
Personalized versus Generic Mood Prediction Models in Bipolar Disorder

Marios Constantinides

Department of Computer Science
University College London
United Kingdom
m.constantinides@cs.ucl.ac.uk

Aleksandar Matic

Telefonica Alpha
Barcelona, Spain
aleksandar.matic@telefonica.com

Lars Vedel Kessing

Psychiatric Center Copenhagen
Rigshospitalet, Copenhagen, Denmark
lars.vedel.kessing@regionh.dk

Jonas Busk

Department of Applied Mathematics and
Computer Science
Technical University of Denmark
jbusk@dtu.dk

Maria Faurholt-Jepsen

Psychiatric Center Copenhagen
Rigshospitalet, Copenhagen, Denmark
maria.faurholtjepsen@regionh.dk

Jakob E. Bardram

Copenhagen Center for Health Technology
Technical University of Denmark
jakba@dtu.dk

ABSTRACT

A number of studies have been investigating the use of mobile phone sensing to predict mood in unipolar (depression) and bipolar disorder. However, most of these studies included a small number of people making it difficult to understand the feasibility of this method in practice. This paper reports on mood prediction from a large ($N = 129$) sample of bipolar disorder patients. We achieved prediction

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

© 2018 Association for Computing Machinery.

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2018 International Symposium on Wearable Computers (UbiComp/ISWC'18 Adjunct), October 8–12, 2018, Singapore, Singapore*, <https://doi.org/10.1145/3267305.3267536>.

accuracies of 89% and 58% in personalized and generic models respectively and we shed light on the "cold-start" problem in practice. Moreover, we showed that the accuracy depends on the labeling strategy of euthymic states. The paper discusses the results, difference between personalized and generic models, and the use of mobile phones in mental health treatment in practice.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *User models; Empirical studies in ubiquitous and mobile computing*;

KEYWORDS

Mobile Sensing; Bipolar Disorder; Depression; Personalized and Generic Models

ACM Reference Format:

Marios Constantinides, Jonas Busk, Aleksandar Matic, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2018. Personalized versus Generic Mood Prediction Models in Bipolar Disorder. In *Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2018 International Symposium on Wearable Computers (UbiComp/ISWC'18 Adjunct)*, October 8–12, 2018, Singapore, Singapore. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3267305.3267536>

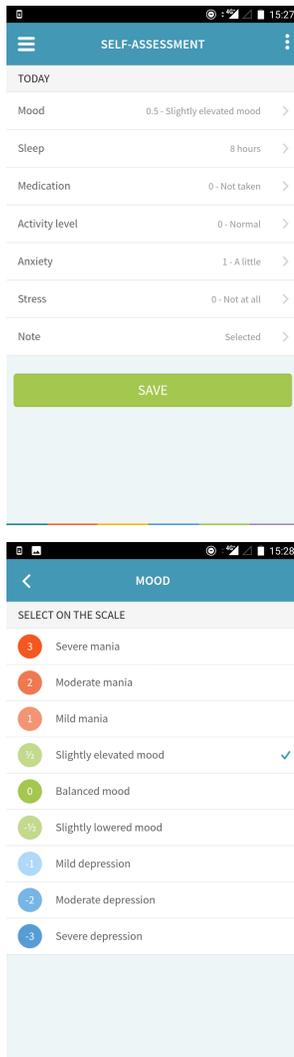
ACM COPYRIGHTS & PERMISSION

Accepted extended abstracts and papers will be distributed in the Conference Publications. They will also be placed in the ACM Digital Library, where they will remain accessible to thousands of researchers and practitioners worldwide. To view the ACM's copyright and permissions policy, see: http://www.acm.org/publications/policies/copyright_policy

INTRODUCTION

Mental health problems account for one-fifth of the disease burden worldwide, representing the third most common reason to visit a health center [10, 14]. The majority of relapse in recurrent episodes require frequent hospitalization which negatively impacts patients' cognitive and psychosocial state, and in turn their general quality of life. Additionally, each relapse involves considerable direct and indirect costs to the healthcare systems. In particular, Bipolar Disorder (BD) is associated with high morbidity and disability, along with significant costs associated with BD patients' treatment due to hospitalization [2]. In this regard, mental health research has recently focused on finding new methods for early detection of mood swings to implement prompt interventions that prevent mental crises and hospitalization.

The last decade witnessed a proliferation of mobile sensing systems designed for monitoring and quantifying human behaviors, particularly in mental health. The extremely high penetration



Sidebar 1: User screens from the Mon-senso smartphone app used for data collection in the MONARCA II trial.

of smartphones and their sensing capability to unobtrusively monitor a range of daily activities, in combination with advancements of Machine Learning (ML) techniques showed a great promise to automatically recognize behavioral patterns associated with mood swings and to detect the right time for intervention. A number of studies leveraged phone sensors to recognize mood in BD patients [3, 6, 7, 9, 13]. However, the automatic monitoring of BD patients is still in its infancy; the evidence is limited to studies with very few participants and controlled experimental settings. To further validate the approach of automatic monitoring of BD patients, deepen the understanding of its potential and limitations and ultimately pave the way towards its practical implementation, it is essential to bring evidence from studies with more participants and different trials.

In this paper, we report from a Randomized Clinical Trial (RCT) of 129 patients in which we collected self-reports on daily mood states and smartphone sensor data. We developed models that rely on sensor data to detect daily emotional states, defined as euthymia, depression and mania. The models were implemented in two ways: (a) generic - without relying on the historical data about the test users, and (b) personalized - including the historical data about the test user. We found that the personalized models outperformed the generic and baseline models while the generic models performed at baseline level. The main contributions of this paper are the following: (i) we evaluated the accuracy of inferring daily emotional states and we discuss the categories of the most predictive features, (ii) we compared the accuracy of models without prior knowledge about the user (referred to as generic models) to models that include historical data about the user (referred to as personalized models) and we shed light on the "cold-start" problem i.e. on the need of acquiring prior data about the new user, and (iii) we analyzed the model accuracy with respect to different daily emotion labeling strategies.

RELATED WORK

A pilot study of the MONARCA system indicated that objective measures of physical and social activity collected via smartphones from 17 patients with BD for 3 months correlated with clinical ratings of depression and mania [5]. To discriminate between affective states, previous work on MONARCA I showed correlations between several objective measures with features extracted from call and SMS logs, and clinical ratings, in a study with 61 patients monitored for a period of 6 months [7]. The same work showed significant correlations between self-reported mood and clinical ratings for depression and mania, suggesting self-reported mood is a valid indicator of symptoms of BD.

Several other studies have investigated the use of smartphone data for predicting daily mood. Breda et al. [13] explored ML techniques for predicting mood in the context of depression from data collected via smartphones from 27 participants. They found that regression models based on sensor data outperformed models based on previous mood alone. Canzian et al. [3] found significant correlations between depressive mood and mobility patterns extracted from smartphone location traces from 28 users, and used personalized Support Vector Machine (SVM) models to successfully

classify mood scores. Chin et al. [4] developed a system capable of predicting negative emotions based on mobile phone usage patterns from 28 participants using a combination of personalized models and timeslot features. Abdullah et al. [1] showed that automated sensing data from 7 BD patients' smartphones can be used to infer Social Rhythm Metric (SRM) entries, a clinically-validated marker of stability and rhythmicity for individuals with BD, and that performance improved with personalized models.

BACKGROUND

¹https://cordis.europa.eu/project/rcn/93747_en.html

The EU funded project MONARCA I ¹, a pioneering project in the use of mobile technologies in mental health treatment, provided the proof of concept mobile sensing framework for behavioral analysis and crisis prediction while also enabling a direct communication channel between patients and clinicians. The follow up RCT suggested that smartphone-based self-monitoring results in more sustained depressive symptoms than patients using a smartphone for normal communicative purposes, but on the other hand the patients exhibited fewer manic symptoms during the trial. Frequently reporting symptoms on the smartphone may raise the awareness of recent history of depressive moods which may further drive depressive states. For this reason, the goal of the MONARCA II, the study presented in this paper, was to investigate the automatic detection of BD symptoms through the objective collected data on phone usage, social activity, physical activity, and mobility. In addition to collecting sensor data and logs, the mobile system used in this study (called Monsenso) also integrates subjective and objective measures of illness.

The MONARCA II trial uses a randomized controlled single-blind parallel-group design. Patients with BD, according to ICD-10, who have been previously treated at the Copenhagen Clinic for Affective Disorder in Denmark are included and randomized to either daily use of the Monsenso system including a feedback loop between patients and clinicians (the intervention group) or to the use of a smartphone for normal communicative purposes (the control group) for a 9-month trial period. The trial started in September 2014 and completed in January 2018. The outcomes are: differences in depressive and manic symptoms; rate of depressive and manic episodes (primary); automatically generated objective data on measures of illness activity; number of days hospitalized; psycho-social functioning (secondary); perceived stress; quality of life; self-rated depressive symptoms; self-rated manic symptoms; recovery; empowerment and adherence to medication (tertiary) between the intervention group and the control group during the trial. The study received ethical permission, and the trial is registered at ClinicalTrials.gov ².

²ClinicalTrials.gov NCT02221336. Registered 26th of September 2014.

Our study builds upon the previous work of MONARCA I, and sets to deepen the understanding of the value of smartphone-based systems in treating BD. If the system turns out to be effective in reducing depressive and/or manic symptoms and the rate of episodes, it will be used as the foundation building the treatment tool for BD applied on a large scale.

		Predicted	
		Euthymia	Depression/Mania
Actual	Euthymia	77.2%	35.3%
	Depression/Mania	22.8%	64.7%

Sidebar 2: Personalized model: Euthymia vs. Depression/Mania states (A).

		Predicted	
		Euthymia	Depression/Mania
Actual	Euthymia	63.7%	62.1%
	Depression/Mania	36.3%	37.9%

Sidebar 3: Generic model: Euthymia vs. Depression/Mania states (A).

		Predicted	
		Euthymia	Depression/Mania
Actual	Euthymia	89.6%	51.3%
	Depression/Mania	10.4%	48.7%

Sidebar 4: Personalized model: Euthymia vs. Depression/Mania states (B).

		Predicted	
		Euthymia	Depression/Mania
Actual	Euthymia	86.69%	85.3%
	Depression/Mania	11.8%	14.7%

Sidebar 5: Generic model: Euthymia vs. Depression/Mania states (B).

DATA COLLECTION AND PRE-PROCESSING

Sensor data collection

The sensor data collection included battery, light, location, screen, proximity, communication logs and step count samples. For each sensor data, we estimated average levels, median, SD, minimum and maximum, entropy, first and last observation in each day. It is important to mention that some sensors data was used to make inferences for other behavioral features, as explained in the next section. For example, we used the light sensor to make inferences for the sleep behavior.

Data Imputation

The data collection resulted in data gaps, however, not to a large and unexpected extent. To mitigate the drawbacks of missing values, for each data category we discarded rows with more than 20% of missing values (i.e. NaN). For the remaining rows that included missing values, we used a common imputation strategy of inserting mean values for the datasets of sensors data, call logs and locations.

Target variable

The Mosenso system collected participants' daily mood using a built-in questionnaire (Sidebar 1) in which participants reported their mood at the end of each day on a scale of -3 (depression) to 3 (mania) [8]. In this study [8], the patients expressed the need to have a score describing mild depressive or manic states, therefore the authors added ± 0.5 rating. We investigated classification analysis with the following four labeling strategies, in the rest of the paper denoted to as A, B, C, D for a better readability. Euthymia vs. Depression/Mania states: (A) discriminate the score of [0] from the other scores [-3, -2, -1, -0.5, 0.5, 1, 2, 3], (B) discriminate the score of [-0.5, 0, 0.5] from the other scores [-3, -2, -1, 1, 2, 3]. Euthymia vs. Depressive states: (C) discriminate the score of [0] from the other scores [-3, -2, -1, -0.5], (D) discriminate the score of [-0.5, 0] from the other scores [-3, -2, -1].

DATA ANALYSIS

Feature Extraction

The feature extraction process was thoroughly planned to extract behavioral indices that reflect patients' sleep behavior patterns, social activities, physical activity and mobility. These dimensions found in previous studies to be associated with dimensions of mental health state [3, 4, 13].

We aggregated the behavioral features in the three levels, namely daily (features computed over a 24-hour period), weekly (aggregated features in a "moving window" of one week of data), intra-daily (partitioning days into four time-periods - morning (5am-11am), afternoon (12pm-4pm), evening (5pm-9pm), and night (10pm-4am)). The intra-daily time-periods were selected based on typical daily activities in Denmark. The behavioral features extraction yielded a set of 370 features, of descriptive

		Predicted	
		Euthymia	Depression
Actual	Euthymia	79.9%	33.1%
	Depression	20.1%	66.9%

Sidebar 6: Personalized model: Euthymia vs. Depressive states (C).

		Predicted	
		Euthymia	Depression
Actual	Euthymia	70.9%	67.5%
	Depression	29.1%	32.5%

Sidebar 7: Generic model: Euthymia vs. Depressive states (C).

		Predicted	
		Euthymia	Depression
Actual	Euthymia	90.7%	50.8%
	Depression	9.3%	49.2%

Sidebar 8: Personalized model: Euthymia vs. Depressive states (D).

		Predicted	
		Euthymia	Depression
Actual	Euthymia	89.5%	89.5%
	Depression	10.5%	10.5%

Sidebar 9: Generic model: Euthymia vs. Depressive states (D).

statistics of the collected data e.g. average, median, minimum, maximum, and entropy. Physical activities were represented with a set of features extracted by applying step-counter algorithm locally on the phone, and extracting descriptive statistics in the post-analysis. Social activities were mainly modelled through the call and SMS logs i.e. Call Detail Records (CDRs). The features reflected the patterns of incoming and outgoing calls, communications in different time frames (such as night vs. day, weekend vs. working days, etc.) and also characteristics of the contact networks (such as strengths of social ties). Though indirectly, social activities are also reflected through mobility patterns, to which we dedicated a special focus. The mobility patterns features included, for instance, unique places visited, number of stationary unique locations (that have never been visited before), time spent at the location identified as home, maximum distance between the home location and any other stationary location visited. The sleep related features are based on inferences made from the light sensor data (for inferring time of going to bed) and any detected interaction with the phone (for both inferring bed and wake-up time). Similarly to [11], we estimated the time-to go to bed using the light sensor data and any logs of movement/interaction with the phone, whereas the latter information was used to estimate either the interrupted sleep or the wake-up time.

Detecting Depressive and Manic States

We approached the goal of automatic detection of depressive and manic states (at a daily level) through a classification problem. We mapped the extracted features to the classes defined from the daily self-reported mood scores reported by the patients and tested several supervised learning classifiers. Similarly to recent studies [12] the XGBoost classifier provided the highest accuracy. We implemented two types of models, namely *personalized* and *generic*. For testing the generic model, the cross-validation was performed by leaving one user out (i.e. leaving out all the instances that belong to the test user). In this way, a generic model correspond to a model that in practice would not require any prior information (i.e. training samples) about a user in order to predict her/his depressive, manic, or euthymia days. On the other hand, in order to evaluate the personalized model we applied cross-validation by leaving one daily instance out, i.e., in the training set we were keeping all the available points from the other users as well as the test user except a test (daily) point for the test user. In practice, developing such model (referred to as personalized) would require some prior knowledge about the patient, e.g. requiring a user to collect data for several days and labeling emotional state before the model can be developed.

RESULTS

We evaluated the accuracy in inferring daily mood status, i.e. mental health state of patients, through a classification task that discriminates euthymic states from a) depressive and manic states, b) depressive states. Table 1 presents model performances along with the corresponding baselines.

	<i>Our Model</i>	<i>Accuracy</i>	<i>kappa</i>	<i>Baseline Model</i>
Euthymia vs.				
Depression/Mania states	Personalized (A)	73.15%	0.406	54.79%
	Personalized (B)	87.75%	0.215	84.38%
	Generic (A)	57.95%	0.013	57.37%
	Generic (B)	86.69%	0.008	86.57%
Euthymia vs.				
Depressive states	Personalized (C)	77.29%	0.404	61.88%
	Personalized (D)	89.12%	0.213	86.17%
	Generic (C)	66.91%	0.018	66.28%
	Generic (D)	88.49%	0	88.49%

Table 1: Classification results. The baseline values assume a random model considering the class distribution.

Detecting Euthymia vs. Depression/Mania states

Personalized vs. Generic Models. The personalized model that detects euthymia over depressive/manic states yielded an overall accuracy of 73.15% using labeling approach (A) (Sidebar 2), whereas expanding the euthymic class to include ± 0.5 (B) increased the accuracy to 87.75% while performing closer to baseline (Sidebar 4). The corresponding generic model yielded an overall accuracy of 57.95% using labeling approach (A) (Sidebar 3) and increased to 86.69% (Sidebar 5) when expanding the euthymic class to include ± 0.5 (B). However, none of the generic models perform above baseline level.

Detecting Euthymia vs. Depressive states

Personalized vs. Generic Models. Similar results were observed for the personalized model that detects depressive states. The model achieved 77.29% accuracy in labeling approach (C) (Sidebar 6) and increased to 89.12% with labeling approach (D) (Sidebar 8), while closer to baseline. Using labeling (C), the generic model performed with an overall accuracy of 66.91% (Sidebar 7), and increased to 88.49% when using labeling (D) (Sidebar 9). Both of the generic models performs at baseline level.

DISCUSSION

In this paper we presented the accuracy of classifying daily mood states. In our modelling approach, we investigated both personalized and generic models in conjunction with two labeling approaches of the ground truth information of daily mood reports that included or excluded ± 0.5 from the euthymic state label (± 0.5). The results suggest that the personalized models yielded a higher accuracy than

the generic models in each case. Such finding suggests that there is indeed a cold-start problem in building predictive models of daily mood in bipolar disorder patients, and that the models can benefit from requiring a user to use the sensing application for several days and report mood states in order for the model to be individually calibrated.

In addition to the personalized vs. generic models, our results corroborate the different criteria for grouping participants' mood responses. Aligned with prior work [8], our results indicate the importance of such design of self-reporting scale. The models that used the definition of euthymia including ± 0.5 achieved higher accuracy than models that considered only 0 as a neutral state, but performed closer to baseline. Our analysis focused also on understanding the categories of the most predictive features among the four categories including sensor, location, sleep and CDR groups, as previously explained. Interestingly, it turned out that the features that describe how the patients use their phone (such as screen time, proximity sensor, battery levels) are more predictive than the ones describing real-world behavior (such as mobility and communication patterns), unlike in the previous studies [9]. Over the last decade, the technologies were evolving which inevitably impacted their use (e.g. nowadays the users rely more on Internet supported instant messages and calls than on telco-network). This further underlines the need to recurrently conduct this and similar studies.

This study suffered from several limitations. We encountered the problem with missing values, which may be improved in the future by providing reminders to the users of the Monsenso system to keep the device sensors (e.g. location) active. In addition, in our analysis we were unable to evaluate the accuracy of predicting manic states due to their low prevalence. In the future work, we will place a particular focus on the reproducibility of the prediction models with respect to the most predictive features. In addition, we will investigate different ways of building personalized and generic models to pave the way towards the implementation of the prediction models in practice.

CONCLUSION

This paper reported the results from a Randomized Clinical Trial involving 129 Bipolar Disorder patients. Using smartphone sensor data coupled with self-reported daily mood we developed and evaluated predictive models tuned to discriminate a) euthymic from depressive and manic states, b) euthymic from depressive states. We applied two types of models, one that rely on prior knowledge about the patient (i.e. requiring a user to collect data for several days before calibrating the model) and one that does not require any inputs from the patient. We called these two models personalized and generic, respectively. We found out that personalized models outperformed the generic ones that performed close to a baseline. Additionally, we found that the labeling strategy, reflecting the way patients report their emotional state, affects model performance. The study underlines the need for repetitive clinical trials and further exploration of predictive modelling approaches to pave the way of, ultimately, their implementation in clinical practice.

REFERENCES

- [1] 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.
- [2] Joette Gdovin Bergeson, Iftekhar Kalsekar, Yonghua Jing, Min You, Robert A Forbes, and Tony Hebden. 2012. Medical care costs and hospitalization in patients with bipolar disorder treated with atypical antipsychotics. *American health & drug benefits* 5, 6 (2012), 379.
- [3] 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 (UbiComp '15)*. ACM, New York, NY, USA, 1293–1304. <https://doi.org/10.1145/2750858.2805845>
- [4] Galen Chin-Lun Hung, Pei-Ching Yang, Chia-Chi Chang, Jung-Hsien Chiang, and Ying-Yeh Chen. 2016. Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study. 5 (08 2016), e160.
- [5] Maria Faurholt-Jepsen, Mads Frost, Maj Vinberg, Ellen Margrethe Christensen, Jakob Eyvind Bardram, and Lars Vedel Kessing. 2014. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry research* 217 1-2 (2014), 124–7.
- [6] Maria Faurholt-Jepsen, Maj Vinberg, Mads Frost, Ellen Margrethe Christensen, Jakob Bardram, and Lars Vedel Kessing. 2014. Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones—the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. *BMC Psychiatry* 14, 1 (2014). <https://doi.org/10.1186/s12888-014-0309-5>.
- [7] Maria Faurholt-Jepsen, Maj Vinberg, Mads Frost, Ellen Margrethe Christensen, Jakob Eyvind Bardram, and Lars Vedel Kessing. 2015. Smartphone data as an electronic biomarker of illness activity in bipolar disorder. *Bipolar disorders* 17 7 (2015), 715–28.
- [8] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2013. Supporting Disease Insight Through Data Analysis: Refinements of the Monarca Self-assessment System. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 133–142. <https://doi.org/10.1145/2493432.2493507>
- [9] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2015. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics* 19, 1 (2015), 140–148.
- [10] Harvard School of Public Health. 2011. The Global Economic Burden of Non-communicable Diseases. http://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf
- [11] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I Hong. 2014. Toss'n'turn: smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 477–486.
- [12] Didrik Nielsen. 2016. Tree boosting with XGBoost. *NTNU Norwegian University of Science and Technology* (2016).
- [13] Ward van Breda, Johnno Pastor, Mark Hoogendoorn, Jeroen Ruwaard, Joost Asselbergs, and Heleen Riper. 2016. Exploring and Comparing Machine Learning Approaches for Predicting Mood Over Time. In *Innovation in Medicine and Healthcare 2016*, Yen-Wei Chen, Satoshi Tanaka, Robert J. Howlett, and Lakhmi C. Jain (Eds.). Springer International Publishing, Cham, 37–47.
- [14] World Health Organization. 2013. Mental health action plan 2013 - 2020. http://www.who.int/mental_health/publications/action_plan/en/