taken Alcohol?	28	ActivityAll-240
14	Elapsed hour of day	29	ActivityAny-300
15	Time spent in current location	30	EMA Index

their wide acceptability as a factor in Human Computer Interaction [25, 103]. Although stress is detectable from sensors, affect is not yet detectable reliably in the field setting from physiological sensors. Hence, we use EMA responses. We include response to our EMA items, *"Cheerful?", "Happy?", "Frustrated/Angry?", "Nervous/Stressed?", "Sad?", "Facing a problem?", "Thinking about things that upset you?", "Difficulties seem to be piling up?"*, and *"Able to control important things?"*. Response in Likert scale 1-6 is used as feature.

EMA: Fatigue (3 features): Each EMA prompt resulted in some level of fatigue on the recipient [96]. We find that responses to the first half of the EMA's are more consistent than the second half EMA for the day ($p = 0.056$, $n = 30$, paired t-test on Cronbach's alpha). Therefore, we add EMA index of the day as a feature. Our EMA questionnaire contained items such as *"Energetic?"* and *"Sleepy?"*. Subjective responses of these items in 1-6 Likert scale are also used as features.

5.7.2 Feature Selection

As reported in the preceding, a total of 99 features were computed. But, to avoid overfitting of the model, we select a subset of the features for modeling availability. We base our feature selection on two complementary methods.

Correlation based Feature Subset Selection: Our goal is to find features that are highly correlated to the class available vs. unavailable, and not correlated with each other. We used Hall's [74] method to find the optimal non-correlated feature set regardless of the underlying machine learning algorithm.

Wrapper for Feature Subset Selection: Correlation based feature selection may discard some features that are useful for a particular machine learning algorithm. Therefore, we also use Wrapper [101] based feature selection to find an optimal feature subset for the SVM machine learning algorithm [146].

By taking a union over the features selected by correlation based feature selection and Wrapper applied to SVM, we obtain a total of 30 features. Table 4 lists these features ordered according to their information gain [48]. We make several observations. First, we observe that most of the features selected are either already detectable from sensors (1-4, 8, 11-30) or are potentially detectable in near future from sensors (9-10). But, three features (5-7) are hard to detect automatically today. An inward looking camera in smart eyeglasses could potentially detect some of these in near future as well. Second, we observe that stress features (16-19, 21-26) figure quite prominently in this list, indicating a significant role of stress in predicting availability. Finally, we observe that driving is not included in the list of selected features, though intuitively it appears relevant. We hypothesize that features ranked 1, 2, 4, and 8 contain information about driving and as such driving may not be needed as a separate feature.

Table 5: Confusion Matrix for predicting availability using SVM model on RBF kernel built on **Representative Dataset**. Overall Accuracy is 74.7% against a base accuracy of 50%, with a kappa of 0.494.

	SVM Classified as	
	Available	Unavailable
Available	**134 (78.8%)**	36 (21.2%)
Unavailable	50 (29.4%)	**120 (70.6%)**

5.7.3 Model

Due to its well-accepted robustness, we train a *Support Vector Machine (SVM)* [146] model with RBF kernel to predict availability of users. To evaluate the model, we use both the standard 10 fold cross-validation and leave-one-subject-out to evaluate between subject generalizability. As described earlier, we use two diverse methods to label EMA's as *available* and *unavailable* to generate training data. We present the performance of the model on each of these labeling methods.

Representative Dataset: Based on the missed, explicitly delayed, or delayed above grace period (124.1s) we mark 170 EMA's as triggered when participants were unavailable. We mark the n quickest answered EMA from each participant as available, where n is the number of EMA prompts for which that particular participant was unavailable. This provides us with 340 instances as training data for modeling with 170 instances coming from each class[2].

Using this dataset we get an overall accuracy of 74.7% (against a base accuracy of 50%) with kappa of 0.494 for 10-fold cross-validation. From the confusion matrix in Table 5, we find that for 78.8% cases, the classifier is able to predict availability versus 70.6% in the case of unavailability. We get a precision of

[2]We note that although only a small subset of EMA's (340 out of 2717) is used in model development, SVM can produce posterior probability of availability for any EMA. Hence, the applicability of the model is not limited to the data used in training.

Table 6: Confusion Matrix for predicting availability using SVM model on RBF kernel built on **Democratic Dataset**. Overall Accuracy is 69.2% against a base accuracy of 50%, with a kappa of 0.383.

	SVM Classified as	
	Available	Unavailable
Available	**135 (75.0%)**	45 (25.0%)
Unavailable	66 (36.7%)	**114 (63.3%)**

0.749, a recall of 0.747, an F-measure of 0.747, and area under the curve of 0.747. For leave-one-subject-out, we get a weighted average accuracy of 77.9%.

Democratic Dataset: In this dataset, we take 12 samples from each participant, which leads to similar 360 samples from 30 participants. The 6 quickest responded EMA's are considered *available* and 6 slowest responded ones (including explicitly delayed ones) are considered as *unavailable*.

For this labeling, the SVM model achieves an accuracy of 69.2% with a kappa of 0.383, slightly lower than the Representative model. However, from the confusion matrix in Table 6, we find that for 75.0% cases, classifier is able to predict availability. We get a precision of 0.694, a recall of 0.692, an F-measure of 0.691, and area under the curve of 0.692. For leave-one-subject-out, we get an accuracy of 76.4%.

5.8 Limitations and Future Work

Being the first work to inform the timing of sensor-triggered just-in-time intervention (JITI), this work has several limitations that open up interesting opportunities for future works.

First, several features used to predict availability are not yet reliably detectable via sensors today. For a model to be automated in informing the timing of JITI, all features need to be inferred from sensors. Second, this work used data from wearable sensors. Since wearing sensors involves user burden, it is more desirable to use only those sensors that are available on the phone. But, some features are not feasible to obtain today from phone sensors (e.g., stress) and hence represents interesting future works. Third, the type of sensors available on the phone is growing richer rapidly. Several sensors such as proximity sensor, acoustic sensor, and phone orientation and other data in the phone (e.g., calendar, task being performed on the phone, etc.) that may inform the current context of a user were not used in this work. Using these and other sensors emerging in phone may further improve the prediction accuracy. Similarly, using additional sensors on the body and those in instrumented spaces such as office, home, and vehicle (e.g., cameras) can also be used wherever available to further improve the prediction accuracy.

Fourth, this work used micro-incentive to improve compliance in responding to EMA prompts and used it to accomplish a high level of motivation. Although the work presented in this dissertation can inform the timing of delivering randomly prompted self-reports in scientific studies, it remains an open question how well the micro-incentive captures the motivation level expected in users who choose to use JITI due to certain health condition or due to a wellness or fitness motivation.

Fifth, given that filling out a 42-item EMA requires significant user involvement (i.e., 2.4 minutes to complete), the results of this work may be more applicable to JITI that involve similar engagement. Its applicability to lighter JITI may need further investigation. We note, however, that if the user is found to be unavailable for a more involved active JITI (e.g., when driving), passive intervention could be delivered in the meantime (e.g., by playing music [138]).

Sixth, the analysis in this work used only the unanticipated (i.e., randomly prompted) EMA's to simulate the triggering of a sensor-triggered JITI, but the participants also filled out EMA's that resulted from their self-initiation. Although these self-initiated EMA's were voluntary, they may add to the burden and fatigue of participants. It remains open whether the results of a future study that only uses randomly prompted EMA's may be any different than the one reported here.

Seventh, we used response delay as a metric for objectively assessing availability. Although we label significant delay in response as unavailable, it is not a gold-standard truth. In future, we can investigate other objective metrics (e.g., phone in airplane mode) and compare with each other.

5.9 Chapter Summary

Sensor-triggered just-in-time-interventions (JITI) promise to promote and maintain healthy behavior. But, critical to the success of JITI is determining the availability of the user to engage in the triggered JITI. This dissertation takes a first step to inform the timing of delivering JITI. We propose a novel objective metric to measure a user's availability to engage in a JITI and propose a model to predict availability in the natural environment based on data collected in real-life. Our findings indicate that availability of a user depends not only on user's ongoing activity or physical state, but also on user's psychological state. Our results can inform the design of JITIs and opens up numerous opportunities for future works to improve the accuracy, utility, and generalizability of our model.

Chapter 6

Identifying Stress Episodes Based on Field Stress Data

6.1 Introduction

In addition to assessing the availability of a person, the success of Just-In-Time Interventions (JITIs) critically depends on being able to mine the time series of noisy sensor data and generate triggers for intervention at the most opportune moments. We will use stress intervention as a running example.

Repeated exposures to stressors cause adverse effect on health, such as, depression, heart disease, diabetes, and addictions, such as, smoking, alcohol, and opioid [10, 11, 43, 55, 56, 80, 124, 125, 153]. It is estimated that productivity lost due to stress in USA is $300 billion per year [6]. Now because of the widespread use of wearable sensors we can now develop just-in-time stress intervention.

To trigger a stress intervention, we need to locate significant stress episodes in the sensor data stream. This introduces several challenges. First, stress measurements obtained from sensors usually have to be inferred from physiological data, which by their very nature rapidly varying, similar to real-time tracking of stock prices. Second, unlike stock-price data, the time series of stress is discontinuous due to factors such as sensor detachment and wireless losses [136, 151]. Third, sensor measurements are frequently confounded by physical activity (23% of the time [151]), that needs to be filtered out for an accurate assessment of stress.

Another set of challenges concerns the triggering of the intervention. First, the decision to trigger must be made quickly so the intervention can be effective. Hence, simple methods that can be efficiently implemented on mobile devices are needed. Second, too-frequent prompts of an intervention can lead to alarm fatigue [96] and render the system useless. Ideally, the intervention policy should be personalized to the tolerance level of the individual and the frequency of intervention (e.g., once per day) desired by the user. For example, a rapid but transient physiological arousal due to a loud noise should not trigger a JITI, whereas stress aroused in a newly abstinent smoker while passing through a corner of the building where they used to smoke should be predicted and prevented; otherwise the smoker may lapse to smoking.

In this chapter, we take first steps towards the development of such JITI and develop time-series-pattern mining methods to detect significant stress episodes in discontinuous ambulatory data. The goal of this work is to establish the foundation on which a just-in-time stress intervention can be developed.

For model development and application, we use data collected in a 4-week field study in 38 opioid-dependent polydrug users receiving opioid agonist maintenance treatment, all of whom were in a larger trial investigating individual and environmental influences on drug use. Each participant wore wireless physiological sensors for 10+ hours per day, from which we obtained a continuous measure of stress [89].

In brief, we first developed methods to deal with physical activity and discontinuities in the time-series data. We then determined that data missing due to physical activity could be considered Missing At Random (MAR) [53]. We applied the cStress model [89], imputed the missing data, and validated the output of cStress (together with its imputation) against self-reported stress and found good agreement. Next, we trained a stock prediction method called Moving Average Convergence Divergence (MACD) [17] to locate the time of an increase in stress in rapidly varying continuous time-series data. We estimated the probability distribution of the likelihood of stress assessments and the probability distribution of stress durations (in the smoothed time series) to personalize the algorithm for each individual. The threshold on stress likelihood can correspond to tolerance level, and the duration can be selected to meet the daily intervention frequency preference. If, in a candidate window, the likelihood of stress crosses the high likelihood threshold and remains elevated for a threshold duration, then this window represents significant stress episode.

6.2 Related Works

The first category of related works are the ones on stress monitoring. Assessment of stress and physiology can be obtained episodically when a user interacts with a device or continuously via sensors on the body or in the user's environment. Examples of the former include capturing ECG from a smartphone camera (during gaming [75]) or from electrodes embedded on smartphone jackets (e.g., Alivecore), hand arm dynamics from the computer mouse [173], and pressure from pressure sensitive keyboard and mouse [81]. Physiology can be obtained continuously from wearable physiological sensors [59]. Stress detection can be done from a variety of physiological parameters including ECG and respiration [89, 145], electrodermal response [117], photoplethysmography from fingertip [112], or near-infrared spectroscopy from forehead [82]. Our method can be applied to stress measurements obtained from any of the above methods.

The second category of works are those that assess interruptibility, workload, or availability to decide when to deliver a prompt for intervention, self-report, or phone call [65, 92, 93, 175]. A recent work [157] proposed a model that uses stress, time, location, and the current context to determine the availability or interruptibility of users, in their natural environment, to respond to randomly triggered self-report prompts. It found that users are least available at work and during driving, and most available when walking outside. These works are complementary to ours. Once a trigger for intervention has been generated by our model, it should be delivered to the user only when they are determined as being physically, cognitively, and socially available.

The third category includes works on stress interventions. An example is a reflective intervention called AffectAura [120] that logs physiological state using audio, visual, sensors, and user activities and aims to support reflection via visualization. Visualization is replaced by a wearable butterfly in [113] that helps users reflect on their stress level and regulate it. Textiles have been designed that can actuate in response to stress [50]. These complementary works indicate interesting intervention possibilities, if appropriate methods such as ours can reliably detect stress episodes in real-life.

The fourth category of related works are sensor-triggered JITIs that have emerged in other contexts. For example, [38] presented a JITI to prevent emotional food intake. Another example is [149] that proposed a system where earpieces (to monitor chewing and swallowing), augmented-reality glasses (for capturing food consumed), and a physiological sensor (for heart rate) are connected to a mobile-phone application that processes the data and gives feedback to the user. Sensor-triggered JITIs have also been proposed for preventive maintenance of a plant (see [42] for a review) and for GPS-based vehicle navigation [9, 18]. But, none of these methods can be used directly to mine the time series of stress to find significant stress episodes.

The closest related works are those that aim to discover or predict stress episodes from time series of physiological data. MoodLight [117] finds episodes of arousal from electro-dermal activity (EDA) in the lab environment and regulates the color of a desk lamp to reflect the user's stress level. When users reduce their stress level, the light color changes to blue. In [95], the authors present a method to predict the time series of heart-rate variability (HRV) using a first-order Hidden Markov Model. The algorithm was tested in a simulated patient environment using a beta distribution ($\alpha = 0.1$ and $\beta = 1$). In contrast to these works, our model addresses real-life challenges of discontinuity and rapid variability.

6.3 Data Description

We used data collected as part of a larger outpatient study of relationships among stress, addictive behaviors, and daily activities. The parent study, and this substudy, were approved by the Institutional Review Board (IRB), and all participants provided written informed consent. The participant demographics, study setup, and the data we collected appear below.

6.3.1 Devices and Sensor Measurements

Participants carry a smart phone and wear AutoSense sensor suite under their clothes that enables us to collect respiration, ECG, and accelerometer data. A description of this AutoSense sensor suite is available in Chapter 2 and Section 2.2.1.1.

6.3.2 Field Study Procedure

Participants were trained in the proper use of the devices. They were shown how to remove the sensors before going to bed and how to put them back on correctly the next morning. They were also asked to take them off during showers and any contact sport. Participants received an overview of the smartphone software's user interface. Once the study coordinator felt that participants understood the technology, they left the research clinic and went about their normal lives. Participants were asked to wear the sensors during their waking hours, complete self-reported questionnaires when prompted, and record instances of drug use and craving on the phone.

Participants were asked to return to the research clinic daily. The study coordinator uploaded the data collected the previous day and reviewed the physiological measurements to ensure that sensors were working and were being worn properly. On the final day, participants returned study equipment and completed an Equipment and Experience Questionnaire. Finally, participants were debriefed on their experiences and comfort with the study.

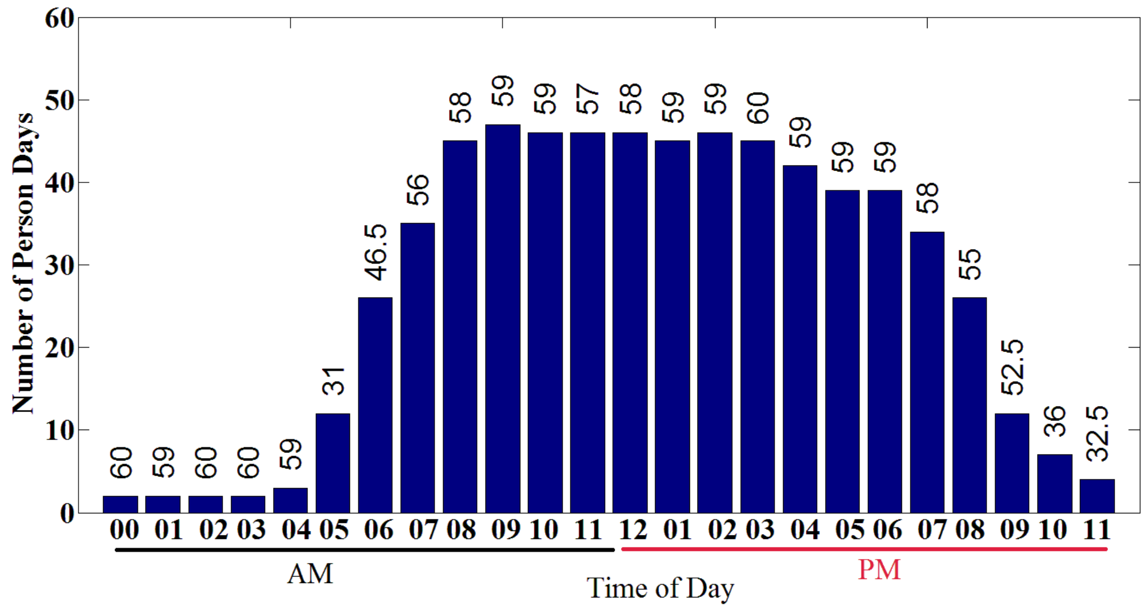


Figure 9: Time of day when participants wear the wearable sensors and contributed in the data collection campaign. X-axis indicates time of day, Y-axis indicates number of person days. The number over each histogram indicates average minutes of data collected on that hour of day.

We recruited 38 polydrug users (age 41 ± 10 years, 11 female, 6 dropped out) who agreed to wear the sensor suite. Because drug use does not occur every day in all these users, we conducted the study for four weeks to maximize the likelihood of capturing real-life drug use events.

Compensation: Participants received $10/day for wearing the sensors (and $5 bonus for 14+ hours of wearing), carrying the smartphone, and completing device-prompted questionnaires consisting of 32 items. In total, participants were paid up to $380 plus bonus (if any) for four weeks of participation.

Self-report: The smartphone initiated Ecological Momentary Assessment (EMA) questionnaires at random times. The 32-item EMA asked participants to rate their subjective assessment of affect on a 6-point scale. In addition, participants were asked about the presence of drug and smoking cues.

Data Collected: Participants wore the physiological sensors and carried the smartphone for 12.52 hours each day in their daily, free-living condition. Due to sensor detachment, displacement, loosening, and wireless loss between phone and the sensor, some of the ECG data were not of acceptable quality. We identified unacceptable ECG data using a method proposed in [151] and excluded them. An average of 10.54 hours per day of acceptable ECG data (10,447 hours of data in total) were obtained; these were the data we used for stress inference. We observed that most of the participants wore the sensors and contributed data between 6:00 AM to 8:00 PM of a day(see Figure 9). A total of 5,755 EMA responses were collected (5.8/day), with a compliance rate of 88.0%.

6.4 Stress Inference from Physiological Data

In this section, we describe the procedure we used to infer physiological stress from wearable sensors. We adapt a recent model called *cStress* proposed in [89] and summarized in Section 2.2.2.1. The model infers stress from electrocardiogram (ECG) and respiration data for each minute. We modified the model to generate stress measurements every 5 seconds from overlapping minute-windows to get a smoother time series.

6.4.1 Stress Likelihood & Stress Density

The cStress model provides a continuous measure of stress, scaled to be between 0 and 1, for every 5 seconds of overlapping one-minute sensor data. This time-series of 5-second probability-like measures of stress, for a particular participant, is referred to hereafter as "*stress likelihood*."

To assess stress within intervals longer than a minute, we use a different measure, called "*stress density*," which accounts for likely variation in contexts and activities (e.g. morning vs. afternoon, driving vs. home). We define stress density as the area under the stress-likelihood time-series divided by the length of the interval.

6.4.2 Need for Personalization

We next analyze the variability in stress densities across participants and across different days for the same participant. Figure 10(a) shows the stress density for each participant in increasing order. There is wide between-person variation. The two participants with highest stress densities have each more than twice the density of the two participants with lowest stress densities. Figure 10(b) shows daily stress for the participant with maximum overall stress density. Here, for 4 (out of 27) days, that participant had three times lower stress density than he/she had on average. On the other hand, the most stressful day has a stress density twice the overall average. These observations demonstrate that the frequency (or even the content) of stress interventions may need to be calibrated to each person and for each day.

6.5 Reducing the Impact of Confounding Factors

Although physiology is affected by several kinds of events in daily life, the main confounder for stress assessment is physical activity. To isolate data affected by activity, we first detect physical activity from chest-worn 3-axis accelerometer data, using an existing model [151] described in Chapter 2 and Section 2.2.2.2. Second, we estimate the time it takes for physiology to recover from the effect of a just concluded activity episode. Both data are then excluded.

Physiological readings generally return to baseline within 2 minutes after physical activity (unless the activity is especially intense) [60]. However, the majority of activity episodes in our daily life are of short durations. Although our participants were physically active 22.7% of their sensor-wearing time, 95% of their activities lasted less than 2.1 minutes. Discarding 2 minutes of data after each activity episode would result in excluding 35.0% of additional data (for a total of 57.7% of all data). We, therefore, need a more systematic person- and

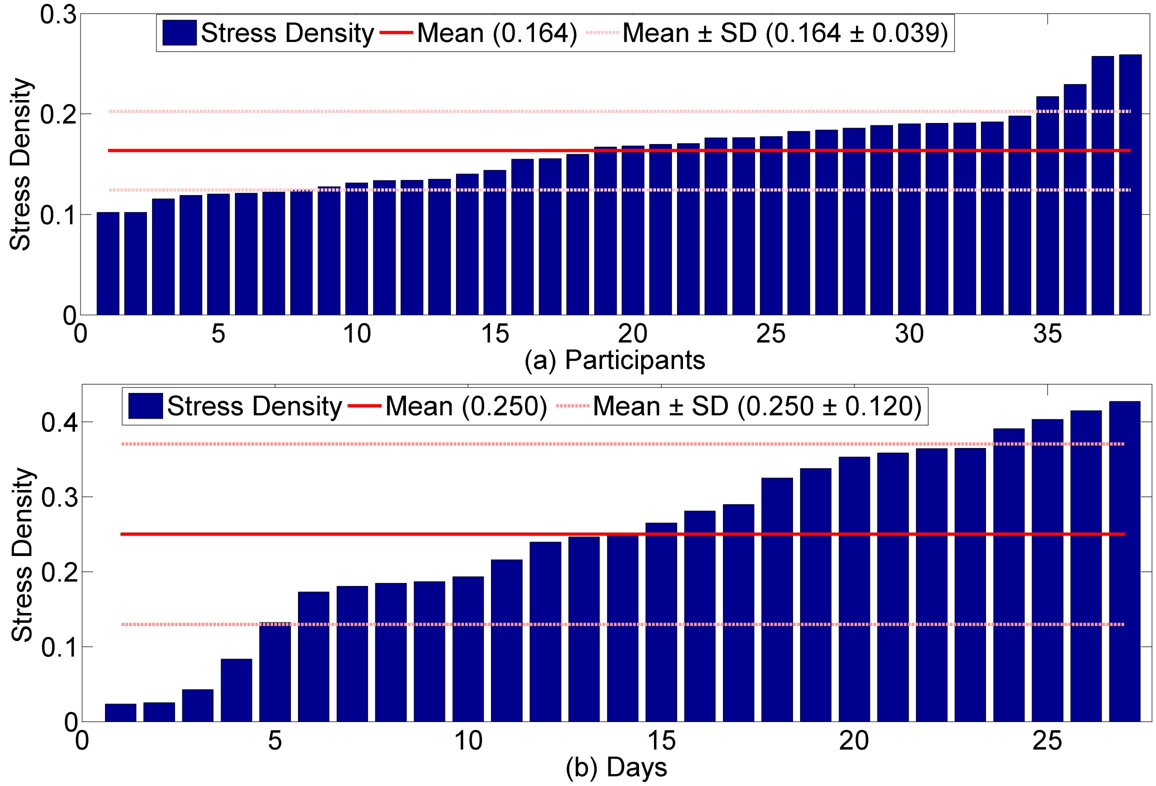


Figure 10: (a) Stress density for each participant. We observe wide between person variation here. (b) Day wise stress for the participant with highest stress density. We observe wide between day-to-day variation for this (and other) participants.

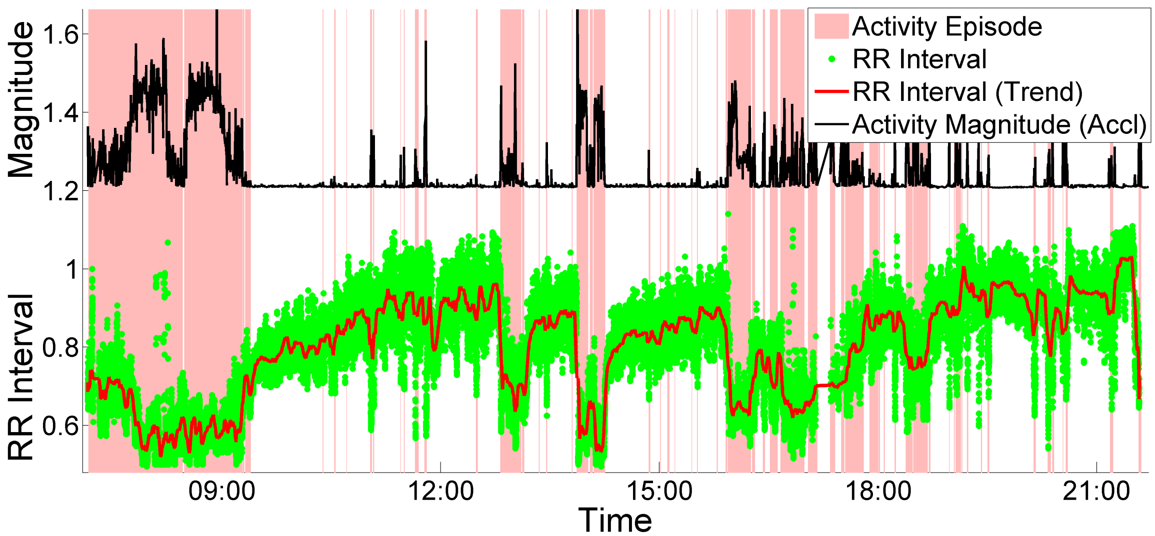


Figure 11: R-R interval decreases (and heart rate increases) due to physical activity and recovers exponentially after the conclusion of activity.

situation-specific method to **estimate recovery time**. We consider two approaches — a data based method and a model based method.

6.5.1 Data Based Approach

To estimate the time it takes for physiology (e.g. heart-rate) to recover after each episode of physical activity, as detected using accelerometry, we can simply record the heart-rate before physical activity, designating it as the resting heart-rate, and then compute the time it takes for the heart-rate to return to the resting heart-rate after the end of physical activity. Heart-rate (HR) is defined as the number of beats per minute.

A key weakness of this direct approach for computing the recovery time is that, in the field setting, the HR may take a very long time to recover to the most recent resting HR (see Figure 11), due to confounding factors, such as caffeine intake, during or after the physical activity episode, that typically raise the HR, resulting in a higher resting HR.

6.5.2 Model Based Approach

To address this weakness, we developed an alternate, model-based approach, which learns a participant-specific HR recovery rate that can be used to estimate the time during which the heart-rate should recover, given the most recent peak heart-rate during physical activity and resting heart-rate before physical activity. An additional benefit of the model is that it summarizes the data succinctly in one parameter. Finally, computation of the recovery rate in the natural environment could serve as an indicator of cardiovascular fitness, similar to the 6-minute walk tests [34, 152] done in clinics.

Estimation of Recovery Rate: According to [69, 87], heart-rate after an arousal (e.g., activity) recovers exponentially (see Equation (6.1)). Figure 12, which plots one participant's heart-rate during a physical activity episode, illustrates this exponential recovery. In Equation (6.1), HR_{Rest} is the resting heart-rate before the physical activity episode, HR_{Peak} is end-of-activity heart rate at time t_0, and HRR is heart rate during the recovery period at time t. The constant τ represents the exponential recovery rate. Whilst there is a possibility that it can vary across time, our model makes a simplifying assumption of a constant participant-specific recovery rate.

After we have learned the recovery rate for a particular participant, we can use Equation (6.2) to estimate the recovery duration once physical activity is over.

$$HRR = HR_{Rest} + (HR_{Peak} - HR_{Rest})e^{-\frac{t-t_0}{\tau}} \tag{6.1}$$

$$t - t_0 = \tau \ln \frac{HR_{Peak} - HR_{Rest}}{HRR - HR_{Rest}} \tag{6.2}$$

To learn the recovery rate parameter τ for each participant, we first identify and isolate clean episodes where there is at least a 2-minute rest period (detected by accelerometry), needed to compute HR_{Rest}, followed by an activity period of at

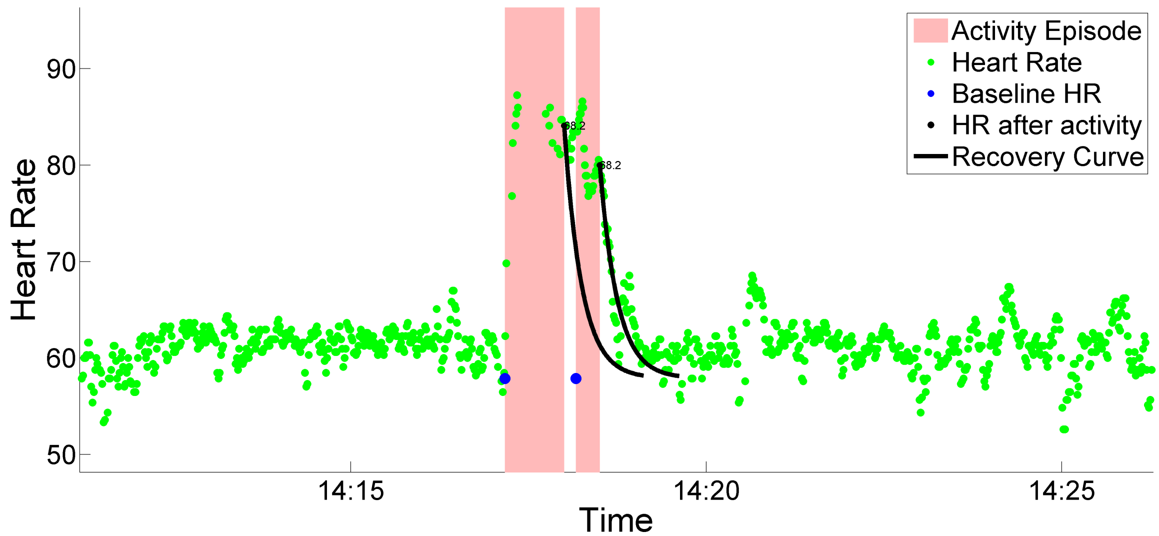


Figure 12: Exponential recovery parameter τ is learned for each participant. Black curves show 99% exponential recovery (Equation 6.1). In this case, before the heart rate fully recovers from the first episode, another activity episode occurs. Hence, baseline heart rate is carried forward.

least 2 minutes to represent a significant activity episode, and lastly at least a 2-minute stationary period so we can compute the latency to recover. Next, for each such episode, we derive HR_{Rest} as the median HR of the last one minute of the initial rest period, and HR_{Peak} as the median HR of the last 10 seconds of the activity period. Finally, we compute the times required for the HR to drop 10%, 20%, up to 90% of the total increase in HR from rest to peak — $[HR_{Peak} - HR_{Rest}]$. With these quantities defined for all episodes, Equation (6.2) can be used to learn τ using least-squares regression.

We computed the recovery rate τ for each participant. The mean of recovery rates across all 38 participants $\bar{\tau}$ is 19.8 seconds (SD=6.3). Participants' mean 95% recovery duration of 59.3 seconds (SD=18.9), is consistent with the literature [60].

Isolating and Excluding Activity Confounds: Figure 12 shows an example of the effect of activity on heart rate in daily life. For any such activity episode, we compute HR_{Rest} and HR_{Peak}. Then, we use Equation (6.2) and the learned value of τ to estimate the time interval ($t - t_0$) required for the heart-rate to return to resting heart-rate. Rather than requiring HRR to return to HR_{Rest} exactly, we consider the heart-rate that has dropped down to the line $HR_{Rest} + \sigma_{HR}$ as fully recovered, where σ_{HR} is the standard deviation of all heart-rates during stationary intervals. Adding σ_{HR} to HR_{Rest} allows for any natural variations in the resting heart-rate throughout the day.

Using this model, in addition to the entire physical activity interval, the estimated recovery interval $(t - t_0)$ that follows is excluded from analysis, i.e., considered missing for the purpose of stress inferencing. With this approach, only 7.4% of data (as opposed to 35%) are excluded due to recovery from physical activity, in addition to 22.7% that are directly affected by physical activity (for a total of 30.1% of all data).

6.6 Missing Data Imputation

Standard methods for finding trends in time-series data [17, 35] require continuous data streams. To apply these methods, we needed a method to impute the missing data. Missing data in time series of stress assessments can be due to unavailability of data or due to presence of confounder such as physical activity. Before imputation, we need to rule out the possibility that the data are Missing Not At Random (MNAR) [53]. We use the self-report item "Nervous/Stressed?" (Likert 1-6) to check the assumption of independence. To address participant biases, we use the z-score of self-report responses. We find no significant difference in self-reported stress during stationary moments and moments of physical activity ($p = 0.984$ on Wilcoxon signed-rank test, paired two-tail, $n = 31$). We also find no significant difference in self-reported stress between stationary and missing data periods ($p = 0.841$ on Wilcoxon signed-rank test, paired two-tail, $n = 24$). Therefore, we conclude that our missing data in stress assessments are not MNAR. They can be either Missing Completely At Random (MCAR) or Missing At Random (MAR) [53].

We believe that our missing data should be considered Missing At Random (MAR) [39] because stress can be explained by other known contextual variables [61, 70, 150] such as day of the week, time of day, previous stress levels, and the slope and intercept of previous time-series samples. We use these variables to impute the missing data using the K-Nearest Neighbor method proposed in [77, 169, 179].

We note that although we impute missing data to have a continuous time-series of stress assessments, JITI can be programmed so that it provides an intervention only when there are non-imputed sensor-inference data (data-loss <50%) with no confounding physical activity.

6.7 Field Validation of Stress Assessment

The previously-described cStress model captures the instantaneous physiological response to stress. Although this model was validated in both lab and field settings [89], before using it on our dataset obtained from polydrug users, we validate it against their field self-reports. We use the same approach described in [89] to map cStress output to self-report ratings.

Figure 13 summarize the F1 scores across participants. They range from 0.130 to 0.917 with a median of 0.717. Although the F1 scores are acceptable for majority of the participants, there are 5 participants whose low F1 score seem to suggest poor agreement between self-reported stress and the model output. We, therefore, analyze the consistency of their self-reports, because they may be subject to consistent bias or careless responding.

We use Cronbach's alpha [27] to assess the consistency of the self-reported responses. Cronbach's alpha measures the internal consistency of

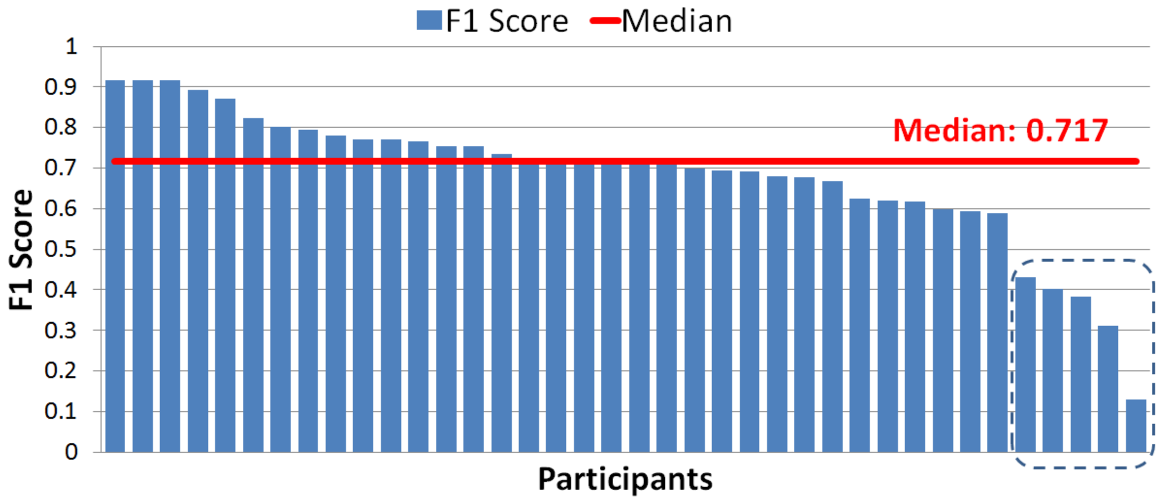


Figure 13: F1 score between self-report and sensor assessment range from 0.130 to 0.917 with median 0.717. Bottom 5 have unacceptable self-report consistency score with median cronbach's alpha score 0.335 while overall consistency score is 0.843.

items that intended to measure the same psychological construct. In most studies, an alpha score of 0.7 or higher is regarded as acceptable [27].

We compute the Cronbach's alpha using 5 affect items of self-report — "*Cheerful?*", "*Happy?*", "*Frustrated/Angry?*", "*Nervous/Stressed?*", and "*Sad?*" (The two positive items, "*Cheerful?*" and "*Happy?*", were reverse-coded). The overall consistency score across all participant's self-reports is 0.843. We compute Cronbach's alpha for the 5 participants from Figure 13 who show poor F1 score. They have unacceptable self-report consistency scores with a median Cronbach's alpha of 0.335. Furthermore, the participant with the smallest F1 score (0.13) answered "3" on item "Nervous/Stressed?" in 173 out of 177 self-reports, suggesting a bias toward neutral self-assessment. These observations also demonstrate the value of an objective sensor-based model of stress.

The above test not only demonstrated the validity of the cStress model in

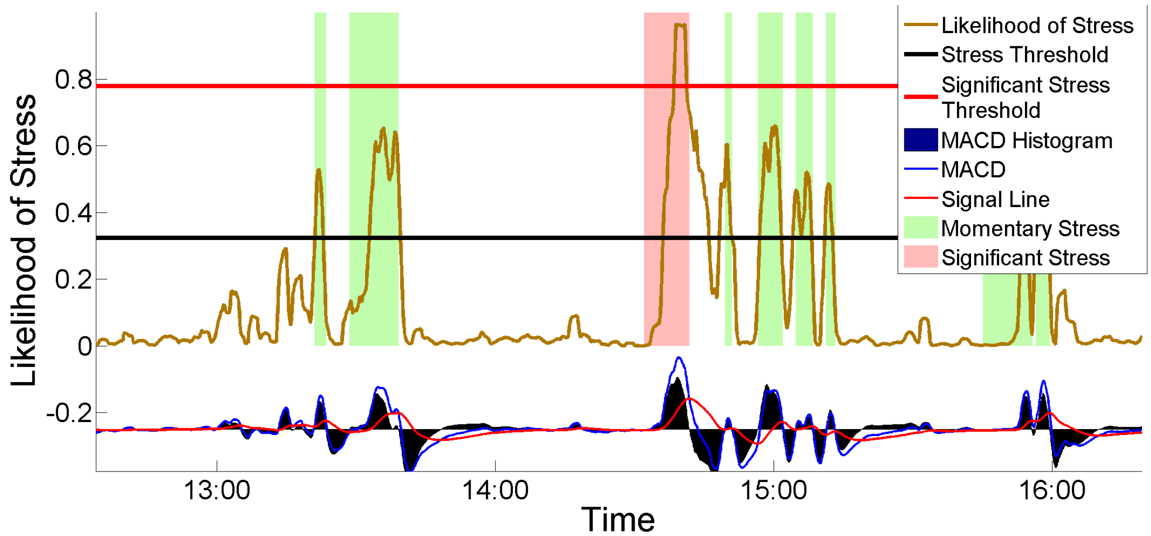


Figure 14: Timing of just-in-time stress intervention for momentary and significant stress episode. Starting of a rectangular region indicates timing of potential proactive interventions as generated by MACD.

our independent data set, but it also shows the effectiveness of the imputation process since this validation was done on the imputed time series.

6.8 Locating Stressful Episodes

There are two types of JITIs. Proactive JITIs are intended to precede and prevent an adverse event, such as an escalation of moderate stress to severe stress. Reactive JITIs follow an adverse event and are intended to mitigate its effects. Although we did not implement a JITI in the current project, we developed our assessment methods with that goal in mind. For either type of JITI, we need a method to determine from a time series of stress data whether a significant stress episode is occurring and if so, when it starts and ends.

To find significant stress episodes in our rapidly varying time-series data, we adapt a stock-prediction model. Such a model operates on a similar dataset, where there exist time series of stock prices and the objective is to predict the precise moments of buy or sell events, based on prior observations. Methods such as the Relative Strength Index (RSI) [182] and Bollinger Band [30] estimate whether stock is in an oversold or overbought condition and provide a buy or sell signal, respectively. “Oversold” means there are fewer people who can sell the stock relative to the number wishing to buy, indicating that the stock is undervalued and will eventually increase in price. The reverse is true for stocks that are overbought.

However, the assumptions that apply to stock prices do not hold for stress levels. If someone is extremely relaxed it does not imply that his/her stress level will go up as a consequence. Fortunately, this assumption is not built into the method we use, called Moving Average Convergence Divergence (MACD) [17], which has recently been used to detect trends in physiological data [87]. 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 stress-likelihood time series, MACD can provide a signal for a proactive intervention when the stress likelihood is going up and a reactive intervention when the stress likelihood is going down.

MACD is computed as follows:

$$\begin{aligned} M &= EMA(L; w_{slow}) - EMA(L; w_{fast}) \\ S &= EMA(M; w_{signal}), \end{aligned} \tag{6.3}$$

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 the sign of the difference between M and S indicates whether the trend is positive or negative.

Before applying MACD, it is important to address the fact that the stress-likelihood time series is rapidly varying and that it may contain inaccuracies as it is the output of a machine learning model that is rarely perfect. To account for this, we first smooth the stress-likelihood time series using a simple moving average with a 2 minute window length, a duration we selected based on visual inspection.

We tune the window length parameters, w_{slow}, w_{fast}, and w_{signal}, used in Equation (6.3), seeking to maximize $\frac{gain}{N}$, 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. Using a grid search with progressive zoom, with initial grids covering the range from 5 seconds to 30 minutes for each parameter, we found that the optimal window lengths are: $w_{slow} = 7.5$ minutes, $w_{fast} = 1.67$ minutes, and $w_{signal} = 14.2$ minutes.

Figure 14 shows a typical example of stress-likelihood time series, with colored boxes highlighting the positive-trend intervals, chosen by the MACD rule using the optimal window length parameters. As the figure illustrates, this approach is able to detect starts for good-quality positive-trend intervals in the stress-likelihood time series. Additionally, we show that stress densities for the minute after the detected positive-trend interval starts are significantly greater than those for the preceding minute ($p < 0.001$ on Wilcoxon signed-rank test, paired one-tail, $n = 15,434$). As an added bonus, we can use the MACD rule to comprehensively mark the start and end of each stress episode, defined as the interval containing a positive-trend interval and an immediately following negative-trend interval.

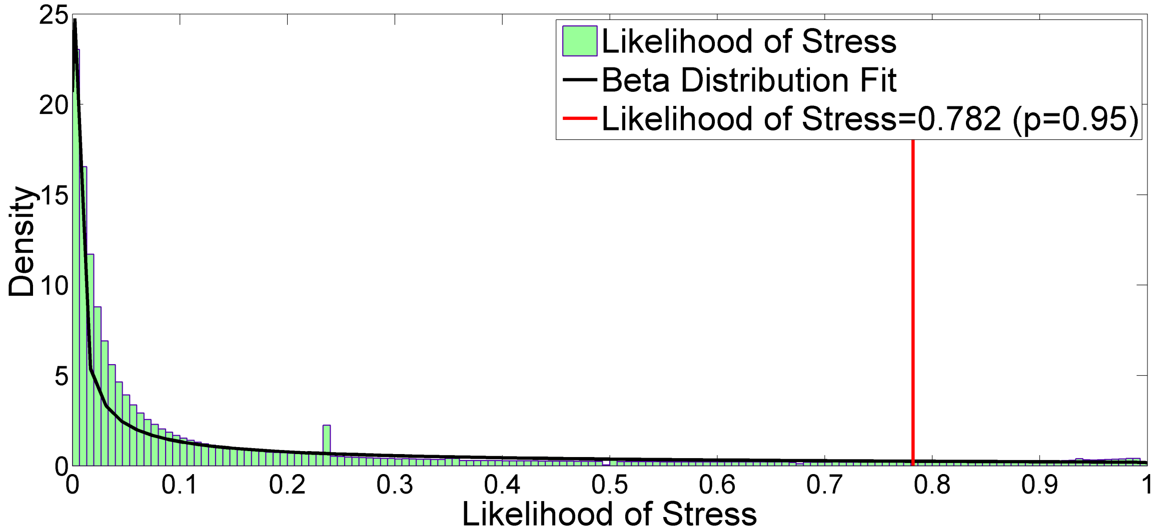


Figure 15: The likelihood of stress follow beta distribution with shape parameter $\alpha = 0.222$ and $\beta = 1.027$. The threshold for significant stress is 0.782 (p=0.95).

6.9 Parameterization of Episode Identification

As soon as we identify a stress episode we need to take a decision about whether we should provide an intervention. Hence, we need to learn parameters to identify stress episodes. We can do so via investigating the field study data or via investigating a lab study.

6.9.1 Approach 1: Based on Field Study

Defining Significant and Momentary Stress Episode: We define two types of stress episodes: Significant Stress Episode (SSE) and Momentary Stress Episode (MSE). MACD divides the stress-likelihood time series into smaller variable length, increasing and decreasing episodes. An episode in the time series is defined as an increasing trend, immediately followed by a decreasing trend. There are 15,434 such episodes. However, in some episodes, stress likelihood does not cross the binary stress classification threshold (from cStress). Such instances are discarded, leaving 9,087 episodes for further analysis. Significant stress episodes are those that have a high likelihood of stress and persist for a significant duration. All others are momentary.

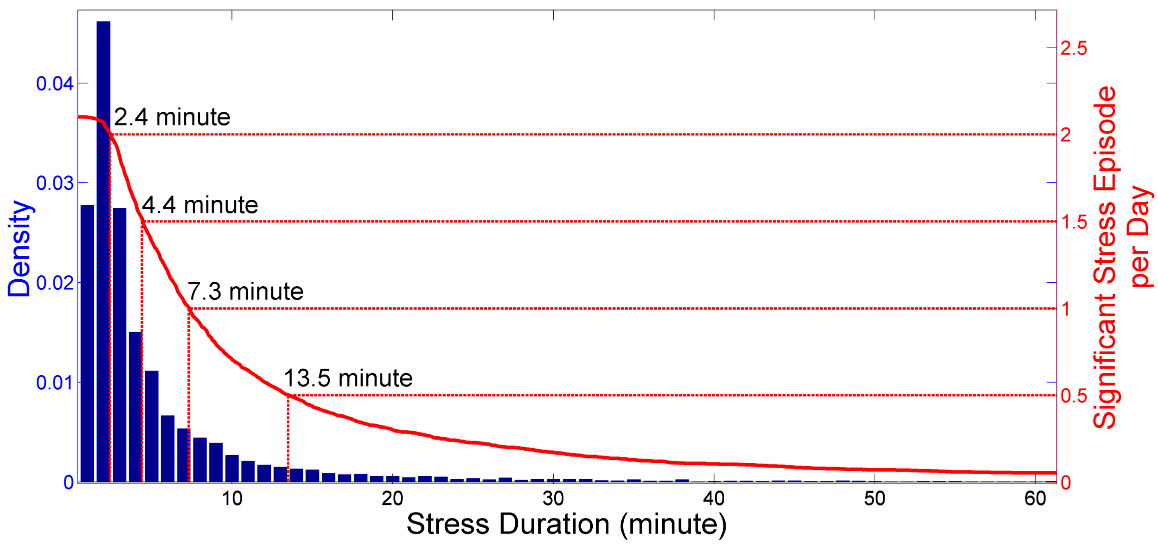


Figure 16: A momentary stress episode with high likelihood of stress (95th percentile) (see Figure 15) and a duration of more than duration threshold is marked as a significant stress episode. Duration threshold of 7.3 minute leads to an average of one significant stressful episode per day (in 10+ hours of sensor wearing time).

To decide which stress likelihoods are significantly high, we calculate a stress-likelihood threshold ν based on the 95^{th} percentile of stress-likelihood values. To address the between-participant differences, we calculate participant-specific thresholds, based on each participant's stress likelihoods only. All stress episodes with likelihoods above this threshold are marked as SSE candidates.

Figure 15 is a histogram of all stress likelihoods pooled together. As it shows, the stress likelihoods are right skewed and follow the Beta distribution with parameter estimates $\alpha = 0.222$ and $\beta = 1.027$. We had sufficient data for every participant, from which ν's could be easily found. If sufficient data are not available for a participant (e.g., when a participant has just begin providing data), we can compute ν based on the estimated parameters of the Beta distribution. In particular, the likelihood threshold ν can be calculated using the inverse Beta Cumulative Distribution Function (CDF), $F^{-1}_{Beta}(p = 0.95|\alpha = 0.222, \beta = 1.027)$.

Table 7: In total there are 9,087 stress episodes with an expected count per day of 9.2. A duration threshold of 13.5 minutes labels 498 significant stress episodes, with an expected daily count of 0.5.

Significant Stress Episode			Momentary Stress Episode	
Duration (minute)	Total Count	E(count) per day	Total Count	E(count) per day
13.5	498	0.5	8,589	8.7
7.3	997	1.0	8,090	8.2
2.4	1,992	2.0	7,095	7.2

Figure 16 illustrates how duration threshold, λ, informs the selection process for SSE candidates. We first select the desired number of significant stress episodes per day, d, and then, we can simply select the λ that corresponds to d episodes per day. The durations of SSE candidates follow the LogNormal distribution, with estimated parameters $\mu = 2.064$ and $\sigma = 0.871$. Out of 9,087 stress episodes, 2,082 contains high stress likelihood (2.1/day). Researchers who are in the designing phase of a stress intervention with no access to data, can calculate λ using the following formula:
$E(SSE/day) = (1 - F_{log_{Norm}}(\lambda|\mu = 2.064, \sigma = 0.871)) * 2.1$, where $F_{log_{Norm}}(\lambda|\mu, \sigma)$ is the LogNormal CDF.

The rule for identifying the SSEs is as follows — all those stress episodes that have stress likelihoods greater than the threshold of ν and persist for duration greater than λ. We identify other stress episodes as MSEs. Figure 14 shows several examples of SSEs and MSEs.

Table 7 summarizes descriptive statistics for SSEs and MSEs. In total, there are 9,087 stress episodes, with an expected daily frequency of 9.2. A duration threshold of 13.5 minutes labels 498 (or 0.5/day) as significant stress episodes.

6.9.2 Approach 2: Based on Lab Study

Previous approach involved very frequent stress assessments (every five seconds), which is not feasible to implement on a smartphone with limited computational capacity and battery life. In addition, 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, we can use 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. In addition, the presented method should also be sensitive to the resource limitations of mobile phones, so it can be deployed in a real-life. In Chapter 7 we will discuss about this lab study based approach.

6.10 Chapter Summary

This chapter proposes a method to identify stress episodes from a discontinuous time series of mobile sensor data. In addition, presented work makes several methodological contributions. First, our method of estimating the recovery time of physiology from a physical activity episode could possibly be used as a measure of cardiovascular fitness outside of controlled settings for heart patients. Second, missing data in the stress-likelihood time series is not Missing Not At Random (MNAR), which enables us to do imputation and obtain a continuous time series. Third, Validation of sensor inferred markers is the field setting is challenging due to lack of gold standard truth. Self-reported assessment is commonly used for this validation [89]. Presented work show that lack of agreement between self-reported stress and sensor inferred stress can subject to inconsistent self-report. Fourth, proposed a method to find trends in the stress-likelihood time series and identify stress episodes based on the distribution of stress likelihood and stress duration.

Chapter 7

Identifying Stress Episodes Based on Lab Stress Data

7.1 Introduction

The stress episodes identification approach proposed in Chapter 6 is parameterized based on a field study data. It is not clear whether these parameters will be directly applicable in other settings (e.g., other population). This chapter discusses about an alternate approach where we learn the model parameters based on a laboratory study data and validate it in an independent smoking cessation field study data.

Smoking cessation is an important health issue because smoking causes the largest number of deaths, accounting for one in every five death [67, 128, 128]. Smoking is very difficult to treat as most smokers trying to quit eventually lapse. Stress is one of the major triggers for smoking lapses [22, 49, 164], and it is usually elevated in early phases of smoking cessation, which is when most lapses occur [13, 49]. But, individuals who continue to be abstinent experience a gradual decrease in their stress level [45].

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). Hence, participants are already being interrupted many times in the day (e.g., 10+ EMA prompt). Therefore, engaging in a stress intervention that can also be perceived as an interruption, such as, breathing exercise or meditation, may add further to their already heavy daily burden. 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 in this chapter 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.

Stress is prevalent among nicotine dependent individuals, especially during their abstinence. We applied the 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.

7.2 Overview

Figure 17 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.

7.3 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

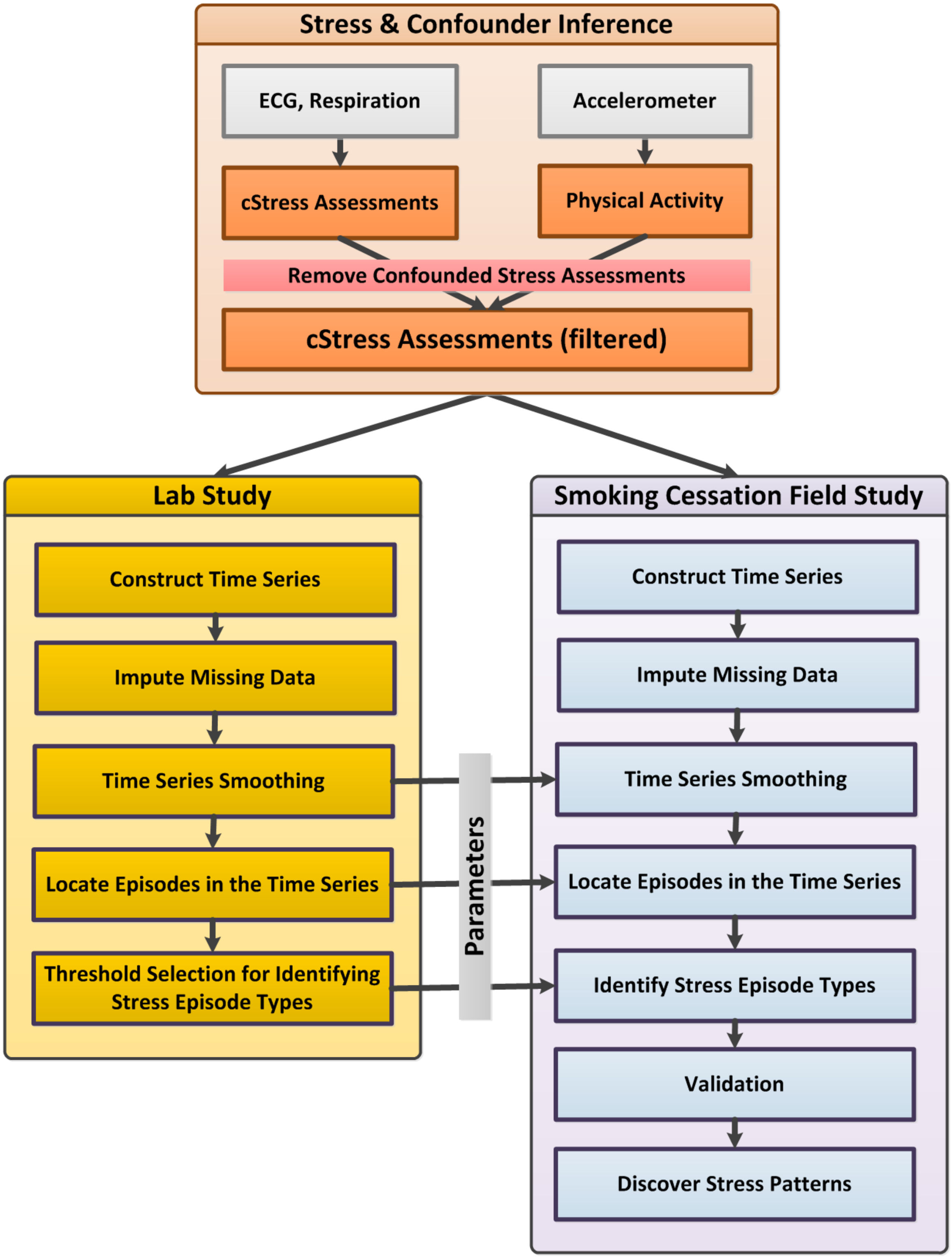


Figure 17: Overview of the approach. First, we infer stress from ECG and respiration data, and confounder physical activity from accelerometer. Second, we remove physical activity confounded stress assessments. Third, we develop our *stress* episode identification model on lab study and apply the model on smoking cessation field study. Finally, we discover stress patterns from the smoking cessation field study.

of lab stress study. The data description of smoking cessation field study is presented in Section 7.6.

7.3.1 Wearable Sensor Suite

The sensors worn by the all participants in both studies are part of a large suite of wearable biosensors, called AutoSense [59]. These unobtrusive sensors are worn mostly under the clothes, and include a two-lead electrocardiograph (ECG), 3-axis accelerometer, and respiration sensors, among others. A description of this AutoSense sensor suite is available in Chapter 2 and Section 2.2.1.1.

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.

7.3.2 Lab Stress Study

We use ground-truth labeled data collected in a lab study that was reported in [89, 145]. 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 [12]. 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.

7.4 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* [89] summarized in Chapter 2 and Section 2.2.2.1.

As stated previously, 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 referred to as *stress likelihood*. To assess stress within intervals longer than a minute, we use a different measure, called *stress density*, from [158]. 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).

7.4.1 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 [151]. 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 [136, 151], or excluded due to the presence of physical activity confounds introduce discontinuity in the stress likelihood time series. In Chapter 6 (also in [158]), missing data was imputed using via k-nearest neighbor method [77, 169, 179] 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.

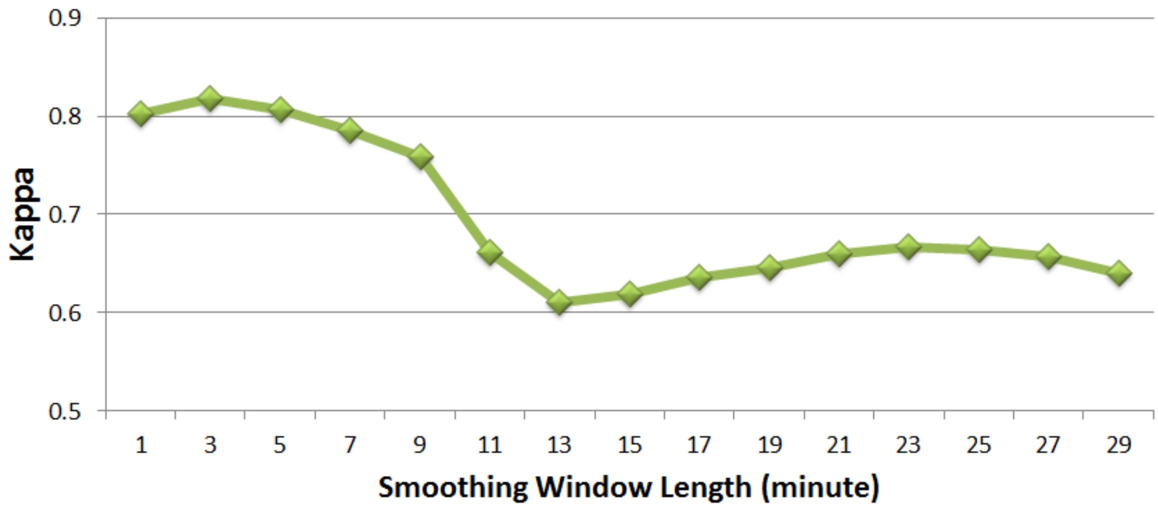
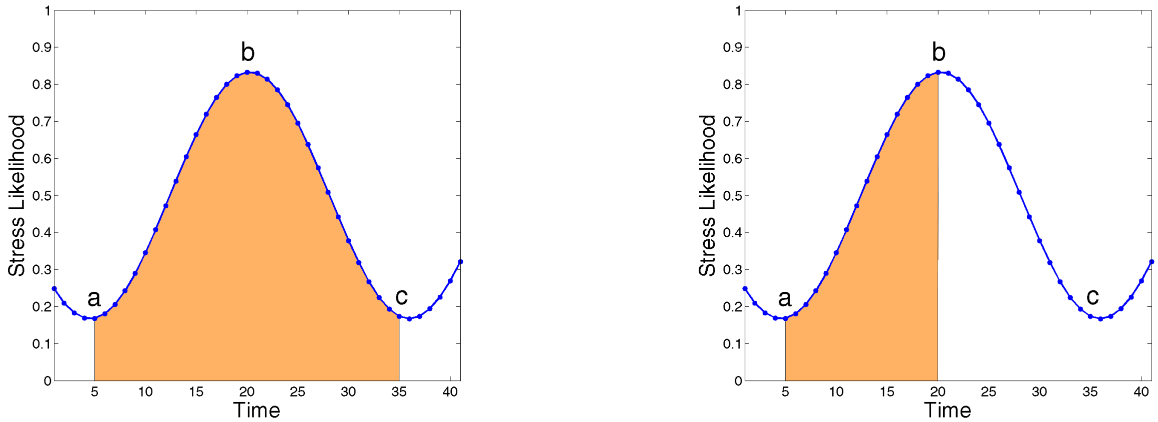


Figure 18: Classification performances for different smoothing window length applied on stress likelihood time series in the lab study. We get the best performance with a kappa of 0.817 for a window length of 3 minutes.

7.4.2 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 as proposed in Chapter 6. However, in compare to doing a visual inspection to find the optimal window length parameter of this simple moving average, we can learn this parameter based on the lab study. In order to find the optimal window length, we compare the original labels (derived from the lab stress protocol) with each one minute assessment in the smoothed *cStress*-based classification. Figure 18 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.



(a) Intervention at 'c' (b) Intervention at 'b'

Figure 19: 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*. Similarly for an intervention at 'b' we compute stress density upto 'b'.

7.5 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 19 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*'). As an alternate approach, using a similar approach – computing stress density at 'b' we can provide an intervention when it is highly likely that the person is stressed. An example of an appropriate intervention can be the recommendation of a breathing exercise [102], allowing the person to be better prepared for subsequent stress occurrences.

In this chapter, we primarily discuss about 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.

7.5.1 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. To identify episodes and find trends in the time series we follow the similar Moving Average Convergence Divergence (MACD) based approach proposed in Chapter 6 and Section 6.8. However, rather than using the field study data to learn three window length parameters <w_{slow}, w_{fast}, w_{signal}> of MACD we use lab study data. We found that the optimal window lengths are: w_{slow} = 19 minutes, w_{fast} = 7 minutes, and w_{signal} = 2 minutes, which maximize the metric $\frac{gain}{N}$. In the lab time series using the specified parameters, we obtained 119 episodes across 21 participants.

7.5.2 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.

Out of 119 episodes in the lab study, 24 are *unknown* due to missing data or poor quality data. Figure 20 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 8 summarizes the calculation of precision and recall for *stressed* and *not-stressed* class. Table 9 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 10 summarizes these results.

In an alternate approach when we provide intervention when it is highly likely that the person is stressed (see Figure 19 b), we compute stress density at 'b'. Based on a threshold of 0.36 we are able to identify *stressed* with a precision of 95.7% and a recall of 0.88% and *not-stressed* with a precision of 86.7% and a recall of 95.1%.

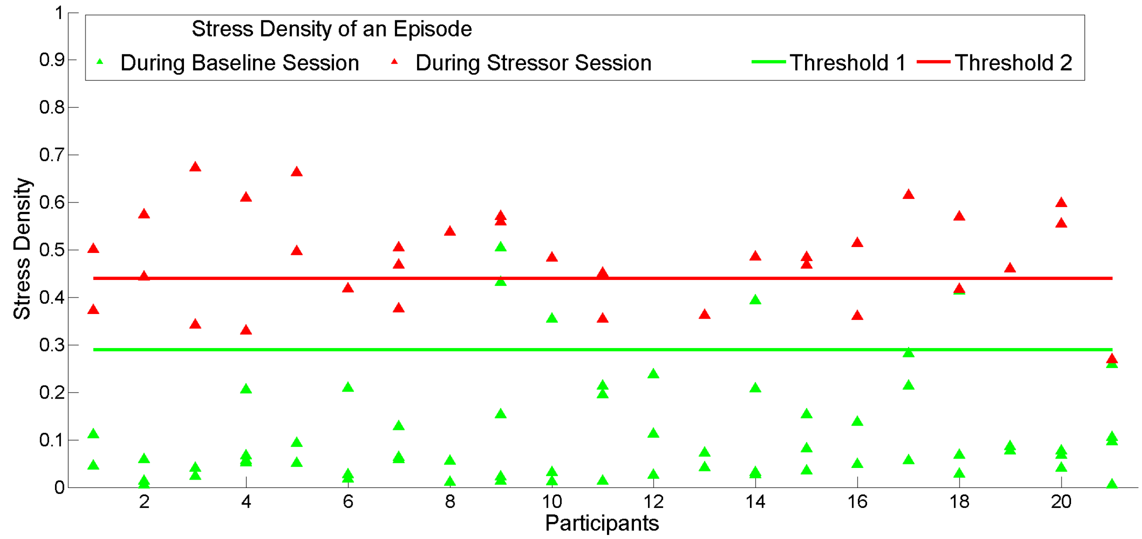


Figure 20: Stress density of each session in the lab study. Discarding episodes with stress density between two thresholds (0.29 and 0.44) ensures both precision and recall of *stressed* and *not-stressed* class above 95% with episodes discarded due to being *unsure* is minimum.

7.6 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.

7.6.1 Data Description

Participants: We use data collected in a smoking cessation study that was reported in [155]. 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

Table 8: Computation of *stress* episodes classification performance metric — precision and recall from Figure 20

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 9: 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

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.

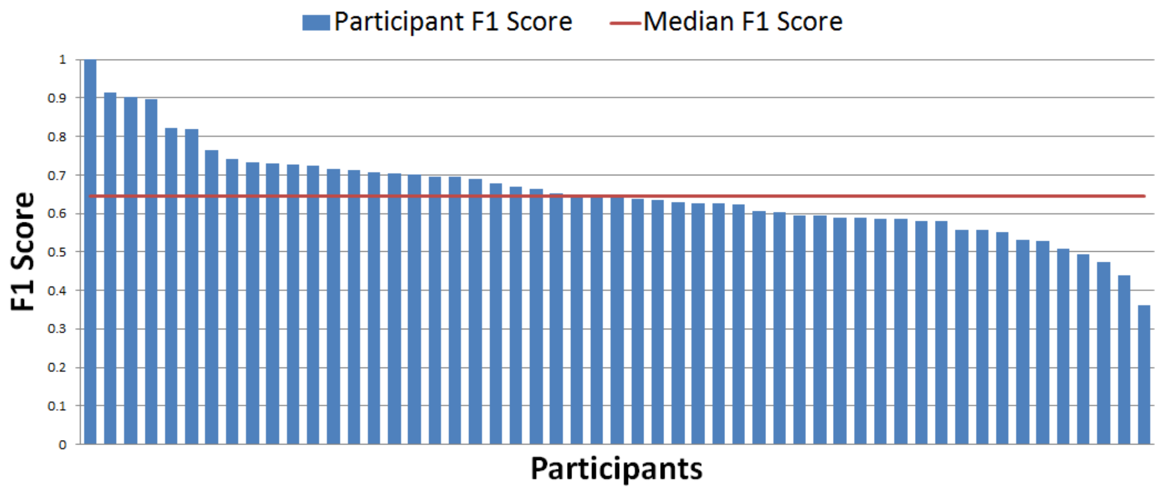


Figure 21: F1 score between self-report and sensor assessment range from 0.36 to 1.00 with median 0.65.

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) [167] 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.

7.6.2 Validation of Stress Assessments in the Smoking Cessation Study

The *cStress* model was validated against lab study and independent field studies [89, 158] as described earlier. To validate the *cStress* assessments in this new data set, we followed the similar approach presented by Hovsepian et al. [89]. First, we check the consistency of self-reports as they are subject to bias and careless responding [158].

We use Cronbach's alpha [27] 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 [27] 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 21 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 [89] and 0.72 in [158].

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 [158] (0.76 vs. 0.84). In general, the median F1 score of 0.72 in [158] 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 21). 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 agreement 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 [89]) 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 Chapter 6(also in [158]), but using a heavy-weight and potentially more accurate method. In contrast, we can 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 Chapter 6(also in [158]), 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.

7.6.3 Stress Patterns Observed in the Smoking Cessation Study

We apply the approach proposed in Section 7.4 and Section 7.5 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 7.5.2, 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 10, we find 28.3 *not-stressed*, 2.7 *unsure*, and 1.5 *stress* episodes per day on average. Figure 22 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.

7.7 Chapter Summary

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 chapter, we

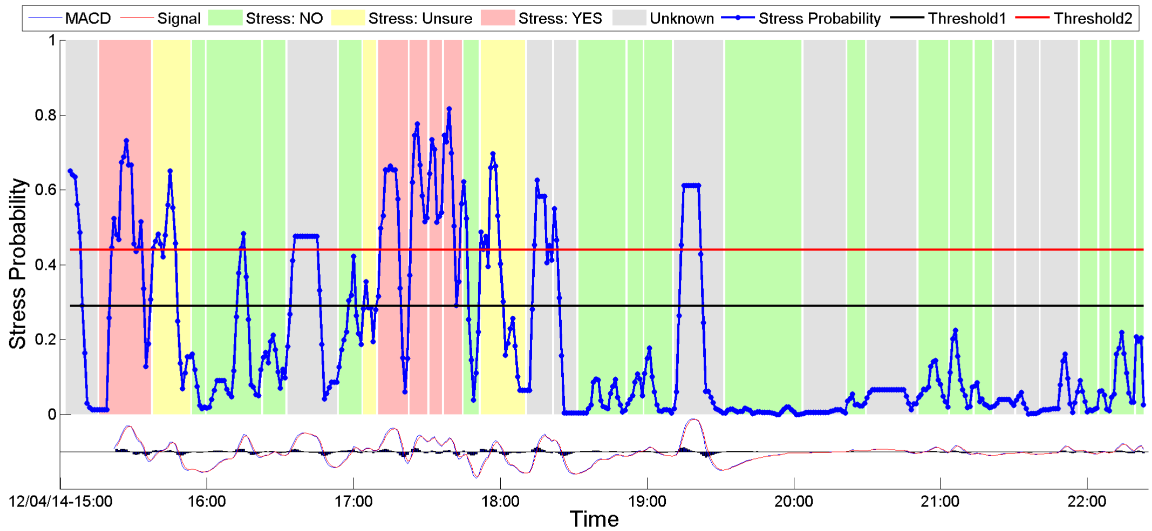


Figure 22: Time series of stress likelihood of one participant on pre-quit day.

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.

Table 10: *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

Chapter 8

Applications of Our Model

To demonstrate the utility of our model, we investigate four possible applications of the model. First, triggering of the self-report for self-reflection. Second, observe the patterns of stress which will help intervention designer. Third, provide proactive and reactive intervention based on context. Fourth, generalize the model for other possible interventions.

8.1 Application 1: Triggering of the Self-Report

First application for identifying stress episodes will be to initiate a self-report at the conclusion of a stress episode. We can ask the user about the reason for past stress experience. For example, is this stress experience related to work, family relationships, health, housing, crime, etc? Recording this response information in the smart phone or in the cloud will help the user to revisit those recordings. Such self-reflections will enable the user to find patterns of stress in his daily life [94, 98, 113, 120, 162]. For instance, via self-reflection user finds that Monday morning at work is stressful. Information like this will enable the user to better prepare for a stressful Monday (e.g., meditation at morning). Obtaining this information also enables intervention designers to find patterns of stress and provide intervention at appropriate moments.

8.2 Application 2: Patterns of Stress

8.2.1 Role of Prior Stress

We analyze the relationship between durations of successive stress episodes. Figure 23 is a scatter plot of the duration of the current stress episode versus the duration of the preceding stress episode. We observe a healthy correlation of 0.42. This correlation can be explained by theory and evidence [84, 85, 133] suggesting a spiral process where current exposure to stressors can lead to subsequent reactivity to other stressors by attenuating the

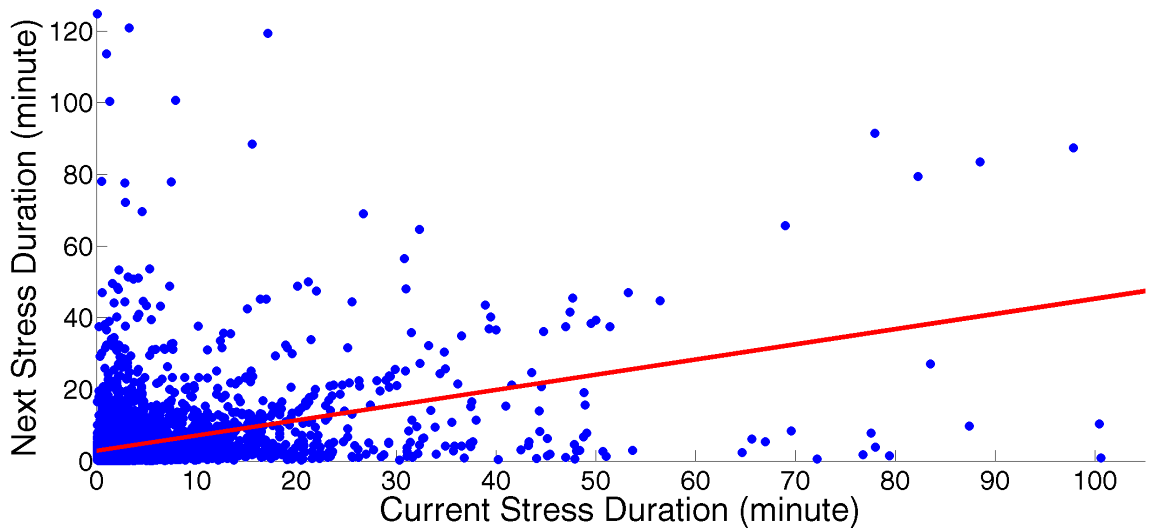


Figure 23: Next stress duration as a function of current stress duration. A healthy correlation of 0.4243 is observed here.

state coping capability of the person. For example, stressors such as facing financial troubles may decrease the person's stress coping capacity. This may lead the person to respond with subsequent stress to an event or an environment that would, in other circumstances, be easy to deal with, such as being in a noisy environment.

8.2.2 State Transition Probability

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 24 shows the estimated transition probabilities between these types of episodes for the field study of 53 participants.

Stress episodes more likely to be of similar kinds in successive episodes. From Figure 24, 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 [158] as well that there is a correlation between the durations of successive *stress* episodes. This can be explained by theory and evidence [84, 85, 133]

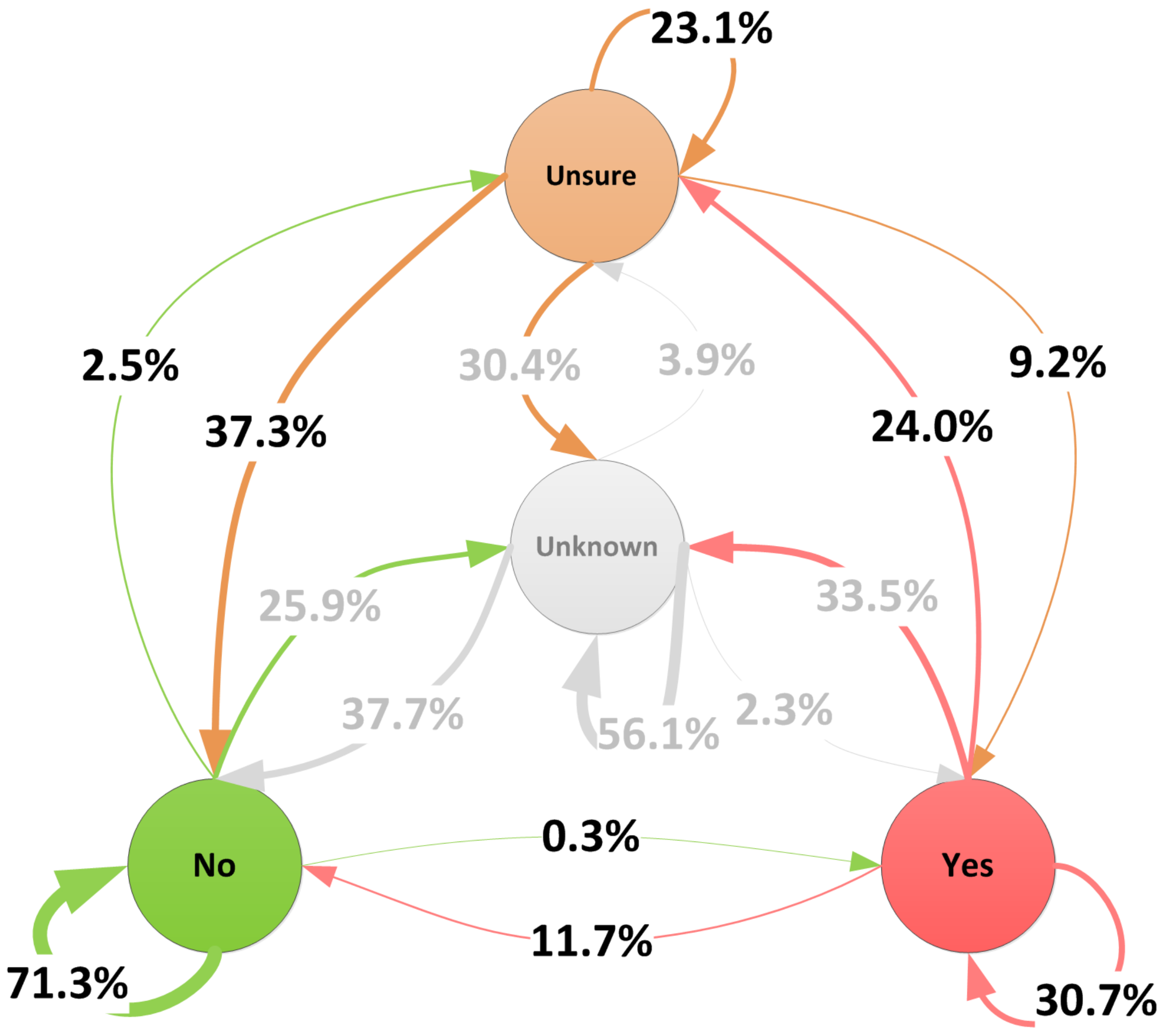


Figure 24: State transition probabilities between different *stress* episode types, *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*.

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

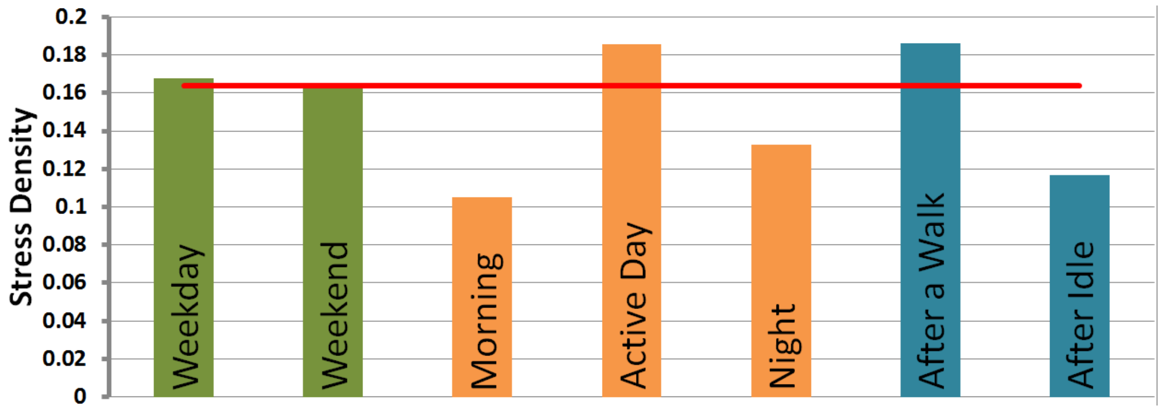


Figure 25: Role of time and activity level on stress density. Here, morning is before 8 AM, day time is 8 AM to 7 PM, and night is after 7 PM. Horizontal (red) line represents the overall stress density.

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.

8.2.3 Temporal Effect on Stress

We do not observe any significant difference in stress level between weekdays and weekends (0.168 vs. 0.163, $p = 0.744$ on Wilcoxon signed-rank test, paired two-tail, $n = 38$). Most of our participants did not have full-time jobs; this may explain the absence of a difference.

As hypothesized in [105], we observe that in our sample, stress varies by time of day. It is low in the mornings, rises during the middle portion of the day, and subsides again at night. These differences were significant in pairwise comparisons of midday versus morning (0.186 vs. 0.105, $p < 0.001$ on Wilcoxon signed-rank test, one-tail, $n = 38$) and midday versus night (0.186 vs. 0.133, $p = 0.001$ on Wilcoxon signed-rank test, one-tail, $n = 38$), and not morning versus night (0.105 vs. 0.133, $p = 0.055$ on Wilcoxon signed-rank test, one-tail, $n = 38$). These are expected observations, as the active day is likely spent looking for work and drugs and being exposed to drug cues and potential conflicts. Some of these events may occur during evening and night times as well, but are less likely than during the daytime.

8.2.4 Effect of Activity on Stress

Even after we remove the confounding periods of moderate to high physical activity, we still find that stress density for the next 15 minutes after a walk is higher than usual, as shown in Figure 25. In contrast, stress density was lower in the 60 minutes following 60 minutes of inactivity, (which generally happen at home) (0.186 vs. 0.117, $p = 0.001$ on Wilcoxon signed-rank test, paired one-tail, $n = 38$).

This observation seems to contradict the common belief that physical activity such as walking helps to reduce stress [51]. This apparent contradiction could be because our participants' physical activities usually corresponds to transportation (e.g., walking and public transport). Upon conclusion of these episodes, they may be exposed to cues, unpleasant environments, work challenges, etc. They could also have been engaged in jobs that required significant physical activity. This observation prompted us to investigate the role of environmental context in stress.

8.2.5 Environmental Effect on Stress

To analyze the effect of environment on stress, we use the Neighborhood Inventory for Environmental Typology (NIfETy) [71] as a measure of environmental

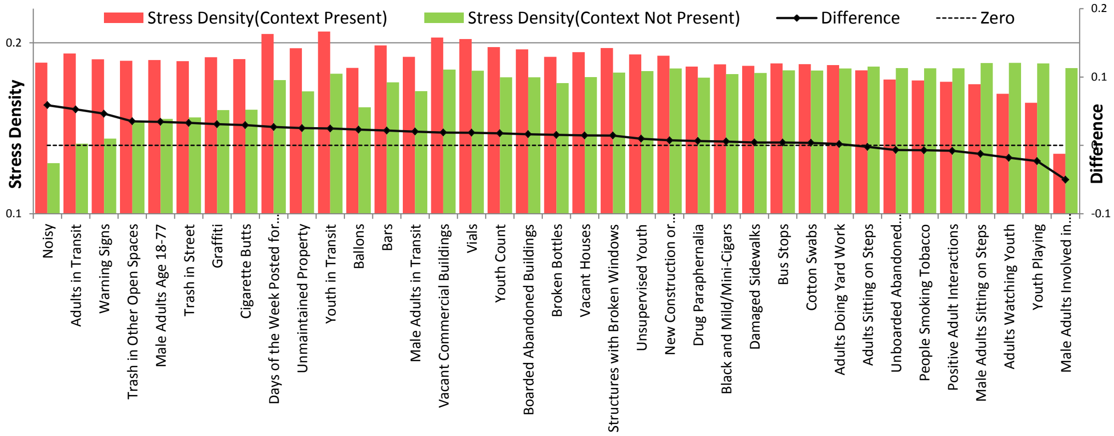


Figure 26: Effect on stress density across different location contexts detected with $\kappa > 0.7$. Noisy environment is highly associated with stress.

disorder. GPS data is mapped to this index. The collection of NIfETy data has occurred in several waves, starting in 2005. We use data from Wave Eight, because they were collected close in time to our participants' provision of GPS data. During Wave Eight, trained NIfETy raters sampled 528 individual georeferenced blockfaces in the city where the study was conducted. The raters noted the presence or absence of each of 77 variables, which were divided *a priori* into five categories: (1) Social Disorder, (2) Physical Disorder, (3) Drug Paraphernalia, (4) Adult Activity, and (5) Youth Activity.

Method: To estimate probable NIfETy ratings for the areas between the 528 rated city blockfaces, we develop a model that incorporated data from remote-sensing-derived maps of surface imperviousness and landcover [185]. The remote-sensing data consist of 180,000 pixel values measured as an image across the city. Next, we use a distance matrix to measure the distance between all NIfETy blockfaces and the centroid coordinate location for individual pixels in the remote sensing image of the city. We complete the distance measurements iteratively, where the first matrix is the distance from each of the 180,000 pixels to the closest NIfETy blockface. The second iteration is the distance from each pixel to second-closest NIfETy blockface. This process is replicated with the distance matrix for all 528 NIfETy blockfaces, so that we have 528 distance layers for each of the 180,000 pixels. These layers are then rasterized for the city and sampled for each NIfETy location.

Next, we develop a RandomForest based classifier [33] to predict a dichotomous outcome (i.e., 0 = “absent” or 1 = “present”) for each of the 77 NIfETy variables, using the 2 remote sensing layers, coordinate location, and the 528 distance values. We reason that with the distance values included, the machine-learning model would generate predictions similar to those of Kriging, a common geospatial interpolation method that uses distance alone to make its predictions [58]. By adding remote-sensing data to our model, we account for real-world physical environments in the city.

We then generate a citywide map of inferred probabilities for each of the 77 NIfETy variables at each pixel. We use Cohen’s kappa to compare model-inferred probabilities to actual ratings at the NifETy blockfaces (representing a gold standard). Only NIfETy values with a kappa greater than 0.4 are used in our analysis here (n=61) as predictors of stress ratings. The posterior probability computed by the Random Forest model is used to infer the binary labels: “absent”/“present”, using 0.5 as the binary threshold.

Findings: Figure 26 presents the stress densities across 37 different location contexts, for which the classification $\kappa > 0.7$, distinguishes between cases where the context is present and absent. We observe that *noisy* locations; the presence of *graffiti*, *cigarette butts*, *trash in street*, and *bars* are associated with high stress likelihood. Bars may be a potent cue for drugs and hence may elevate stress in our population. In contrast, locations where the NIfETy raters had seen *male adults involved in positive interaction* and *youth playing* are associated with lower stress than average.

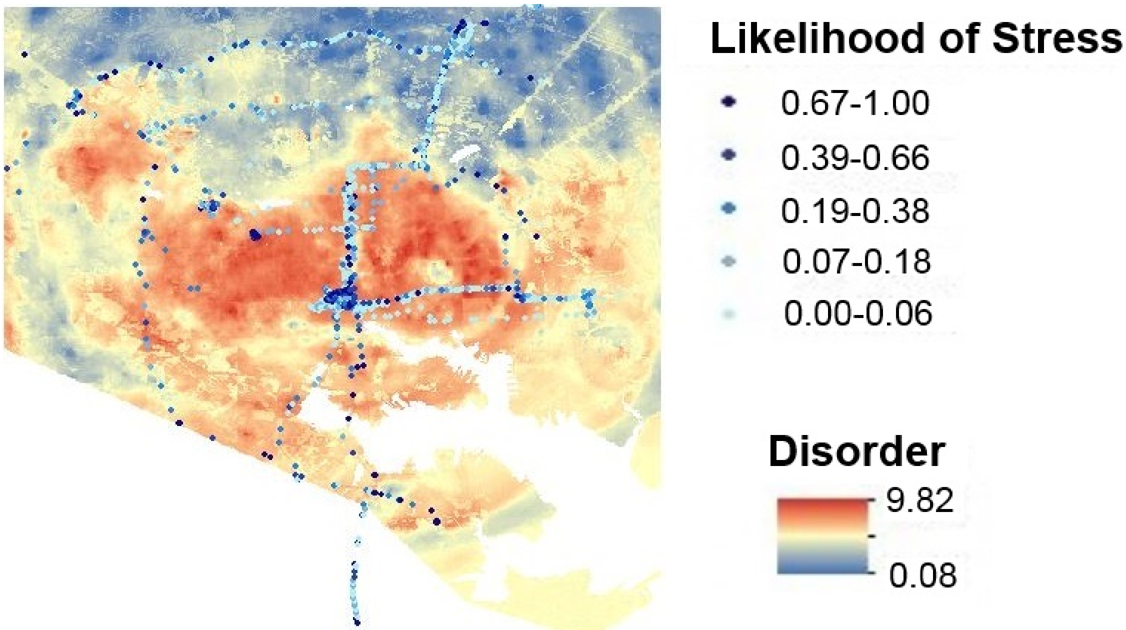


Figure 27: The likelihood of stress for one participant overlaid on the disorder map. Disorder here is the aggregated posterior probability value for top 10 NIfETy variables (see Figure 26) with $\kappa > 0.70$.

This suggests that geolocation tracking can help inform the timing of JITIs, that might, for example, propose a relatively less stressful route. As an example, Figure 27 shows one participant's stress assessments overlaid on disorder map of the city. Disorder here is the aggregated posterior probability value for the top 10 NIfETy variables with $\kappa > 0.70$. The figure suggests that people are more likely to be stressed in some specific parts of the city with high disorder score.

8.3 Application 3: Intervention

As the third application of identifying the stress episode, we can provide just-in-time intervention. Stress intervention can be proactive or reactive by nature.

8.3.1 Reactive Stress Intervention

As shown in Figure 28 we can provide a reactive intervention at 'c' or at 'd.' As soon as an episode is over (at 'd') we can compute the area under the curve. Divide the area by time and compute stress density. If stress density is over a

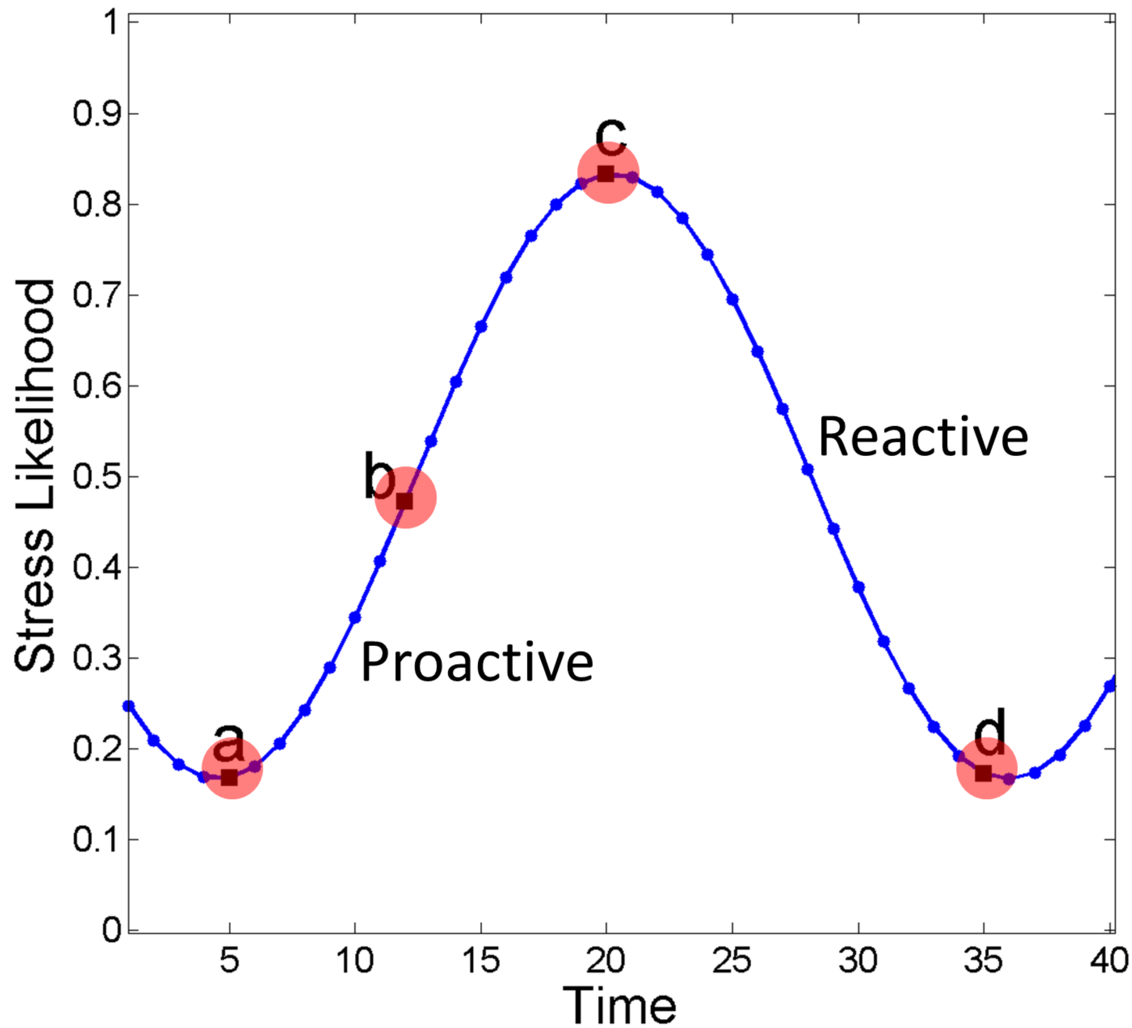


Figure 28: Timing for proactive or reactive intervention.

specific threshold we can provide stress intervention. We can pick this threshold based on a targeted precision and recall (e.g., 95%). Chapter 7 and Section 7.5.2 discusses about the threshold selection process. A person experiencing stress may not be available for receiving intervention (see Chapter 5). We observed from Figure 24 that a person recently experienced a stress episode is vulnerable for future stress occurrences. In addition, at ‘d,’ when stress likelihood is least, it is highly likely that the person is available. So it makes sense to provide a active intervention that requires significant user engagement. Meditation and breathing exercise are examples of such active interventions.

On the other hand, at ‘c’ when it is highly likely that the person is experiencing stress, we can follow the same approach and compute stress density. We can provide intervention if stress density is above some cutoff. However, when experiencing a stress episode, a person is less likely available (see Chapter 5). So we can provide a passive intervention which does not require active user engagement. Changing ambient light [117] or playing music [138] are examples of such passive interventions.

8.3.2 Proactive Stress Intervention

As another application of our model, we employ it to train a classifier for predicting significant stress episodes. As described earlier, we use the MACD method to identify and locate stress episodes. All stress episodes, momentary or significant, are considered candidate windows during the training process. Our goal in this prediction task is to determine early on, as soon as an MSE is detected, whether it will become an SSE, which essentially becomes a MSE/SSE classification task. For this task, we identify and compute 173 candidate features, and then train a model with 100 selected features.

Feature Computation: We compute 173 features to train a MSE/SSE classifier. These features are based on the observations and findings presented earlier.

Time and Day (3 features): As shown in Figure 25, there are temporal factors that affect stress, such as time of day. Therefore, we include the following features: “time of day,” “hour of day,” and “weekday”.

Previous Stress Episode (3 features): As shown in Figure 23, durations of adjacent stress episodes are correlated. Hence, we include the features “duration of previous stress episode,” “time since previous episode,” and “time required to cross binary stress threshold.”

Slope and Intercept (22 features): We use the slope and intercept of a best-fit line, fitted to past stress likelihood values. The rationale behind the inclusion of this feature was an assumption of a “calm before the storm.” In addition, a fast ramp-up of the stress likelihood has a good potential to break into an SSE. To compute these features, we use the slope and intercept associated with the crossing of the binary stress threshold. We also use the slope and intercept of prior 30 sec, 1 min, 2 min, etc., up to 10 min.

Prior Stress Density and Skewness (30+30 features): Figure 23 suggests that the prior stress density is correlated with the current stress density. Hence, we compute the stress densities of the previous N minutes, where N increases from 1 to 30. We also compute the skewness of the previous N minutes, varying N from 1 to 30.

Location (61 features): Figure 26 shows the apparent effect of location on stress density. We use 61 NIfETy scores out of 77 which are detected with performance $\kappa > 0.4$.

Physical Activity (24 features): Figure 25 shows that there is a significant association between the post-walk period and a high stress likelihood. Inspired by [157], we use 24 aggregated features of activity (All-N, Any-N, Duration-N, and

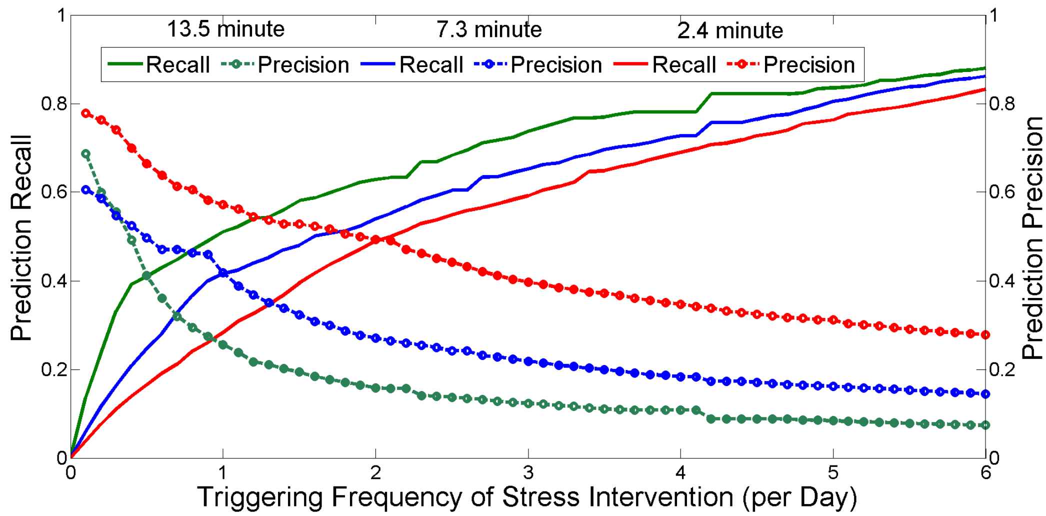


Figure 29: Trade-off analysis for triggering frequency of stress intervention. The x-axis represents model-proposed triggering frequency of stress intervention per day and two y-axes represent precision and recall for predicting SSEs.

Change-N) over windows of varying size N — 5 min, 10 min, 15 min, 20 min, 25 min, and 30 min.

Feature Selection: To improve the generalization performance of the classifier, we perform feature selection and retain only the top 100 features with the highest information gain [48]. This ensures approximately one feature for every 100 samples (total 9,087 samples).

Model: We train a RandomForest learning algorithm [33] to discriminate between MSEs and SSEs. To address the issue of imbalanced class sizes, we use a cost-sensitive classification approach [52], assigning a higher cost to misclassifications of actual SSEs. For evaluation, we use leave-one-subject-out validation.

Table 11 summarizes the performance of our model. The model is able to predict SSEs with a duration of 13.5 minutes with accuracy of 94.8% and $\kappa = 0.444$. Figure 29 shows the tradeoff analysis. The x-axis represents a triggering frequency of stress intervention per day and the two y-axes represent precision and recall for predicting SSEs. Researchers designing an intervention

Table 11: Performance of the model for predicting Significant Stress Episodes for duration thresholds of 13.5, 7.3, and 2.4 minutes.

Duration (minute)	E(count) per day	Accuracy	Kappa
13.5	0.5	94.8%	0.444
7.3	1.0	88.3%	0.428
2.4	2.0	77.7%	0.495

can use this information to find a triggering frequency that will achieve specific values of precision and recall.

8.3.3 Phone Implementation

Finally, the proposed intervention timing design considerations are presented to behavioral scientists. Out of the proactive and reactive approach, behavioral scientists preferred the reactive one because of two primary reasons. First, duration threshold that we used in identifying significant stress episodes requires some domain knowledge to select, which is unavailable during the designing phase of the intervention. Second, when we are providing intervention, we need to have high confidence that the user is stressed. A false positive intervention trigger will increase user burden.

As shown in Figure 28 we can provide a reactive intervention at 'c' or at 'd.' Out of the two, behavioral scientists preferred the intervention triggering timing at 'c' which will enable us to provide intervention when we have high confidence that the user is currently experiencing stress. Proposed method is also 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 [4]. This implementation is being used in an ongoing stress intervention study at Northwestern Medical School under the supervision of Dr. Bonnie Spring.

8.4 Other Applications

In this dissertation about just-in-time intervention, we use stress intervention as a running example. We address confounding events, handle missing data, construct the time series, validate the stress assessments, identify episodes in the time series, discover patterns in the time series, and build models to provide proactive or reactive intervention. This proposed method can be generalized to design intervention for other adverse health conditions, such as, craving for smoking, food, or drug. As a specific example, a recent work [40] developed a computational model to estimate cigarette craving using mobile sensor data. This Conditional Random Field (CRF) based model construct a time series of craving probabilities via inferring cigarette-craving for each minute. We can devise the similar time series pattern mining method that is described in this dissertation to find episodes in the time series, identify patterns, and build models to provide proactive or reactive intervention.

8.5 Chapter Summary

Identification of the timing of intervention is a critical step for the success of just-in-time intervention. Proposed approach of stress episodes identification opens up enormous opportunities for the behavioral scientists to design context sensitive intervention content and modality. It is now possible for the users to self-reflect, or the behavioral scientist to observe stress patterns and provide proactive or reactive intervention.

Right now, it is 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, whether we should provide an intervention when somebody are not stressed so that they can better tolerate future stress episodes. Which one leads to better and efficacious intervention? These are things that can be investigated via conducting a microrandomized trial.

Chapter 9

Conclusion and Future Directions

9.1 Summary and Key Contribution

Sensor-triggered just-in-time-interventions (JITI) promise to promote and maintain healthy behavior. But, critical to the success of JITI is determining the availability of the user to engage in the triggered JITI.

Scheduled and context-sensitive interventions do not adequately support just-in-time-interventions as they do not consider a user's cognitive, physical, and physiological availability to engage in a triggered intervention. This dissertation takes a first step to inform the timing of delivering JITI. We propose a novel objective metric to measure a user's availability to engage in a JITI and propose a model to predict availability in the natural environment based on data collected in real-life. Findings indicate that availability of a user depends not only on user's ongoing activity or physical state, but also on user's psychological state.

Context-sensitive just-in-time interventions have been possible for quite some time for applications such as traffic-aware navigation. GPS sensors have also made it possible to explore interventions that are based on geofencing. Identifying the appropriate timing of intervention is a critical component in a just-in-time stress intervention. This dissertation presents the first approach to analyze the time-series of stress data for determining the timing of just-in-time stress intervention. Given the wide prevalence of stress and its adverse impacts on health, job performance, and quality of life, stress management is useful for everyone. This work opens up numerous opportunities to now design efficacious interventions for helping dealing with daily stress in work life, social life, or otherwise. For the specific population addressed here — outpatients undergoing treatment for addiction — stress management in real-world circumstances will be most valuable if it is linked to prevention of drug or nicotine craving and relapse.

During analysis of stress episodes in smoking cessation data, we have found that there is a relationship between successive *stress* episodes which is informative for intervention delivery. Stress is one of the major triggers for smoking and also responsible for lapses during cessation. Providing stress intervention for such population can help them maintain smoking cessation during abstinence.

In addition to showing how time-series data can be mined for determining the timing of interventions, presented work makes several methodological contributions. First, presented method of estimating the recovery time of physiology from a physical activity episode could possibly be used as a measure of cardiovascular fitness outside of controlled settings for heart patients. Second, Missing data is prevalent among wirelessly transmitted physiological signals. In addition, there are confounding activities (e.g., physical activity) for a specific biomarker (e.g., stress). This causes discontinuity in the time-series. To analyze trends in the series and to compute statistical features, we need a continuous time series. Imputation of missing data is an important step in such cases. But we can't do imputation if missing data is Missing Not At Random (MNAR). We found that missing stress assessments are not MNAR. Hence, it is possible to impute missing stress markers. Third, Validation of sensor inferred markers is the field setting is challenging due to lack of gold standard truth. Self-reported assessment is commonly used for this validation [89]. Presented work show that lack of agreement between self-reported stress and sensor inferred stress can subject to inconsistent self-report. Fourth, 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. This work also proposes a method to mine time-series sensor data on human health status and explore the tradeoffs between intervention frequency and probability of capturing the event of interest.

In summary, This dissertation presented an approach to determine the timings of just-in-time-intervention from sensor based measurements in the context of smoking and opioid 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.

9.2 Future Directions

Being the first work to inform the timing of sensor-triggered just-in-time intervention (JITI), this dissertation has several limitations that open up interesting future research directions.

9.2.1 Availability Assessment

- Some features used to predict availability are not yet reliably detectable via sensors today. These features includes the affect items, such as, being happy or being energetic. We estimated these affect items from self-report. For a model to be automated in informing the timing of JITI, all features need to be inferred from sensors.

- The type of sensors available on the phone or on smart watch is growing richer rapidly. Several sensors such as proximity sensor, acoustic sensor, and phone orientation and other data in the phone (e.g., calendar, task being performed on the phone, etc.) that may inform the current context of a user were not used in this work for assessing availability. Using these and other sensors emerging in phone may further improve the prediction accuracy. Similarly, using additional sensors on the body and those in instrumented spaces such as office, home, and vehicle (e.g., cameras) can also be used wherever available to further improve the prediction accuracy. For example, task-evoked pupillary response [92] can be used as a metric of mental workload, where head-mounted eye tracker measures the size of pupil that indicates cognitive load. Electromyogram (EMG) can measure the intensity of subvocalization [144] which is significantly different when someone is involved in a difficult programming related task. Similarly, Electroencephalogram (EEG) can be used to detect a user's cognitive state [41, 116].

- This work use 42-item EMA task for assessing significant user involvement (i.e., 2.4 minutes to complete). The results of this work may be more applicable to JITI that involve similar engagement. Its applicability to lighter JITI may need further investigation. On a side note, however, that if the user is found to be unavailable for a more involved active JITI (e.g., when driving), passive intervention could be delivered in the meantime (e.g., by playing music [138]).

- We used response delay as a metric for objectively assessing availability. Although we label significant delay in response as unavailable, it is not a gold-standard truth. In future, we can investigate other objective metrics (e.g., phone in airplane mode) and compare with each other.

9.2.2 Just-in-Time Stress Intervention

- We have inferred physical activity from chest worn accelerometer sensor which can capture whole body movement. It is possible that there are other intense hand (or leg) activities which are not captured from chest worn accelerometer sensors and can be confounder for stress inference. Detection of such physical activities from other modalities (e.g., smart watch) will enable us to isolate those confounded stress inferences.

- 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.

- 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.

- Presented model for generating stress intervention triggers can be supplemented with visual-exposure (via smart eyeglasses), digital traces (e.g., appointments on a smartphone calendar), and social exposures (e.g., twitter, facebook, etc.) to improves its accuracy and context sensitivity.

- 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.

- This dissertation have presented the relationship between *stress* episodes among the nicotine dependent individuals who are going through abstinence. Detection of the first lapse during abstinence [155] 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.

REFERENCES

[1] APA's Stress in America. http://www.apa.org/news/press/releases/stress/index.aspx, Accessed: September 2015.

[2] Zephyr. http://www.zephyranywhere.com/products/bioharness-3/, Accessed: February 2015.

[3] Gartner Says Smartphone Sales Accounted for 55 Percent of Overall Mobile Phone Sales in Third Quarter of 2013. http://www.gartner.com/newsroom/id/2623415, Accessed: March 2016.

[4] mCerebrum: An Open Source Software Suite for Mobile Sensor Data. https://md2k.org/software/, Accessed: October 2016.

[5] ANT Radio. http://www.thisisant.com/, Accessed: January 2016.

[6] APA Center for Organizational Excellence: Creating a Psychologically Healthy Workplace. http://www.apaexcellence.org/resources/creatingahealthyworkplace/, Accessed: November 2016.

[7] FitBit. http://www.fitbit.com/, Accessed: March 2016.

[8] U.S. Smartphone Use in 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/, Accessed: March 2016.

[9] Abbott, H., and Powell, D. Land-vehicle navigation using gps. *Proceedings of the IEEE 87*, 1 (1999), 145–162.

[10] Al'Absi, M. *Stress and addiction: Biological and psychological mechanisms*. Academic Press, 2011.

[11] Al'Absi, M., and Arnett, D. Adrenocortical responses to psychological stress and risk for hypertension. *Biomedicine & pharmacotherapy 54*, 5 (2000), 234–244.

[12] Al'Absi, M., Bongard, S., Buchanan, T., Pincomb, G. A., Licinio, J., and Lovallo, W. R. Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. *Psychophysiology 34*, 3 (1997), 266–275.

[13] al'Absi, M., Hatsukami, D., Davis, G., and 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 (2004), 267–278.

[14] Ali, A. A., Hossain, S. M., Hovsepian, K., Rahman, M. M., Plarre, K., and Kumar, S. mpuff: automated detection of cigarette smoking puffs from respiration measurements. In *ACM IPSN* (2012), 269–280.

[15] Alpers, G. Ambulatory assessment in panic disorder and specific phobia. *Psychological Assessment 21*, 4 (2009), 476.

[16] Anam, A. I., Alam, S., and Yeasin, M. Expression: A dyadic conversation aid using google glass for people with visual impairments. In *ACM UbiComp Adjunct* (2014), 211–214.

[17] Appel, G. *Technical analysis: power tools for active investors*. FT Press, 2005.

[18] Aswani, A., and Tomlin, C. Game-theoretic routing of gps-assisted vehicles for energy efficiency. In *American Control Conference (ACC)*, IEEE (2011), 3375–3380.

[19] Atallah, L., Lo, B., King, R., and Yang, G. Sensor placement for activity detection using wearable accelerometers. In *IEEE BSN* (2010), 24–29.

[20] Avrahami, D., Fogarty, J., and Hudson, S. Biases in human estimation of interruptibility: effects and implications for practice. In *ACM CHI* (2007), 50–60.

[21] Avrahami, D., Gergle, D., Hudson, S., and Kiesler, S. Improving the match between callers and receivers: A study on the effect of contextual information on cell phone interruptions. *Behaviour & Information Technology 26*, 3 (2007), 247–259.

[22] Baer, J. S., and Lichtenstein, E. Classification and prediction of smoking relapse episodes: an exploration of individual differences. *Journal of consulting and clinical psychology 56*, 1 (1988), 104.

[23] Bangerter, A., Chevalley, E., and Derouwaux, S. Managing third-party interruptions in conversations: Effects of duration and conversational role. *Journal of Language and Social Psychology 29*, 2 (2010), 235–244.

[24] Bekkering, H., Neggers, S. F., Walker, R., Gleißner, B., Dittrich, W. H., and Kennard, C. The preparation and execution of saccadic eye and goal-directed hand movements in patients with parkinson's disease. *Neuropsychologia 39*, 2 (2001), 173–183.

[25] Bessiere, K., Newhagen, J., Robinson, J., and Shneiderman, B. A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Computers in Human Behavior 22*, 6 (2006), 941–961.

[26] Bickmore, T. W., Caruso, L., and Clough-Gorr, K. Acceptance and usability of a relational agent interface by urban older adults. In *ACM CHI Extended Abstracts* (2005), 1212–1215.

[27] Bland, J., and Altman, D. Statistics: notes cronbach's alpha. *BMJ 314*, 7080 (1997), 572–572.

[28] Boca, F., and Noll, J. Truth or consequences: The validity of self-report data in health services research on addictions. *Addiction 95*, 11s3 (2000), 347–360.

[29] Bohannon, R. Comfortable and maximum walking speed of adults aged 20-79 years: Reference values and determinants. *Age and Ageing 26*, 1 (1997), 15–19.

[30] Bollinger, J. Bollinger on bollinger band.

[31] Bonnet, C., Rusz, J., Megrelishvili, M., Sieger, T., Matoušková, O., Okujava, M., Brožová, H., Nikolai, T., Hanuška, J., Kapianidze, M., et al. Eye movements in ephedrone-induced parkinsonism. *PloS one 9*, 8 (2014), e104784.

[32] Bouchard, S., Bernier, F., Boivin, É., Morin, B., and Robillard, G. Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers. *PloS one 7*, 4 (2012), e36169.

[33] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[34] Brooks, G. C., Vittinghoff, E., Iyer, S., Tandon, D., Kuhar, P., Madsen, K. A., Marcus, G. M., Pletcher, M. J., and Olgin, J. E. Accuracy and usability of a self-administered 6-minute walk test smartphone application. *Circulation: Heart Failure 8*, 5 (2015), 905–913.

[35] Brown, R. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation, 2004.

[36] 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 (2003), 272–276.

[37] Burns, M., Begale, M., Duffecy, J., and Gergle, D. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research 13*, 3 (2011).

[38] Carroll, E., Czerwinski, M., Roseway, A., Kapoor, A., Johns, P., Rowan, K., and Schraefel, M. Food and mood: Just-in-time support for emotional eating. In *IEEE ACII* (2013), 252–257.

[39] Chandola, T., Brunner, E., and Marmot, M. Chronic stress at work and the metabolic syndrome: prospective study. *Bmj 332*, 7540 (2006), 521–525.

[40] Chatterjee, S., Hovsepian, K., Sarker, H., Saleheen, N., al'Absi, M., Atluri, G., Ertin, E., Lam, C., Lemieux, A., Nakajima, M., et al. mcrave: continuous estimation of craving during smoking cessation. In *ACM UbiComp* (2016), 863–874.

[41] Chen, D., and Vertegaal, R. Using mental load for managing interruptions in physiologically attentive user interfaces. In *ACM CHI Extended Abstracts* (2004), 1513–1516.

[42] Choudhary, A. K., Harding, J. A., and Tiwari, M. K. Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing 20*, 5 (2009), 501–521.

[43] Chrousos, G., and Gold, P. The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis. *JAMA 267*, 9 (1992), 1244.

[44] Cinaz, B., La Marca, R., Arnrich, B., and Tröster, G. Towards continuous monitoring of mental workload. In *ACM UbiComp* (2010).

[45] Cohen, S., and Lichtenstein, E. Perceived stress, quitting smoking, and smoking relapse. *Health Psychology 9*, 4 (1990), 466.

[46] Consolvo, S., Everitt, K., Smith, I., and Landay, J. A. Design requirements for technologies that encourage physical activity. In *ACM CHI* (2006), 457–466.

[47] Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., et al. Activity sensing in the wild: a field trial of ubifit garden. In *ACM CHI* (2008), 1797–1806.

[48] Cover, T. M., and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

[49] Cummings, K. M., Jaén, C. R., and Giovino, G. Circumstances surrounding relapse in a group of recent exsmokers. *Preventive Medicine 14*, 2 (1985), 195–202.

[50] Davis, F., Roseway, A., Carroll, E., and Czerwinski, M. Actuating mood: design of the textile mirror. In *International Conference on Tangible, Embedded and Embodied Interaction* (2013), 99–106.

[51] Davis, M., Eshelman, E., and McKay, M. *The relaxation and stress reduction workbook*. New Harbinger Publications, 2008.

[52] Domingos, P. Metacost: A general method for making classifiers cost-sensitive. In *ACM KDD* (1999), 155–164.

[53] Donders, A., van der Heijden, G., Stijnen, T., and Moons, K. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology 59*, 10 (2006), 1087–1091.

[54] Ehrenreich, B., Righter, B., Rocke, A., Dixon, L., and Himelhoch, S. Are mobile phones and handheld computers being used to enhance delivery of psychiatric treatment?: a systematic review. *The Journal of Nervous and Mmental Disease 199*, 11 (2011), 886–891.

[55] Enoch, M.-A. Pharmacogenomics of alcohol response and addiction. *American Journal of Pharmacogenomics 3*, 4 (2003), 217–232.

[56] ENOCH, M.-A. Genetic and environmental influences on the development of alcoholism. *Annals of the New York Academy of Sciences 1094*, 1 (2006), 193–201.

[57] Enright, P. L. The six-minute walk test. *Respiratory care 48*, 8 (2003), 783–785.

[58] Epstein, D., Tyburski, M., Craig, I., Phillips, K., Jobes, M., Vahabzadeh, M., Mezghanni, M., Lin, J., Furr-Holden, D., and Preston, K. Real-time tracking of neighborhood surroundings and mood in urban drug misusers: application of a new method to study behavior in its geographical context. *Drug and alcohol dependence 134* (2014), 22–29.

[59] Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., and Shah, S. Autosense: Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *ACM SenSys* (2011), 274–287.

[60] Esco, M., Olson, M., Williford, H., Blessing, D., Shannon, D., and Grandjean, P. The relationship between resting heart rate variability and heart rate recovery. *Clinical Autonomic Research 20*, 1 (2010), 33–38.

[61] Evans, G., Wener, R., and Phillips, D. The morning rush hour predictability and commuter stress. *Environment and Behavior 34*, 4 (2002), 521–530.

[62] Fathi, A., Li, Y., and Rehg, J. M. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV*. 2012, 314–327.

[63] Fisher, R., and Simmons, R. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *IEEE ICMLA*, vol. 1 (2011), 436–441.

[64] Fogarty, J., Hudson, S., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. Predicting human interruptibility with sensors. *ACM TOCHI 12*, 1 (2005), 119–146.

[65] Fogarty, J., Hudson, S., and Lai, J. Examining the robustness of sensor-based statistical models of human interruptibility. In *ACM CHI* (2004), 207–214.

[66] Fonner, K. L., and Roloff, M. E. Testing the connectivity paradox: Linking teleworkers' communication media use to social presence, stress from interruptions, and organizational identification. *Communication Monographs 79*, 2 (2012), 205–231.

[67] 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 (2008), 1226.

[68] Free, C., Knight, R., Robertson, S., Whittaker, R., Edwards, P., Zhou, W., Rodgers, A., Cairns, J., Kenward, M., and Roberts, I. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. *The Lancet 378*, 9785 (2011), 49–55.

[69] Freeman, J., Dewey, F., Hadley, D., Myers, J., and Froelicher, V. Autonomic nervous system interaction with the cardiovascular system during exercise. *Progress in cardiovascular diseases 48*, 5 (2006), 342–362.

[70] Fritz, C., Sonnentag, S., Spector, P., and McInroe, J. The weekend matters: Relationships between stress recovery and affective experiences. *Journal of Organizational Behavior 31*, 8 (2010), 1137–1162.

[71] Furr-Holden, D., Smart, M., Pokorni, J., Ialongo, N., Leaf, P., Holder, H., and Anthony, J. The nifety method for environmental assessment of neighborhood-level indicators of violence, alcohol, and other drug exposure. *Prevention Science 9*, 4 (2008), 245–255.

[72] George, D., and Mallery, M. Using spss for windows step by step: a simple guide and reference, 2003.

[73] Gopalsamy, C., Park, S., Rajamanickam, R., and Jayaraman, S. The wearable motherboardâĎć: The first generation of adaptive and responsive textile structures (arts) for medical applications. *Virtual Reality 4*, 3 (1999), 152–168.

[74] Hall, M. A. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[75] Han, T., Xiao, X., Shi, L., Canny, J., and Wang, J. Balancing accuracy and fun: Designing camera based mobile games for implicit heart rate monitoring. In *ACM CHI* (2015), 847–856.

[76] Harr, R., and Kaptelinin, V. Interrupting or not: exploring the effect of social context on interrupters' decision making. In *ACM NordiCHI* (2012), 707–710.

[77] Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., and Botstein, D. Imputing missing data for gene expression arrays, 1999.

[78] Haug, S., Meyer, C., Dymalski, A., Lippke, S., and John, U. Efficacy of a text messaging (sms) based smoking cessation intervention for adolescents and young adults: study protocol of a cluster randomised controlled trial. *BMC Public Health 12*, 1 (2012), 1.

[79] Hayes, T., Cobbinah, K., Dishongh, T., Kaye, J., Kimel, J., Labhard, M., and Leen, T. A study of medication-taking and unobtrusive, intelligent reminding. *Telemedicine and E-Health 15*, 8 (2009), 770–776.

[80] Henry, J. P. Neuroendocrine patterns of emotional response. *Emotion: theory, research and experience 3* (1986), 37–60.

[81] Hernandez, J., Paredes, P., Roseway, A., and Czerwinski, M. Under pressure: sensing stress of computer users. In *ACM CHI* (2014), 51–60.

[82] Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., and Fantini, S. Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *ACM CHI*, ACM (2009), 2185–2194.

[83] Ho, J., and Intille, S. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *ACM CHI* (2005), 909–918.

[84] Hobfoll, S. E. Conservation of resources: A new attempt at conceptualizing stress. *American psychologist 44*, 3 (1989), 513.

[85] Hobfoll, S. E., Vinokur, A. D., Pierce, P. F., and 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 (2012), 217.

[86] Hong, J., Ramos, J., and Dey, A. Understanding physiological responses to stressors during physical activity. In *ACM UbiComp* (2012), 270–279.

[87] Hossain, S., Ali, A., Rahman, M., Ertin, E., Epstein, D., Kennedy, A., Preston, K., Umbricht, A., Chen, Y., and Kumar, S. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In *ACM IPSN* (2014), 71–82.

[88] Hossain, S. M., Ali, A. A., Rahman, M. M., Ertin, E., Epstein, D., Kennedy, A., Preston, K., Umbricht, A., Chen, Y., and Kumar, S. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In *ACM IPSN* (2014), 71–82.

[89] Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., and Kumar, S. cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *ACM UbiComp* (2015), 493–504.

[90] Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *ACM CHI* (2003), 257–264.

[91] Iqbal, S., and Bailey, B. Effects of intelligent notification management on users and their tasks. In *ACM CHI* (2008), 93–102.

[92] Iqbal, S., Zheng, X., and Bailey, B. Task-evoked pupillary response to mental workload in human-computer interaction. In *ACM CHI Extended Abstracts* (2004), 1477–1480.

[93] Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., and Bailey, B. P. Towards an index of opportunity: understanding changes in mental workload during task execution. In *ACM CHI* (2005), 311–320.

[94] Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., and Whittaker, S. Echoes from the past: how technology mediated reflection improves well-being. In *ACM CHI* (2013), 1071–1080.

[95] Jaimes, L., Llofriu, M., and Raij, A. A stress-free life: just-in-time interventions for stress via real-time forecasting and intervention adaptation. In *ICST BODYNETS* (2014), 197–203.

[96] Kapoor, A., and Horvitz, E. Experience sampling for building predictive user models: a comparative study. In *ACM CHI* (2008), 657–666.

[97] Kaushik, P. *The design and evaluation of a mobile handheld intervention for providing context-sensitive medication reminders*. PhD thesis, Massachusetts Institute of Technology, 2005.

[98] Kendall, L., Morris, D., and Tan, D. Blood pressure beyond the clinic: Rethinking a health metric for everyone. In *ACM CHI* (2015), 1679–1688.

[99] Kim, S., Chun, J., and Dey, A. K. Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. In *ACM CHI* (2015), 487–496.

[100] Klein, J., Moon, Y., and Picard, R. W. This computer responds to user frustration:: Theory, design, and results. *Interacting with computers 14*, 2 (2002), 119–140.

[101] Kohavi, R., and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence 97*, 1-2 (1997), 273–324. Special issue on relevance.

[102] Konrad, A., Bellotti, V., Crenshaw, N., Tucker, S., Nelson, L., Du, H., Pirolli, P., and Whittaker, S. Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction. In *ACM CHI* (2015), 3829–3838.

[103] Krediet, I. *Work and Emotions: The Role of Interruptions*. PhD thesis, Tilburg University, 1999.

[104] Krumm, J., and Rouhana, D. Placer: Semantic place labels from diary data. In *ACM UbiComp* (2013), 163–172.

[105] Kudielka, B., Schommer, N., Hellhammer, D., and Kirschbaum, C. Acute hpa axis responses, heart rate, and mood changes to psychosocial stress (tsst) in humans at different times of day. *Psychoneuroendocrinology 29*, 8 (2004), 983–992.

[106] Kulesza, T., Amershi, S., Caruana, R., Fisher, D., and Charles, D. Structured labeling for facilitating concept evolution in machine learning. In *ACM CHI* (2014), 3075–3084.

[107] Kumar, S., Nilson, W., Pavel, M., and Srivastava, M. Mobile health: Revolutionizing healthcare through trans-disciplinary research. *IEEE Computer 46*, 1 (2013), 28–35.

[108] Lee, M.-H., Cha, S., and Nam, T.-J. Patina engraver: Visualizing activity logs as patina in fashionable trackers. In *ACM CHI* (2015), 1173–1182.

[109] Lin, J. J., Mamykina, L., Lindtner, S., Delajoux, G., and Strub, H. B. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *UbiComp 2006: Ubiquitous Computing*. Springer, 2006, 261–278.

[110] Liu, L., Karatas, C., Li, H., Tan, S., Gruteser, M., Yang, J., Chen, Y., and Martin, R. P. Toward detection of unsafe driving with wearables. In *ACM WearSys* (2015), 27–32.

[111] Looije, R., Cnossen, F., and Neerincx, M. Incorporating guidelines for health assistance into a socially intelligent robot. In *IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)* (2006), 515–520.

[112] Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y., and Kameyama, K.-i. Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In *ACM CHI* (2015), 857–866.

[113] MacLean, D., Roseway, A., and Czerwinski, M. Moodwings: a wearable biofeedback device for real-time stress intervention. In *ACM PETRA* (2013), 66.

[114] Marije, R., Hogenelst, K., and Schoevers, R. Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review 32*, 6 (2012), 510–523.

[115] Mark, G., Gudith, D., and Klocke, U. The cost of interrupted work: more speed and stress. In *ACM CHI* (2008), 107–110.

[116] Mathan, S., Whitlow, S., Dorneich, M., Ververs, P., and Davis, G. Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *International Conference of the Augmented Cognition Society* (2007).

[117] Matthews, M., Snyder, J., Reynolds, L., Chien, J. T., Shih, A., Lee, J. W., and Gay, G. Real-time representation versus response elicitation in biosensor data. In *ACM CHI* (2015), 605–608.

[118] Mayberry, A., Hu, P., Marlin, B., Salthouse, C., and Ganesan, D. ishadow: design of a wearable, real-time mobile gaze tracker. In *ACM MobiSys* (2014), 82–94.

[119] Mayberry, A., Tun, Y., Hu, P., Smith-Freedman, D., Ganesan, D., Marlin, B. M., and Salthouse, C. Cider: Enabling robustness-power tradeoffs on a computational eyeglass. In *ACM MobiCom* (2015), 400–412.

[120] McDuff, D., Karlson, A., Kapoor, A., Roseway, A., and Czerwinski, M. Affectaura: an intelligent system for emotional memory. In *ACM CHI* (2012), 849–858.

[121] McEwen, B. Protection and damage from acute and chronic stress. *Ann NY Acad Sci 1032* (2004), 1–7.

[122] McEwen, B. Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences 840*, 1 (2006), 33–44.

[123] McEwen, B. Physiology and neurobiology of stress and adaptation: Central role of the brain. *Physiological Reviews 87*, 3 (2007), 873–904.

[124] McEwen, B., and Stellar, E. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine 153*, 18 (1993), 2093.

[125] 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 (2004), 1–7.

[126] McFarlane, D. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction 17*, 1 (2002), 63–139.

[127] Minami, H., McCarthy, D., Jorenby, D., and Baker, T. An ecological momentary assessment analysis of relations among coping, affect and smoking during a quit attempt. *Addiction 106*, 3 (2011), 641–650.

[128] Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L. Actual causes of death in the united states, 2000. *Journal of the American Medical Association (JAMA) 291*, 10 (2004), 1238–1245.

[129] Molitor, R. J., Ko, P. C., and Ally, B. A. Eye movements in alzheimer's disease. *Journal of Alzheimer's Disease 44*, 1 (2015), 1–12.

[130] Montoliu, R., Blom, J., and Gatica-Perez, D. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications 62*, 1 (2013), 179–207.

[131] Mundt, C. W., Montgomery, K. N., Udoh, U. E., Barker, V. N., Thonier, G. C., Tellier, A. M., Ricks, R. D., Darling, R. B., Cagle, Y. D., Cabrol, N. A., et al. A multiparameter wearable physiologic monitoring system for space and terrestrial applications. *Information Technology in Biomedicine, IEEE Transactions on 9*, 3 (2005), 382–391.

[132] Musthag, M., Raij, A., Ganesan, D., Kumar, S., and Shiffman, S. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *ACM UbiComp* (2011), 435–444.

[133] Nahum-Shani, I., Hekler, E., and Spruijt-Metz, D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. *Health Psychology*.

[134] Natarajan, A., Parate, A., Gaiser, E., Angarita, G., Malison, R., Marlin, B., and Ganesan, D. Detecting cocaine use with wearable electrocardiogram sensors. In *ACM UbiComp* (2013), 123–132.

[135] Naughton, F., Prevost, A. T., Gilbert, H., and Sutton, S. Randomized controlled trial evaluation of a tailored leaflet and sms text message self-help intervention for pregnant smokers (miquit). *Nicotine & Tobacco Research 14*, 5 (2012), 569–577.

[136] Ni, K., Ramanathan, N., Chehade, M., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M., and Srivastava, M. Sensor network data fault types. *ACM TOSN 5*, 3 (2009), 25.

[137] Nielsen, P., Le Grice, I., Smaill, B., and Hunter, P. Mathematical model of geometry and fibrous structure of the heart. *American Journal of Physiology-Heart and Circulatory Physiology 260*, 4 (1991), H1365–H1378.

[138] Nirjon, S., Dickerson, R., Li, Q., Asare, P., Stankovic, J., Hong, D., Zhang, B., Jiang, X., Shen, G., and Zhao, F. Musicalheart: A hearty way of listening to music. In *ACM SenSys* (2012), 43–56.

[139] Oorschot, M., Kwapil, T., Delespaul, P., and Myin-Germeys, I. Momentary assessment research in psychosis. *Psychological Assessment 21*, 4 (2009), 498–505.

[140] Pan, J., and Tompkins, W. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering 32*, 3 (1985), 230–236.

[141] Pandian, P., Mohanavelu, K., Safeer, K., Kotresh, T., Shakunthala, D., Gopal, P., and Padaki, V. Smart vest: Wearable multi-parameter remote physiological monitoring system. *Medical Engineering & Physics 30*, 4 (2008), 466–477.

[142] Parate, A., Chiu, M.-C., Chadowitz, C., Ganesan, D., and Kalogerakis, E. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *ACM MobiSys* (2014), 149–161.

[143] Paredes, P., and Chan, M. Calmmenow: exploratory research and design of stress mitigating mobile interventions. In *ACM CHI Extended Abstracts* (2011), 1699–1704.

[144] Parnin, C. Subvocalization-toward hearing the inner thoughts of developers. In *IEEE ICPC* (2011), 197–200.

[145] 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* (2011), 97–108.

[146] Platt, J. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[147] Poppinga, B., Heuten, W., and Boll, S. Sensor-based identification of opportune moments for triggering notifications. *IEEE Pervasive Computing 13*, 1 (2014), 22–29.

[148] Preuveneers, D., and Berbers, Y. Mobile phones assisting with health self-care: a diabetes case study. In *ACM MobileHCI* (2008), 177–186.

[149] Purpura, S., Schwanda, V., Williams, K., Stubler, W., and Sengers, P. Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. In *ACM CHI* (2011), 423–432.

[150] Ragsdale, J., Beehr, T., Grebner, S., and Han, K. An integrated model of weekday stress and weekend recovery of students. *International Journal of Stress Management 18*, 2 (2011), 153.

[151] 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., al'Absi, M., and Kumar, S. Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors. In *ACM BCB* (2014), 479–488.

[152] Rasekaba, T., Lee, A., Naughton, M., Williams, T., and Holland, A. The six-minute walk test: a useful metric for the cardiopulmonary patient. *Internal medicine journal 39*, 8 (2009), 495–501.

[153] Rosmond, R., and Björntorp, P. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism 47*, 10 (1998), 1187–1193.

[154] Sadler, K., Robertson, T., and Kan, M. It's always there, it's always on: Australian freelancer's management of availability using mobile technologies. In *ACM MobileHCI* (2006), 49–52.

[155] Saleheen, N., Ali, A. A., Hossain, S. M., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., and Kumar, S. puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *ACM UbiComp* (2015), 999–1010.

[156] Sapolsky, R. M. *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping-now revised and updated*. Macmillan, 2004.

[157] Sarker, H., Sharmin, M., Ali, A., Rahman, M., Bari, R., Hossain, S., and Kumar, S. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *ACM UbiComp* (2014), 909–920.

[158] 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., and Kumar, S. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *ACM CHI* (2016), 4489–4501.

[159] Sauro, K. M., and Becker, W. J. The stress and migraine interaction. *Headache: The Journal of Head and Face Pain 49*, 9 (2009), 1378–1386.

[160] Schmitt, L. M., Cook, E. H., Sweeney, J. A., and Mosconi, M. W. Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunctions in cerebellum and brainstem. *Molecular autism 5*, 1 (2014), 1.

[161] Serre, F., Fatseas, M., Debrabant, R., Alexandre, J., Auriacombe, M., and Swendsen, J. Ecological momentary assessment in alcohol, tobacco, cannabis and opiate dependence: A comparison of feasibility and validity. *Drug and Alcohol Dependence 126*, 1 (2012), 118–123.

[162] Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J. G., Vhaduri, S., Preston, K., and Kumar, S. Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In *ACM UbiComp* (2015), 505–516.

[163] Sharry, J., McDermott, M., and Condron, J. Relax to win treating children with anxiety problems with a biofeedback video game. *Eisteach 2* (2003), 22–26.

[164] Shiffman, S. Relapse following smoking cessation: a situational analysis. *Journal of consulting and clinical psychology 50*, 1 (1982), 71.

[165] Shiffman, S. Ecological momentary assessment (ema) in studies of substance use. *Psychological Assessment 21*, 4 (2009), 486.

[166] Shiffman, S., Paty, J. A., Gnys, M., Kassel, J. A., and Hickcox, M. First lapses to smoking: Within-subjects analysis of real-time reports. *Journal of Consulting and Clinical Psychology 64*, 2 (1996), 366.

[167] Shiffman, S., Stone, A., and Hufford, M. Ecological momentary assessment. *Annual Review of Clinical Psychology 4* (2008), 1–32.

[168] Sinha, R. Chronic stress, drug use, and vulnerability to addiction. *Annals of the New York Academy of Sciences 1141*, 1 (2008), 105–130.

[169] Speed, T. *Statistical analysis of gene expression microarray data*. CRC Press, 2004.

[170] Srivastava, A., Sharma, R., Sood, S. K., Shukla, G., Goyal, V., and Behari, M. Saccadic eye movements in parkinson's disease. *Indian journal of ophthalmology 62*, 5 (2014), 538.

[171] Stone, A., and Broderick, J. Real-time data collection for pain: Appraisal and current status. *Pain Medicine 8*, s3 (2007), S85–S93.

[172] Strecher, V. J. Computer-tailored smoking cessation materials: a review and discussion. *Patient education and counseling 36*, 2 (1999), 107–117.

[173] Sun, D., Paredes, P., and Canny, J. Moustress: detecting stress from mouse motion. In *ACM CHI* (2014), 61–70.

[174] Sun, M., Burke, L., Baranowski, T., Fernstrom, J., Zhang, H., Chen, H.-C., Bai, Y., Li, Y., Li, C., Yue, Y., et al. An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle. *Journal of healthcare engineering 6*, 1 (2015), 1–22.

[175] Tan, C. S. S., Schöning, J., Luyten, K., and Coninx, K. Investigating the effects of using biofeedback as visual stress indicator during video-mediated collaboration. In *ACM CHI* (2014), 71–80.

[176] Thomaz, E., Essa, I., and Abowd, G. D. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *ACM UbiComp* (2015), 1029–1040.

[177] Torres, S. J., and Nowson, C. A. Relationship between stress, eating behavior, and obesity. *Nutrition 23*, 11 (2007), 887–894.

[178] Toscos, T., Faber, A., An, S., and Gandhi, M. P. Chick clique: persuasive technology to motivate teenage girls to exercise. In *ACM CHI Extended Abstracts* (2006), 1873–1878.

[179] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. Missing value estimation methods for dna microarrays. *Bioinformatics 17*, 6 (2001), 520–525.

[180] Weiser, M. The computer for the 21st century. *Scientific american 265*, 3 (1991), 94–104.

[181] Wenze, S. J., and Miller, I. W. Use of ecological momentary assessment in mood disorders research. *Clinical Psychology Review 30*, 6 (2010), 794–804.

[182] Wilder, J. W. *New concepts in technical trading systems*. Trend Research Greensboro, NC, 1978.

[183] Wilhelm, F., Roth, W., and Sackner, M. The lifeshirt: an advanced system for ambulatory measurement of respiratory and cardiac function. *Behavior Modification 27*, 5 (2003), 671–691.

[184] Wilson, K., Hopkins, R., deVries, M., and Copeland, J. Research alliance and the limit of compliance: experience sampling with the depressed elderly. *The Experience of Psychopathology: Investigating Mental Disorders in Their Natural Settings* (1992), 339–346.

[185] Xian, G., and Homer, C. Updating the 2001 national land cover database impervious surface products to 2006 using landsat imagery change detection methods. *Remote Sensing of Environment 114*, 8 (2010), 1676–1686.

[186] Zhang, L., Li, X.-Y., Huang, W., Liu, K., Zong, S., Jian, X., Feng, P., Jung, T., and Liu, Y. It starts with igaze: Visual attention driven networking with smart glasses. In *ACM MobiCom* (2014), 91–102.

[187] Zhang, Z., Pi, Z., and Liu, B. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on Biomedical Engineering 62*, 2 (2015), 522–531.