Predicting the level of stress using smartphone data

Rifat Zahan[1], Xuguang Yang[2], Saman Rahbar[3]

1 rifat.zahan@usask.ca Computational Epidemiology and Public Health Informatics Lab (CEPHIL)

2 nathan.yang@cs.usask.ca Bioinformatics Lab

3 saman.rahbar@usask.ca Advanced Research in Intelligent Systems (ARIES) Lab

1, 2, 3 Department of Computer Science/University of Saskatchewan, Saskatoon, Saskatchewan, Canada

Abstract

Moderate level of stress has been found helpful, but frequent and chronic stress are associated with immune disorder, depression and different types of mental illness that negatively impact the productivity of work or study. Monitoring stress at work or school and studying the associated factors have become quite popular in health-related research. Massive use of smartphones has created a strong ground to study the human behaviour. The amount, type, and duration of app usage are affected by the user state. The above information can be associated with the self-reported stress state of a smartphone user. In this project, we focused on investigating what factors were associated with stress level and on finding the optimal model to predict stress levels among university students. A cohort of students returned their cellphone usage data along with some survey responses. Individual cumulative link mixed model revealed that school weeks were significantly associated with stress level. Our proposed model was established upon the school weeks and utility-related app usage. The results showed that the stress scores decreased as the time spent on utility-related apps increased, and that stress levels were elevated as the school week went on.

Introduction

Nowadays, feeling stressful is a common phenomenon among all social members, and people experience stress at different levels throughout their lives [1]. According to *Stallman & Hurst*, while moderate stress contributed to the academic performance of students, high-level stress constantly caused mental and physical illness, such like depression, anxiety, chronic pain, *etc.*. Additionally, their study showed that stress might have resulted in some issues underlying a student's everyday life: poor exam grades, being late to school and losing items [2]. Stress among university students hence has become a heated research subject recently. Although stress can be assessed via different approaches, self-reported stress levels, combined with smartphone data, may shed light on interpreting human behavior [2]. In the meantime, human behavior studies may also benefit from the massive use of smartphones [3].

In this study, two research questions were raised: (1) What factors are significantly associated with stress level among university students? and (2) What would be the best model to predict stress level with smartphone data (*e.g.*, time spent on social networks, utility apps, browsing, *etc.*) and survey data (*e.g.*, transportation modes, time spent on campus, types of food purchased, involvement in physical activities, *etc.*)?

Related work

Thomée et al. surveyed 4,156 young adults (20-24 years old) with 1-year follow-up to investigate whether there were associations between psychological aspects of mobile phone use and mental health symptoms. The results of their study showed that high frequency of mobile phone use was a risk factor for mental health outcomes, and that the risk for reporting mental health symptoms was greatest among those who stressfully accessed mobile phones [4]. *Mehrotra* & *Musolesi* discussed how the emerging technologies such like social media and smartphone data made human behavioral research possible [5]. In their systematic review, *Þórarinsdóttir et al.* examined the current stress studies using smartphone data and discovered that most of them rarely investigated stress as their primary objective. Besides, they observed that within those studies focusing on stress, the number of participants were not sufficient to analyze the correlation between smartphone data and stress [1]. For instance, *Adams et al.* had only 7 participants involved in their study, to compare the minimally invasive techniques for measuring daily stress [6]. *Bogomolov et al.* applied tree classifiers based on random forest algorithm to address daily stress recognition using smartphone data, weather conditions and personalities. The process of their feature selection retained 32 features out of an initial pool of 500. A 72.28% accuracy was achieved by their multifactorial model, outperforming other candidate models (*e.g.* SVM and Generalized Boosted Model) [7]. While they took different factors into consideration to study the stress, their model was not adequate to analyze stress in a more precise way, *i.e.*, numerically predicting stress level.

Ferdous et al. applied SVM to predict the stress levels, based on individual app usage behaviours [3]. Notably, their categorization of smartphone applications shed light on analysis of app usage patterns for future researchers. Albeit the high accuracies achieved by their model, the fact that only 28 participants were involved limited the validity of their study.
Samaha & *Hawi* investigated relationships among smartphone addiction, stress, academic performance and satisfaction with life [8]. Though 300 university students participated in their study, they didn't primarily focus on exploring the association between smartphone data and stress levels.

While exploring the impact of various commuting modes on stress and mood upon arrival at work, *Brutus et al.* found that cyclists were less stressed compared to drivers [9]. *Legrain et al.* applied ordered logistic regression to survey potential relation between university students' stress and transportation choices. Similar to what *Brutus et al.* found, their results showed that driving was the most stressful mode of transportation compared to others, namely walking and public transits [10].

Ikeda et al. applied multiple regression analysis and found that weight management through physical exercise reduced mental stress among Japanese university students [11]. In their study, *Burg et al.* discovered that the relationship of stress to exercise could be uni- or bi-directional, depending on individuals [12]. In that sense, exercise might not necessarily alleviate stress for everyone and mental stress might affect people's exercise participation.

Sano et al. incorporated SVM with wearable sensors, personality traits and mobile phones to recognize academic performance, sleep quality, stress level and mental health. Particularly, they found associations between stress and personality. Besides, they reported their classification accuracies using objective data from wearable devices and smartphones ranged from 67-92% [13].

Data and variables

This study was comprised of 117 students from a Canadian university. All participants were provided with our data collection app installed on their phones with informed consent. The data was collected during February 7, 2017 to March 6, 2017. A daily survey question would be popped up whenever participants arrived to the campus, asking them to report stress levels (scaling between 0 and 10, such that 0 was not stressed at all and 10 stressed out). The daily survey was triggered whenever a participant's phone was connected to a campus WiFi. At the same time, participants were asked in another survey about what type of transportation method they used to come to school. Participants were also asked to fill out a survey whenever they fell, participated in physical activity or purchased food. To facilitate this study, the selected datasets needed some pre-processing.

Stress level

Daily stress scores were obtained upon the time of arrival, but only the median scores were kept if there were multiple responses to the entry survey from a single person. If a participant did not respond to a survey timely, a label "Expired" would be put within the duration record. Also the participants had the option to cancel a survey if they did not want to respond and a label "Cancelled" would be added to that record. Those "Expired" and "Cancelled" records were removed from further analysis. Taking survey compliance into account, some thresholds, $\delta = 5, 10, 12, 15$ were considered for data sampling. The participants who responded less than δ times were not analyzed. Since the duration of the survey was 29 days, we needed a sufficient amount of data to conduct further analysis. Notably, compliance of the survey directly impacts the quality of the data analysis.

Physical activity

The "walk/run" exercise records were removed from the study. Because there was a chance that participants might consider walking from their on-campus residence to school (7-10 minutes) as an exercise. According to any physical activity guidelines, the recommended amount of time to exercise is no less than 30 minutes per day for 4 days a week, which implies about 17 minutes per day of moderate to vigorous physical activity [14]. Further, any exercise record that occurred more than 10 hours ago was eliminated, as they did not contribute to our study.

Fast-food consumption behaviour

People mentioned in their user-triggered surveys about what type of food they purchased. We wanted to see if there was any association between fast-food consumption and stress level. So we classified the food purchase responses as "fast-food" and "no fast-food".

Time spent on campus

WiFi data was used to calculate participants' time spent on campus. Given a set of survey responses and WiFi records, we calculated the on campus dwell time in hours. The time spent on campus was calculated based on the duration of campus WiFi connection. Only those with accuracies less than or equal to 50 meters were considered. It's because to calculate each participant's on campus dwell time, we need more precise records to determine on campus status. To implement this filtering a simply SQL query statement using python was used to retrieve all the records we needed. In addition, records were sorted based on the record times. Further, to comply with the survey consent and to facilitate further data aggregation, only the de-identified filtered results were locally stored on individual basis.

App usage data

To get a clear understanding of the applications used by the participants, we divided the apps into five major categories according to the study conducted by *Ferdous et al*, which are given in Table 1.

Table 1. Major categorization of the apps used by the respondents[3]

Socializing	Entertainment	Utility	Browser
Facebook	Audio & Video	Calendar	Chrome
Twitter	Book Reader	Map	Firefox
Google+	FM Radio	Navigator	Email
Dropbox	TV Apps	Flashlight	Search App
Pinterest	Music Instruments	Weather	
Flipboard	YouTube	Camera	
Whatsapp	News Apps	Voice Recorder	
GTalk		Trip Advisor	
Skype		Calculator	
Viber		Banking	
		Online Shopping	
Games	Card games, puzzles, fun games		

During the process of aggregating app usage data, we first eliminated background app records. On top of that, the records that indicated certain apps used less than 5 minutes were removed. Further, on daily and individual bases, time spent on each category of apps was calculated.

Methodology

From our literature review, it can be seen that most of the stress-related researches considered either regression analysis or support vector machine (SVM) to predict stress levels using other covariates, whereas only few studies adopted user-specific models. Based on the repeated observations on the stress and smartphone use patterns of different participants in different days, there was a possibility that variation existed within the perceived stress among individuals. For example, stress level of 10 reported by one individual might be same as that of 7 reported by another. In recent years, scientists have been working on developing mixed-effects model where the response variable is ordinal[15]. This type of models is called *Cumulative Link Mixed Model (CLMM)*. In light of our data structure, we applied the CLMM, which is an extended version of Ordinal Logistic Regression, where we considered individual-specific variation.

Representation of Ordinal Logistic Regression Model

Since logit link is widely used in terms of ordinal data, in this regression setting, the following continuous variable is considered:

$$Y^* = -\boldsymbol{\beta}\mathbf{X} + \epsilon,$$

where $\mathbf{X}$ is the vector of explanatory variables and ϵ is an error component with cumulative distribution function F. When user-specific random variations lie within the data, one needs to consider a link between unobserved variable Y^* and the observed response variable Y. There is a concept of threshold that assumes $Y \in \{1, 2, \ldots, g\}$ is represented as a rough version of Y^* by,

$$Y = r \iff \alpha_{r-1} < Y^* \leq \alpha_r, \tag{1}$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_k = \infty$ are called cut-points or thresholds [15]. Based on the above settings, the *cumultaive model* is given by:

$$P(Y \leq r|\mathbf{X}) = F(\alpha_r + \boldsymbol{\beta}\mathbf{X}), \qquad r = 1, 2, \ldots, g-1, \tag{2}$$

where $\alpha_1, \ldots, \alpha_r, \boldsymbol{\beta}$ are the parameters of the model. The model in Equation 2 is the *categorical regression model*. In such scenario, the usual choice of F is based on the logistic function, $\frac{1}{1+\exp[-\alpha_r+\boldsymbol{\beta}\mathbf{X}]}$, which is defined as *ordinal logistic regression model*, *proportional odds model* or *logistic cumulative model* [15,16]. The simple interpretation and easy computation of logistic regression make it widely applied in various areas of research. The link function for this model is given by,

$$g_r(\pi_1, \pi_2, \ldots, \pi_q) = \log\left(\frac{\pi_1 + \pi_2 + \ldots + \pi_r}{1 - (\pi_1 + \pi_2 + \ldots + \pi_r)}\right)$$

Cumulative Link Mixed Model (CLMM)

When the participant-specific random-effects can be incorporated in the *ordinal logistic regression model*, it becomes CLMM. To incorporate users' random effects into the model, the observed stress levels are denoted by $Y_{i,t}$, where $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T_i$. There are n participants whose number of observations for stress levels $T_1, T_2, \ldots, T_n$ may vary. Therefore, the unobserved variations associated with participant i and time t is given by,

$$Y^*_{i,t} = -\boldsymbol{\beta}\mathbf{X}_{i,t} - \zeta_i + \epsilon_{it},$$

where $\epsilon_{i,t}$ has a cumulative distribution function F and $\zeta_i \sim N(0, \sigma^2)$ is a individual-specific random component. Notably, $\epsilon_{i,t}$ and ζ_i are independent from each other. Now the threshold from Equation 3 can be redefined as:

$$Y^*_{i,t} = r|\zeta_i \iff \alpha_{r-1} < Y^*_{i,t} \leq \alpha_r, \tag{3}$$

which yields the *cumulative link mixed model (clmm)* as follows:

$$P(Y_{i,t} \leq r|\mathbf{X}, \zeta_i) = F(\alpha_r + \boldsymbol{\beta}\mathbf{X} + \zeta_i), \tag{4}$$

meaning that the original thresholds $\alpha_1, \alpha_2, \ldots, \alpha_{g-1}$ in user i are now simultaneously shifted to new thresholds $\alpha_1 + \zeta_i, \alpha_2 + \zeta_i, \ldots, \alpha_{g-1} + \zeta_i$. In that case, the thresholds vary among participants. The original thresholds $\alpha_1, \alpha_2, \ldots, \alpha_{g-1}$ represent an average of stress scores across participants.

Using the Maximum Likelihood Estimation (MLE) the parameters can be estimated by Expectation-Maximization (EM) algorithm [15].

Results

A total of 416 responses were received from 26 participants. We examined how the median stress scores changed among different groups within covariates for all the individuals. The distribution of stress level among each individuals are given in Fig 1. It is worth noting that the distribution among individuals did not follow the same pattern, implying that the stress levels varied among individuals. As a result, a stress level reported by a particular participant is not applicable to indicate another one's stress status.

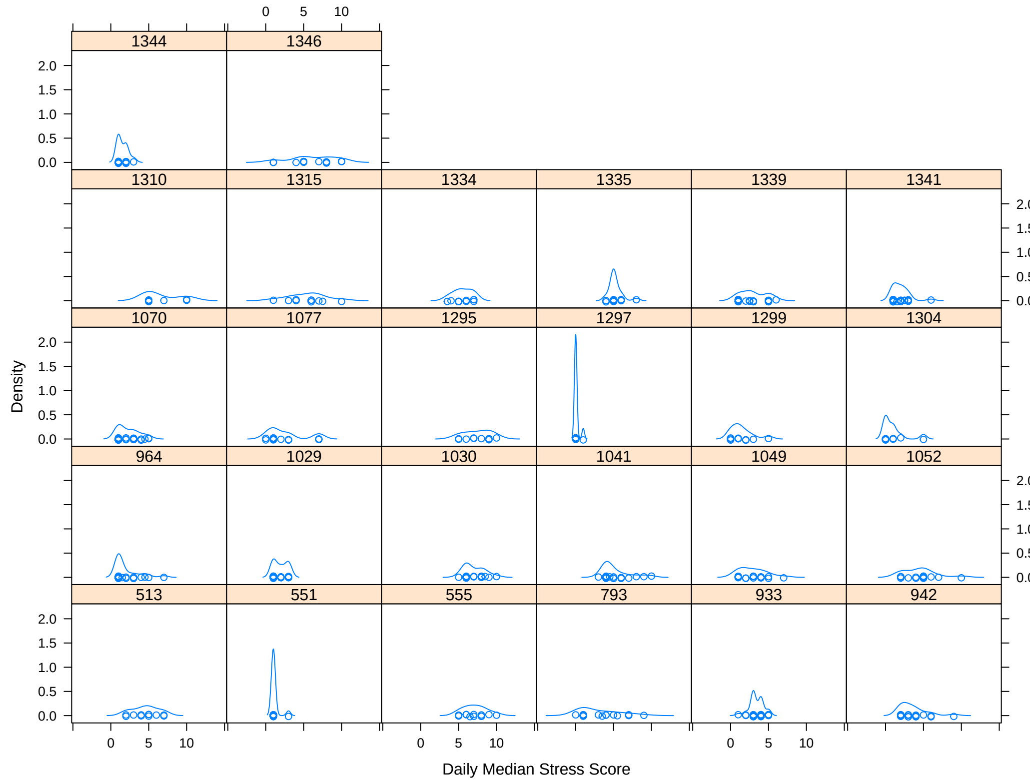

Figure 1. Distribution of median stress scores among different individuals.

Our preliminary analysis revealed that the largest daily amount of time spent on smartphone apps was about 2.5 hours. Additionally, the relationship between stress level and other covariates for different individuals was analyzed. The scatterplot in Fig 2 visualized the relationship between stress level and daily hours spent on apps based on different individuals, showing a pattern of different app usage behaviours. The same pattern was perceived from individual app use (*i.e.*, social network, entertainment and browser) and covariates (*i.e.*, school weeks, dwell time on campus, transportation upon arrival). The usage of game apps were not considered, since the majority of participants did not spend time on playing smartphone games.

The differences among individuals imply that it's necessary to apply a prediction model that takes the random effects of participants into account. The Cumulative Linear Mixed Model (CLMM) with random effect (i.e., participant) shows great promise to fit in that situation. `R` package `ordinal`[17] is useful for analyzing this type of data where the response variable is ordinal and varies among individuals.

The association between daily median stress levels and other covariates were examined via the CLMM, where each individual was considered as a random effect component in the model. The results are displayed in Table 2.

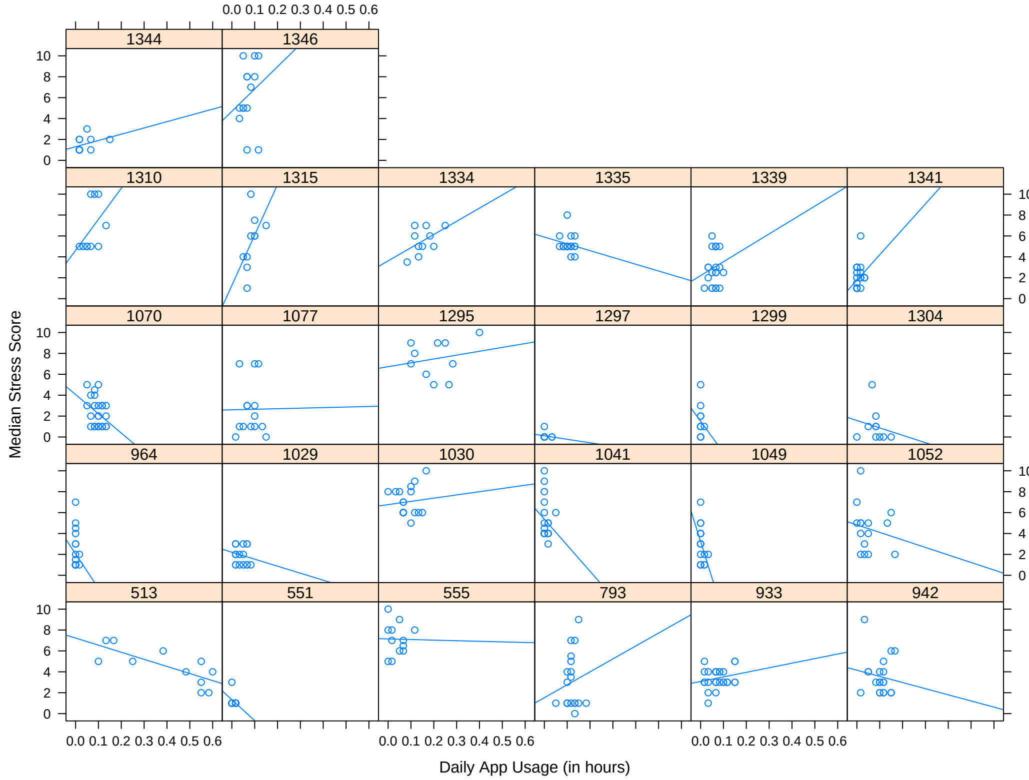

Figure 2. Relationship between median stress scores and hours spent on apps for different individuals.

Table 2. Association between stress level and individual covariates.

Covariates	Estimate	Standard Error	z-statistic	p-value
Week of the year				
Week 6	-	-	-	-
Week 7*	-0.803	0.282	-2.85	0.004
Week 8	-0.174	0.297	-0.58	0.560
Week 9	-0.052	0.276	-0.19	0.850
Week 10	0.066	0.386	0.17	0.865
Hours Spent on Campus				
Less than an hour	-	-	-	-
One hour	0.016	0.496	0.03	0.975
Two hours	0.262	0.476	0.55	0.582
Three hours	-0.152	0.451	-0.34	0.736
Four hours	-0.129	0.457	-0.28	0.777
Five hours	-0.315	0.472	-0.67	0.504
Six hours	0.099	0.446	0.22	0.825
Seven hours	0.704	0.538	1.31	0.191
Eight hours or more	0.145	0.497	0.29	0.769
Transportation upon Arrival				
Bus	-	-	-	-
Car/Carpool	-0.637	0.457	-1.39	0.164
Walk/Run**	-0.608	0.369	-1.64	0.099
App Usage				
Social Media	0.002	0.002	1.54	0.125
Entertainment	0.001	0.001	0.83	0.404
Utility	-0.001	0.001	-1.61	0.108
Browser	-0.001	0.003	-0.10	0.922
Smartphone Usage	-0.693	1.979	-0.35	0.726

∗Note: Statistically significant at 5% level of significance.
∗∗Note: Statistically significant at 10% level of significance.

As shown in Table 2, only Week 7 of the study had a significant association with stress level at 5% level of significance, the rest of the factors were not. Therefore, the backward elimination regression method was applied to look for a model that could best describe the relationship between stress level and the other covariates. To accomplish this, the covariates whose p-values were larger than 0.05 were removed. Further, datasets with poor compliance were not considered. In our study, the threshold response rate 10 (out of 29, which is the number of days of the study) was found to be suitable in our model. The results of the final model are presented in Table 3.

Table 3. Final model to predict stress level controlling for several factors.

Covariates	Estimate	Standard Error	z-statistic	p-value
Week of the year				
Week 6	-	-	-	-
Week 7*	-0.843	0.284	-2.97	0.003
Week 8	-0.163	0.298	-0.55	0.585
Week 9	0.011	0.278	0.04	0.968
Week 10	0.123	0.386	0.32	0.750
Utility Apps Use Time (min)*	-0.001	0.007	-2.25	0.024

∗Note: Statistically significant at 5% level of significance.

From the final model (Table 3), it can be told that the median stress levels were relatively lower during the winter break (week 8) and the week before, compared to the start of the study week. Also the stress level increased as time passed by and the highest median stress level was observed by the end of the study. Nevertheless, only Week 7 was found to be statistically associated with stress level at 5% level of significance. Apart from the above discovery, Table 3 also indicates that as people spent more time in utility apps, their stress levels decreased.

Conclusion

This study identified some factors associated with self-reported stress level among university students. The study was conducted in a Canadian university from early February to early March. Daily survey questions gave us the ground truth about the perceived stress level whenever a participant entered into the campus. We set up different factorial designs to get the optimal model. The preliminary analyses revealed that there existed a variation among participants in terms of their responses to the stress survey. In that case, one individual model for all the participants may be limited in sensitivity. Keeping this in mind, we employed *cumulative link mixed model* to investigate the relationship between stress level and individual covariates. We then validated the optimal model, by examining whether adding other covariates could assist to more precisely predict stress level. During the process, we found that including week of the study and daily amount of time spent on utility-related apps strengthened our model. The final model revealed that as the time spent on utility-related apps increased, the stress level decreased and that as the school week went on, the stress level increased. The stress levels were comparatively lower during the week of winter break and the week before the break, as opposed to the other weeks.

There are some limitations in this study: (1) diagnostic checking of the models is not finished by the time of writing; and (2) cross-validation should be conducted in future work to validate the models.

In future work, we will check for diagnostics of the models. Besides, modern Machine Learning algorithms (e.g., SVM, HMM and ANN) will be incorporated into user-specific models to predict stress level. For instance, *Ferdous et al.* applied user-centric SVM model to predict stress level and achieved significant accuracies[3]. Since the stress level could be hidden states and the covariates be observed states, it shows some promise to incorporate the user-specific ordinal HMM to predict stress level, which has gained popularity in the realm of data science.

In conclusion, this is a pilot study of understanding the factors associated with stress and assessing self-reported stress levels among university students.

References

1. Þórarinsdóttir H, Kessing LV, Faurholt-Jepsen M. Smartphone-Based Self-Assessment of Stress in Healthy Adult Individuals: A Systematic Review. J Med Internet Res. 2017;19(2):e41.

2. Stallman HM, Hurst CP. The University Stress Scale: Measuring domains and extent of stress in university students. Australian Psychologist. 2016;51(2):128–134.

3. Ferdous R, Osmani V, Mayora O. Smartphone app usage as a predictor of perceived stress levels at workplace. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). IEEE. IEEE Conference Publications; 2015. p. 225–228.

4. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. BMC public health. 2011;11(1):66.

5. Mehrotra A, Musolesi M. Sensing and Model Human Behavior Using Social Media and Mobile Data. arXiv preprint arXiv:170201181. 2017;.

6. Adams P, Rabbi M, Rahman T, Matthews M, Voida A, Gay G, et al. Towards Personal Stress Informatics: Comparing Minimally Invasive Techniques for Measuring Daily Stress in the Wild. ICST; 2014.

7. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland AS. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM '14. New York, NY, USA: ACM; 2014. p. 477–486.

8. Samaha M, Hawi NS. Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. Computers in Human Behavior. 2016;57:321–325.

9. Brutus S, Javadian R, Panaccio AJ. Cycling, car, or public transit: a study of stress and mood upon arrival at work. International Journal of Workplace Health Management. 2017;10(1).

10. Legrain A, Eluru N, El-Geneidy AM. Am stressed, must travel: The relationship between mode choice and commuting stress. Transportation Research Part F: Traffic Psychology and Behaviour. 2015;34:141 – 151.

11. Ikeda T, Ikeda T, Aoyagi O. The Relationship among Stress Response, Weight Management, and Physical Exercise in Japanese University Students. Journal of Sports Science. 2016;4:163–169.

12. Burg MM, Schwartz JE, Kronish IM, Diaz KM, Alcantara C, Duer-Hefele J, et al. Does Stress Result in You Exercising Less? Or Does Exercising Result in You Being Less Stressed? Or Is It Both? Testing the Bi-directional Stress-Exercise Association at the Group and Person (N of 1) Level. Annals of Behavioral Medicine. 2017; p. 1–11.

13. Sano A, Phillips AJ, Yu AZ, McHill AW, Taylor S, Jaques N, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN); 2015. p. 1–6.

14. Warburton DE, Katzmarzyk PT, Rhodes RE, Shephard RJ. Evidence-informed physical activity guidelines for Canadian adults. Applied Physiology, Nutrition, and Metabolism. 2007;32(S2E):S16–S68.

15. Tutz G, Hennevogl W. Random effects in ordinal regression models. Computational Statistics & Data Analysis. 1996;22(5):537–557.

16. Kleinbaum DG, Klein M. Logistic Regression: A Self-Learning Text. Springer Publishers New York, Inc.; 2010.

17. Christensen RHB. ordinal—regression models for ordinal data. R package version. 2010;22.