PREDICTING SYMPTOM TRAJECTORIES OF SCHIZOPHRENIA USING MOBILE SENSING

Photo, istockphoto.com

Rui Wang, Weichen Wang *Dartmouth College* **Min Hane Aung** *Cornell University*
Dror Ben-Zeev, Rachel Brian *University of Washington*
Andrew T. Campbell *Dartmouth College* **Tanzeem Choudhury** *Cornell University*
Marta Hauser, John Kane *The Donald and Barbara Zucker School of Medicine at Hofstra/Northwell*
Emily A. Scherer *Dartmouth College* **Megan Walsh** *Northwell Health*

Editors: Nic Lane and Xia Zhou

Continuously monitoring schizophrenia patients' psychiatric symptoms is crucial for in-time intervention and treatment adjustment. The Brief Psychiatric Rating Scale (BPRS) is a survey administered by clinicians to evaluate symptom severity in schizophrenia. The *CrossCheck* symptom prediction system is capable of tracking schizophrenia symptoms as measured by BPRS using passive sensing from mobile phones. We present results from a randomized control trial, where passive sensing data, self-reports, and clinician administered 7-item BPRS surveys are collected from 36 outpatients with schizophrenia. We show that our system can predict a symptom scale score based on a 7-item BPRS within ±1.45 error on average. Finally, we discuss how well our predictive system reflects symptoms experienced by patients by reviewing a case study.

Schizophrenia is a severe chronic psychiatric disorder associated with high individual and societal costs. Psychosis is not considered as a fixed state. Rather, the majority of people with schizophrenia fluctuate between full or partial remission and episodes of symptomatic relapse. Psychotic symptoms may change over months, weeks, or even days, and can be affected by both external conditions and internal states [2]. In the case of relapse, patients may find themselves suffering severe hardship if not helped, such as homelessness, incarceration, and victimization. Patients are often hospitalized as a consequence of schizophrenia relapse.

Clinicians need to track schizophrenia patients' symptom states to identify risks and adjust treatment as necessary. At the CrossCheck study [1,2,3] partner hospital, Zucker Hillside Hospital, in New York City, schizophrenia outpatients regularly schedule clinical visits with their clinicians. The time between visits varies from once a week to once a month, depending on the patients' symptom severity and risk. Clinicians use a battery of mental health tests to evaluate the patients' symptom states and adjust their treatment accordingly. However, clinicians are not aware if a patient experiences deteriorated symptoms between visits. Because of this gap of knowledge in outpatient management between visits, clinicians are more likely to miss the optimal time to intervene to treat patients who are increasingly symptomatic and experiencing increased risk of relapse. The high cost of hospital visits and face-to-face assessments further prohibits patients from more frequent visits with their clinicians to adjust treatment or provide intervention.

In order to address these shortcomings, we developed the CrossCheck symptom prediction system to monitor patients' trajectory of psychiatric symptoms. The system predicts patients' weekly 7-item Brief Psychiatric Rating Scale (BPRS) total scores using passive sensing and self-reported ecological momentary assessment (EMA) responses from smartphones. Weekly predictions track participants' overall psychiatric symptoms and level of risk for relapse. The 7-item BPRS is a subset of the original 24-item BPRS, which measures psychiatric symptoms associated with schizophrenia. The clinical team at our study partner hospital determines the 7 BPRS items (grandiosity, suspiciousness, hallucinations, unusual thought content, conceptual disorganization, blunted affect, and mannerisms and posturing) to be the strongest predictors of deterioration in symptoms. The 7-item BPRS is administered by a trained clinician at our study partner hospital. The scored 7-item BPRS survey serves as a clinical indicator of treatment for patients who have moderate to severe disease.

In this paper, we present 7-item BPRS prediction results from an ongoing randomized control trial (RCT), in which passive sensor data, self-reports and clinically administered 7-item BPRS reports are collected from 36 outpatients with schizophrenia, who have been recently discharged from hospital over a period ranging from 2 to 12 months. We show that our system can predict 7-item BPRS using a combination of passive sensing data and self-reported EMA. Importantly, we also show that we can predict 7-item BPRS scores based purely on passive sensing data from mobile phones.

CrossCheck STUDY

The CrossCheck study is an ongoing randomized controlled trial [6] conducted in collaboration with a large psychiatric hospital, Zucker Hillside Hospital, in New York City [3]. The study aims to recruit 150 participants for 12 months using rolling enrollment. The participants are randomized into one of two study arms: smartphone (n=75) or treatment-as-usual (n=75) [3]. Figure 1 shows an overview of the CrossCheck symptom prediction system. This study is approved by the Committee for Protection of Human Subjects at Dartmouth College and Institutional Review Board at Northwell Health System.

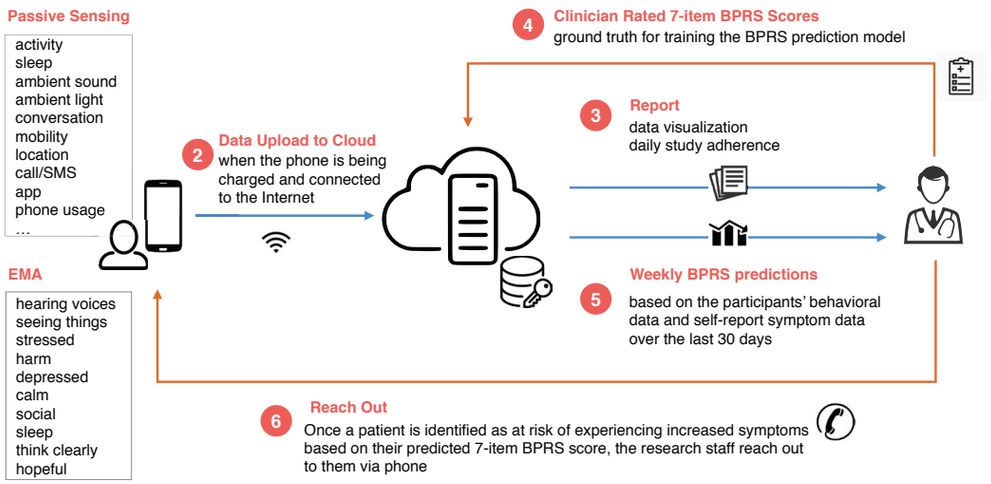


FIGURE 1. System overview of the CrossCheck symptom prediction system.

We use the study hospital's electronic medical record to identify potential study candidates. A candidate is a patient who is 18 or older, meets DSM-IV or DSM-V[1] criteria for schizophrenia, schizoaffective disorder or psychosis, and has had psychiatric hospitalization, daytime psychiatric hospitalization, outpatient crisis management, or short-term psychiatric hospital emergency room visits within 12 months before study entry. The study candidates are randomly assigned to either the smartphone arm or treatment-as-usual arm. Participants in the smartphone arm are given a Samsung Galaxy S5 Android phone equipped with the CrossCheck app. Participants' personal phone numbers are migrated over to the new phone and they are provided with an unlimited data plan for data uploading.

The CrossCheck app [3] is built based on our prior sensing work [4,7,11]. The app continuously infers and records participants' physical activities (e.g., stationary, in a vehicle, walking, running, cycling), sleep (duration, bedtime, and rise time), and sociability (i.e., the number of independent conservations a participant is around and their duration). The app also collects audio amplitude, accelerometer readings, light sensor readings, location coordinates, application usages, and call logs. The app uses a built-in MobileEMA component [11] to administer self-reported EMAs [5]. To protect participants' privacy, the app does not collect phone numbers, content of text messages or any conversation content. We remotely erase the CrossCheck data on the phone and reset it if the phone is lost. The app uploads the data to the secured CrossCheck data analytics service in the cloud when participants are charging their phones, with Wi-Fi or cellular data services. See [3] for more detailed discussion of the implementation of the phone application and system.

Participants schedule monthly visits with their clinicians. During their visits, clinicians administer the 7-item BPRS. The 7-item BPRS score ranges from 7 to 49. Higher scores are associated with more severe symptoms. The clinician-rated 7-item BPRS scores are used as the ground truth for training the 7-item BPRS prediction model.

The CrossCheck system predicts participants' BPRS scores every week. The prediction system sends out a 7-item BPRS prediction report every week to research staff by email. The prediction report shows the participant's predicted 7-item BPRS scores for the last 3 weeks. Our research staff use the predicted 7-item BPRS scores over the preceding 2-week period to identify any patients who may be potentially at risk. A patient at risk is one whose predicted 7-item BPRS score is above 12 or experiences an increase of 10% or more since their last predicted 7-item BPRS score. The research and clinical teams determined the rising risk threshold criteria (i.e., the score cut off and percent change) by studying the historical BPRS scores from patients who experienced relapse; that is, we analyzed scores in time periods prior to relapse to determine the cut-off and, in addition, because some patients' data prior to relapse showed a lower cutoff but large increasing percent changes, we also determined the additional criteria of the 10% change or greater between two predictions as a red flag.

CrossCheck DATASET

The CrossCheck dataset comprises the participants' monthly 7-item BPRS scores rated by their clinicians, behavioral features extracted from passive sensing, and symptom features extracted from self-report EMAs. We use 30 days of sensing and self-report EMA data to predict a 7-item BPRS score. The 30-day time frame is called the 7-item BPRS prediction time frame. The 30-day time frame matches the interval of clinician rated 7-item BPRS, which is 30 days on average. The passive sensing features summarize the level of behaviors (e.g., the average conversation duration per day in the 30-day time frame) and behavior changes (e.g., increase or decrease in conversation duration and the dynamics – for example direction and steepness – of change) in the 7-item BPRS prediction time frame. To compute a feature for the prediction time frame, we first compute the daily feature time series from the raw

[1] Diagnostic and Statistical Manual of Mental Disorders (DSM–5): https://www.psychiatry.org/psychiatrists/practice/dsm

sensing data. We then compute the 30-day features from the daily feature time series. In what follows, we discuss the construction of the dataset in detail.

The CrossCheck app collects a wide range of behavioral passive sensing data from the phone. This data captures physical activity, sociability (based on speech and conversational data), mobility, sleep, phone usage, and characteristics of the ambient environment in which participants dwell. We compute daily features from the passive sensing data, for example, conversation duration during a day. Patients periodically respond to a set of short questions related to their symptoms and functioning using their phones. The CrossCheck app administers a 10-item EMA [3] every Monday, Tuesday and Friday. We use each item's score as a self-report feature to capture symptoms and affects. We also calculate the EMA negative score, positive score, and sum score from the responses [3].

We compute mean and slope for each of the passive sensing features and self-report EMA items during the 30-day 7-item BPRS prediction time frame. In order to ensure that we have enough data to compute the mean and slope for each passive data, we include only the data in our analysis that satisfies the following criteria. We define a "good day" as a day with more than 19 hours of the sensing data. In order to avoid missing data and skewing the time series features in the prediction time frame, we need to control the data completeness in the 30-day time frame. We include time frames with more than 20 good days of the sensing data. We use 116 7-item BPRS records and corresponding features from 36 participants for evaluating the 7-item BPRS prediction performance.

PREDICTION MODEL AND RESULTS

We use Gradient Boosted Regression Trees (GBRT) [9] to predict the 7-item BPRS scores. GBRT is an ensemble method that trains and combines several weak regression trees to make accurate predictions. In order to understand the prediction accuracy of the three different feature setups, we train three models with (i) using both the passive sensing features and the EMA features; (ii) using just the passive sensing features; and (iii) using just the EMA features. We evaluate the prediction accuracy with leave-one-record-out cross validation and leave-one-subject-out cross validation. The leave-one-record-out cross validation leaves one 7-item BPRS example out from the dataset as the testing example and uses the rest of the examples for training the model. The results from the leave-one-record-out cross validation show the prediction accuracy of predicting an existing participant's 7-item BPRS score. The participant's previous clinician-rated 7-item BPRS scores are available to the system to improve the prediction accuracy by incorporating the data in the training examples. The leave-one-subject-out cross validation trains the model with data from subjects other than the testing subject and tests on the testing subject's data. The results from the leave-one-subject-out cross validation shows the prediction accuracy of predicting a new participant who just joined the study when their clinician-rated 7-item BPRS scores are not available to the system.

TABLE 1. Prediction Performance

		passive sensing + EMA	passive sensing	EMA
leave-one-record-out	MAE	1.45	1.59	1.62
	Pearson's r	0.70*	0.63	0.62*
	GEE coeff	1.05*	1.11	0.81*
leave-one-subject-out	MAE	1.70	1.80	1.90
	Pearson's r	0.61*	0.48	0.50*
	GEE coeff	0.99*	0.93	0.81*

* $p < 0.0001$

We use mean absolute error (MAE), the Pearson's r, and generalized estimating equations (GEE) [8] to evaluate the prediction performance. MAE describes the bias of the predictions. The Pearson correlation treats the predicted BPRS scores as independent variables. The Pearson's r describes how well the predictions capture the outcome's variance. GEE focuses on estimating the average response over the population [8]. It is a more robust method to evaluate correlations between repeated measures. The GEE coefficient shows the direction of the correlation and the p-value indicates the statistical significance of the coefficient.

Table 1 shows the mean absolute error, the Pearson's r, and GEE coefficient for all models predicting the BPRS score. The leave-one-record-out cross validation with both passive sensing and EMA features achieves the best result with MAE = 1.45, meaning we can predict the 7-item BPRS score with on average ±1.45 error. The predicted 7-item BPRS scores strongly correlate with the 7-item BPRS ground truth with $r = 0.70$, $p < 0.0001$. Using only passive sensing or the EMA feature obtains slightly poorer MAE. The result shows that our existing system can accurately predict patients' 7-item BPRS scores. The result gives us confidence to track symptoms every week. The prediction performance for leave-one-subject-out cross validation using only passive sensing or EMA features is MAE = 1.80, $r = 0.48$, $p < 0.0001$ and MAE = 1.90, $r = 0.50$, $p < 0.0001$ (4.5% of the scale), respectively. When comparing, using both passive sensing and EMA features, this results in a 0.1 and 0.2 increase in absolute errors, respectively. Again, passive sensing features outperform EMA features in terms of MAE. In both cross validations, we see that combining the passive sensing and EMA features performs better than just using passive sensing features, which in turn outperforms EMA features.

Figure 2 shows the average within-individual prediction error of the two models with the different feature setups and cross-validation methods. The order of the patients shown in the plots is determined by their average clinician-rated BPRS scores. We observe that our models archive lower prediction errors on patients with lower clinician-rated BPRS scores but higher errors on patients with higher clinician-rated BPRS scores. Most of the clinician-rated BPRS scores are between 7 and 12. Therefore, the dataset is unbalanced

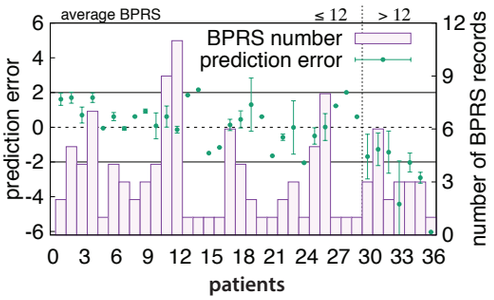
and skews to lower BPRS scores (≤ 12). The GBRT models are undertrained for higher BPRS scores (> 12). As a result, the models underestimate high-BPRS-score patients' scores (i.e., patients with average BPRS > 12). The prediction models need more high-BPRS-score patients' data to improve the prediction performance.

Clinicians may use the predicted BPRS scores to assess patients' symptoms. Our research team sets the score cutoff for symptom deterioration to 12. For patients with a BPRS score less than 12, a positive error may lead to a false positive of symptom deterioration. For example, if a patient is rated as 11 but the predicted BPRS score is 12.5, the patient would be falsely labeled with symptom deterioration. Conversely, for patients with BPRS scores higher or equal to 12, a negative error may lead to a false negative of symptom deterioration. However, if a patient is rated a higher score, it allows a larger margin of error. For example, if a patient would be rated as 20, the same as patient 36 as shown in Figure 2(a), a prediction error of -6 (i.e., predicted BPRS score is 14) is still a true positive. Since our models tend to underestimate high-BPRS-score patients' BPRS score, clinicians could use a lower symptom deterioration cutoff amongst the high-BPRS-score patients to reduce the false negative rate.

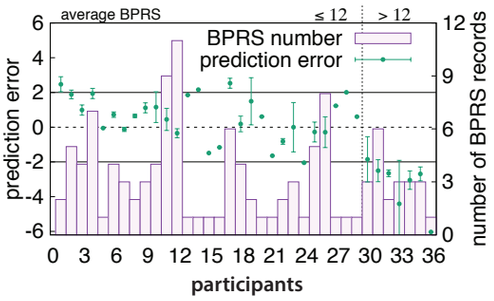
PATIENT CASE STUDIES

The CrossCheck prediction model is retrained each week if new clinician-rated 7-BPRS scores are available. Each week our research staff reviews the weekly prediction scores of all patients using the smartphone. Once we identify that a patient might be at risk, our research staff outreach and contact the participant and the clinical team at the hospital. In what follows, we provide insights into the life of one patient at the time our system indicates increasing symptoms. We show through anecdotal information from research staff and the clinical team reaching out to patients when the prediction system indicates rising risk.

The patient is a 55-year-old African American male diagnosed with schizophrenia, paranoid type. He was clinically flagged on Aug. 22, 2016, based on an elevated predicted 7-item BPRS score of 12.86. When our research staff at the hospital contacted the patient on Aug. 24, 2016, he endorsed symptom decompensation over the past three months with increasing intensity over the past three weeks. He discussed negative thoughts he had had about his deceased mother who had passed away five years earlier. The patient also said he believed these thoughts were present to make him feel "emotionally sick." The on-site researcher and patient discussed coping mechanisms. Once the researcher determined that the patient was not in any imminent danger, the researcher encouraged him to share all these symptoms with his treatment team and then brought the call to an end. The researcher contacted the clinical team to inform them of the symptoms reported by the patient. The patient's psychiatrist reviewed the new information and told the researcher that the patient had been experiencing difficulty scheduling his next outpatient medication management appointment. The psychiatrist immediately reached out to the patient's case manager to coordinate an in-person visit, which occurred less than a week after the initial research outreach. The psychiatrist adjusted his medication accordingly during the clinical visit. This case shows that the predictive system, outreach and clinical assessment all concur strongly.



(a) Passive sensing + EMA leave-one-record-out.



(b) Passive sensing + EMA leave-one-subject-out.

FIGURE 2. The average within-individual prediction error of the six models. The patients are ordered by their average rated BPRS scores. The vertical dashed line separate patients with average BPRS score ≤ 12 and patients with BPRS score > 12. The horizontal lines label the region with prediction error more than -2 and less than 2. Patients with higher-rated BPRS scores get worse predictions. This is because the dataset is skewed to patients with lower BPRS scores.

CONCLUDING REMARKS

The CrossCheck system discussed in this paper shows promise in using mobile phones and passive sensing to predict symptoms of schizophrenia for people living out in the community. The system and models show extremely good performance using passive sensing and self-reports as well as when just using passive sensing. A system based purely on passive sensing opens the way for continuous assessment of symptoms and risk as people go about their everyday lives.

We also recognize the limitations of our work. We only had 116 BPRS clinician-scored surveys to train our model. Outpatients do not experience severe symptoms often and thus report lower 7-item BPRS scores. Therefore, the dataset is unbalanced and skews to lower BPRS scores. The unbalanced dataset causes our prediction models to underestimate the BPRS scores of patients with higher clinician-rated BPRS scores. However, we showed that clinicians may adjust the score cutoff for symptom deterioration and leverage the changes in predicted BPRS scores to reduce the false negatives. To further advance BPRS prediction, we need to collect more data, especially from patients with more severe symptoms. We would need to apply re-sampling techniques, such as SMOTE [10], to balance the dataset. Another possible limitation is that all patients live in a large, dense city and the models may not generalize to other locations, such as patients living in rural communities.

The CrossCheck symptom prediction system accurately captured the changing conditions of these patients as reported by the research and clinical teams that reached out to them or interacted with them during

subsequent clinical visits, respectively. These results look very promising. For more detailed discussion of the CrossCheck BPRS prediction system, see [1]. ■

Acknowledgments

The research reported in this article is supported by the National Institute of Mental Health, grant number R01MH103148.

Rui Wang is a PhD student in the Department of Computer Science at Dartmouth College. His research focuses on mobile sensing, behavioral modeling and data mining.

Weichen Wang is a PhD student in the Department of Computer Science at Dartmouth College. He received his BE degree in Software Engineering from Nanjing University and his ME degree from the Graduate School of Information, Production and Systems of Waseda University. Now a member of DartNets Lab, his research interests include ubiquitous computing, mobile sensing, behavioral modeling, social computing, mobile health and data mining.

Min Hane Aung holds a PhD in Machine Learning for Health Applications. A visiting lecturer at the University College London Interaction Centre and a data science consultant for the smartphone-based mental health monitoring start-up company HealthRhythms Inc., he was formerly a research associate within the People-Aware Computing group at Cornell University. A member of the United Kingdom's Institution of Engineering and Technology, his research is in the development of intelligent systems for the understanding of behavior using ubiquitous sensing techniques. Principal application areas include mental health and well-being, chronic pain management, and general functioning and affect.

Dror Ben-Zeev is a professor of Psychiatry and Behavioral Sciences at the University of Washington and a licensed clinical psychologist, specializing in the development and evaluation of technology-based approaches in the study, assessment, treatment, and prevention of mental illness. He serves as the director of the mHealth for Mental Health Program and co-director of UW's Behavioral Research in Technology and Engineering (BRiTE) Center.

Rachel Brian is the research project director for the mHealth for Mental Health Program at the University of Washington. She develops research infrastructure for studies, manages and trains research staff, develops recruitment and reporting systems, oversees study logistics, ensures the protection of human subjects, and is responsible for institution and funder documentation and reporting.

Andrew T. Campbell is a professor of computer science at Dartmouth College. His group developed a number of early-sensing applications for smartphones and is now focused on turning smartphones into cognitive phones. Before joining Dartmouth, he was an associate professor of electrical engineering at Columbia University.

Tanzeem Choudhury is an associate professor in the Department of Computing and Information Science at Cornell University. She directs the People-Aware Computing Group, which works on inventing the future of technology-assisted well-being. Choudhury received a PhD from the MIT Media Lab. tanzeem.choudhury@cornell.edu.

Marta Hauser is a licensed psychologist. She received her PhD from the Charité Medical University, Berlin, Germany, and is now the director of outreach at the Vanguard Research Group. Prior to this, she held positions in the research and clinical departments of two large health systems and practiced psychotherapy in faculty practice.

John M. Kane, M.D. is a professor and chairman of the Department of Psychiatry at The Donald and Barbara School of Medicine at Hofstra/Northwell. He received his BA from Cornell and his MD from The New York University School of Medicine. His research interests include disease management for chronic and severe mental illnesses, such as schizophrenia.

Emily A. Scherer received her PhD in biostatistics from Boston University and is an associate professor in the Department of Biomedical Data Science. Her interests include modeling intensively collected data in mental health and substance-use research.

Megan Walsh is a senior administrative manager for the Department of Psychiatry Research at Northwell Health. She received her MA in Mental Health Counseling from Brooklyn College and her MBA in Health Services Management from Hofstra University. Prior to her current role, Megan worked clinically with people with severe mental illnesses.

REFERENCES

[1] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. "Predicting symptom trajectories of schizophrenia using mobile sensing." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT).* 1, 3, Article 110 (September 2017), 24 pages.

[2] D Ben-Zeev, R Brian, R Wang, W Wang, A T Campbell, M S Aung, M Merrill, V W Tseng, T Choudhury, M Hauser, and others. 2017. "CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal* (2017).

[3] Rui Wang, Min S.H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. 2016. "CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia." In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 886–897. DOI:http://dx.doi.org/10.1145/2971648.2971740

[4] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. "Automatic detection of social rhythms in bipolar disorder." *Journal of the American Medical Informatics Association* 23, 3 (2016), 538–543.

[5] Dror Ben-Zeev, Gregory J McHugo, Haiyi Xie, Katy Dobbins, and Michael A Young. 2012. "Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group." *Schizophrenia Bulletin* 38, 3 (2012), 396–404.

[6] Thomas C Chalmers, Harry Smith, Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman, and Alexander Ambroz. 1981. "A method for assessing the quality of a randomized control trial." Controlled Clinical Trials 2, 1 (1981), 31–49.

[7] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiao chao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. "Bewell: A smartphone application to monitor, model and promote well-being." In 5th International ICST Conference on Pervasive Computing Technologies for Healthcare. 23–26.

[8] Kung-Yee Liang and Scott L Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73, 1(1986), 13–22. DOI:http://dx.doi.org/10.1093/biomet/73.1.13

[9] Jerome H. Friedman. 2001. "Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. http://www.jstor.org/stable/2699986

[10] Nitesh V. Chawla, Kevin W Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[11] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. "StudentLife: Assessing mental health, academic Performance and behavioral trends of college students using smartphones." In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, 3–14. DOI:http://dx.doi.org/10.1145/2632048.2632054