Supporting Disease Insight through Data Analysis: Refinements of the MONARCA Self-Assessment System

Mads Frost*, Afsaneh Doryab, Maria Faurholt-Jepsen***, Lars Kessing***, Jakob E. Bardram***
* IT University of Copenhagen, Denmark, {madsf, bardram}@itu.dk
** Carnegie Mellon University, Pittsburgh, USA, adoryab@cs.cmu.edu
*** Psychiatric Center Copenhagen, Denmark,
{maria.faurholtjepsen, lars.vedel.kessing}@regionh.dk

ABSTRACT
There is a growing interest in personal health technologies that sample behavioral data from a patient and visualize this data back to the patient for increased health awareness. However, a core challenge for patients is often to understand the connection between specific behaviors and health, i.e. to go beyond health awareness to disease insight. This paper presents MONARCA 2.0, which records subjective and objective data from patients suffering from bipolar disorder, processes this, and informs both the patient and clinicians on the importance of the different data items according to the patient's mood. The goal is to provide patients with a increased *insight* into the parameters influencing the nature of their disease. The paper describes the user-centered design and the technical implementation of the system, as well as findings from an initial field deployment.

Author Keywords
Pervasive Healthcare, Personal Health Monitoring, Mental Illness Management, Bipolar Disorder, Smartphone, Data analysis

ACM Classification Keywords
J.3 Life and Medical Sciences: Health; H.5.2 Information Interfaces and Presentation: User-centered design.

General Terms
Design; Human Factors; Algorithms

INTRODUCTION
The management of mental health and well-being through phone-based monitoring systems, tracking daily life and routines, is a promising, rapidly growing area in pervasive healthcare. Smartphones are capable of capturing multiple dimensions of human behavior, encompassing physical, mental and social aspects of well-being [12, 5]. Many Smartphone applications take advantage of persuasive visualizations and features that can help with adjustment of behaviors to improve adherence and consistency. For example, *UbiFitGarden* [6], and *Bewell* [12] collect behavioral data, such as physical activity from phone sensors and provide visual feedback such as an ambient display to promote healthy behavior.

In the clinical domain, systems are moving from reactive response to acute conditions to a proactive approach, characterized by early detection of conditions, prevention, and long-term management. The goal is to make patients and clinicians aware of the current state of the illness with the help of technology [4]. An example is the *Health Buddy* [11], which is used for monitoring patients with schizophrenia who were recently admitted for suicidal behavior. It presents patients with a series of pre-programmed questions about symptoms of depression and suicide, allowing mental health service providers to monitor the patients symptoms. The *Mobile Mood Diary* [16] uses a mobile phone to allow patients to report mood, energy, and sleep levels, which can then be accessed on a website. The *Mobilyze!* system [5] is an example of an intervention system that uses machine learning to predict the cognitive state of the patients from phone sensors and environmental context.

However, most of these health monitoring systems use the collected data solely for visualization purposes, and provides little insight into the nature of the disease. Bipolar and other mental diseases are chronic disorders and patients need to learn to cope with their illness the rest of their lives. There is a great individual variability in the illness in terms of how it affects the individual patient, thus the treatment requires an ongoing process of experimenting with different combinations of medications, combined with learning how to cope with, and reduce, symptoms through awareness and insights into healthy behaviors and routines (e.g. good sleeping habits, avoidance of alcohol, reducing stress, etc.). Therefore, helping patients identify patterns in their behavior and recognizing factors impacting their mental state would provide them with a greater insight into the nature of the disease and helps them cope better. The need for disease insight has also been recognized in other chronic illnesses, such as diabetes [14], but so far the patients have been responsible for making inferences, not the system.

This paper presents an approach to personal health technologies that aims at providing patients with an *insight* into their disease. This is done by collecting self-assessment and sensor

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.
ACM 978-1-4503-1770-2/13/09.
http://dx.doi.org/10.1145/2493432.2493507

data on a Smartphone and using this to analyze and present the underlying patterns and correlations to the patient. The paper presents a system implementation of such a disease insight approach designed for patients suffering from bipolar disorder. This work extends previous work on the MONARCA 1.0 system [3, 2] and we thus call this new version MONARCA 2.0. The next section will describe MONARCA 2.0 in more detail and outline the differences between version 1.0 and 2.0. The rest of the paper describes the design and technical implementation of MONARCA 2.0 and outlines how semi-automatic data sampling is used to find correlations between disease symptoms (mood) and data patterns such as activity, sleep, and phone usage. The paper describes how such patterns are used to predict and forecast disease symptoms (mood) and how this is presented to the user to support and improve disease insight. The paper reports on the results of a 6 months trial with 6 patients. This study revealed which data parameters had the highest correlation ranking and hence highest prediction power, as well as provided insight into the usefulness of the system.

BACKGROUND & DESIGN OF MONARCA 2.0

Bipolar disorder is a mental illness characterized by recurring episodes of both depression and mania, and is associated with a high risk of relapse and hospitalization [13]. It is difficult for patients to reflect on their own mood and behavior, and they may only recognize symptoms if they understand the illness and know what to look for. The treatment of bipolar disorder involves management of a patient's daily life and routines. Clinical research has found that routine is the most effective way to reduce symptoms of depression and mania, and prevent relapses which have extreme consequences for the patient's quality of life [8, 9].

MONARCA 1.0 was designed to provide patients with an awareness of how their life and mental state progress [3, 2]. MONARCA 1.0 is a 'classic' personal health technology consisting of a Smartphone app used for collecting data and presenting it to the patient, as well as a web portal that provides access for both patients and clinicians to the data stored in a server-based infrastructure. The main focus of MONARCA 1.0 was on collecting self-assessed data on core parameters such as mood, sleep, medicine compliance, stress, and self-reported activity level. Automatic collection of accelerometer data and phone usage was also collected. This data was presented to the patient in 'raw' format and was hence not subject to any data analysis. The data was used by patients to gain a recounted awareness of the development of their disease and was shared with the clinicians in charge of their treatment. A field deployment of the system showed that patients found MONARCA very useful and easy to use, and the system had a high (87%) 'compliance' rate i.e. patients used the system on a regular basis and the system hence had a high data quality. However, the study also showed areas for improvement, which lead to the design of MONARCA 2.0.

Designing MONARCA 2.0

The design of MONARCA 2.0 continued the user-centered design process applied previously [10, 2, 15], involving patients and clinicians affiliated with a psychiatric clinic at the national university hospital in Denmark. 6 patients and 3 clinicians participated in collaborative design workshops; two-hour sessions held every three weeks for 4 months. All participants had previously participated in the first field trial of the MONARCA system, and had all continued to use the system after the trial period on a regular basis.

Self-assessed Item	1	2	3	4	5	6	7	8	9	10
Activity	5	1	1	2	1	0	0	0	0	0
Sleep	0	3	2	3	1	0	0	1	0	0
Stress	1	2	1	1	1	0	1	2	0	1
Warning Signs	1	0	3	2	0	1	1	0	0	2
Mixed Mood	1	2	0	0	2	0	1	2	2	0
Irritability	1	1	0	0	2	3	1	1	0	1
Alcohol	1	0	1	0	0	2	3	2	1	0
Cognitive Problems	0	0	2	0	2	1	0	1	4	0
Medicine Changed	0	1	0	1	1	0	2	1	3	1
Medicine Taken	0	0	0	1	0	3	1	0	0	5

Table 1. The results of applying Chi-squared correlation evaluator to rank the self-assessed data items according to the patients' mood score. Activity is ranked as the highest for 5 patients, and sleep is ranked as the second highest for 3. In general, Activity, Sleep, Stress, and Warning Signs are the 4 highest ranked items.

The patients involved in the design process thus had extensive experience in using the first version of the MONARCA 1.0 system and has valuable input on how to improve the user interface, the self-assessment data forms, the visualization, etc. – all of which has been incorporated into the design of MONARCA 2.0, as we shall present below.

During the design of MONARCA 1.0, it was important to limit the amount of data items that the patient should self-report in the app. Therefore, a lot of effort went into minimizing the list of self-assessment data items [2]. But during the trial of MONARCA 1.0, a significant amount of self-assessed data had been collected. A natural question then to ask is; which data items were most important in correlation with the main disease parameter, i.e. the mood score? By analyzing the collected data from the trial phase, we could gain some insight into this question, and use this insight in the further design of the system.

Analyzing Self-Assessed Data

In order to understand the correlation between the self-assessed data and the mood score, we analyzed the data from 10 bipolar disorder patients who had used the MONARCA 1.0 system from May 2011 to March 2012. This analysis had two goals; firstly to reveal which items correlate with the mood score, and secondly to uncover how accurately we can estimate and forecast the emotional state (mood score) of the patient, based on self-assessed data. The self-assessed data set included the following items:

- Mood – Highly depressed (-3) to highly manic (3)
- Sleep – Amount of sleep, reported in half hour intervals
- Medicine Taken – Yes/No
- Medicine Changed – Yes/No
- Activity – Highly inactive (-3) to highly active (3)
- Warning Signs – Number of active warning signs
- Mixed Mood – Yes/No
- Irritable – Yes/No

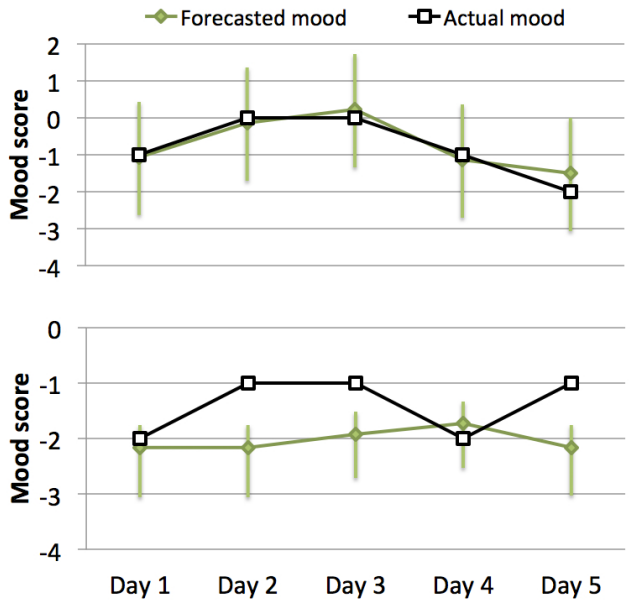

Figure 1. A five days mood forecast for two patients. The horizontal black line is the patient's self-reported mood score and the green line is the predicted values by the model. The vertical green line depicts the 95% confidence interval. The top graph (i) shows a highly accurate forecast, while the forecast in the bottom graph (ii) is less accurate.

- Cognitive Problems – Yes/No
- Stress – No stress (0) to highly stressed (5)
- Alcohol – Number of alcoholic drinks

To answer the first question, we applied the Chi-Squared method to rank the correlations between the mood score and the self-assessed items. The result is shown in Table 1, and reveals that Activity and Sleep are the highest ranked items followed by Stress and Warning Signs Active.

To answer the second question, we applied machine learning techniques to our data set, utilizing best performing learners including linear regression, SVM, additive regression, and model trees. We found that we are able to assess the mood of patients with an average mean absolute error (MAE) of 0.5 compared to the actual mood reported by patients in their self-assessment. For example, if a patient's reported mood score is 1, the inferred value by the model range between 0.5 and 1.5. We also demonstrate that using time series techniques and considering a 95% confidence interval, we can on average estimate the tendency in the mental state with a min MAE of 0.36 and a max MAE of 0.77. Figure 4 shows the forecast results from two patients – one where the forecast was particularly good (top), and one which were less accurate (bottom). The performance of each forecast model depends on the quality of the patients' data, but for 9 out of the 10 patients the majority of the outcome falls within a 95% confidence interval. The explicit details of the machine learning techniques applied in this analysis is submitted for review in a separate publication [7].

DESIGN OF MONARCA 2.0

The initial field trial, the user-centered design phase, and the data analysis suggested that the original MONARCA design could be improved in several ways. In particular, the data analysis of self-assessed data seemed to suggest that it is possible to (i) find correlations between mood and self-assessed data, and (ii) automatically infer the mood as well as estimate the tendency in the emotional state of a patient.

Both of these can be useful in providing patients with a greater *insight* into their disease. The correlation information can help give patients an insight into how their behavior impacts their mood state, both on a past and current basis. We define the features that have the highest correlation with mood as *Impact Factors*, since they are features that affects the patient's mood state. The estimation information can provide patients with insight on the temporal unfolding of their disease. This kind of mood estimation can result in reducing – or possibly even preventing – extreme manic and depressive episodes by faster interventions through the monitoring system.

Another important outcome of the data analysis was that self-reported activity was the highest ranking parameter correlating with mood. With this in mind, it seemed that a natural next step was to ask if activity monitoring could be done automatically by sampling movement and usage data form sensors in the phone, which may provide an indication of a patient's activity level.

In summary, the main design goals of MONARCA 2.0 as compared to MONARCA 1.0 was to improve and incorporate the following components:

User Interface – the user interface of the system (both the phone and the web portal) had to be improved and upgraded, partly based on feedback from the trial of MONARCA 1.0 and the user-centered design process, and partly based on incorporating the new features related to impact factor analysis and mood forecasting.

Data Sampling – the data sampling component of the system had to be significantly improved in order to collect and process a much larger set of data from the phone and its sensors.

Impact Factor Analysis – an impact factor analysis component should continuously calculate correlations between mood and all the collected data, including self-reported as well as automatically sampled data.

Mood Forecasting – a mood forecasting component should continuously train a model that is used for mood prediction on a 5-day horizon.

The following sections describe these four components and explain how self-reported and automatically sampled data is used to pinpoint impact factors and predict mood, and how this is presented and used by the patients and clinicians.

MONARCA 2.0 USER EXPERIENCE

The user experience of MONARCA 2.0 has undergone significant improvements as compared to the first version. The most important improvements to the patient's mobile phone app are; better support for self-reporting of data; support for retrospective reporting of data; and the presentation and management of the new impact factor feedback. On the website the main improvement is that the overview of patients now

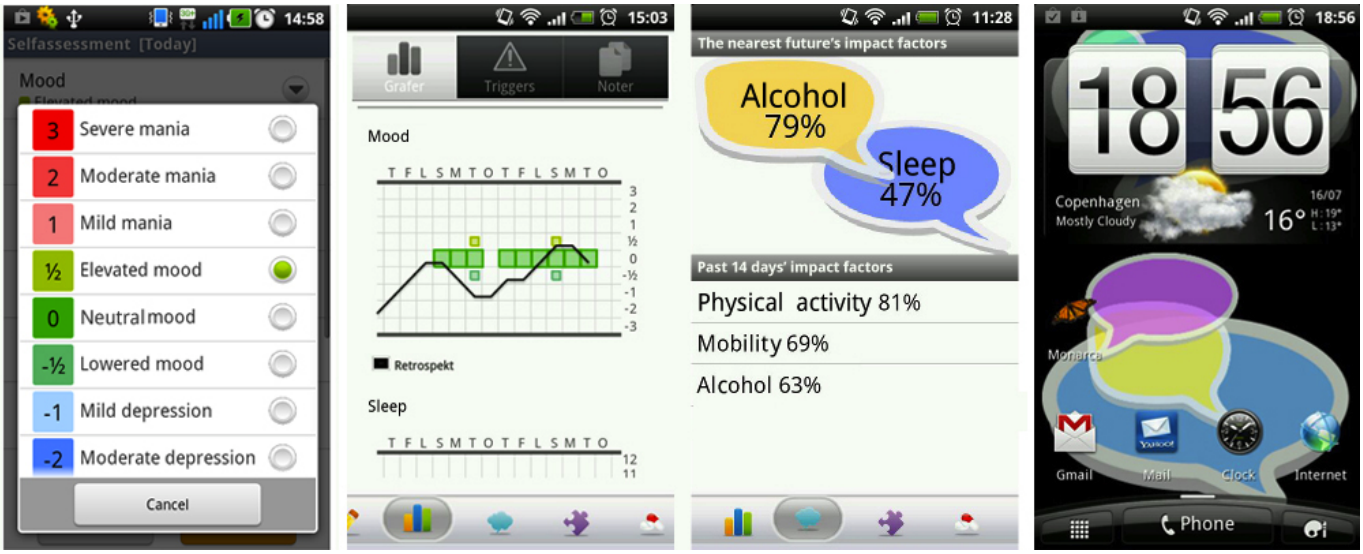

Figure 2. The main MONARCA 2.0 user screens. (i) The new 1/2 point mood scale. (ii) Graphing of mood, mixed mood, and retrospect. (iii) The 'Impact Factors' screen, showing current and past impact factors and their % of magnitude. (iv) The 'Live Wallpaper' displaying impact factors through speech bubbles.

shows the 5-day mood forecast for each patient, as well as a general update of the design with a new css template. Figure 2 shows the main new screens on the Android phone and Figure 3 displays the improved web portal showing the clinician overview screen with impact factors and mood forecasts for a set of patients.

Changes in self-assessment

The design process revealed the need for improving the self-assessment support in the system. There was a request for having a more fine-grained mood assessment by adding 1/2-point mood scores to the existing mood-scale, and having mixed-mood on a scale instead of a yes/no feature. Moreover, the support for personalization was extended allowing for the addition of custom fields to the self-assessment form. Each of these features will be described in detail in the following sections.

Half-point mood scales – the transition from a neutral mood state (0) to 'mild mania' (+1) or 'mild depression' (-1) was too coarse grained. It did not reflect the fact that the patients were able to sense a change that was not yet severe enough to be considered a manic or depressed state, but still significant enough for them to track. Thus, we added the elevated (+1/2) and lowered (-1/2) mood state, as seen in Figure 2(i).

Mixed mood scale – the severity of mixed mood was difficult for the patients to express through the previous simple yes / no option. Thus, mixed mood was transformed into a scale, visually represented by the same preference dialog as the 1/2 point mood scale. The data visualization is depicted in Figure 2(ii) – mixed mood values are represented as small rectangles while the main mood score is still represented as the larger rectangle, enabling the clinicians to understand the span of the mixed mood.

Custom user-defined SA fields – given the great individual variability in bipolar disorder, we experienced the need for enabling SA customization, where the patients could add their personal items to track in the SA. Patients during the design sessions mentioned the need for tracking e.g. anxiety, cups of coffee, minutes of work-out, etc., and thus we have created a SA management feature. Self-assessment comes with a predefined set of items, as previously listed. First the 5 mandatory - *mood, sleep, medicine intake, activity, mixed mood*. The rest of the items can be reordered or excluded from the daily SA, so that if e.g., a patient never drinks, the alcohol field is removed. Furthermore, the patients have the ability to define up to 3 new items of the types Yes/No, Range -3 to 3, Range 0 to 10. The custom items are graphed in the visualization screen, and will show up in the clinician's interface as well. The limitation of a max of 3 custom fields and 3 predefined types is based on the notion of keeping the system simple, as the patients should be able to grasp and cope with the system even when in a severe manic or depressed state.

Retrospect

The retrospect feature allows user to assess their mood in retrospect. The subjective perception of mood can be influence by the mood itself, so in some cases, the lapse of time can help patients assess their mood more accurately. This is especially seen in cases of hypomania. The retrospect feature aims at facilitating this by allowing patients to re-assess a previous mood, adding a retrospect score to the system up to two weeks back in time. The retrospect score in graphed in the mood chart as black line in Figure 2(ii).

Impact Factors

The impact factors screen provides both graphical and textual views of the current and past impact factors, as seen in Figure 2(iii). The current impact factor icons are drawn along with their corresponding text in the upper half of the screen, and past impact factors are displayed below as a simple list view. There can be up to 2 current and 5 past impact factors displayed in the screen. When selecting any of the current or

past impact factors, the user is taken to a sub-dialog screen, which displays a detailed textual description of that particular impact factor as well as strategies and actions for self-help, tailored based on the outcome of the forecast component. In this way, the patient gets suggestions about different strategies and actions for self-help according to the assessed mood state.

We designed a Live wallpaper to give the patients a visual insight of their impact factors without forcing them to enter the MONARCA application. It is a mechanism for providing daily feedback to users regarding the impact factors that have the biggest impact on their mood, as explained earlier. The impact factors are visualized of the patient's phone using animated speech bubbles in different colors and sizes, which moves calmly around in the background of the phone's home screen. An example is shown in Figure 2(iv). Each color is color coded according to the colors used in the graphs in the visualization screen. The relative size of the icon on the screen correlates to the magnitude of the impact attributed to that particular factor. Colored speech bubbles were chosen because they are socially neutral, and they symbolize the system trying to say something to the patient. Furthermore, they convey information to the user without compromising their privacy, as would be the case with text. If the patients press the bubbles on the screen, they are immediately taken to the impact factor screen inside the MONARCA application.

Mood Forecasting for patients

The amount of data the system collects through subjective and objective sampling of behavior data from a patient, provides us with the possibility to not only report what happened and why, but also to build models that may predict what will happen – at least to a certain degree. Being able to inform patients about what their future mood state might be if their current behavior continued unchanged, could provide significant insight, both for patients and clinicians, allowing them to be proactive and prevent possible mood swings.

During the design, we explored the area of presenting forecasts to patients, and had in-depth discussions on how mood forecasting could play a role in the feedback to the patient and clinicians. An mock-up Android user interface presenting the mood forecast for patients using a weather forecast metaphor was proposed, but in the end it was rejected by the clinicians mainly due to ethical concerns. The main challenge was that mood forecasting could end up as a self-fulfilling prophecy; patients could become depressed by a forecasted depression. And this again could have significant impacts on the life and wellbeing of a patient, and potentially be life-threading for suicidal patients. The mood forecast hence never became a part of the Android phone UI used by the patients, but was only shown on the clinician's website.

Clinicians Web Portal

The information regarding impact factors and forecasts is presented to clinicians in the web portal. Both are integrated into the overview screen, as seen in Figure 3, and they are also accessible in the detailed patient information, where clinicians

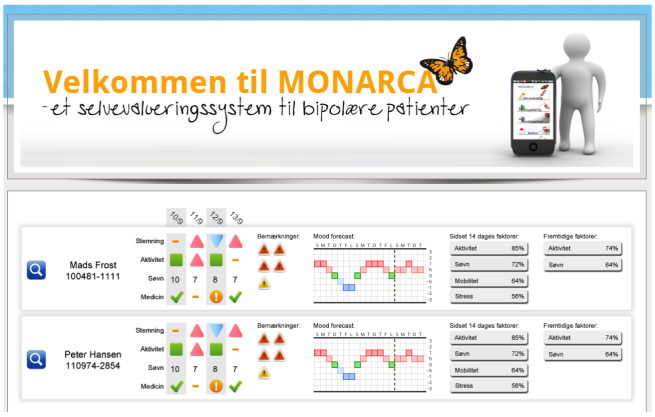

Figure 3. Patient data on the clinician's website. Each line is a patient (name and ID number in the left column), showing mood, activity, sleep, and medicine data for the past 4 days. Then triggers and early warning signs activated, the mood forecast for the next 5 days, and to the far right is the past and present impact factors. An enhancement of the forecast can be seen in Figure 5.

can review the information provided to the patients on strategies and actions for self-help.

TECHNICAL SYSTEM DESIGN

Figure 4 shows the overall architecture and process flow of MONARCA 2.0. This section describes the technical design of the system and the details of the sub-components.

System Architecture

MONARCA 2.0 uses the same technical architecture as MONARCA 1.0 [3] consisting of a server running a CouchDB as the main database, a web application server running the web site, and an Android phone app to be used by the patients. Data sampling and logging takes place on the phone and is transmitted to the CouchDB running on the MONARCA server. On the server, data processing and inference have been implemented as a separate service. This service runs every night, extracts the collected data from the CouchDB, processes the data, and submits the output back into the CouchDB. The processed data, including the calculation of impact factors and the 5-day forecast is then accessible from the phone and the website, as shown in Figure 2 and 5.

Data Logging and Processing

MONARCA 2.0 collects the same self-reported data items as MONARCA 1.0, i.e. the 10 items listed in Table 1. We call this data set the *subjective* data set. MONARCA 2.0 was designed to also collect what we call *objective* data from the phone, which include sensor data from e.g. accelerometers, cell tower ids, and communication logs from the phone.

We use the Funf Open Sensing Framework [1] and integrated this into the MONARCA 2.0 phone app in order to acquire and pre-process the raw sensing inputs. In Funf, the collection and upload of a wide range of data types is done through so-called probes. Each probe is responsible for collecting data from the on-phone sensors, e.g., accelerometer or GPS as well as other information resources such as media files stored on the device, call-logs, application usage, browsing history,

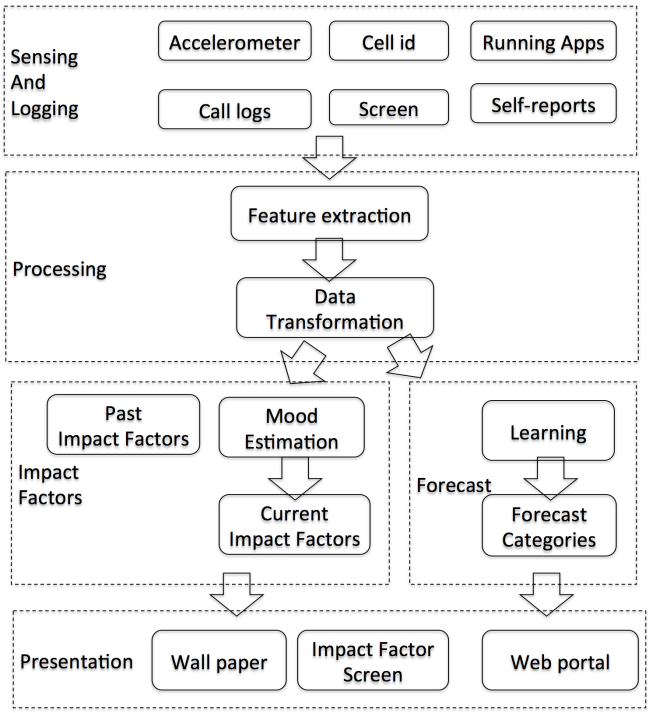

Figure 4. The overview of MONARCA 2.0's main components and process flow.

etc. The data sampling is implemented as a background service, running even if the MONARCA application is not active.

To balance resource consumption (i.e. battery) with optimal sensing frequency, we did a series of iterative tests during the design phase. The following list describes the final design of the data acquisition probes.

- `ActivityProbe` - records how active the user is. It uses the Funf `AccelerometerProbe` data to calculate how many intervals the variance of a device's acceleration is above (hi) or below (low) a certain threshold. It is configured to run every 5 minutes for 20 seconds at 20 Hz.
- `CellProbe` - records ids for the cell tower currently connected to. Configured to run every 5 minutes.
- `ScreenProbe` - records when the screen turns on/off. No configuration needed as the probe acts as a listener of the screen's state.
- `RunningApplicationsProbe` - records the list of currently running applications. Runs every 5 minutes.
- `ApplicationsProbe` - records which applications are installed/uninstalled on the device. Runs every 5 minutes.

Feature Extraction

The Funf probes were used to generate four new 'objective' features: (i) Social Activity, (ii) Physical Activity, (iii) Mobility, and (iv) Phone Usage. These objective features are used in two ways; first they are shown in the visualization screen in the Smartphone application (Figure 2(ii)), providing a status from the past 14 days. Second, they are added to the list of feature attributes that are used for the mood prediction and forecast.

Social Activity

The social activity feature is calculated based on incoming and outgoing calls and text messages. The social incoming (si) and social outgoing (so) feature is generated from the number of incoming and outgoing calls (ic, oc), their duration (id, od), number of incoming and outgoing messages (is, os). These features are then used to build the social activity (sa) feature.

$$sa = si + so, \text{ where}$$
$$si = ic * w + id + is * w$$
$$so = oc * w + od + os * w$$

Incoming and outgoing calls, incoming and outgoing messages are numbers, while durations are calculated in seconds. To balance the weight of the features, we multiply them by a constant value – in our case 10. The value can be calculated using aggregate functions or probability methods. We choose the constant value for simplicity, and since this formula is the same for all data instances, the results are consistent.

Physical Activity

To measure the overall daily physical activity for each patient, we first calculate the level of high (ha) and low activity (la) based on the measurements from the Funf framework which include high and low activity intervals (hai, lai) as well as total activity intervals (ti). We then compute the overall activity rate (ar) by subtracting the low activity rate from the high activity rate which will provide a number between -1 and 1.

$$ar = ha - la, \text{ where}$$
$$ha = hai/ti \ \& \ la = lai/ti$$

Mobility

The mobility feature, called mobility rate (mr), is computed from two raw location features; the number of changes in cell ids (cc) and the total number of identified cell ids (ct) during the day.

$$mr = cc/ct$$

Phone Usage

To measure the phone usage (pu), we look at how many seconds the screen has been turned on (tst), the number of changes in the screen (cs), the number of changes in the running applications on the phone (cra), and the number of changes in the installed applications (cia). We boost the the last 3 by again multiplying them by a constant weight value (w) – in our case 10.

$$pu = tst + cs * w + cr * w + cia * w$$

—- o —-

The generated features produce different values depending on the type and the value of the raw features. A min/max normalization method was used to balance the weight of each feature before training phase. In total, there is now a feature set with 14 different features, consisting of the 10 original subjective and the 4 new objective features. The combined list can been seen in Table 2.

Impact Factor Component

Impact Factors are specific features from the feature list previously mentioned, which the data analysis points out as having a big influence on a patient's mood. This is done to try and provide insights for both the patients and clinicians on what impacts the patients mood, as it can be difficult to spot through simple graphs, which were the only data feedback the patients got in the version 1.0 of the system. Thus, 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*. These factors are shown to the patients in the Android app (see Figure 2(iii)) and to the clinicians on the web site (see Figure 5). By calculating the current impact factors, we inform patients of what features they should be aware of or react to immediately, while the past impact factors serve to provide a retrospective insight into what has influenced their mood historically.

To identify the impact factors (both current and past) and score their impact, we apply three different methods on our data; first we find correlations between each feature and the mood, then we measure the significance of the features wrt. the mood, and finally we measure the information gained from each feature wrt. the mood.

Current Impact Factors

We keep the mood score as continuous values and use prediction methods to estimate the current mood. In our pre-design analysis, we experimented with both individual as well as unified models built from all patients data. Our observation was that although the performance of the individual models varies from patient to patient depending on the size and quality of their data set, in general, they perform slightly better than the unified models. 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, in our system, each learner is trained on the data for each patient and individual models are built per learner.

Based on the performance of the learners in the pre-analysis, we choose a combination of basic and meta methods to estimate the mood scores. We use K-nearest neighbors and model trees as well as a set of regression based learners such as linear regression, SVM for regression, and additive regression. Please, note that we do not use the output of the mood estimators directly. The estimated values from the models are only used to identify a mood range that is used for ranking the impact factors, as we will describe below.

To estimate the mood, we use the data collected until the day before $(t-1)$ as our training set, and the data collected on the current day (t) as a test set. We then apply the trained models on the data, compute the residuals from each model, and choose the output of the one where the range between the actual and estimated mood is lowest, and store the range between the two. If the actual mood score is missing, we choose the range between minimum and maximum predicted values. The training set is then filtered based on the mood range, i.e., only instances with mood scores in the mood range are kept.

The new data set is used for parameter ranking. We calculate and normalize the Chi-squared correlation values, the information gain and the significance scores of the parameters. The parameters that are common in at least two evaluators with ranking higher that 25% are selected as the current impact factors. The significance scores provide us with the magnitude of the impact attributed to each individual factor.

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 data set from the past 14 days instead of only the current day. If there is not enough data from the past two 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 (one month period in total). Features that are common in at least two evaluators with ranking higher that 25% are selected as past impact factors.

Forecast Component

To estimate the tendency of the mental state, we formulate the problem as a time series forecasting where the value of the variable mood is predicted at a time interval – in our case 5 days. The main difference between the mood estimation, used in the impact factor component, and the mood forecast is that in mood estimation, the mood score is predicted from the models that are built on data which contain actual mood values, while in the forecast component, the data does not contain the actual mood values from the self-assessment.

We address the temporal dependency between data points via additional lagged features which values are computed from the past data points. After transformation, we apply learning algorithms similar to the ones used in the mood estimation, to predict the tendency in the mood state in the form of a 5 days forecast on a daily basis. The mood is forecast on a daily basis 5 days ahead in time by looking at the pattern of the data from the past 14 days. The forecast is shown to the clinicians in their overview screen (Figure 3), where an enhanced section can be seen in Figure 5.

Forecast categories – Based on the forecast mood scores, we calculate the forecast categories which later are used in giving feedback to the patients, as explained in the impact factor part of the design section above. The categories are determined as follows:

1. If at least 2 days values are over 0.5 and none under -0.5, then forecast category $= Manic$
2. If at least 2 days values are under -0.5 and none above 0.5, then forecast category $= Depressed$
3. If values both above 0.5 and below -0.5, then forecast category $= Mixedmood$
4. Else, forecast category $= Neutral$

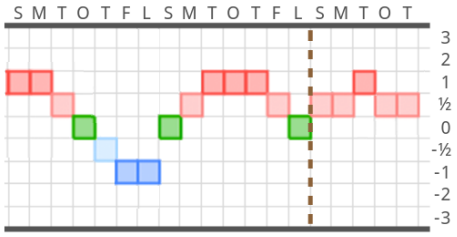

Figure 5. Enhancement of the Mood Forecast from the clinician's patient overview web page, seen in Figure 3. The dotted line represents today; mood scored to the left are self-reported historic data, whereas the 5 days mood scores on the right are forecast.

6 MONTHS FIELD DEPLOYMENT

In order to evaluate MONARCA 2.0, it was deployed for a small 6-month field trial from March to August, 2012, involving 6 patients. The purpose of this study was to verify the redesign of the system, and to investigate if the new data mining functionality would find relevant impact factors and make sensible forecasting. This should prepare for a larger trial with more patients. The use of the system was approved by the Danish National Committee on Health Research Ethics and the security and data handling was approved by the Danish Data Protection Agency. Informed consent was obtained from all patients.

In this section we discuss the findings from this initial field deployment of MONARCA 2.0, focusing on (i) the general system usage and performance, (ii) the analysis of the data collected and its ability to identify impact factors and forecast mood, and (iii) the patients' and clinicians' feedback on usability and usefulness of using MONARCA 2.0 based on a set of interview during the trial period.

System Usage and Performance

During the field trial, the system collected self-reported data in 511 days and sensor data in 563 days. This gives an total 55.6% uptime of the Android app. In total 1,043 mb of data was collected. In order to gauge the battery consumption of the system, we measured and compared the battery performance over a 24 hour period on 1) an out-of-the-box HTC Desire S phone, 2) a phone with MONARCA 1.0 installed, and 3) a phone with MONARCA 2.0 installed. During the 24 hours, the consumption was respectively 12%, 32%, and 68% of the total battery power. For the measurements to be comparable, the phone was not used in the measurement period. This means that energy consumption will be higher when actually used. But the energy consumption is sufficiently low for the patients to use the phone during a normal day of ca. 16 hours without having to recharge the phone. In the trial, there were a few cases where patients ran out of power, but only when they had used the phone excessively for phone calls. In general, the energy consumption did allow the patients to use the phone throughout a day without the need for recharging.

In the trial of MONARCA 1.0, we tested the adherence rate of the patients' self-reporting, i.e. to what degree a patient would fill in the self-report each day. In the original study we found an adherence rate of 87%, taking into consideration the days where the system was actually working [2]. When performing the same analysis of MONARCA 2.0, we found an adherence rate of 91%, which is slightly better but equivalent. It should be noted that these high adherence rates are in itself a major achievement of the system, since self-reporting typically have very low rates of adherence.

Data features	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Activity	3	1	1	1	0	0	0	0	0	0	0	0	0	0
Stress	1	1	2	0	1	0	1	0	0	0	0	0	0	0
Sleep	0	1	0	3	2	0	0	0	0	0	0	0	0	0
Phone Usage*	1	0	3	0	0	0	1	0	0	0	0	1	0	0
Social Activity*	0	2	0	1	0	0	1	0	1	0	1	0	0	0
Irritability	0	0	0	0	2	2	0	1	1	0	0	0	0	0
Cognitive Problems	0	1	0	1	0	1	1	0	1	0	0	0	1	0
Physical Activity*	1	0	0	1	0	0	1	0	1	1	0	1	0	0
Alcohol	0	0	0	1	0	1	1	2	0	0	0	0	1	0
Warning Signs	0	0	0	0	1	1	0	1	1	1	0	0	1	0
Mobility*	0	0	0	0	0	1	0	2	0	1	1	1	0	0
Mixed Mood	0	0	0	0	0	0	1	1	1	0	0	0	0	3
Medicine Changed	0	0	0	0	0	0	0	1	0	0	2	1	1	1
Medicine Taken	0	0	0	0	0	0	0	0	0	0	0	2	2	2

Table 2. Ranking of the correlation between Impact Factors (features) and the mood score. The objective features are marked with *.

Analyzing the use of the web site, we found that none of the patients logged in. This confirms previous findings that patients do not want to use a web interface; they prefer to have all features on the Smartphone. The clinicians monitoring the patients logged in on a regular basis with 286 logins.

Data Sampling and Analysis

The data (both subjective and objective) collected during the trial was subject to two types of analysis. First, we investigated how data features correlate with the mental state of a patient. Second, we analyzed the performance of the mood forecasting, with a specific focus on how accurate the mood can be inferred using only the objective data set.

Analyzing Impact Factors

We repeated the experiment done during the design phase using the Chi-Squared method on the new data set, now including both objective and subjective features. We applied the method on each individual patient's data, and analyzed the rankings with respect to the mood score as the class.

As shown in Table 2, (self-reported) Activity, Stress, Sleep and Phone Usage are among the 4 highest ranked parameters. For example, Activity is ranked in the top 4 for all 6 patients, and Phone Usage is ranked in the top 4 for 4 out of 6. Although the parameters of Activity, Stress, and Sleep still are amongst the highest ranking, the table also shows that 2 out of 4 objective features, namely Phone Usage and Social Activity are among the parameters that are highly correlated with the participants' mood score.

We repeated the method of inferring mood from the features with the same set of learners used in the design phase analysis. This time, we created two data sets for each patient. The first one include all 14 subjective and objective features, while the second contained only the 4 objective features. We ran the cross-validation on both data sets with the selected learners and compared the output results. From the mood estimation model built with both objective and subjective features, we observed an average min MAE (mean absolute error) of 0.40,

while this value from the model built with only objective features is 0.45. Hence, although the combination of objective and subjective features gave slightly better results, we still got a pretty close estimation of the mood using only objective features.

Analyzing Mood Forecast

In order to analyze the mood forecast, we first built models with both subjective and objective features and then compared it with the models built only from objective features. We used the same set of learners as in the design phase analysis, and analyzed how the base learners performed on each data set. The main metric is again the mean absolute error (MAE) between the actual and the forecast value. In order to compare the performance of the two models – the one built with subjective and objective and the one with only objective data – we looked at the MAEs calculated for the 5 days, and computed the minimum and maximum values between them. This helped us determine the closest (minMAE) and the furthermost (maxMAE) predicted mood scores in each model.

We observed that the forecast mood values in the models with only objective data are closer to the actual reported mood scores. In other words, the mean absolute error in 5 days forecast is on average lower than the corresponding value in the models including both subjective and objective data. Figure 6 shows that both minimum and maximum MAEs are lower or equal in the objective models for at least 4 out of 6 patients.

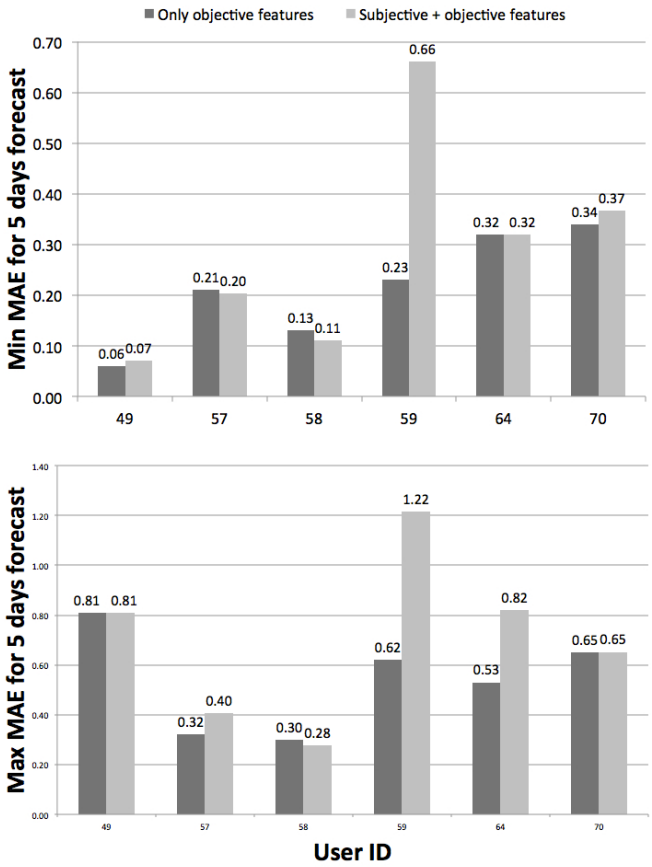

Figure 6. Minimum MAE and maximum MAE for the 5-day mood forecast. For most patients, both minMAE and maxMAE are lower in the 5 days forecast model with only objective features compared to the corresponding model built with both the objective and subjective features.

Feedback from patients and clinicians

When interviewing patient, they reported that the redesign had improved the overall usability and usefulness of the system. For example, they found the improved self-assessment form highly useful, especially the fact that they could add 1/2-point mood score. As stated by patient $P57$; *"the 1/2 point scale allows me to keep track of little details that mean a lot to me; these small changes can be early indicators that something is under way."* Also the personalization of the self-assessment form by adding additional individual features were reported to be key for the patients to manage their disease. However, the limitations in the scale were a limiting factor. As $P59$ stated; *"I would like to keep track of the number of cigarettes I smoke a day, but I cannot enter more that 10. It is annoying that you can't define your own scale."*

Both patients and clinicians appreciated the new objective sensor-based information available in MONARCA 2.0. Patients especially mentioned the benefits of the new objective features. For example, patients reported that it gave them an insight into the circumstances of their disease to see the visualization of the correlation between e.g. social interaction and mood. However, some patients were not completely convinced of the accuracy of the collected data. For example, $P64$ reported that his mobility level was constant whether he was staying in his apartment or traveling long distances with the train.

The structure of the impact factor screen was deemed intuitive by the patients, and the use of colors consistent with the visualization screen made it very coherent. The output fostered a process of reflection, which at times challenged the patients' own insight into their illness, informing them of interconnections which the patients were not aware of. Given the impact factors were built based on the patients' objective and subjective data, there were issues with getting meaningful output when the system did not work properly.

All patients reported that the Live Wallpaper was easy to comprehend and provided a subtle overview of the impact factors generated by the system. The patients did however express difficulties with interacting with the bubbles if their interface were filled with shortcuts and widgets. The phone is becoming a highly personalized tool for many users, and some of the patients reported that the MONARCA wallpaper did not allow them to have other things there. For example, $P58$ reported that she would like to have her newborn baby there, just like her friends.

The clinicians' reactions to the forecast were mixed in the beginning of the trial. They seemed to be hesitant to take actions based on an inferred forecast. For instance, when a patient's forecast pointed towards a depressive state, they did not know if they should call the patient, change their medication, or wait a few days to verify the actual change in the state. They ended up using the forecast as an of indicator to watch, but basically relied on their own clinical experience in handling patients.

CONCLUSIONS

We presented the refinements of the MONARCA system focusing on how this system was designed to convey a *disease insight* to patients. This was based on an approach where the system helps patients to identify patterns in their behavior as well as recognizing factors impacting their mental state. MONARCA 2.0 was tested in a small 6 months field deployment involving 6 patients. This evaluation showed that the system was stable and performed well in real use. The data collected was sufficient to identify the factors impacting the mood of patients, and the subsequent analysis showed that data features related to activity, stress, sleep, and phone usage were those with the highest correlation with the mood score. Patients and clinicians involved in the study reported a high degree of satisfaction with the usefulness and usability of the system.

Through the analysis of the objective sensing from patients' phones during the trial, we observed that by using only these features in our models, we were able to closely estimate the current and future mood state. We also observed that the objective features are strong indicators of the mood. It shows that the new components are a promising approach towards an increased disease insight among bipolar patients. However, we evaluated the system only with a small set of patients, and with an average phone uptime of 55.6%. It is therefore difficult to draw definitive conclusions and should be viewed as a promising initial field trial. Currently we are launching a larger study involving more patients as well as studying the clinical effect of using the MONARCA system.

ACKNOWLEDGEMENTS

This work has been done in close collaboration with a group of clinicians and patients from the Affective Disorder Clinic at the University Hospital of Copenhagen. MONARCA is funded as s STREP project under the FP7 European Framework program.

REFERENCES

1. Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput. 7*, 6 (Dec. 2011), 643–659.
2. Bardram, J. E., Frost, M., Szántó, K., Faurholt-Jepsen, M., Vinberg, M., and Kessing, L. V. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. 2627–2636.
3. Bardram, J. E., Frost, M., Szántó, K., and Marcu, G. 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, NY, USA, 2012), 21–30.
4. Bonato, P. Advances in wearable technology and its medical applications. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2010* (2010), 2021–2024.
5. Burns, M., Begale, M., Duffecy, J., Gergle, D., Karr, C., Giangrande, E., and Mohr, D. Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res 13*, 3 (2011).
6. Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., and Landay, J. A. Activity sensing in the wild: a field trial of ubifit garden. In *In Proc. ACM CHI 2008*, CHI '08, ACM (New York, NY, USA, 2008), 1797–1806.
7. Doryab, A., Frost, M., and Bardram, J. E. Mood prediction and behavior patterns in bipolar patients (analysis and design implications). *Under review for International Journal of Human-Computer Studies Special Issue: Emotional Wellbeing* (2013).
8. Frank, E. *Treating Bipolar Disorder: A Clinician's Guide to Interpersonal and Social Rhythm Therapy*. The Guilford Press, 2005.
9. Frank, E., Kupfer, D., Thase, M., Mallinger, A., Swartz, H., Fagiolini, A., Grochocinski, V., Houck, P., Scott, J., Thompson, W., and Monk, T. Two year outcomes for interpersonal and social rhythm therapy in individuals with bipolar i disorder. *Arch Gen Psychiatry 62* (2005), 996–1004.
10. Greenbaum, J., and Kyng, M., Eds. *Design at Work - Cooperative Design of Computer Systems*. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, 1991.
11. Kasckow, J., Zickmund, S., Rotondi, A., Mrkva, A., Gurklis, J., Chinman, M., Fox, L., Loganathan, M., Hanusa, B., and Haas, G. Development of telehealth dialogues for monitoring suicidal patients with schizophrenia: Consumer feedback. *Community Mental Health Journal* (2013), 1–4.
12. Lane, N. D., Choudhury, T., Campbell, A., Mohammod, M., Lin, M., Yang, X., Doryab, A., Lu, H., Ali, S., and Berke, E. 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 (2011).
13. LV Kessing, PK Andersen, P. M., and Bolwig, T. Recurrence in affective disorder. i. case register study. *British Journal of Psychiatry 172* (1998), 23–28.
14. Mamykina, L., Mynatt, E. D., and Kaufman, D. R. Investigating health management practices of individuals with diabetes. In *In Proc. ACM CHI 2006*, CHI '06, ACM (New York, NY, USA, 2006), 927–936.
15. Marcu, G., Bardram, J. E., and Gabrielli, S. A Framework for Overcoming Challenges in Designing Persuasive Monitoring Systems for Mental Illness. In *Proceedings of Pervasive Health 2011*, IEEE Press (2011), 1–10.
16. Matthews, M., Doherty, G., Sharry, J., and Fitzpatrick, C. Mobile phone mood charting for adolescents. *British Journal of Guidance & Counselling 36*, 2 (2008), 113–129.