A Step Towards Quantifying When an Algorithm Can and Cannot Predict an Individual's Wellbeing

Orianna DeMasi
University of California, Berkeley
Berkeley, CA 94720, USA
odemasi@berkeley.edu

Benjamin Recht
University of California, Berkeley
Berkeley, CA 94720, USA
brecht@berkeley.edu

UbiComp/ISWC'17 Adjunct ,, September 11–15, 2017, Maui, HI, USA
ACM 978-1-4503-5190-4/17/09.
https://doi.org/10.1145/3123024.3125609

Abstract
Researchers are exploring the ability to infer complex signals, such a mental wellbeing, from easily collected smartphone behavioral data. Rather than focusing on improving overall accuracy of such an approach, we seek to understand when we are and are not capable of predicting an individual's wellbeing. In particular, we consider the ability to predict daily wellbeing from smartphone GPS location data as a case study. We hypothesize that user characteristics, such as behavioral variability, level of depression symptoms, and amount of labeled data, are related to improvements in prediction accuracy. Our preliminary results indicate that there may be a relationship between an algorithm's ability to successfully predict an individual's wellbeing reports and the individual's location behavior variability. While further work is needed to improve model accuracy and confirm this relationship in a larger study, our work is a step in the necessary direction of understanding which individuals can be monitored with smartphone data.

Author Keywords
Mental health; Depression; Mobile health (mHealth); Mobile sensing; Supervised learning.

ACM Classification Keywords
J.3 [Computer Applications]: Health; I.5.0 [Computer Methodologies]: Pattern Recognition (General).

Introduction

Mental health disorders can be devastating to those who suffer from them and are widespread. Collectively, it is thought that mental health disorders, such as depression, are so widespread that they are a major contributor to the global disease burden [12]. Improving mental health is particularly challenging as disorders can last for a lifetime and it is difficult to collect data on and monitor individuals over such long timescales.

The recent development and adoption of personal electronics provides an exciting opportunity for mental health, as personal electronics are a frequent source of highly personal data. It has been shown that data from personal electronics, such as smartphones, can be used to infer behavioral signals, such as sleep [7] and activity [11] without any user input. In addition to physical and social behaviors, researchers have begun exploring whether personal electronics can also sense mental wellbeing from passively collected data, such as smartphone GPS location and mobility [5, 6, 10, 14]. By not needing any user input, these devices may be a sustainable way to collect data on and track individuals' behavior over longer periods than are sustainable with paper journaling.

The possibility of automatic wellbeing tracking over long periods of time offers exciting opportunities for mental health research and treatment. However, the hope of tracking every individual with a smartphone may be naïve due to the large variance in individuals' behavior. Some individuals, say those who use their phone often, have active lifestyles, or have irregular schedules may be easier to track than individuals who, for example, often leave their phone at home or turn it off.

In this work, we explore the potential for understanding, and eventually predicting, whether an individual's wellbeing can be tracked by a smartphone application through a user study. As an example, we use features of location and mobility from GPS coordinates to predict individuals' daily wellbeing. The features of location and mobility that we use are inspired by previous research that sought to diagnose depression from similar features [14]. These features were chosen due to their apparent relevance to detecting depression and their reproducibility with the collected data.

We begin by calculating location and mobility features for each participant. Using these features, we utilize machine learning algorithms to predict each individual's wellbeing and then quantify the model improvement with GPS data over a simple baseline approach. We then look at the relationship of prediction improvement with user characteristics to see if user behavior is broadly related to an algorithm's ability to model their wellbeing. The user characteristics that we consider are data quality (as measured by median GPS accuracy) and quantity, behavioral variability, depressive symptoms, and emotional variability.

We find some significant positive correlations of user characteristics with prediction improvement. In particular, we find a positive correlation of the number of data points with prediction improvement, a negative correlation of baseline accuracy (i.e., how constant a user reports their state to be) with prediction improvement, and a positive correlation of location or behavioral variance with prediction improvement. By considering location variability as a course measure for behavioral variability, this result indicates what one would expect from a statistical perspective – more varying features are better able to model signals than features that rarely vary. From a psychological perspective, this result indicates that users with unfluctutating behavior are more difficult to model, perhaps because changes are outliers. Notably, we do not find significant relationships of depressive

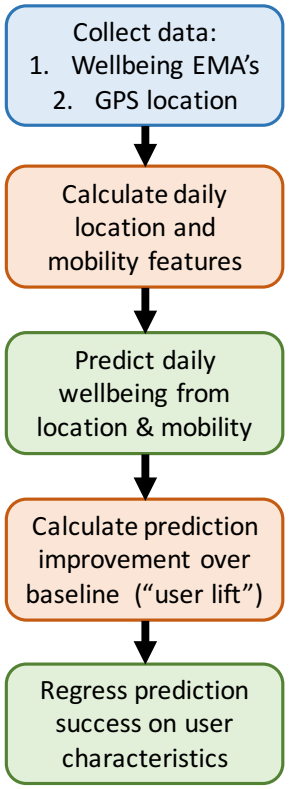

Figure 1: Overview of data flow from collection and processing through the final analysis of whether user characteristics are related to the improvement in prediction accuracy.

symptoms with prediction improvement. However, these relationships are not present across models, which could be the result of the task being difficult for models. While our preliminary results indicate some promise in being able to understand which individuals' wellbeing are easier to predict, further work and a larger study are needed to confirm the relationships of prediction improvement with user characteristics.

User Study and Data Collection

To explore whether there is a relation between user characteristics and success in predicting their wellbeing, we ran a user study. For this study, we recruited undergraduates with Android phones who spoke English as a native language on the University of California, Berkeley campus. While we recruited 107 participants, only 87 installed our custom Android application and 60 took the exit survey at the end of the study period.

The study ran for eight weeks, consisted of three phases, and collected two types of data: active user input and passive smartphone sensor data. The first phase of the study was an entry survey which asked user profile information, such as personality, demographics, and the Beck's Depression Inventory (BDI) [4]. The second phase was the daily collection of user input data, ecological momentary assessments (EMA's) of user wellbeing, and passive collection of smartphone sensor data. The final phase of the study was an exit survey which collected personality, the BDI, reflection on personal behavior during the study, and study feedback.

Users were queried four times a data for their wellbeing along two axes of the Circumplex model: mood and energy level [13]. These axes were labeled with words such as "good" and "bad" or "high energy" and "low energy", respectively, and implemented as two 9-point Likert scales. The words labeling the scales were selected by users from short lists of antonyms.

In addition to data from other sensors, data were collected of users' GPS location using the Funf Open Sensing Framework [1]. A user's location was recorded every five minutes. While a multi-modal approach is ultimately desired, we focus in this study only on the location data as an exploratory study of this approach. We chose GPS location for passively collected data as prior studies have indicated positive results with such an approach [5, 6, 10, 14]

Data Processing

The first stage of data processing aggregated multiple wellbeing observations that were made each day into a daily measure of wellbeing. Then the passively collected smartphone sensor data was processed into features describing individuals' daily mobility and location. Finally, features or "user characteristics" were calculated on each user that sought to quantify behaviors that may account for variability in prediction accuracy between users.

User Wellbeing

Users' wellbeing scores, which were solicited with four EMA's per day, were averaged to give daily levels. The two wellbeing dimensions measured, energy and mood, were considered separately. The means of the daily mood an energy levels during the course of the study were taken as study-means. These study-mean levels were used to determine when a user was having a particularly good day in terms of mood or energy. A particularly good mood day was when the mood level was above the study-mean mood score and similarly for energy. Using this approach, we accumulated two wellbeing measures for each day that a user responded to any wellbeing prompts: whether the

user was reporting an above-mean mood level for that day and whether the user was reporting an above-mean energy level.

Daily Location and Mobility
Here we focus on using GPS location and mobility features as predictors of daily user wellbeing. Location was intended to be collected every five minutes. However, some individuals' locations were collected at a higher frequency. For these users, we downsampled data to roughly five minute intervals. The features we used to describe daily location and mobility are adapted from a previous study that used similar features to quantify user behavior during an entire study period [14]. We selected these features due to our ability to reproduce them, given our regular sampling approach, and their success on a related task.

Before constructing daily features, we used a preprocessing stage to determine frequented locations. The preprocessing used K-Means clustering [2] to cluster all of a user's stationary location coordinates that were recorded during the entire study period. Points were determined to be stationary if the calculated gradient was less than 1km/hr. We chose the number of clusters for each user be such that the largest distance from any coordinate to the center of its assigned cluster was about 3km. We labeled the "home" location to be that which the user spent the most time at during the study period between the hours of midnight to 6am.

For each day of the study period when a user had sufficient GPS readings we used the cluster centers from the preprocessing stage and calculated the following measures:

- The sum of the variance of the latitude and the variance of the longitude coordinates, on a log scale.
- The number of locations (clusters) visited.
- The location entropy, i.e., $-\sum_i p_i \log p_i$ where p_i is the probability of the user being in location i at any point during that data.
- The fraction of time that the user spent at what we presume is their home location.
- The fraction of time the user was moving.
- The average distance that a user traversed between location readings (normalized by the time between readings).
- The "circadian rhythm", which we calculated as the euclidian distance between the vector where entry j is the fraction of time that a user spent at location j on an average day, and the day's vector where each entry i is the fraction of time that day that the user spent at location i.
- The radius of the minimum size circle that surrounded all of the user's locations for the day.
- The fraction of observations during which the user was moving (as determined by the calculated location gradient).
- The fraction of observations that were "GPS" rather than "Network", which could indicate the fraction of time that the user spent outside.

User Characteristics
We hypothesized that it is plausible that how well a user's location and mobility behavior reflects – and thus is predictive of – their wellbeing could be related to the following five dimensions:

1. How reliable a phone is at measuring location.
2. How much data a model has to learn from.
3. How depressed a user is.
4. How much a user's daily location pattern fluctuates.
5. How much a user's emotional wellbeing fluctuates.

R^2 = 0.241
Adj. R^2 = 0.101
F-stat = 1.717
p =.165

Feature	coef	p-val
Intercept	-25.49	.574
GPS radius	4.05	.324
No. days	0.34	.206
Avg. BDI	0.06	.819
Loc. var.	0.92	.208
Base acc.	-0.03	.967

Model 1: Linear model relating user lift (from L1-penalized logistic regressions) to user characteristics. On average, daily predictions were 3.77% less accurate than a constant baseline model.

We quantified these potential sources of variability with the following measures:

- The median radius of confidence reported by the GPS sensor (on a log scale).
- The number of labeled data points we have for a user (i.e., days with GPS location and user wellbeing).
- The expression of depressive symptoms (as measured by the BDI and averaged between the entry and exit responses).
- The sum of the variance of longitude coordinates and variance of the latitude coordinates during the course of the study (on a log scale).
- The user's baseline accuracy: the percent of wellbeing observations that would be correctly predicted if the user were always predicted to be at their most commonly reported state.

The radius of confidence or "inaccuracy" of the GPS location data, is the radius of the circle that the sensor estimates the true location falls into with high confidence. The second to last measure, location or behavioral variance, is related to the daily location variance described previously. Instead of being calculated on the coordinates for a single day, it was calculated on all coordinates from the entire study period. This feature characterized a user's behavior during the study period instead of during a single day.

R^2 = 0.647	
Adj. R^2 = 0.582	
F-stat = 9.906	
p < .001	

Feature	coef	p-val
Intercept	51.27	.032
GPS radius	2.59	.214
No. days	0.34	.015
Avg. BDI	-0.2	.130
Loc. var.	0.8	.035
Base acc.	-1.25	.002

Model 2: Linear model relating user lift (from L2-penalized logistic regressions) to user characteristics. On average, daily predictions were 1.97% less accurate than a constant baseline model.

Data Analysis

We performed two stages of analysis to explore whether user characteristics relate to how successfully an algorithm can predict an individual's wellbeing. First we used standard machine learning procedures to predict daily user wellbeing from the location and mobility features. Second we related the success of these models to the user characteristics described above. An individual's emotion and location had to be observed for at least 14 days of the study for them to be included in the analyses. To quantify success, we needed to account for variability in how regularly individuals reported a single wellbeing measure.

Individual Wellbeing Baseline Models

Individuals reported different levels of emotional variance, e.g., some individuals always reported the same mood while others report different mood levels. As a result, certain individuals are "easy" to predict with high accuracy, as predicting that they are always at the same state will usually be correct. However, from an algorithm's perspective, these individuals are challenging. It is difficult for an algorithm to predict the individual's wellbeing better than a baseline model that always guesses that the individual is always at the same state. To account for individuals' base level of difficulty, we considered the "baseline accuracy", which is the percent of observations that would be correctly predicted if an individual were always predicted to be at their most frequently reported state. The "baseline error" is the percent of observations that would be incorrectly predicted by always assuming that an individual is at their most commonly reported state.

Wellbeing Prediction

In the first stage of analysis, we attempted to predict whether a user was having a particularly good day (in terms of mood or energy level) from their location and mobility measures. For these predictions, we used a variety of standard machine learning models: logistic regression (with L1 and L2 penalties), random forest classifiers, and support vector machines (SVM's) [3, 9]. Models were trained on each individual's data (personal models) with leave-one-out cross-validation. Model hyperparameters were trained with 10-fold cross-validation on the training set.

Characterizing Prediction Improvement with User Lift
To characterize prediction improvement over a naïve approach that uses no features, we considered ***user lift*** to be the difference of model accuracy with the baseline accuracy described above [8]. User lift quantifies for each user how much better a machine learning model is than guessing.

Relating User Characteristics to Prediction Success
To better understand when users' daily wellbeing may be predicted by an algorithm, we related different algorithms' user lift for each individual to the user's above mentioned user characteristics. We related prediction improvement, i.e., user lift, to user characteristics with a multivariate linear regression. This model was chosen for interpretability.

Results

Of the individuals who participated in our field study, 33 had enough data to be included in our analyses. This limited number was in part due to compatibility issues that we encountered with the smartphone application and in part due to limited user participation. The level of depressive symptoms for each participant was quantified as their average report (between entry and exit surveys) to 20 questions of the BDI. The mean level reported across included participants was 12.68 (standard deviation: 10.66). Of the participants included in the analyses, 29.63% reported levels above 15, which could indicate mild levels of depressive symptoms.

Predicting Daily Wellbeing
In general, we found that models did not have appreciably higher prediction accuracy than the baseline approach, i.e., predicting users to be at their most common state all the time. This result is reflected in negative average user lift for predicting daily energy. It is also reflected by low correlation of model accuracy across individuals, as can be seen in Figure 2. The model with maximum average user lift for predicting daily energy was the support vector machine, which still had negative user lift (i.e., improvement over the baseline approach) of -1.50%. We also noted models were worse at predicting whether an individual's mood was particularly good than predicting whether an individual's energy was particularly high. However, on individual users some of the models performed considerably better than the constant baseline approach. This variance of performance between individuals motivates the second stage of analysis.

Explaining Prediction Improvement
Correlation between user characteristics was fairly low, as seen in Figure 3. The user characteristics that were most correlated were the variance in location coordinates and the total number of observations. The multivariate regression models relating user lift of daily energy predictions to user characteristics are summarized in Models 1 - 4. These models explore the improvement of predicting energy and not mood. While user lift for mood prediction did vary between individuals, the overall average user lift was better for energy prediction. As a result, we proceed with understanding the error of predicting energy and will investigate mood further after better models have been developed.

Despite accounting for little of the overall variability between individuals, user characteristics had significant relationships with the user lift from L2-penalized logistic regressions and random forest models ($p < .01$) as well as from SVM's ($p < .05$). When random forests or an L2-penalized logistic regression are used as the prediction model, we see a positive correlation of location variance with prediction improvement. This indicates that individuals who displayed more physical behavioral variance were easier to successfully predict than those with little variation. For the L2-penalized logistic regression, we also note a significant positive correlation of the number of data points with user lift and a

R^2 = 0.402
Adj. R^2 = 0.291
F-stat = 3.633
p =.012

Feature	coef	p-val
Intercept	-4.67	.884
GPS radius	2.63	.366
No. days	0.22	.249
Avg. BDI	-0.02	.931
Loc. var.	1.38	.011
Base acc.	-0.25	.630

Model 3: Linear model relating user lift (from random forests) to user characteristics. On average, daily predictions were 5.16% less accurate than a constant baseline model.

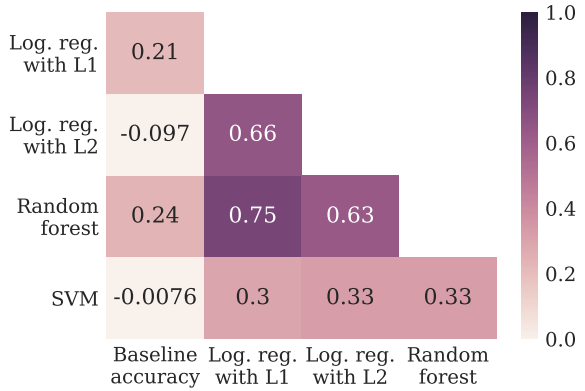

Figure 2: Correlation of how well different models and a constant baseline model predict individuals' daily energy. Correlation is calculated between average prediction accuracy on individuals from different models.

R^2 = 0.330
Adj. R^2 = 0.206
F-stat = 2.656
p =.045

Feature	coef	p-val
Intercept	45.04	.252
GPS radius	-5.0	.160
No. days	0.43	.063
Avg. BDI	-0.04	.859
Loc. var.	0.47	.455
Base acc.	-0.91	.147

Model 4: Linear model relating user lift (from support vector machines) to user characteristics. On average, daily predictions were 1.50% less accurate than a constant baseline model.

negative correlation with the baseline accuracy. These relationships indicate that individuals who have more data are easier to learn, and those who report little fluctuation in state are harder to predict more accurately than a baseline model, which is already quite accurate.

An interesting consistency between models is a lack of significant relationship between the reported expression of depressive symptoms, as measured by the BDI, with model improvement. Higher BDI scores indicates increased depressive symptoms. No significant relationships indicate that individuals with higher depressive symptoms are either easier or more difficult to predict.

Discussion

In this work, we explored the potential to explain when individuals' wellbeing can and cannot be predicted by location data from their smartphone. We have focused preliminarily on the example of predicting perceived energy level from GPS location and mobility data and relating prediction improvement to user characteristics, such as emotional variability, location variance, level of depressive symptoms, and amount of data collected.

In general, it was difficult to learn models that made better predictions than predicting each individual to always be at their most common state. Daily mood was particularly difficult and insufficiently accurately predicted. There were improvements in prediction accuracy when using location data to predict energy, but this did not seem to remain consistent across models, as indicated by low correlation between average model accuracy (Figure 2). The variability in model improvement between individuals motivated us to compare prediction improvement with user characteristics.

When we related model improvement to user characteristics, we found a significant positive correlation between location variance, which we used as a coarse measure of behavioral variance, and model improvement (user lift). From a statistical perspective, this indicates that users' wellbeing can be better learned from more varying and potentially descriptive features. From an applied perspective it indicates that more active individuals might be easier to monitor with this approach. Additional characteristics were also found to be significant, but were dependent upon which model was used. Depressive symptoms where notably not found to have a significant relationship with model improvement regardless of model.

There are limitations to this work, including a small sample size. A study with a larger cohort size is needed to validate the above mentioned relationships (and lack of relationships). We have also restricted our first step to explore GPS data, but other sensors should be included, as some sen-

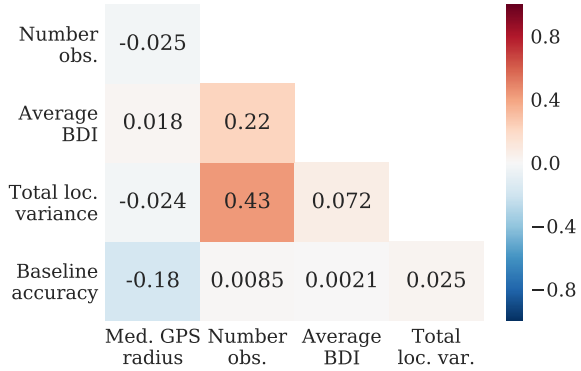

Figure 3: Correlation between user characteristics. "Med." denotes the median and the Beck's Depression Inventory (BDI) is a measure of depressive symptoms. In general, there isn't high correlation between user characteristics.

sors may be more predictive for different individuals. The limited predictive capability of location and mobility features that we found could have also constrained our ability to explain model improvement by having little model improvement in general. It is possible that with more descriptive features (or a multi-modal approach) daily wellbeing prediction would be more accurate and thus the resulting model improvement would have stronger or different relationships with user characteristics. Finally, additional user characteristics should be considered, which may improve the quantification of user variability and reveal stronger relationships between model improvement and user characteristics. In particular, different measures of depressive symptoms, other than the BDI, may better capture depressive symptoms that may influence predictive capability.

In future work we would like to explore a larger study population and incorporate more descriptive features for daily wellbeing prediction. Such features could include those from other sensors, such as accelerometer activity. As a result of including more daily features, we would also like to explore different user characteristics that describe the behaviors measured by other sensors. For example, when exploring the benefit of using accelerometer activity measures to predict daily wellbeing, we would like to address if the model improvement is related to the user's general activeness and variability. With sufficiently descriptive daily features that generate better models, we could also explore the relationship of user characteristics with predicting daily mood in addition to daily energy.

This preliminary work is a case study in trying to understand model discrepancies for wellbeing prediction, a problem that has been generating optimism for medical applications. Larger studies with multi-modal prediction approaches are still needed to improve monitoring accuracy. However, these studies may consider including an analysis, such as we have presented, to understand for which individuals such a monitoring approach (i.e., with a smartphone) is plausible and for whom it is unrealistic. Smartphone monitoring may be attractive for its ease of use, but it is imperative to have accurate monitoring for individuals suffering from mental health disorders. Understanding when smartphones are unable to monitor individuals, as we have attempted to do, may eventually help achieve such necessary reliability.

REFERENCES

1. Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7, 6 (2011), 643–659.

2. David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

3. Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. 2007. Support vector regression. *Neural Information Processing-Letters and Reviews* 11, 10 (2007), 203–224.

4. Aaron T Beck, C Ward, M Mendelson, and others. 1961. Beck depression inventory (BDI). *Arch Gen Psychiatry* 4, 6 (1961), 561–571.

5. Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

6. Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1293–1304.

7. Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tonmoy Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 145–152.

8. Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2017. Meaningless comparisons lead to false optimism in medical machine learning. (2017). (Under review).

9. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

10. Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. 2015. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 222–228.

11. Oscar D Lara and Miguel A Labrador. 2013. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE* 15, 3 (2013), 1192–1209.

12. World Health Organization. 2017. Depression [fact sheet no. 369]. (2017). (Accessed April 2017), URL=http://www.who.int/mediacentre/factsheets/fs369/en/.

13. J.A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (1980), 1161–1178.

14. Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.