Pers Ubiquit Comput
DOI 10.1007/s00779-014-0826-8

ORIGINAL ARTICLE

Impact factor analysis: combining prediction with parameter ranking to reveal the impact of behavior on health outcome

Afsaneh Doryab · Mads Frost · Maria Faurholt-Jepsen · Lars V. Kessing · Jakob E. Bardram

Received: 17 March 2014 / Accepted: 31 July 2014

Abstract An increasing number of healthcare systems allow people to monitor behavior and provide feedback on health and wellness. Most applications, however, only offer feedback on behavior in form of visualization and data summaries. This paper presents a different approach—called *impact factor analysis*—in which machine learning techniques are used to infer the progression of a primary health parameter and then apply parameter ranking to investigate which behavioral data have the highest 'impact' on health. We have applied this approach to improve the MONARCA personal health application for patients suffering from bipolar disorder. In the MONARCA system, patients report their daily mood score and by analyzing self-reported and automatically sensed behavioral data with this mood score, the system is able to identify the impact of different behavior on the patient's mood. We report from a study involving ten bipolar patients, in which we were able to estimate mood values with an average mean absolute error of 0.5. This was used to rank the behavior parameters whose variations indicate changes in the mental state. The rankings acquired from our algorithms correspond to the patients' rankings, identifying physical activity and sleep as the highest impact parameters. These results revealed the feasibility of identifying behavioral impact factors. This data analysis motivated us to design an impact factor inference engine as part of the MONARCA system. To our knowledge, this is a novel approach in monitoring and control of mental illness, and we argue that the impact factor analysis can be useful in the design of other health and wellness systems.

Keywords Health and behavior · Machine learning · Mental health · Bipolar disorder

1 Introduction

The management of mental health and well-being through monitoring systems is a promising and rapidly growing area in pervasive healthcare. Self-monitoring is a central part of treatment of mental disorders, due to its reactive effects on those behaviors being monitored [20]. Within clinical assessment, self-monitoring procedures are popularized by behavior therapists, particularly within behavioral self-control procedures.

Bipolar disorder is a mental disorder where self-monitoring plays a vital role in controlling the disease. Bipolar disorder is characterized by recurring episodes of both depression and mania, with treatment aiming to reduce symptoms and prevent recurrence throughout a patient's life time. An important goal in treatment of bipolar

A. Doryab (✉)
Carnegie Mellon University, 5000 Forbes Ave,
Pittsburgh, PA 15213, USA
e-mail: adoryab@cs.cmu.edu

M. Frost · J. E. Bardram
Pervasive Interaction Technology Laboratory (PIT Lab),
IT University of Copenhagen, Rued Langgaards Vej 7,
2300 Copenhagen, Denmark
e-mail: madsf@itu.dk

J. E. Bardram
e-mail: bardram@itu.dk

M. Faurholt-Jepsen · L. V. Kessing
Psychiatric Center Copenhagen, Department O, 6233,
University Hospital of Copenhagen, Blegdamsvej 9,
2100 Copenhagen, Denmark
e-mail: maria.faurholtjepsen@regionh.dk

L. V. Kessing
e-mail: lars.vedel.kessing@regionh.dk

Published online: 21 September 2014

disorder is to predict and prevent episodes of mania or depression. This is done by training patients to recognize their own early warning signs, i.e., indicators that they are headed toward an episode [3]. The training is resource-intensive, and its success varies highly from patient to patient. Therefore, a high degree of self-awareness is important in early detection of signs and behavioral changes.

Most existing applications, however, mainly focus on monitoring of behavioral traits and only offer feedback on behavior in form of visualization and historical data summaries. Many smartphone applications take advantage of persuasive visualizations and features that can help with adjustment of behaviors to improve adherence and consistency (e.g., [5, 7, 14, 15, 17]). In clinical care, the patients or therapists decide on the appropriate treatment by observing the assessed historical data. As such, existing applications lack the ability to identify warning signs and predict future episodes. Even the patients or therapists might not be able to identify the warning signs or behavioral factors that have most impact on the emotional state of the patient.

We present an approach for identifying warning signs or behavioral factors that have the most impact on the health outcome of mentally ill patients. We call this approach *impact factor analysis* since the aim is to identify the behavioral factors that has the highest impact on the health of a patient. Impact factor analysis applies machine learning techniques to infer and forecast the progression of a primary health parameter and then apply parameter ranking algorithms to investigate which behavioral data have the highest 'impact' on health.

The impact factor analysis method has emerged from our machine learning approach reported in this paper. We performed an exploratory analysis of the data collected from MONARCA smartphone application (Fig. 1), which is a personal monitoring system designed for the treatment of bipolar disorder patients [1, 2, 9]. The insights gained from the analysis reported in this paper led to the design of an impact factor inference engine, which was implemented in version 2.0 of the MONARCA smartphone application [9]. It is important to notice that this analysis was done before the design of MONARCA 2.0 system. The contribution of this paper is to demonstrate the feasibility of data-driven methods in design for health and wellness. We show how our machine learning analysis provided us insights into the possibility of inferring the mental state of patients from their smartphone data and providing the overview of behavioral factors related to their condition. The analysis will be inspiring and useful for other researchers interested in behavior tracking especially in health and wellness domain.

In this paper, we report on (1) automatically predicting the mental state in bipolar patients from their past data and (2) identifying the impact of different behavioral parameters (e.g., sleep and activity) on the patients' mood. We apply a smartphone as the platform as it offers a set of built-in sensing channels that can be used for collecting behavioral data without any efforts required from the patients. For example, sensors such as accelerometer, GPS, light, and microphone can be used to track the patient's activities. The occurrence of opposite extremes of behavior in bipolar patients is likely to be apparent in physical and social activities, which becomes possible to monitor and infer automatically.

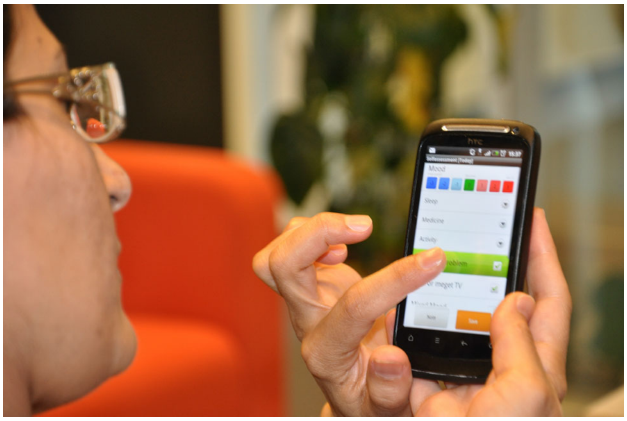

Fig. 1 The MONARCA mobile application

2 Related work

A number of different commercial and research projects have applied smartphone technology for health and well-being monitoring and feedback. In the Apple App Store, there are now more than 5,000 health monitoring apps available, and lately, Apple announced the Apple Health app that provides an overview of health and fitness by gathering data from different sources and visualize it in a personal dashboard [13]. The use of such technology to collect and reflect on their personal information has been described as being part of a new research fields of personal informatics [16, 19] or quantified-self [6, 12]. A number research projects have been investigating approaches and technologies for such this kind of personal informatics. *Health mashups* [4] gives self-tracking users a continuous feed of information based on an aggregation of data from various fitness devices, personal diaries, and context logging. The goal is to design a mobile system that helps people understand how context impacts their well-being over time and to encourage them to dig deeper into how various aspects affect each other. The work in [8] extends prior work on analyzing and summarizing self-tracking data, with the goal of helping self-trackers identify more meaningful and actionable findings. They develop a set of cuts over location and physical activity data and visualize those cuts using a variety of presentations.

A category of apps including *UbiFitGarden* [7] and *Bewell* [15] collect behavioral data, such as physical activity from phone sensors, and provide visual feedback such as an ambient display to promote healthy behavior. Other personal health systems such as *Health Buddy* [14], *Mobile Mood Diary* [17], and *Mobilyze!* [5] also use visualization graphs to help patient monitor and control their mental disease. *MyCompass* [11] is used for monitoring and managing stress and anxiety. *EmotionSense* [18] seeks to sense individual emotions as well as activities, verbal and proximity interactions among members of social groups.

Health community websites and online tools are other ways for people with similar disease and illness to share experience and seek support and help. Websites such as *curetogether.com*, *patientslikeme.com*, and *mentalhelp.net* let users log an enormous range of conditions, symptoms, and feelings.

The vast majority of this substantial amount of commercial and research-based solutions are, however, primarily focused on the collection and visualization of personal data, and it can be difficult for users to understand and interpret this data [6]. This again hinders users in understanding and getting an insight into which behavioral factors in their life actually influence their health and well-being. The objective with the impact factor analysis method is exactly to provide users with this kind of insight and thereby help them interpret and understand the many different data being collected. Although we present the impact factor analysis with one case in bipolar disorder, we believe that this approach can be generalized to help making sense of behavioral data in other personal informatics health applications designed to provide users with an insights into long-term health and well-being.

3 Exploratory analysis of self-assessment data

A central issue regarding mental illness is that many patients are unable of recognizing early warning signs in their disease, i.e., symptoms that indicate an oncoming episode. Designing for this group of patients poses challenges as it is unclear what behavior parameter should be monitored. It is also difficult for patients to reflect on their own mood and behavior, and their families and others around them may only recognize symptoms if they understand the illness and know what to look for. In our research, we were motivated to find out how the power of machine learning and data mining can help people suffering from bipolar disorder with an insight into the unfolding of the disease. Basically, we were interested in answering the following questions:

1. How closely can we estimate and forecast the state of bipolar patients from their past data using machine learning? Can we build a general model to fit all bipolar patients or should a model be built for each individual patient?
2. Can we identify the behavior parameters that reveal changes in the mental state of the patients? Are the impact parameters common among all patients or different from patient to patient?

The answers to these questions can provide valuable intervention insights for clinicians, patients, and researchers. The accurate prediction of the mental state can result in reducing—or possibly even preventing—extreme mania and depression episodes by faster interventions. It can also provide patients with insight on the temporal unfolding of their disease. Identification of the important behavior parameters can help clinicians and patients identify the warning signs and gain insight into how the patients behavior impacts their mental state, both on a past and current basis. For example, decrease in the sleep hours during the past week can be a sign of an upcoming episode. The findings can also help the research team improve and extend the design of the MONARCA system by improving the data sampling strategies and better data interpretation to, e.g., automatically infer the mood, sleep quality, and activity instead of asking the patients for self-reports.

3.1 Data collection

The data analysis in this study is based on self-reported data collected from the MONARCA application, including the following items:

- Mood—highly depressed (−3) to highly manic (3)
- Sleep—amount of sleep, reported in half hour intervals
- Activity—highly inactive (−3) to highly active (3)
- Medicine taken—yes/no
- Medicine changed—yes/no
- Mixed mood—yes/no
- Cognitive problems—yes/no
- Irritable—yes/no
- Warning signs—number of personalized active warning signs
- Alcohol—number of alcoholic drinks
- Stress—no stress (0) to highly stressed (5)

The data were collected from ten bipolar patients using the system between May 2011 and March 2012 in the Affective Disorder Clinic at the University Hospital of Copenhagen, Denmark. The use of the system was approved by the Regional Ethics Committee in The Capital Region of Denmark (H-2-2011-056) and The Danish Data Protection Agency (2013-41-1710). The participants were a diverse set of males and females in the age range of 20–51 who were considered stable with an initial HAMD mood score below 14. HAMD is the Hamilton rating scale for depression, which is widely used by healthcare professionals [10]. A total of 1193 self-reports was collected, with an average of 119 days per patient, which gave us a big enough dataset for the analysis.

3.2 Mood estimation

To answer the first question, we formulated the problem of detecting the current emotional state as a machine learning problem where the value of the mood variable is estimated based on the model built from the training data. The outcome of the model can be evaluated as:

$$Actual\ Value = Predicted\ Value + Residual$$

where the residual is the error between the actual and predicted value. To measure the success of our predictors, we evaluated the mean absolute error (MAE), which is calculated as:

$$MAE = \sum_{i=1}^{N} \frac{|R_i|}{N}$$

where R_i is the residual at point i and N is the number of data points that are being predicted. Since we are interested in the closeness of the estimated and the actual values, the absolute difference is more suitable than the squared error.

The classification test on our data with several methods resulted in high misclassifications making it infeasible to apply to our dataset. Therefore, we treated the daily mood scores as numeric and applied regression-based learners implemented in Weka to our dataset. The Weka API (also supported in Android) seemed a proper choice as we could later customize and implement the same set of learners into a mood inference engine to be used by the MONARCA system. We tested different methods to obtain an understanding of which learning algorithms perform best and give us closest estimations. The best performing learners were Linear Regression, Additive Regression, SVM, and Model Trees.

An important aspect of the human behavior modeling is to identify the generalizability of the proposed model. In our case, however, each patient might have a different behavior pattern, and therefore, the models built from a patient's data can more closely estimate the mood of that particular person than a unified model built for all patients. To find the most suitable approach, we tried unified vs. individualized models as described in the next sections.

3.2.1 *Individualized learning model*

We created a set of models for each patient by performing learners on each patient's data individually. We used tenfold crossvalidation, which trains the model on all but one-tenth of the samples and validates the model on the remaining samples. Table 1 shows the distribution of mean absolute error across all patients obtained from applying different learning algorithms.

In individualized learning models, we observed an average mean absolute error (MAE) of 0.5 across all patients, with a standard deviation of 0.22, a minimum MAE between 0.15 and 0.79. The standard deviation revealed that the mood estimation models work better for some patients than others, but the MAE rates are low enough to suggest that the mood value can be closely inferred from the self-assessments; it takes an error of about 0.5 to move from the center of one mood label halfway toward another label. Figure 2 shows the mood estimation for the patient ID 59 who had the most swings among our participants (Mean mood 0.56 and std 1.15). The diagram compares the actual and predicted scores. The estimated values are bigger than one standard deviation for 24.3 % of the instances and 1.87 % with an error more than two standard deviation (error bigger than 2.3). It means that in the worse case, our predictor estimated the depressed or manic state as normal or mild manic-depressed, but never reported an extreme depressed state as extreme manic and vice versa. The closeness of estimation also depended on the amount of training data and the variation of mood scores. We expect to get even closer scores to the reported mood with more data.

3.2.2 *Unified learning model*

Another approach in mood estimation was to form a unified model built from an aggregate of all of the patients' data. If successful, this model could be used as an initial model for a new patient.

To test the feasibility of a unified model, we performed a leave-one-patient-out crossvalidation where at each iteration the model was trained on nine patients and tested on the tenth one. Basically, we removed a patient's data from our dataset and performed the same set of learning methods on the remaining data to create a model. We then applied each model to the patient's data and computed the mean absolute error of the estimated mood. After training, we found that the unified model performed surprisingly well for some patients, with a minimum MAE between 0.28 and 1.64. However, for some participants, the MAE was quite high, as our dataset had an average MAE of 0.71, with standard deviation of 0.39. The maximum

Table 1 The results of applying different machine learning methods to estimate the mood of patients from their self-assessed data

Patient ID	RepTree	M5P	M5Rules	AdditiveReg	SMOReg	LinearReg	MAE min	MAE avg
48	0.69	0.78	0.79	0.79	0.76	0.8	0.69	0.77
49	0.18	0.18	0.18	0.2	0.15	0.2	0.15	0.18
57	0.28	0.27	0.28	0.31	0.21	0.29	0.21	0.27
58	0.54	0.52	0.3	0.57	0.46	0.52	0.3	0.49
59	0.84	0.79	0.79	0.84	0.85	0.81	0.79	0.82
61	0.44	0.4	0.4	0.42	0.29	0.43	0.29	0.40
64	0.81	0.67	0.67	0.69	0.79	0.68	0.67	0.72
66	0.52	0.4	0.43	0.4	0.42	0.42	0.4	0.43
67	0.62	0.54	0.54	0.69	0.58	0.56	0.54	0.59
70	0.32	0.3	0.3	0.33	0.26	0.31	0.26	0.30
MAE avg	0.52	0.49	0.47	0.52	0.48	0.50	0.43	0.50
MAE std	0.22	0.21	0.22	0.22	0.26	0.21	0.23	0.22

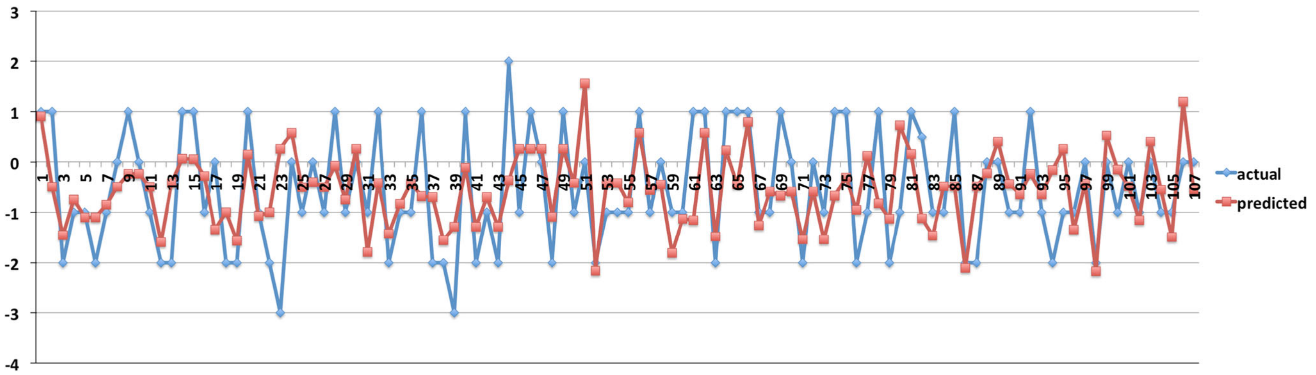

Fig. 2 Mood prediction results for a patient who has the most swings in the mood scores. The diagram compares the actual and predicted scores. Only 24.3 % of the instances have predicted values bigger than one standard deviation and in two occurrences, the error is bigger than two std

Table 2 The results of applying different machine learning methods to estimate the mood of patients from the unified model

Patient ID	RepTree	M5P	M5Rules	AdditiveReg	SMOReg	LinearReg	MAE min	MAE avg
48	1.67	1.70	1.66	1.66	1.73	1.64	1.64	1.68
49	0.32	0.47	0.45	0.45	0.34	0.37	0.32	0.40
57	0.63	0.57	0.62	0.63	0.51	0.42	0.42	0.56
58	0.52	0.50	0.76	0.76	0.67	0.47	0.47	0.62
59	0.86	0.79	0.88	0.90	0.93	0.89	0.79	0.88
61	0.40	0.41	0.39	0.39	0.45	0.28	0.28	0.39
64	0.81	0.78	0.80	0.84	0.81	0.92	0.78	0.83
66	0.50	0.65	0.59	0.61	0.66	0.42	0.42	0.57
67	0.82	0.92	0.76	0.76	0.98	0.81	0.76	0.84
70	0.36	0.35	0.36	0.37	0.46	0.32	0.32	0.37
MAE avg	0.69	0.71	0.73	0.74	0.76	0.65	0.62	0.71
MAE std	0.40	0.39	0.37	0.37	0.40	0.42	0.41	0.39

The method was leave-one-patient-out crossvalidation where at each iteration the model was trained on nine patients and tested on the tenth one

average MAE was 1.68 (obtained from the patient ID 48), which is bigger than 4 standard deviations from the average MAE. Table 2 summarizes the distribution of estimates across all patients.

3.3 Parameter ranking

The second question in our analysis was finding behavioral parameters that highly relate to the mood, thus revealing

Table 3 The results of applying chi-square correlation evaluator to rank the parameters

Chi-squared method	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Activity	4	1	1	2	1	0	0	0	0	0
Mixed mood	1	2	0	0	1	1	0	3	1	0
Irritable	1	1	0	0	2	2	1	1	0	1
Stress	1	2	0	1	1	0	1	1	1	1
Alcohol	1	0	1	0	0	2	1	2	1	1
Active warning signs	1	0	3	2	0	0	0	1	0	2
Sleep	0	3	2	1	2	0	0	0	1	0
Medicine taken	0	0	0	1	0	2	2	0	0	4
Medicine changed	0	0	0	1	1	0	2	0	2	3
Unable to concentrate	0	0	2	0	1	1	0	1	2	2

Activity is ranked as the number one impact factor for four participants, and sleep is ranked as the number two impact factor for three. In general, activity, sleep, active warning signs, stress, and mixed mood are among the five highest ranked parameters

the change in the state of the patient. This information can help both clinicians and patients to keep track of which parameters frequently affect the mood or are likewise affected by the mood and hence should be observed, controlled, or even changed.

To find the impact of different parameters on the mood values, we apply three different attribute evaluation techniques to rank the parameters based on correlations, information gain, and their significance. More specifically, our methods include:

- Correlation-based evaluation—to measure the worth of a parameter by computing the value of the chi-square value with respect to the class.
- Information gain evaluation—to compute the worth of a parameter by measuring the information gain with respect to the class.

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)$$

- Significance evaluation—to rate the worth of an attribute by computing the probabilistic significance as a two-way function (attribute–classes association and classes–attribute association).

We apply these methods on each individual patient's data and report the rankings wrt. the mood parameter as the class. As shown in Tables 3, 4, 5, and 7, activity, sleep, stress, and mixed mood are among the five highest ranked parameters resulted from all three methods. For example, activity is ranked as the number one impact parameter for four out of nine participants, and sleep is ranked as the number two for three out of nine. To find out how much these rankings agree with the participants general observation of themselves, we ask them to rank the parameters in the order they perceive the both-sided impact of the parameters and their mood state. Nine out of ten participants did the rankings. We compared their rankings with the output of our three methods and observed that the participants list of five top ranked parameters highly agree with the lists resulted from our ranking methods (see Tables 6, 7). The only difference is in the significance evaluation where alcohol is ranked higher than active warning signs. These observations encourage us to take a step toward incorporating parameter ranking algorithms in our system to automatically and continuously infer the behavior factors highly related to the mental state of bipolar patients.

4 Design implications

Our analysis gave us insights into new possibilities offered by machine learning to improve monitoring, treatment, and control of bipolar disorder. The following sections present the design implications resulted from the data analysis presented in previous sections.

4.1 Mental state inference

Overall, our exploratory analysis suggested that it is possible to automatically infer the emotional state of patients. Having the insight of patient's status by clinicians or patient's relatives, can prevent extreme manic and depressive episodes. To estimate the daily mood of patients, we used multiple approaches. Individualized models reported closer estimations to the actual reported mood for most patients, while the unified all-patient models performed slightly worse. However, the all-patient model can be used to estimate the mood scores until it collects enough training data from a new patient.

4.2 Impact factors identification

From the high agreement between the results obtained from parameter ranking methods and self-ratings, it seems feasible to give patients an overview of the behavior parameters that reveal changes in their mood state. As mentioned before, most patients have difficulties recognizing their

Table 4 Information gain method ranks activity as the highest ranked parameter followed by sleep, active warning signs, stress and mixed mood

Information gain method	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Activity	4	2	0	3	0	0	0	0	0	0
Mixed mood	2	0	2	0	0	2	1	1	0	1
Sleep	1	3	1	2	1	0	0	1	0	0
Irritable	1	1	0	0	2	0	3	1	0	1
Stress	1	2	0	1	1	1	0	1	1	1
Medicine taken	0	0	0	1	0	0	4	0	0	4
Medicine changed	0	0	0	0	2	0	0	4	0	3
Unable to concentrate	0	0	2	1	0	1	0	1	3	1
Alcohol	0	0	2	0	0	4	0	0	1	2
Active warning signs	0	1	2	1	2	0	0	0	1	2

Table 5 With the significance evaluator, the activity is the highest ranked parameter followed by sleep, mixed mood, stress, and alcohol consumption

Significance method	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Activity	3	1	0	3	1	1	0	0	0	0
Medicine taken	1	0	0	0	0	2	1	1	0	4
Medicine changed	1	0	1	0	1	0	1	1	0	4
Mixed mood	1	1	2	1	0	0	0	3	1	0
Irritable	1	1	0	2	0	1	2	0	0	2
Alcohol	1	2	0	1	1	0	3	0	0	1
Active warning signs	1	1	1	0	1	0	1	1	1	2
Sleep	0	1	2	0	3	2	0	0	1	0
Unable to concentrate	0	1	1	0	0	1	0	1	3	2
Stress	0	1	2	0	2	1	1	0	1	1

Table 6 Participants in the study ranked the parameters in the self-reports

Patient self-rates	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Activity	4	0	1	1	1	1	1	0	0	0
Sleep	2	3	1	2	0	0	0	1	0	0
Mixed mood	1	1	1	1	0	1	1	0	2	1
Active warning signs	1	0	0	0	4	2	0	1	0	1
Alcohol	0	1	1	0	0	0	2	1	0	4
Irritability	0	0	1	2	0	1	0	0	1	4
Unable to concentrate	0	1	1	0	1	0	2	4	0	0
Medicine taken	0	1	1	0	0	2	1	0	2	2
Medicine changed	0	0	0	1	1	0	2	1	1	3
Stress	0	2	2	1	2	1	0	0	1	0

Each row shows the parameter, and each column presents the placement of the parameter according to the ranking. For example, in the first row, the activity parameter is ranked as number one impact factor by four participants. sleep has been ranked as number one by two and as number two by three participants

warning signs. Indication of the impact factors can help them identify the start of mania or depression and react immediately to avoid it. For example, if a patient sees his sleep as the highest ranked parameter for the past few days, he may pay more attention to his sleep habits and try to adjust the amount and quality of his sleep. The parameter ranking can also be used in the feedback loop to provide relevant suggestions and actions to take. For example, in case of low activity level, the system can send messages and information that encourage the patient to be more active. The rankings provide insights for both the patients and clinicians on what impacts the patients mood.

Table 7 The rankings from three methods are compared to the self-rated parameters by participants

Rank	Patients	NP	Chi-squared	NP	Information gain	NP	Significance	NP
1	Sleep	8	Activity	9	Activity	9	Activity	8
2	Activity	7	Sleep	8	Sleep	8	Sleep	6
3	Stress	7	Active W. S.	6	Active W. S.	6	Mixed mood	5
4	Active W. S.	5	Stress	5	Stress	5	Stress	5
5	Mixed mood	4	Mixed mood	4	Mixed mood	4	Alcohol	5
6	Irritable	3	Irritable	4	Irritable	4	Irritable	4
7	Unable to con.	3	Unable to con.	3	Unable to con.	3	Active W. S.	4
8	Alcohol	2	Alcohol	2	Med. changed	2	Med. changed	3
9	Med. taken	2	Med. changed	2	Alcohol	2	Unable to con.	2
10	Med. changed	2	Med. taken	1	Med. taken	1	Med. taken	1

The five highest ranked parameters are mainly common (NP = number of patients)

This insight can be difficult to spot through simple historical graphs, which is the main data visualization in the MONARCA system.

4.3 Automatic behavior tracking from phone channels

One of the motivations for this data analysis was further to investigate what behavioral parameters can be collected without any efforts from the patient's side, i.e., automatically. This goal is particularly important as we expect a lower adherence to self-reports when a patient enters an manic or depressive episode. Moreover, despite the importance of self-reports and their impact on the patients' treatment, researchers and practitioners agree that they cannot substitute for actual objective behavioral data coming from everyday observance. Therefore, we are interested in acquiring as much patient-related data as possible from other channels than the patients themselves. Based on our results, it seems that the perceived level of activity among patients is the strongest impact parameter. The ranking of our three methods highly agreed with the patient's self-rated list. We also observed a significant correlation in the data between mood and activity $(r = 0.3, p < 0.001)$, which confirms our observations in the data. In the depressed state, the mean resided the value of -0.3, while this score was 0.16 in normal and 0.71 in the manic state. These rates indicate that on average depressed patients had a lower activity level as opposed to a higher level in normal and manic state. These results point in the direction of designing for focusing on collection of activity data also from other channels than from the patient him- or herself. Sensors such as accelerometer and GPS as well as phone usage and communication logs can indicate the activeness level of a patient and could provide valuable input to the mood prediction algorithm.

5 Impact factor inference engine

Guided by our analysis, we design an impact factor inference engine capable of (1) inferring the current mental state of the users based on the collected data from smartphones and (2) identifying the past and current behavior parameters that highly relate to the mental state of the users. The engine consists of two main parts: one residing in the phone and the other in the server. The phone-side software collects sensor inputs and self-reports, and the server side is responsible for feature extraction and processing as well as training a predictive mood model and infer impact parameters.

The flowcharts in Figs. 3 and 4 present the overall and detailed steps of calculating the impact factors. As the relation between the parameter impact and the mental state changes over time, we on a daily basis compute the impact factors related to the current mood—the *current impact factors*, as well as features that have had an impact on the mood over the past 14 days —the *past impact factors*. By providing the current impact factors, we inform users of what features they should be aware of or react on immediately, while the past impact factors serve to provide a retrospective insight into what actually influenced their mental state, trying to inform the users of what to be aware of in the future.

5.1 Current impact factors

As our exploratory analysis revealed, the individual models for each patient perform slightly better than the unified models built from all patients data. The main reason is that each patient has a different behavior pattern, and therefore, a model built from a patient's data can more closely predict the mood of that particular person. Hence, to infer current impact factors, we first use the data collected for each patient until the day before $(t - 1)$ to build models for

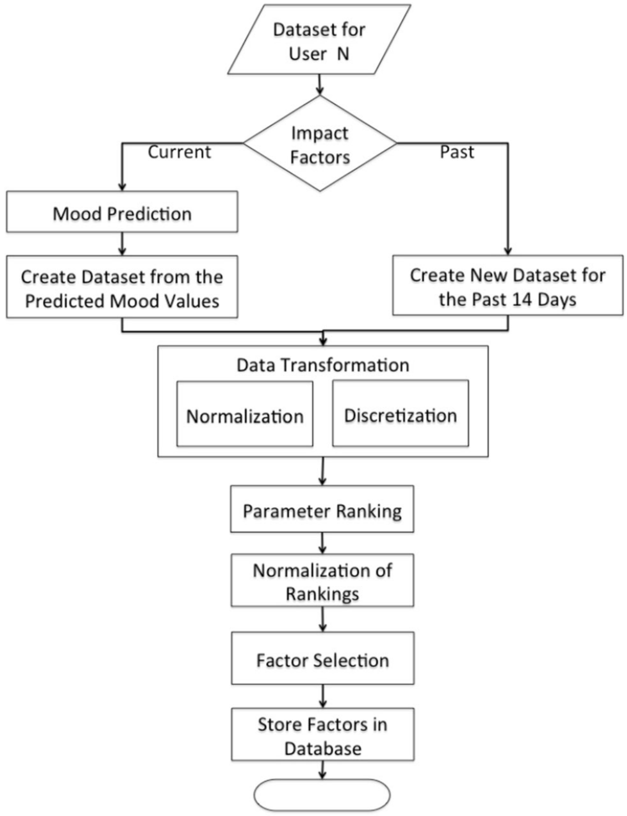

Fig. 3 The overall process of inferring the impact factors

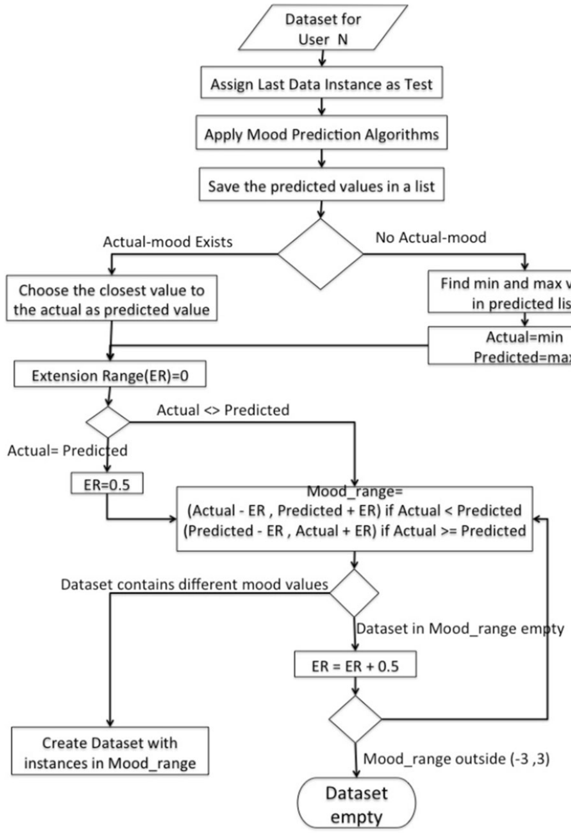
Dataset for User N
Assign Last Data Instance as Test
Apply Mood Prediction Algorithms
Save the predicted values in a list
Actual-mood Exists
No Actual-mood
Choose the closest value to the actual as predicted value
Find min and max values in predicted list
Actual=min Predicted=max
Extension Range(ER)=0
Actual <> Predicted
Actual= Predicted
ER=0.5
Mood_range=
(Actual - ER , Predicted + ER) if Actual < Predicted
(Predicted - ER , Actual + ER) if Actual >= Predicted
Dataset contains different mood values
Dataset in Mood_range empty
ER = ER + 0.5
Create Dataset with instances in Mood_range
Mood_range outside (-3 ,3)
Dataset empty

Fig. 4 The detailed steps in calculating the impact factors

mood estimation. We train a set of algorithms including LinearRegression, SMOReg, AdditiveRegression, M5P, and Bagging on the dataset and apply the built models on the data from the current day (t) to estimate the mood score. The choice of learning algorithms is based on the performance results we got in the predesign analysis. The estimated value is then compared to the self-reported score (actual) and a mood range between those two values is identified. In case the actual and predicted values are equal, the window is extended by 0.5 to find data instances that are close to the actual value. The algorithm repeats until at least two instances with two different values are found. If the actual value (i.e., self-reported mood) is missing, the range between the minimum and maximum predicted value is chosen to be used for creation of the new dataset. Otherwise, the actual and predicted values which are closest to each other are chosen. The dataset is then filtered based on the mood range, i.e., only instances with mood scores in the mood range are kept (see the flowchart in Fig. 4).

The new dataset is used for parameter ranking with the chi-square correlation, information gain, and the significance algorithms that were explained in previous sections. The parameters that are common in at least two evaluators with ranking higher that 25 % are selected as the current impact factors.

5.2 Past impact factors

The overall method for calculating the past impact factors is the same as the current factors. The difference is, that for each patient, we create a dataset from the past 14 days instead of only the current day. If there is not enough data from the past 2 weeks, the algorithm is terminated. In case of mood scores with equal values throughout the 14 days, the time window is extended until two different mood scores are found. The window limit is set to 16 days (1 month period in total). Parameters that are common in at least two evaluators with ranking higher that 25 % are selected as past impact factors.

5.3 Objective features

As mentioned earlier, our analysis motivated us to explore the power of objective data in detecting the impact factors. Hence, in our design, we integrate more sensing inputs into the impact factor engine. The mobile phones include accelerometer and GPS as well as other information resources such as media files stored on the device, call logs, text messages, application usage, and browsing history. From the raw sensor data, we generate four behavioral features, namely social activity, mobility, physical activity, and phone usage. The social activity is the aggregation of incoming and outgoing calls and text messages. Physical activity is computed from changes in the acceleration level, and mobility is calculated from the number of changes in cell ids during the day, indicating different places the user visits. Phone usage on the other hand is an aggregation of user interactions with the phone including the number of changes in the screen, the number of changes in the running applications on the phone, and the number of changes in the installed applications. Please note that we only use phone channels and not any environmental or embedded sensors.

6 Conclusions

We presented the impact factor analysis as a novel approach to help provide patients with an insight into what behavioral parameters have an effect on the progression of their disease. Our proposed approach was based on an analysis of the self-assessment data collected from ten bipolar patients who used the MONARCA system for 11 months. We demonstrated that by applying machine learning techniques, we are able to closely measure the mood of patients with an average mean absolute error (MAE) of about 0.5 compared to the actual mood reported by the patients in their self-assessment. For example, if the patient's reported score is 1, the inferred value by the model can be 0.5, 1.5, or a value between them. We then evaluated the impact of behavior parameters with respect to the mood scores and found that the rankings were in high agreement with the self-ratings performed by the participants.

The analysis motivated us to design and implement an impact factor inference engine as a part of the MONARCA system to increase the disease insight among patients by estimating their emotional state and inferring the behavior parameters that impact their internal state. The implementation of the impact factor engine is described in [9]. The focus and contribution of this paper was to demonstrate the feasibility of data-driven design for health and wellness. This approach can be adapted by other researchers in the field to extract knowledge and insights from behavioral data.

Acknowledgments This work has been done in close collaboration with a group of clinicians and patients from the Copenhagen Affective Disorder Clinic at the University Hospital of Copenhagen. MONARCA is funded as a STREP project under the FP7 European Framework program. More information can be found at http://monarca-project.eu/ and http://pit.itu.dk/monarca.

References

1. Bardram JE, Frost M, Szántó K, Faurholt-Jepsen M, Vinberg M, Kessing LV (2013) Designing mobile health technology for bipolar disorder: a field trial of the monarca system, pp 2627–2636. doi:10.1145/2470654.2481364
2. Bardram JE, Frost M, Szántó K, Marcu G (2012) The monarca self-assessment system: a persuasive personal monitoring system for bipolar patients. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, IHI '12. ACM, New York, pp 21–30. doi:10.1145/2110363.2110370
3. Basco MR, Rush AJ (2005) Cognitive-behavioral therapy for bipolar disorder, 2nd edn. The Guilford Press, New York
4. Bentley F, Tollmar K, Stephenson P, Levy L, Jones B, Robertson S, Price E, Catrambone R, Wilson J (2013) Health mashups: presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. ACM Trans Comput Hum Interact 20(5):30:1–30:27. doi:10.1145/2503823
5. Burns M, Begale M, Duffecy J, Gergle D, Karr C, Giangrande E, Mohr D (2011) Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res 13(3). doi:10.2196/jmir.1838
6. Choe EK, Lee NB, Lee B, Pratt W, Kientz JA (2014) Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '14. ACM, New York, pp 1143–1152. doi:10.1145/2556288.2557372
7. Consolvo S, McDonald DW, Toscos T, Chen MY, Froehlich J, Harrison B, Klasnja P, LaMarca A, LeGrand L, Libby R, Smith I, Landay JA (2008) Activity sensing in the wild: a field trial of ubifit garden. In: Proceedings of ACM CHI 2008, CHI '08. ACM, New York, pp 1797–1806. doi:10.1145/1357054.1357335
8. Epstein D, Cordeiro F, Bales E, Fogarty J, Munson S (2014) Taming data complexity in lifelogs: exploring visual cuts of personal informatics data. In: Proceedings of the 2014 conference on designing interactive systems, DIS '14. ACM, New York, pp 667–676. doi:10.1145/2598510.2598558
9. Frost M, Doryab A, Faurholt-Jepsen M, Kessing LV, Bardram JE (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, pp 133–142. doi:10.1145/2493432.2493507
10. Hamilton M (1967) Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 6(4):278–296
11. Harrison V, Proudfoot J, Wee P, Parker G, Pavlovic D, Manicavasagar V (2011) Mobile mental health: review of the emerging field and proof of concept study. J Mental Health (Abingdon, England) 20(6). doi:10.3109/09638237.2011.608746. http://europepmc.org/abstract/MED/21988230
12. http://quantifiedself.com/
13. https://www.apple.com/ios/ios8/health/
14. Kasckow J, Zickmund S, Rotondi A, Mrkva A, Gurklis J, Chinman M, Fox L, Loganathan M, Hanusa B, Haas G (2013) Development of telehealth dialogues for monitoring suicidal patients with schizophrenia: consumer feedback. Community Mental Health J, 1–4. doi:10.1007/s10597-012-9589-8
15. Lane ND, Choudhury T, Campbell A, Mohammod M, Lin M, Yang X, Doryab A, Lu H, Ali S, Berke E (2011) BeWell: a smartphone application to monitor, model and promote wellbeing. In: Proceedings of the 5th international ICST conference on pervasive computing technologies for healthcare (Pervasive Health 2011), pervasive health 2011. IEEE Press
16. Li I, Dey A, Forlizzi J (2010) A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI conference

on human factors in computing systems, CHI '10. ACM, New York, pp 557–566. doi:10.1145/1753326.1753409

17. Matthews M, Doherty G, Sharry J, Fitzpatrick C (2008) Mobile phone mood charting for adolescents. Br J Guid Counsel 36(2):113–129
18. Rachuri KK, Musolesi M, Mascolo C, Rentfrow PJ, Longworth C, Aucinas A (2010) Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on ubiquitous computing, Ubicomp '10. ACM, New York, pp 281–290. doi:10.1145/1864349.1864393
19. Rooksby J, Rost M, Morrison A, Chalmers MC (2014) Personal tracking as lived informatics. In: Proceedings of the 32Nd annual ACM conference on human factors in computing systems, CHI '14. ACM, New York, pp 1163–1172. doi:10.1145/2556288.2557039
20. Wilde MH, Garvin S (2007) A concept analysis of self-monitoring. J Adv Nurs 57(3):339–350. doi:10.1111/j.1365-2648.2006.04089.x