Machine Learning for Smartphone-based Monitoring and Treatment of Unipolar and Bipolar Disorders

Busk, Jonas

Publication date: 2019

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Machine Learning for Smartphone-based Monitoring and Treatment of Unipolar and Bipolar Disorders

Jonas Busk

Kongens Lyngby 2019
Summary

Bipolar disorder is a common mental illness characterized by unusual changes in mood and energy and is regarded as one of the most important causes of disability worldwide. Smartphones provide a unique platform for unobtrusive disease monitoring and management and are almost constant companions to their users. By replacing traditional paper-based self-assessments with a smartphone-based system, users can unobtrusively collect and monitor their own data. Modern smartphones additionally enable pervasive collection of detailed objective data that can track a wide range of human behaviors relevant to monitoring mental illness. Digital data collection has the additional advantage of making data available for immediate analysis by humans and computers, which can support disease monitoring and treatment tasks in and between outpatient consultations at their treatment center. Automated analysis of smartphone-based data can potentially detect early warning signs and predict disease outcomes, which can facilitate early intervention and thus potentially mitigate severity of affective episodes and prevent costly hospitalizations.

The overall objective of the PhD has been to establish methods and algorithms for analysis of behavioral smartphone data from patients with bipolar disorder aiming at pattern recognition and prediction of recurring depressive and manic symptoms by applying data mining and machine learning techniques. This report presents a summary of research conducted during the author’s PhD studies and includes three research manuscripts. We show that by applying hierarchical Bayesian regression models we are able to forecast subjective mood up to seven days ahead based on short self-assessment histories. Using the same hierarchical modelling approach, we can produce daily estimates of clinical severity ratings of depression and mania from self-assessments. We also show how to utilize uncertainty in the estimated severity ratings to compute individual scores indicating risk of relapse. Finally, we show how simple features of objective smartphone data can discriminate between patients with bipolar disorder and healthy control individuals during different affective states.

Based on the current research, we are confident that predictive analysis based on data collected with smartphones has the potential to improve disease monitoring and treatment in bipolar disorder. To accomplish this goal, predictions must be accurate, interpretable and actionable. Modern machine learning techniques have proved unparalleled in revealing complex patterns and achieving high predictive accuracy in a wide range of domains. To unlock the potential of advanced methods and drive research forward, detailed datasets from large groups of patients are needed.


På baggrund af den fremlagte forskning er vi overbeviste om, at prædiktiv analyse baseret på smartphone-data har potentielle til at forbedre monitorering og behandling af bipolar lidelse. For at opnå dette skal prædiktioner være nøjagtige, forståelige og handlingsorienterede. Moderne maskinlæringsmetoder har vist sig at være uovertrufne til at afdække komplekse mønstre og opnå høj præcision på tværs af mange forskellige domæner. For at indfri potentialet fra avancerede metoder og drive forskningen fremad er der behov for at indsamle detaljerede datasæt fra store patientgrupper.
Preface

This thesis has been prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, DTU Compute, at The Technical University of Denmark, DTU, in fulfillment of the requirements for acquiring a PhD degree at The Technical University of Denmark.

The PhD was funded by the Innovation Fund Denmark through the RADMIS project and the Copenhagen Center for Health Technology (CACHET). The PhD project was supervised by Ole Winther, Professor, PhD, DTU Compute; and Jakob E. Bardram, Professor, PhD, DTU Health Tech.

The thesis consist of a summary report and a collection of research manuscripts. The work was conducted at the Section for Cognitive Systems, DTU Compute, between March 2016 and June 2019, and an external research stay at Telefónica Innovación Alpha, Barcelona, Spain in the spring of 2018.

Kongens Lyngby, June 14, 2019

Jonas Busk
I would like to thank my supervisor, Professor Ole Winther, and co-supervisor, Professor Jakob E. Bardram, for their advice and encouragement throughout this journey. I would also like to thank Maria Faurholt-Jepsen and Lars Vedel Kessing from the RADMIS research group for excellent collaboration and feedback. Additionally, I would like to thank Mads Frost and the entire team at Monsenso for welcoming me in their office. I have been privileged to work among brilliant and friendly colleagues at the Section for Cognitive Systems at DTU Compute.

A special thanks to Aleksandar Matic and the research team at Telefónica Innovación Alpha for hosting me on my external research stay and to my new friends in Barcelona.

I am grateful for the funding of this research by the Innovation Fund Denmark and the Copenhagen Center for Health Technology (CACHET) and for the unique opportunity it has given me to grow as a researcher and an individual.

Last but not least, a big thanks to my family and friends for all of their support. I would not have made it without you.
List of publications

Included in thesis


Not included in thesis

Maria Faurholt-Jepsen, Jonas Busk, Mads Frost, Maj Vinberg, Ellen M Christensen, Ole Winther, Jakob Eyvind Bardram, and Lars V Kessing. “Voice analysis as an objective state marker in bipolar disorder”. In: Translational psychiatry 6.7 (2016), e856.

Maria Faurholt-Jepsen, Mads Frost, Klaus Martiny, Nanna Tuxen, Nicole Rosenberg, Jonas Busk, Ole Winther, Jakob Eyvind Bardram, and Lars Vedel Kessing. “Reducing the rate and duration of Re-ADMISsions among patients with unipolar disorder and bipolar disorder using smartphone-based monitoring and treatment – the RADMIS trials: study protocol for two randomized controlled trials”. In: Trials 18.1 (2017), page 277.


# Contents

Summary i

Resumé (Danish) iii

Preface v

Acknowledgements vii

List of publications ix

Contents xi

1 Introduction 1
   1.1 Background and motivation ............................. 1
   1.2 The RADMIS project .................................... 2
   1.3 Contributions and outline ............................. 3

2 Materials and methods 5
   2.1 Data description ....................................... 5
   2.2 Statistical methods .................................... 11

3 Research 19
   3.1 Objectives ............................................. 19
   3.2 Principal results ...................................... 20

4 Discussion 27
   4.1 Main findings .......................................... 27
   4.2 Advantages ............................................ 28
   4.3 Limitations ............................................ 28
   4.4 Perspectives .......................................... 29
   4.5 Conclusions ........................................... 31

A Forecasting Mood in Bipolar Disorder from Smartphone Self-assessments with Hierarchical Bayesian Models 33
B Daily Estimates of Clinical Severity in Bipolar Disorder from Smartphone Self-Assessments 61

C Objective smartphone data as a potential diagnostic marker of bipolar disorder 83

D Stan Models 95

Bibliography 99
CHAPTER 1

Introduction

1.1 Background and motivation

Bipolar disorder, also known as manic-depressive illness, is a common and complex mental illness with an estimated prevalence of 1-2%, and is regarded as one of the most important causes of disability worldwide [1, 2]. It is characterized by unusual shifts in mood and energy with episodes of depression, (hypo)mania and mixed mood, intervened by periods of neutral mood (euthymia) [3]. Depressive episodes are characterized by sadness, decreased activity, disturbed sleep, poor concentration, loss of interest or enjoyment and feelings of hopelessness and worthlessness, while symptoms of mania include elevated or irritable mood, increased energy, inflated self-esteem, pressure of speech, impulsiveness and reduced need for sleep [4]. Suffering from bipolar disorder can result in an inability to cope with life’s ordinary demands and routines and is associated with a high degree of comorbidity, functional impairment and increased risk of suicide [5]. Among people with bipolar disorder, life expectancy is decreased with 8 to 12 years [6, 7]. According to the World Health Organization (WHO), about 60 million people are affected by bipolar disorder worldwide, while an estimated 300 million people are affected by unipolar disorder (depression) and the burden of depression and other mental health conditions is on the rise globally [4]. This imposes an increasing burden on health care services and there is therefore a need for better and more efficient treatment methods with a focus on outpatient treatment and prevention to mitigate major affective episodes and avoid costly hospitalizations.

Treatment of bipolar disorder today consists of a combination of pharmacological and psychological treatment accompanied by psychoeducation to teach strategies for coping with the disorder and managing symptoms [8]. There are currently no objective tests to assess bipolar disorder, so the diagnostic process relies on subjective information and clinical evaluation using rating scales [9]. Periodic clinical evaluations using clinical rating scales, such as the Hamilton Depression Rating Scale (HDRS) [10] and the Young Mania Rating Scale (YMRS) [11], are currently used as the golden standard for assessing the severity of depressive and manic symptoms. This subjective evaluation involves a risk of recall bias, individual observer bias and decreased illness insight mainly during affective episodes [12, 13, 14, 15, 16]. Therefore, in addition to clinical evaluations, patients may be asked to carry out daily self-assessments to track day-to-day behavior in between outpatient visits to the clinic. Frequent sampling of behaviors and experiences in a natural environment can help to reveal dynamic pro-
cesses, increase accuracy and reduce recall bias [17, 18]. It can also be useful for identifying and managing particular behaviors that influence the course of illness and thus provide increased disease insight.

Smartphones provide a unique platform for disease monitoring and management [19, 20] and are ubiquitous in modern society. In 2018 a median of 76% of adults across 18 advanced economies reported to have smartphones [21]. These are computationally powerful devices and nearly constant companions in people’s everyday lives. By replacing traditional paper-based self-assessments with a smartphone-based solution, users can unobtrusively collect and review their data. Data can also be made available to clinicians immediately, who can then monitor a group of patients in real-time and intervene if they observe any alarming behaviors. Modern smartphones additionally enable pervasive collection of detailed objective data from build-in sensors and usage logs that can be utilized to track a wide range of human behaviors, which are potentially relevant to mental illness monitoring, such as physical and social activity, social rhythm, mobility and sleep patterns [22, 23, 24, 25, 26, 27]. Discovery of objective measures of bipolar disorder symptoms that could assist in diagnosis and treatment would potentially be a major breakthrough.

Digital data collection with smartphones has the additional advantage of making data available for immediate automatic analysis by algorithms, that can help support disease monitoring and treatment tasks in and between outpatient consultations. Machine learning techniques can reveal patterns in data that would be otherwise impractical or impossible for humans to find. Machine learning additionally provides tools to derive statistical models from data, which can be used to make predictions about unobserved outcomes [28]. Automated analysis of data from patients with bipolar disorder can potentially help detect early warning signs and predict symptoms of depression and mania and thus support disease monitoring and provide additional disease insights. Accurate predictions could furthermore be utilized to facilitate early intervention strategies and thus reduce the severity of affective episodes and potentially prevent costly hospitalizations.

1.2 The RADMIS project

The PhD is part of the RADMIS project: Reducing the rate and duration of RADMISsions among patients with unipolar and bipolar disorder using smartphone-based monitoring and treatment [29]. The two ongoing RADMIS randomized clinical trials (RCTs) aim to investigate whether the use of a smartphone-based monitoring and treatment system, including an integrated feedback loop between patients and clinicians, reduces the rate and duration of re-admissions more than standard treatment in unipolar disorder and bipolar disorder. Additionally, the study aims to investigate if the use of a smartphone-based system reduces the severity of clinically rated affective symptoms and the number of affective episodes in unipolar disorder or bipolar disorder, respectively. The specific objective of the PhD has been to establish methods and algorithms for analysis of behavioral data aiming at pattern recognition.
and disease prediction of recurring depressive and manic symptoms by applying data mining and machine learning techniques.

The RADMIS research group is comprised of partners from Psychiatric Center Copenhagen (Rigshospitalet), Technical University of Denmark (DTU) and Monsenso ApS, who since 2010 have collaborated closely on the design, implementation, and clinical evaluation of data-driven smartphone-based treatment of mental health. Prior work done by RADMIS researchers has developed and tested a smartphone-based system for the treatment of bipolar disorder (the MONARCA system) [30, 31, 32].

1.3 Contributions and outline

This summary report is based on research conducted during the PhD studies and includes three research manuscripts (see the full list of publications in the front-matter section) enclosed in appendices A, B and C, respectively:

- Article A, titled “Forecasting Mood in Bipolar Disorder from Smartphone Self-assessments with Hierarchical Bayesian Models”, examines the feasibility of forecasting daily subjective mood based on daily self-assessments collected from patients with bipolar disorder with a smartphone-based system. The PhD candidate performed all analyses presented in the manuscript and prepared the first draft. The manuscript has been submitted for publication in the Journal of Medical Internet Research.

- Article B, titled “Daily Estimates of Clinical Severity in Bipolar Disorder from Smartphone Self-Assessments”, studies the feasibility of producing daily estimates of clinical severity ratings based on smartphone self-assessments from patients with bipolar disorder. The PhD candidate performed all analyses presented in the manuscript and prepared the first draft. The manuscript is unpublished.

- Article C, titled “Objective smartphone data as a potential diagnostic marker of bipolar disorder”, investigates the use of objective smartphone data reflecting behavioural activities to classify patients with bipolar disorder and healthy control individuals. The PhD candidate prepared smartphone data for analysis, designed and performed the classification analysis and drafted the description of the predictive classification analysis presented in the manuscript. The article was published in the Australian & New Zealand Journal of Psychiatry.

The report is organized in 4 chapters: Chapter 2 presents data collection and statistical methods central to the research, and does not represent any original work by the author. Chapter 3 presents the main objectives and principal results of the three included research articles. Chapter 4 provides a brief discussion of the overall discoveries presented in Chapter 3, including advantages, limitations, and future perspectives of the research. Finally, the chapter presents conclusions and closing remarks.
CHAPTER 2

Materials and methods

2.1 Data description

Data collection in the RADMIS RCT [29] commenced in 2016 and is still ongoing at the time of writing this report. Therefore the PhD work have primarily relied on data from previous trials lead by RADMIS researchers and in particular the MONARCA II RCT (2014-2018) [32] investigating the effect of smartphone-based monitoring in patients with bipolar disorder. The MONARCA II RCT included patients with a diagnosis of bipolar disorder who had previously been treated at the Copenhagen Clinic for Affective Disorder for a follow-up period of 9 months. The patients were randomized to either using a smartphone-based monitoring system for daily self-monitoring (the intervention group) or to treatment as usual (the control group). Data from the intervention group (N=84) constituted the primary data basis for the analyzes presented in research articles A and B and is described in detail in the following sections.

In the RADMIS trials, data is collected with smartphones using the Monsenso system, which was designed and in close collaboration with clinicians and in an interactive process with patients with unipolar and bipolar disorder. The Monsenso system is developed by Monsenso ApS and is based on the MONARCA system, which was previously developed and tested by RADMIS researchers and applied in the MONARCA II trial [30, 31]. The Monsenso smartphone application is highly configurable and supports user input of daily self-assessments, setting of reminders, delivery of custom push notifications. It is capable of pervasively collecting a wide range of objective data from the device sensors and usage logs and also provides visualizations of the user’s data to support disease monitoring and management. The system includes an integrated feedback loop between patients and clinicians, that enables clinicians to monitor data from a group of patients in real-time through a secure web application. Additionally, the system is designed to support smartphone-based Cognitive Behavioral Therapy (CBT) content [29]. A picture of the Monsenso smartphone application is presented in Figure 2.1.
Figure 2.1: The Monsenso smartphone application used for data collection in the RADMIS trial. The picture shows the screens for providing a self-assessment and the daily mood score. The Monsenso system is based on the MONARCA system, which was developed and tested by RADMIS researchers. The image is used with permission from Monsenso ApS.

2.1.1 Clinical ratings

During the follow-up period of the MONARCA II RCT [32], study participants were periodically evaluated by trained psychiatrists on clinical rating scales for measuring the severity of depression and mania using the Hamilton Depression Rating Scale (HDRS) [10] and Young Mania Rating Scale (YMRS) [11]. Each rating scale consists of a series of items that are scored and added up to summarize the current severity of symptoms with higher scores indicating more severe symptoms. A score of 13 or more on either scale is classified as a depressive or manic episode, respectively. It is possible to score high on both scales at the same time constituting what is known as a mixed episode. Clinical ratings with HDRS and YMRS are treated as valid on the day of the rating as well as the 3 preceding days, thus each rating is attributed a total of 4 days for the purpose of analysis.

2.1.2 Self-assessments

In addition to periodic clinical evaluations, study participants assigned to the intervention group were instructed to complete daily self-assessments on their smartphones through the Monsenso application. The self-assessment consisting of a questionnaire including the items listed in Table 2.1 where four of the items (mood, activity, sleep and medicine) were mandatory. Specifically, mood was scored on a scale from -3, to 3 including -0.5 and 0.5, where negative values indicate various degrees of depres-
### 2.1 Data description

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity*</td>
<td>Level of physical activity.</td>
<td>-3 to 3</td>
</tr>
<tr>
<td>Alcohol</td>
<td>Alcoholic beverages consumed.</td>
<td>0 to 10+</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Level of anxiety.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Irritable</td>
<td>Level of irritability.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Cognitive difficulty</td>
<td>Level of cognitive discomfort.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Medicine*</td>
<td>Medicine adherence</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Mixed mood</td>
<td>Experienced mixed mood.</td>
<td>0 to 1</td>
</tr>
<tr>
<td>Mood*</td>
<td>Experienced mood.</td>
<td>-3 to 3</td>
</tr>
<tr>
<td>Sleep*</td>
<td>Hours of sleep.</td>
<td>#hours</td>
</tr>
<tr>
<td>Stress</td>
<td>Level of stress.</td>
<td>0 to 2</td>
</tr>
</tbody>
</table>

Table 2.1: The ten items of the daily self-assessment questionnaire in the MONARCA II RCT. The four items marked with an asterisk (*) were mandatory. The table is adapted from article A.

Assessment, positive values indicate mania and zero indicates neutral mood (euthymia). In addition to these items, users could save a daily note along with the self-assessment. The smartphone application enabled setting and receiving reminders and users were allowed to provide self-assessments retrospectively for up to 2 days in case forgot to do it daily. Additionally, users could track their own data history in the application.

#### 2.1.3 Sensor data

In addition to self-monitoring, the smartphone application was setup to collect objective data from sensors and usage logs in the device. The data included location coordinates, screen on-off events, step counts, call and SMS logs, battery information and data from light and proximity sensors. These data were collected at different intervals and stored in a raw, semi-structured format. Therefore, they required additional preprocessing and aggregation to for example a daily level before being suitable for interpretation and analysis (see Section 2.1.6).

#### 2.1.4 Voice data

The smartphone application was also configured to extract voice features from phone conversations. A large number of voice features describing characteristics of the voice such as pitch and energy were extracted from the raw audio signal directly on the smartphone device using the openSMILE toolkit [33] and the extracted features were then transmitted to a secure server. Changes in speech have been suggested as important measures of depression and mania in bipolar disorder and voice features are therefore an interesting subject for analysis. The use of voice data to classify depression and mania in bipolar disorder was studied in the author’s master’s thesis and later published [34], but is not included in this report.
2.1.5 Data processing

The Monsenso system managed collecting and transferring data from the participants’ smartphones to a central database on a secure server. The data was then exported from the database in a raw, semi-structured document format and processed on a computer with an encrypted hard drive located in a locked office. In order prepare the data for analysis, it was preprocessed, transformed and stored in a table-like structure by applying a series of custom Python scripts. The subjective self-assessment data was extracted and stored without any further processing, while the objective sensor and phone usage data required more intricate preprocessing (see Section 2.1.6). The clinical ratings were provided by the clinic in a spreadsheet format and merged with the smartphone data to create a coherent dataset. Additional processing and analyzes were performed using the Python programming language and Jupyter Notebooks with data analysis packages NumPy [35], Pandas [36], the Scikit-learn machine learning library [37], the XGBoost Python package [38], probabilistic programming library Stan [39] and visualization package Matplotlib [40].

2.1.6 Behavior extraction from objective data

In order to be suitable for analysis, the raw objective smartphone sensor and usage data needed to be preprocessed and aggregated to a daily level. The aim of preprocessing was to generate features describing relevant human behaviors, such as social and physical activity, smartphone usage, mobility and sleep patterns.

As features describing social activity, the number of incoming and outgoing SMS messages were counted. The number of incoming, outgoing and missed phone calls were also counted and simple statistics describing the distribution of call durations were computed, including minimum, maximum, mean and total duration in minutes. The step counter (pedometer) was used as a measure of physical activity. Because steps were counted cumulatively and occasionally reset to zero, some processing was required to compute step counts in daily intervals. As a measure of smartphone usage, the number of screen-on events where counted and screen-on duration was measured by computing the times between each screen-on event and the next screen-off event. Then statistics describing the distribution of screen-on durations were computed, including minimum, maximum, mean and total duration in minutes.

Probably the richest source of objective smartphone data is location, which was used to compute mobility patterns, inspired by the methods presented in [25, 41]. Location measured by a Global Positioning System (GPS) consists of points described by coordinates accompanied by a timestamp and provide on their own little meaning. Therefore, it is useful to identify spatio-temporal clusters of points representing stops, and sequences of points between the stops representing paths of travel. The length of paths can then be computed and summed to represent the distance traveled. Stops can further be clustered to form locations that are visited several times and time spent at each location can be computed to identify important locations of interest, such as
home and work. Thus, a wide range of behaviors can be extracted from location data, representing mobility but also social rhythm.

Several methods for estimating sleep from smartphone data exist [42, 41], but it remains very difficult without access to additional information. Probably the simplest method is to identify periods of inactivity (no interactions with the smartphone) during normal sleep hours and use that as a proxy for sleep duration. However, this method leads to biased results for people who do not frequently interact with their smartphone and does not work in case of odd circadian rhythms. More sophisticated methods aim to predict sleep duration directly from objective features or estimate bedtime and wake-up time and use that to infer sleep duration.

Simple objective features based on SMS messages, phone calls and screen on-off events were used to classify patients with bipolar disorder and healthy control individuals in article C. The use of more complex objective features has not yet been explored with the MONARCA II dataset and is an interesting future research project.

2.1.7 Challenges

Data collection from smartphones does not come without challenges. Probably the biggest issue was missing data. It is not uncommon for users to forget to or refrain from answering the daily self-assessments. Even with automatic reminders in the app and the option to answer retrospectively up to two days, most users fail to provide complete self-assessment histories for extended periods of time. Furthermore, not all self-assessment items are mandatory, and consequently some self-assessments are incomplete. The objective sensor data is also sometimes missing. This may be because the smartphone was turned off or the user decided to disable one or more sensors, for example to make the battery last longer. Often the sensor data is not sampled with regular frequency but is only collected when there is some activity on the smartphone. For example, screen on-off events are only logged when they happen and location data is only recorded if the location of the device changes. Thus, it can be difficult to determine if sensor data is missing by mistake or because there was no interaction with the smartphone. Additionally, users cannot be expected to carry the smartphone on their person all the time. It is quite normal for people to leave their smartphone on a table or in a bag, at home or at work. It can also be impractical to bring the smartphone to certain activities such as sports. In this case the sensor data is not actually missing, but it does not capture all the user’s activities either. Therefore, we must assume that the objective sensor data only provides a partial picture of the user’s behavior – we simply don’t know what we don’t know. Both subjective and objective data might also be missing because of software error in either the smartphone application or on the server side, or because of network issues (in the latter case, the application is configured to start syncing data when it regains connectivity). When the data is missing due to technical problems, it is often missing in a systematic way that is easy to recognize. The missing data can however not always be recovered, even after the issue is resolved.
A different kind of challenge concerns the subjectivity of the self-assessments and clinical ratings. Users can be biased in their answers and for example rate their mood or physical activity in different ways relative to their own experience and normal behavior. Another well-known phenomenon is recall bias, causing individuals to have difficulties remembering previous events or emotions accurately. Depressive or manic symptoms can affect recall and lead to biased answers and decreased illness insight. Additionally, individuals may experience a certain emotional state for an extended period, wherein other emotions will be transient, but can affect answers at a given point in time. Individual differences can be accounted for in statistical analysis to some extent, but relying on subjective information will inevitably result in some noise in the data.
2.2 Statistical methods

2.2.1 Supervised machine learning

In supervised machine learning the goal is to learn a mapping from inputs $x$ to outputs $y$ given a training set of input-output pairs, $\{(x_i, y_i)\}$ [28]. Each input $x_i$ is typically a vector of numerical values representing a single observation, such as a basic measurement or a complex structure, like an image or a string of text. Similarly, the output $y_i$ can take many forms, but many machine learning methods assume either a categorical or continuous output. When $y_i$ is a categorical variable with values from a finite set of categories or classes the machine learning problem is known as classification and when $y_i$ is a continuous value the problem is known as regression. If $y_i$ is a discrete variable with a relative ordering of the possible outcomes, the problem is sometimes referred to as ordinal regression, which can be seen as an intermediate between classification and regression.

One way of formalizing supervised machine learning is through function approximation. Assuming there exist an unknown function $f$ such that $y = f(x)$, the objective of learning is to estimate the function $f$ from a training set of known examples of $x$ and $y$:

$$\hat{y} = \hat{f}(x).$$

(2.1)

where we use the hat symbol to denote an estimate. The goal of obtaining $\hat{f}$ often is to make predictions on new observations of $x$, known as generalization, with as little error as possible. To generalize to new examples, the estimated function must make assumptions about the mapping from inputs to outputs (such as similar inputs have similar outputs) also known as inductive bias. The expected generalization error can be estimated by computing the prediction error on an independent test set not used for fitting the model.

Many different machine learning methods of varying complexity exist for estimating $f$ and which one(s) to choose depends on the nature of the problem as well as additional requirements, such as interpretability, time and resources. Besides supervised learning there also exists other types of machine learning, such as unsupervised learning and reinforcement learning, but they are outside the scope of this thesis. Please see [28] for more information about machine learning in general.

2.2.2 Bayesian inference

It is often useful to reason about the uncertainty in estimated values in terms of probability theory [28]. We can assume a probabilistic view of supervised learning by considering the outputs $y$ as samples from a probability distribution representing the conditional probability of all possible values of $y$ given the inputs $x$:

$$y \sim p(y|x, \theta).$$

(2.2)
Here we are also explicitly conditioning on the unknown parameters $\theta$ of the sampling distribution. The objective of learning then becomes to estimate $\theta$ as closely as possible given the observed data.

By applying the rules of probability, conclusions about the unknown parameters can be inferred in terms of probability statements, conditional on the observed values of $y$ and implicitly conditional on any observed values of $x$. This process is also known as Bayesian inference. To make inference about the unknown parameters, $\theta$, we begin with a probability model defining the joint probability distribution of $\theta$ and the observed data, $y$, which is equal to the product of the sampling distribution and the prior distribution of the parameters:

$$p(\theta, y) = p(y|\theta)p(\theta).$$

Now we can simply apply Bayes’ rule to infer the posterior distribution of $\theta$ given the observed data, $y$:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta) \, d\theta}. \tag{2.3}$$

Since the term $p(y)$ does not depend on $\theta$, it can be considered as a constant and ignored, resulting in the unnormalized posterior distribution, which is proportional to the posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{2.4}$$

The observed data, $y$, affect the inference only through the term $p(y|\theta)$, which here is known as the likelihood function and is considered as a function of $\theta$ with fixed $y$. Note how with no prior opinion about the value of $\theta$, i.e. with a uniform (constant) prior, $p(\theta) \propto 1$, the posterior distribution is simply proportional to the likelihood of the data given the parameters.

As mentioned above, the goal of supervised learning is to make predictions on unobserved instances of $y$, also known as predictive inference. The probability distribution of an unobserved outcome, $\tilde{y}$, given the observed data, $y$, is defined by the posterior predictive distribution:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y) \, d\theta. \tag{2.5}$$

What this says is that the probability of $\tilde{y}$ given $y$ is a weighted average of the probability of $\tilde{y}$ given all plausible values of $\theta$ given $y$, essentially averaging over all possible values of $\theta$, why this is also called Bayes model averaging.

As mentioned earlier in this section, to apply Bayesian inference we need to define a joint probability model of $\theta$ and $y$. This includes specifying the sampling distribution (likelihood function), $p(y|\theta)$, relying on assumptions about the process that generates $y$, and the prior distribution, $p(\theta)$, indicating our prior belief about $\theta$ before observing the data, which can be informative or not.
2.2 Statistical methods

2.2.3 Hierarchical Bayes for multi-task learning

When analyzing multiple related sets of data, a simple approach is to pool all the data together and learn one common model. A major drawback of this method is the risk of disregarding important differences among the datasets. The other extreme approach is to model each dataset separately. However, you risk missing commonalities that could otherwise help inform the models. When it is reasonable to assume that models are similar across the datasets, it makes sense to fit all the models in parallel while using a shared representation that allows for partial pooling. Especially when the individual datasets are too small to learn meaningful models separately. This idea of improving generalization by utilizing similarities of related tasks as an inductive bias is known as multi-task learning (MTL) [44]. While several MTL methods exist, a popular approach is to use hierarchical Bayes.

In a hierarchical Bayesian model, each group (task), indexed by \( j = 1 : J \), has its own set of parameters, \( \theta_j \), which is considered a sample from a common population distribution with hyperparameters \( \phi \), tying the groups together [43]. This yields the posterior distribution of the hierarchical model:

\[
p(\theta, \phi|y) \propto p(y|\theta)p(\theta|\phi)p(\phi) = \prod_{j=1}^{J} p(y_j|\theta_j)p(\theta_j|\phi)p(\phi).
\]  

(2.6)

In this hierarchical structure, data from each group can contribute to estimating the (hyper)parameters of the population distribution and groups with small sample size can borrow statistical strength from other groups with larger sample size. An illustration of a hierarchical model is presented in Figure 2.2.

The variance of the population distribution determines the amount of pooling in the hierarchical model by controlling how similar the group-level parameters are. A
population distribution with zero variance corresponds to complete pooling while a noninformative uniform distribution leads to a model with no pooling at all. An informative population distribution will promote parameters to be similar, but not identical, corresponding to partial pooling and thereby acting as a means of regularization. By inferring the variance of the population distribution as part of the model, the appropriate amount of pooling can be determined by the data.

There are two posterior predictive distributions of interest in the hierarchical model. One concerns unobserved outcomes for a known group, $j$, while the other concerns unobserved outcomes for an unknown group from the same population. In the first case, predictions are based on posterior samples of the group’s parameters, $\theta_j$. In the second case, predictions are based on a new set of group parameters, $\tilde{\theta}$, sampled from the population distribution given posterior samples of the hyperparameters, $\phi$.

### 2.2.4 Hierarchical linear models

Linear regression is a simple and widely used model in statistics and supervised machine learning [28]. The ordinary linear regression model assumes a linear relationship between inputs and outputs, $y(x) = \alpha + \beta^T x + \epsilon$, with a normal distributed error, $\epsilon \sim \text{Normal}(0, \sigma^2)$. Using the probability notation from earlier, this is equivalent to the following sampling distribution:

$$y | x, \theta \sim \text{Normal}(y | \alpha + \beta^T x, \sigma^2), \quad (2.7)$$

where $\theta = (\alpha, \beta, \sigma^2)$ are the parameters of the model (the intercept, the regression weights and the variance).

We can extend the hierarchical model with linear regression at the group level by using a normal linear model for the likelihood of each group, $j$:

$$y_j | \alpha_j, \beta_j, \sigma^2 \sim \text{Normal}(y_j | \alpha_j + \beta_j^T x_j, \sigma^2), \quad (2.8)$$

and common population distributions:

$$\alpha_j | \mu_\alpha, \tau_\alpha \sim \text{Normal}(\alpha_j | \mu_\alpha, \tau_\alpha) \quad (2.9)$$

$$\beta_j | \mu_\beta, \tau_\beta \sim \text{Normal}(\beta_j | \mu_\beta, \tau_\beta), \quad (2.10)$$

where the unknown hyperparameters, $\phi = (\mu, \tau^2)$, (the population means and variances) and the sample variance, $\sigma^2$, are assigned independent normal priors. In this model, the population variances, $\tau^2$, determine the amount of pooling across groups, which is inferred from the observed data. The hierarchical linear regression model is presented as a Bayesian network in Figure 2.3.

Ordered logistic regression is an ordinal regression method for modelling a discrete variable with a relative ordering of the possible outcomes [28]. For $y$ belonging to one of $K$ ordered categories, we can interpret $y$ as an incomplete measurement of a latent continuous linear predictor, $z = \beta^T x$. By introducing a sequence of $K + 1$ ordered
Figure 2.3: Bayesian network showing the hierarchical linear regression model. Individual regression intercept $\alpha_j$ and weights $\beta_j$ are drawn from population distributions parameterized by $\mu_\alpha$, $\tau_\alpha$ and $\mu_\beta$, $\tau_\beta$. This allows the model to account for individual differences while constraining individual parameters to be similar across the population. The diagram is reused from article A.

Cutpoints, $c$, we assign $y = k$ with high probability if $c_{k-1} < z < c_k$, assuming $c_0 = -\infty$ and $c_K = \infty$. This corresponds to the following sampling distribution:

$$y|x, \theta \sim \text{OrderedLogistic}(y|z, c),$$

where $\theta = (\beta, c)$ are the parameters of the model (the regression weights and the sequence of cutpoints).

Like the hierarchical linear regression model presented above, we can construct a hierarchical ordered logistic regression model by assigning a set of parameters, $\theta_j$, to each group with common population distributions:

$$\beta_j|\mu_\beta, \tau_\beta \sim \text{Normal}(\beta_j|\mu_\beta, \tau_\beta)$$
$$c_j|\mu_c, \tau_c \sim \text{Normal}(c_j|\mu_c, \tau_c), \quad \text{s.t.} \quad c_{k-1} < c_k,$$

where the unknown hyperparameters, $\phi = (\mu, \tau^2)$, (the population means and variances) have normal priors.

In practice, we reparameterized the hierarchical models using an equivalent, non-centered parameterization [45]:

$$\theta_j = \mu + \theta_j' \tau$$
$$\theta_j' \sim \text{Normal}(0, 1),$$

where each $\theta_j'$ is sampled independent of $\mu$ and $\tau$, and $\theta_j$ is computed in a deterministic manner, making sampling more efficient especially for small values of $\tau$. We found this method worked well with our data and made the models converge much faster.
2.2.5 Stan

The hierarchical Bayesian models were fitted using the statistical modeling platform Stan [39]. Stan provides tools for specifying Bayesian models in a probabilistic programming language and performing inference in the models with Hamiltonian Monte Carlo sampling. The key steps of the Stan algorithm after specifying the model are: providing data and model input; computation of the posterior density; parameter tuning and sampling; and convergence monitoring and summaries [43]. The outputs of the inference are posterior samples of model parameters and generated quantities along with an inferential summary for monitoring convergence of the model. Generally, the models were fitted using 4 sampling chains and 5,000 iterations, where the first half was warm-up, resulting in 10,000 posterior samples, which we found to be sufficient for the models to converge. Stan code for the hierarchical Bayesian linear and ordinal regression model is included in Appendix D.

2.2.6 Time series forecasting

Forecasting can be described as the task of predicting the future based on relevant information from the past and the present [46]. Hence it differs from predicting outcomes that occur close in time to the observed predictor variables, which can instead be called detection. Forecasting is only feasible if it is reasonable to assume that there is a relation between the available history of information and the future outcome of interest. In the context of disease monitoring, forecasting is interesting as accurate forecasts of symptoms ahead of time could be utilized as a tool to enable early intervention and possibly help prevent relapse.

Figure 2.4: Time series forecasting is the task of predicting future outcomes given relevant information from the past and the present. The window size, $w$, defines the size of the history of observed predictor variables and the horizon, $h$, is how far into the future the target variable is predicted. The illustration is reused from article A.
In time series forecasting, data is observed at discrete time intervals and the goal is to predict outcomes at later time steps based on a history of observed data. A typical forecasting task is illustrated in Figure 2.4. We denote by \( w \) the size of the data history used in the forecast and by \( h \) the horizon of how far into the future the target variable is predicted. Many different strategies for time series forecasting exist \([47, 46] \). Among the simplest methods is to apply the most recently observed outcome as the forecast. An alternative is to use the mean of previously observed data or the recent history as an estimate. These simple methods can be considered naive forecasts and act as useful baselines for evaluating more complex models.

A widely-used strategy relies on identifying time-dependent patterns, such as seasonality and trend, in the observed data and extrapolate these into the future. This approach can be effective when sufficient data is available to reveal such patterns and when making long-term forecasts. Another strategy is to apply standard supervised learning methods from the machine learning literature to predict the outcome of interest based on relevant, available information, such as previous observations of the target variable along with additional predictor variables. Applying this approach, we can describe a forecast model using the notation of supervised learning:

\[
\hat{y}_{t+h} = \hat{f}(y_t, \ldots, y_{t-w-1}, x_t, \ldots, x_{t-w-1})
\]  

(2.16)

where \( t \) denotes the time step. A benefit of this strategy is that it easy to include additional predictor variables, however, it might not be as good at capturing long-term time-dependent patterns and is therefore often better suited for short-term forecasts.

To properly evaluate the predictive performance of a time series forecast, it is important to only use information available in advance of the forecast time to fit the model \([46] \). Thus, the training error or the error on a randomly sampled test set are not good estimates of the generalization error. This must be considered when designing the experiment used to evaluate the forecast model. Time series cross-validation is a method specifically for evaluating forecasts. In time series cross-validation the data is partitioned into a sequence of consecutive test sets. For each test set, the corresponding training set consists of data observed prior to the test set. Thus, no future data is used to fit the forecast model and the training sets grow larger over time. Finally, the cross-validation error is computed across all the test sets. Since it is not feasible to construct reliable forecasts on very small training sets, the earliest test sets are sometimes left out of time series cross-validation.
3.1 Objectives

The overall objective of the PhD research has been to establish methods and algorithms for analyzing behavioral smartphone data from patients with bipolar disorder and apply machine learning techniques to predict symptom outcomes with the aim of improving disease monitoring and treatment tasks. Among the core research areas have been to utilize daily smartphone based self-assessments, including self-reported mood scores, to develop an online mood forecast as well as to relate self-reported information to clinical severity ratings of depression and mania. Another important task has been to analyze automatically generated smartphone sensor and usage data to generate features reflecting behavioral activities and relate these to disease outcomes of bipolar disorder. Throughout the research, there has been emphasis on applying interpretable methods, as traceability and interpretability are central to adopting new technology in the health industry. Assuming a Bayesian data analysis approach [43] provides methods for reasoning about uncertainty and supports interpretable models such as linear regression. It also provides methods for analyzing data in a hierarchical structure, such as repeated measurements from a group of similar subjects, and has therefore been a common theme throughout the work. This chapter summarizes research conducted during the PhD and presents the principal results of the included research articles enclosed in appendices A, B and C.

In article A we studied the feasibility and technical foundation of forecasting self-reported mood scores based on daily smartphone self-assessment data by applying hierarchical regression models to account for individual differences. We hypothesized that an accurate, online mood forecast can help improve disease monitoring and enable early intervention, and thus help reduce the severity of affective episodes and prevent hospitalizations. Several prior studies have utilized smartphone data to detect current mood [48, 25, 49, 50, 51, 52, 23], but we found that only a few studies examined forecasting future mood [53, 54]. We found that our proposed methods were able to forecast mood several days ahead with low error compared to pooled and separate baseline models. The main results of article A are summarized in Section 3.2.2.

The aim in article B was to study the feasibility of generating daily estimates of clinical severity ratings of depression and mania based on smartphone self-assessment data. We hypothesized that automatic estimates of clinical symptom severity and risk of relapse scores can support disease monitoring and treatment tasks between
outpatient visits to the clinic. We found that our approach of applying hierarchical regression models was able to estimate clinical severity ratings with low error compared to baseline methods and we demonstrated how samples from the posterior predictive distribution of the Bayesian model can be utilized to compute individual risk of relapse scores. To our knowledge, this is the first study that aims to predict scores of clinical severity from smartphone self-assessment data in patients with bipolar disorder. The main results of article B are summarized in Section 3.2.3.

The objective of the pilot study presented in article C was to investigate if objective smartphone sensor and usage data representing different user behaviors can be utilized to discriminate between patients with bipolar disorder and healthy control individuals and thus potentially represent objective diagnostic markers in bipolar disorder. The study included classification of patients with bipolar disorder and healthy control individuals. Interestingly, the study found differences in objective smartphone data between the groups overall and between different affective states. Section 3.2.4 summarizes selected results from article C.

3.2 Principal results

3.2.1 Descriptive statistics

The analyzes in articles A and B were based on data collected during the MONARCA II RCT [32]. The intervention group of the RCT included a large population of patients (N = 84) diagnosed with bipolar disorder, consisting of of 61.9% females and ages between 21 and 71 years (mean = 43.1, SD = 12.4). A total of 280 clinical evaluations an 15,975 daily smartphone-based self-assessments including subjective mood scores were collected during the 9 month follow-up period of the RCT along with smartphone sensor and usage data. The participants in the study presented primarily with low severity of depressive and manic symptoms during the follow-up period, resulting in low severity ratings on the Hamilton Depression Rating scale (HDRS) [10] and the Young Mania Rating Scale (YMRS) [11]. The mean HDRS rating was 7.56 (SD = 6.29) and 20.4% of ratings were greater than or equal to a threshold value of 13. The mean YMRS rating was 2.85 (SD = 4.17) and 5.0% of ratings were greater than or equal to a value of 13. Similarly, the self-reported mood scores had a mean score of -0.14 (SD = 0.48) and most mood scores were close to zero indicating euthymia (-1 < mood < 1 = 89.64%) with only few values indicating depression (mood ≤ -1 = 8.68%) and even less values indicating mania (mood ≥ 1 = 1.68%).

With each clinical rating being valid for up to 3 days prior to the evaluation, there were 764 observations of clinical severity ratings associated with daily self-assessments. The joint and marginal distributions of the clinical severity ratings (HDRS and YMRS) and self-reported mood scores are presented in Figure 3.1. The left plot clearly shows a negative correlation between HDRS ratings and mood (r = -0.40, P < .001), meaning that high HDRS ratings often coincide with a feeling of depressed
3.2 Principal results

Figure 3.1: Joined and marginal distributions of clinical ratings and self-reported mood scores. Data is expected to primarily occupy the areas of the scatter plots with white background: a negative mood score is expected to indicate a high HDRS rating, a positive mood score is expected to indicate a high YMRS rating, and HDRS and YMRS ratings are rarely high at the same time, indicating mixed mood. The plot is reused from article B.

or neutral mood. Similarly, the center plot shows that YMRS ratings and mood have a positive correlation \((r = 0.22, P < 0.001)\), meaning that high YMRS ratings often coincide with a feeling of elevated or neutral mood. The plot to the right shows that HDRS and YMRS have a weak positive correlation \((r = 0.13, P = 0.02)\) and were only rated high at the same time in a few instances, indicating mixed mood.

3.2.2 Mood forecast

We applied pooled, separate and hierarchical regression models to forecast daily mood based on a history of self-reported data (article A). The optimal window size of the data history used in the forecast was identified in a cross-validation experiment, evaluating one-day \((h = 1)\) forecast models with window sizes from 1 to 7 days \((w = 1 : 7)\). The best result in terms of root mean squared error (RMSE) was achieved using \(w = 4\) days, which was then used in subsequent analyzes.

The results of two time series cross-validation experiments comparing forecast models with \(h = 1\) and \(w = 4\) are presented in Table 3.1. The leave-all-out cross-validation experiment (left column) simulates a scenario where a group of subjects start monitoring at the same time, and training sets of increasing size are pooled across subjects. The leave-one-out experiment (right column) simulates a situation where each subject joins monitoring with limited training data which is pooled with data from an existing population (see Appendix A for more details on the experiments). In both experiments the proposed hierarchical models achieved the best results, demonstrating the advantage of partial pooling of the data to include information from the population while accounting for individual differences. The pooled models generally performed better than the separate models in these experiments, most likely because
Table 3.1: Results of mood forecast leave-all-out time series cross-validation (left) and leave-one-out time series cross-validation (right). The hierarchical Bayesian linear regression model achieves the best results. The pooled models are better than the separate models overall. The table is reused from article A.

<table>
<thead>
<tr>
<th>Model</th>
<th>Leave-all-out</th>
<th>Leave-one-out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$ ↑</td>
<td>RMSE ↓</td>
</tr>
<tr>
<td>Last observed</td>
<td>0.342</td>
<td>0.376</td>
</tr>
<tr>
<td>Pooled mean</td>
<td>-0.007</td>
<td>0.465</td>
</tr>
<tr>
<td>Pooled ridge</td>
<td>0.450</td>
<td>0.344</td>
</tr>
<tr>
<td>Pooled XGBoost</td>
<td>0.455</td>
<td>0.342</td>
</tr>
<tr>
<td>Separate mean</td>
<td>0.213</td>
<td>0.412</td>
</tr>
<tr>
<td>Separate ridge</td>
<td>0.345</td>
<td>0.375</td>
</tr>
<tr>
<td>Separate XGBoost</td>
<td>0.302</td>
<td>0.388</td>
</tr>
<tr>
<td>Hierarchical Bayesian linear</td>
<td>0.511</td>
<td>0.324</td>
</tr>
<tr>
<td>Hierarchical Bayesian ordinal</td>
<td>0.495</td>
<td>0.330</td>
</tr>
</tbody>
</table>

the individual training sets were too small to fit meaningful models able to generalize to new data. Overall, the regression models performed better than naïve baseline models, showing the benefit of including information from additional subjective predictor variables in the mood forecast.

Figure 3.2 shows results from forecasting mood 1 to 7 days ahead using data histories of 4 days. As expected in forecasting, the errors increase when forecasting further ahead in time. The proposed hierarchical models achieved the lowest error for

Figure 3.2: Results of forecasting mood up to seven days ahead. The RMSE was evaluated in time series cross-validation experiments for $w = 4$ and $h = 1$ through 7. As expected, the RMSE increases when forecasting further ahead. The proposed hierarchical models achieved consistently lower RMSEs than the baseline models. The plot is reused from article A.
every value of the forecast horizon, $h$. Even at $h = 7$ the best performing regression models achieved lower errors than the naïve baselines. We find it remarkable how a relatively short self-assessment history of 4 days is informative of mood up to a week ahead in time. This verifies that substantial mood changes in bipolar patients often occur gradually over several days or weeks rather from one day to the next.

3.2.3 Daily estimates of clinical severity ratings

In article B, we applied pooled, separate and hierarchical regression models to estimate clinical ratings of the severity of depression (HDRS) and mania (YMRS) symptoms. The predictive performance of the models was evaluated in a $K = 100$ cross-validation experiment, where in each iteration data from one randomly sampled clinical evaluation (corresponding to the day of the evaluation and 3 prior days) from each subject was held out and the remaining data was used to fit the models (see Appendix B for details).

<table>
<thead>
<tr>
<th>Model</th>
<th>HDRS $R^2$ (SD)</th>
<th>HDRS RMSE (SD)</th>
<th>YMRS $R^2$ (SD)</th>
<th>YMRS RMSE (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled naïve mean</td>
<td>-0.02 (0.03)</td>
<td>5.99 (0.37)</td>
<td>-0.04 (0.05)</td>
<td>4.18 (0.70)</td>
</tr>
<tr>
<td>Pooled Ridge</td>
<td>0.37 (0.10)</td>
<td>4.68 (0.48)</td>
<td>0.02 (0.15)</td>
<td>4.03 (0.60)</td>
</tr>
<tr>
<td>Pooled XGBoost</td>
<td>0.44 (0.10)</td>
<td>4.40 (0.41)</td>
<td>-0.04 (0.21)</td>
<td>4.11 (0.53)</td>
</tr>
<tr>
<td>Pooled Bayesian</td>
<td>0.36 (0.12)</td>
<td>4.72 (0.51)</td>
<td>0.00 (0.21)</td>
<td>4.04 (0.56)</td>
</tr>
<tr>
<td>Separate naïve mean</td>
<td>0.47 (0.11)</td>
<td>4.29 (0.47)</td>
<td>-0.00 (0.33)</td>
<td>4.00 (0.53)</td>
</tr>
<tr>
<td>Separate Ridge</td>
<td>0.47 (0.12)</td>
<td>4.30 (0.49)</td>
<td>0.04 (0.30)</td>
<td>3.92 (0.54)</td>
</tr>
<tr>
<td>Separate XGBoost</td>
<td>0.27 (0.15)</td>
<td>5.03 (0.49)</td>
<td>-0.38 (0.50)</td>
<td>4.64 (0.45)</td>
</tr>
<tr>
<td>Hierarchical Bayesian</td>
<td><strong>0.57 (0.10)</strong></td>
<td><strong>3.85 (0.47)</strong></td>
<td><strong>0.12 (0.31)</strong></td>
<td><strong>3.74 (0.46)</strong></td>
</tr>
</tbody>
</table>

Table 3.2: Results of $K=100$ cross-validation experiments showing predicted coefficient of determination ($R^2$) and root mean square error (RMSE) on the HDRS (left) and YMRS (right) datasets, respectively. The table is reused from article B.

The cross-validation results are presented in Table 3.2. Because of low variance in the data, the errors are generally small and the naïve models perform rather well. Still the proposed hierarchical Bayesian regression models achieved the best results when predicting both HDRS and YMRS and were significantly better than the separate mean models in both cases according to independent $t$-tests ($P < .001$). Overall, the separate and hierarchical models performed better than their pooled counterparts, showing the importance of accounting for individual differences in the data. The HDRS models generally achieved lower errors than the YMRS models, however, we believe this is mainly due to low variance in the observed YMRS data and not because YMRS ratings are inherently more difficult to predict than HDRS ratings.

We hypothesized that in practical settings it is more useful to identify individuals with high risk of relapse rather than predicting the value of clinical severity ratings exactly. By utilizing samples from the posterior predictive distribution provided by
the Bayesian approach, we were able to compute the probability that an unobserved severity rating, $\tilde{y}_{ji}$, exceeded a predefined threshold value of 13 on the original clinical rating scales: $\Pr(\tilde{y}_{ji} \geq 13)$. This statistic can be interpreted as the probability that an individual is experiencing severe symptoms and can thus be utilized as a personal score indicating the risk of relapse. We applied this method to the results of the hierarchical model from the cross-validation experiment described above.

The ability to correctly assign high risk of relapse scores to observation of clinical severity ratings greater than 13 was evaluated as a binary classification task. Results are presented with receiver operating characteristic (ROC) curves in Figure 3.3, illustrating the trade-off between true positive rate (TPR) and false positive rate (FPR). In terms of area under the curve (AUC), the hierarchical model performs better than naïve mean baselines at identifying high risk instances in both the HDRS and YMRS case. Note how the pooled mean model will always predict the same value (the mean of the population) in every instance, corresponding to an AUC of 0.5. The separate mean models assign a value to each individual independently, which is slightly better in the HDRS case. The hierarchical regression model includes information from additional predictor variables and benefits from partial pooling of data from the population to account for individual differences and thus achieves the best results.

### 3.2.4 Classification based on objective smartphone data

In article C, we used objective smartphone data reflecting behavioral activities to classify patients with bipolar disorder and healthy control individuals. The data used in this analysis was collected during a pilot study investigating the use of automatically generated objective smartphone data as an electronic diagnostic behavioral marker.
in patients with bipolar disorder, performed prior to the MONARCA II RCT. Data was collected from 29 patients recruited from The Clinic for Affective Disorder, Psychiatric Centre Copenhagen, Denmark, between 2013 and 2014, and from 37 healthy control individuals recruited from the Blood Bank at Rigshospitalet, Copenhagen University Hospital, Denmark, between 2015 and 2016. Patients were evaluated fortnightly on HDRS and YMRS and the healthy control individuals were evaluated once at inclusion. Both groups collected smartphone data throughout the 12 week study period (see Appendix C for additional details about this pilot study).

The objective smartphone data included features reflecting social activity and phone usage: number of outgoing and incoming text messages and phone calls, the duration of phone calls, the number of times the screen was turned on, the duration the screen was turned on. The objective features were first aggregated to a daily level and then data from the day of the clinical evaluation and the 3 prior days was averaged to describe the 4-day period the clinical evaluation referred to. Additional available smartphone data, such as location and voice data, was not included in the analysis at this point due to the need for additional preprocessing. Class labels of euthymia, depression and mania were generated from the HDRS and YMRS severity ratings with a threshold value of 13. The classification accuracy of a Gradient Boosting classifier from the Scikit-learn machine learning library [37] was evaluated in 10-fold cross-validation experiments. To mitigate class imbalance in the dataset, we used random oversampling (sampling of the minority class with replacement) to construct a balanced training set in each iteration.

<table>
<thead>
<tr>
<th>Classes</th>
<th>TPR↑</th>
<th>TNR↑</th>
<th>PPV↑</th>
<th>NPV↑</th>
<th>AUC↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD, overall versus HC</td>
<td>0.92</td>
<td>0.39</td>
<td>0.88</td>
<td>0.52</td>
<td>0.66</td>
</tr>
<tr>
<td>Euthymic state versus HC</td>
<td>0.90</td>
<td>0.56</td>
<td>0.85</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>Depressive state versus HC</td>
<td>0.79</td>
<td>0.50</td>
<td>0.71</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>Manic state versus HC</td>
<td>0.47</td>
<td>0.89</td>
<td>0.69</td>
<td>0.76</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3.3: Classification of patients with bipolar disorder (BD) versus healthy control individuals (HC) based on automatically generated objective smartphone data. The table is adapted from article C.

The results of four different binary classification cross-validation experiments are presented in Table 3.3. The results indicate that objective smartphone data differed between patients with bipolar disorder and healthy control individuals, and that it was possible to discriminate between the groups with rather high true positive rate (TPR) and positive predictive value (PPV) but a low true negative rate (TNR) and negative predictive value (NPV). Interestingly, it was possible to discriminate between the patients with bipolar disorder and healthy control individuals regardless of the affective state.

Please refer to the full articles included in the appendices A, B and C for more detailed results and specific discussion of the research presented in this summary.
4.1 Main findings

The research summarized in this report has explored and analyzed data collected from a large population of patients with bipolar disorder collected through a smartphone-based system. The overarching aim has been to relate the smartphone data to disease outcomes measured both subjectively by the patients and clinically using clinically validated rating scales for severity of depression and mania. In particular, the main outcomes of interest have been to: A) forecast self-reported mood in the coming days based on self-reported data histories, B) estimate clinical severity ratings of depression and mania based on self-assessments and utilize uncertainty in the estimates to compute individual risk of relapse scores, and C) utilize objective smartphone sensor and usage data to discriminate between bipolar patients and healthy control individuals, overall and across different affective states. In all three studies, we found positive evidence that the smartphone-based data related to the respective disease outcomes, suggesting data collected via smartphones indeed contains relevant information about symptoms of bipolar disorder that is worth monitoring and studying. In particular, we found significant correlations between smartphone-based self-reported mood scores and clinical severity ratings of depression and mania in patients with bipolar disorder (article B), which confirms electronically self-reported mood is a valid indicator of symptom severity of depression and mania in patients with bipolar disorder [55, 56, 57]. Together with mood changes, the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) acknowledges activity level as one of the core features of hypomania and mania [58], and physical activity is therefore another interesting target for continuous monitoring.

In the predictive analyzes presented in articles A and B, we were able to forecast self-reported mood up to seven days ahead and estimate clinical severity ratings using hierarchical regression models based on self-reported data. In both cases, our proposed method achieved low error compared to common baseline methods and, unsurprisingly, self-reported mood was found to be the strongest predictor of future mood as well as clinical severity ratings. While these results are innovative and encouraging, the practical benefit in a clinical setting for disease monitoring and for enabling early intervention has still to be examined further. In practical settings, it is not only important to provide accurate predictions, but they should also add new information and be actionable in order to enable early intervention and help direct
resources to where they are most needed. Therefore, medium to long-term prognosis or aggregate risk scores combining multiple data that can be used to identify critical cases might be more useful than providing accurate daily symptom score estimates. However, we believe that understanding how smartphone data relates to daily disease outcomes is an important step on the way towards this goal.

4.2 Advantages

The present research is based on data collected from patients diagnosed with bipolar disorder participating in clinical studies and the clinical evaluations were conducted by experienced researchers with specific knowledge about bipolar disorder. In contrast to other studies that rely on data collected from non-clinical subjects, like students or volunteers recruited online, the patients in our dataset and their symptoms are clinically well characterized. Moreover, the size of the patient population in the MONARCA II RCT [32] is large (N = 84) compared to prior clinical studies.

The smartphone-based system used to collect data (the Monsenso system) was customized for the studies and developed in close collaboration with clinicians with feedback from users to ensure high usability and ease of learning [59, 60]. Smartphones offer a unique platform for disease monitoring as users can ubiquitously record and review data replacing the need for traditional paper-based self-assessments. Thus, smartphones provide an opportunity to collect data unobtrusively outside the clinical setting, which has been shown to increase accuracy and reduce recall bias [17].

We found that our proposed method of using hierarchical Bayesian models to analyze data from a population of patients with bipolar disorder (articles A and B) was advantageous compared to approaches that either pool data across the population or analyze each individual dataset separately. We believe this confirms the need to apply models that are suited to account for individual differences in the data and provide personal predictions pointed out in previous work [25, 51, 54, 24, 61]. Additionally, by the means of parameter sharing, the hierarchical approach is able to deal with the cold start problem that occurs when analyzing small individual datasets and enables reasoning about new individuals belonging to the same population. Thus, we believe hierarchical modelling, and more generally multi-task learning, is an appropriate approach in affective computing when dealing with groups of individuals with common characteristics.

4.3 Limitations

The participants of the MONARCA II RCT [32] experienced a low number of affective episodes during the study period, resulting primarily in low clinical severity ratings of depression and mania symptoms. Similarly, the majority of daily self-reported mood scores were close to zero indicating a neutral mood. The relative low variance in the data led to some limitations in the subsequent analyzes. In the mood forecast
task presented in article A, the low variance of self-reported mood meant that simple baseline models were able to achieve low forecast errors and the regression models had a tendency to extrapolate mood form previous days and gradually regress towards the mean of the data. While this forecast behavior succeeds in achieving low error, it largely fails to identify dramatic changes in mood and its utility in a practical monitoring setting can therefore be limited. Similarly in article B when estimating clinical severity, the naive mean models fitted the data well because of low variance in the data and the generally low severity ratings made it difficult to evaluate the proposed methods ability to correctly identify high risk individuals. In the classification task presented in article C, the low prevalence of affective episodes led to imbalanced classes, which for many machine learning algorithms is a problem and can lead to biased models. Various methods exist to correct for class imbalance, by for example undersampling the majority class or oversampling the minority class, neither of which are great solutions if one class is represented by only a few examples.

As described in section 2.1.7, we experienced problems with missing data, limiting the amount of data available for analysis. The proposed forecasting method in article A relies on a data history of consecutive observations, hence expanding the window size would reduce the number of training examples in cases where an intermediate observation was missing. Clinical severity rating estimation in article B required self-assessments to be available on the days of the clinical evaluations and data from at least 3 clinical evaluations from an individual in order to construct a training set and a test set, which reduced the number of individuals with sufficient amounts of data for analysis.

Finally, besides the many advantages of applying a Bayesian approach, specifying and fitting the hierarchical models with Stan \[39\] can be a challenging and time consuming task. The method of defining the full statistical model and choosing hyperparameters offers a lot of flexibility but is also cumbersome for an inexperienced user. Additionally, fitting complex hierarchical models is computationally intensive and can take hours and sometimes days to complete. This process becomes increasingly challenging when performing cross-validation experiments. Higher level interfaces to Stan providing common models, such as RStanArm \[62\] currently available for the statistical computation environment, R, could perhaps help speed up parts of the process and make iterations faster.

**4.4 Perspectives**

We believe that there are several interesting perspectives and directions for future research in adoption of smartphone-based systems and automatic data analysis in monitoring and treatment of mental disorders including bipolar disorder and depression:

- Predictive analysis of smartphone-based data can be used to detect early warning signs of relapse in clinical treatment. The use of smartphone data and
automatic analyzes could be integrated in a clinical setup, where clinicians can monitor and supervise a group of patients as part of outpatient treatment. This can enable early intervention and potentially reduce the severity of symptoms and degree of treatment and even prevent major affective episodes that could lead to costly hospitalizations. A mood forecast prototype based on the work in article A has already been implemented as part of the RADMIS trial [63] in order to study its practical application. However, symptom prediction and especially forecasting must be used with care and disease prognosis should preferably be made available only to health care professionals rather than being presented directly to patients as prospects of illness might become self-fulfilling.

Another use of machine learning within smartphone-based treatment of mental disorders (and health technology in general) is to encourage behavioral change by providing insights regarding unhealthy behaviors and recommending alternative activities. An important part current treatment of affective disorders is to learn how to recognize personal warning signs and develop personal coping strategies through cognitive behavioral therapy (CBT). This could be supported by generating data driven insights and automatically trigger warnings when detecting alarming behaviors such as reduced physical activity or too little sleep, or by recommending specific exercises.

Sensor technology in modern smartphones enables tracking of an increasingly wide range of human behaviors that are potentially relevant for monitoring mental health [27]. Even more detailed behaviors can be captured with wearable devices such as wristbands and fitness trackers, which are becoming increasingly ubiquitous. Objective behavioral features derived from sensor data could help augment or potentially eliminate the need for momentary self-assessment and thus reduce subjective bias and at the same time improve the user experience of monitoring systems. In particular, diagnosis and monitoring of affective disorders such as depression and bipolar disorder today relies entirely on subjective information and judgment and lacks objective tests. Thus, accurate objective measures have the potential to revolutionize how we assess these mental disorders. We have already examined the use of simple objective smartphone measures in article C and previously explored the use of voice features extracted from phone calls [34]. In future research, we plan to include richer sensor modalities such as location and provide a more thorough analysis of the use of objective voice features in bipolar disorder.

Modern machine learning techniques, such as deep neural networks, require large amounts of training data to fit predictive models, but have proved unparalleled in revealing complex patterns and achieving high predictive accuracy in a wide range of domains [64]. However, most studies involving predictive methods within bipolar disorder have relied on rather simple, classic machine learning and statistical methods. The potential of new advanced methods will only be unlocked with truly large scale studies able to provide data of the neces-
sary quantity and quality to fit complex models. This could however potentially bring new insights to field and improve predictive accuracy.

4.5 Conclusions

We set out to research methods and algorithms for analysis of behavioral smartphone data from patients with bipolar disorder and depression aiming to detect and predict manic and depressive symptoms. This report has presented a summary of research conducted during the author’s PhD studies and presented the principal results of three included research manuscripts.

We have shown that by applying hierarchical Bayesian regression models we were able to forecast subjective mood from 1 to 7 days ahead based on a short history of self-assessments. The proposed hierarchical methods achieved low error compared to pooled and separate baseline models. Using the same hierarchical modelling approach, we were able to successfully produce daily estimates of clinical severity ratings of depression and mania from self-assessments and we showed how uncertainty in the estimated values can be used to produce individual scores indicating risk of relapse. We also found that it is possible to utilize simple features of objective smartphone data to discriminate between patients with bipolar disorder and healthy control individuals during different affective states. Finding relationships between more complex objective features and symptoms of depression and mania has great potential and needs to be studied further.

Based on the results of the present research, we are confident that predictive analysis based on subjective and objective smartphone data has potential to improve disease monitoring and treatment tasks in bipolar disorder and enable early intervention. To accomplish this goal, predictions must be accurate, interpretable and actionable. While this is not achieved in full, we believe that the research presented here contributes in the right direction.

A strength of the research is that it is based on data from patients diagnosed with bipolar disorder participating in a trial conducted by experienced researchers within the field. The smartphone-based system used in the studies was developed through several field trials in close collaboration with clinicians and patients. There has, however, also been challenges along the way. Data collection in the two RADMIS RCTs [29] was delayed and did not commence until late in the PhD and is still ongoing today. Therefore, the PhD research was primarily based on data from previous trials conducted by RADMIS researchers, in particular the MONARCA II RCT [32], which was concluded in 2018. The data analyzes have also been challenged by missing data and low variance in the primary symptom outcome scores, which have made it difficult to produce definite statistical results. Thus, high quality datasets collected from large groups of patients with symptoms are required to drive the research forward.

The application of smartphones in disease monitoring and treatment in outpatient clinics has some immediate benefits. Smartphones are ubiquitous in modern society and provide a unique platform for pervasive data collection and offer an improved
user experience compared to traditional paper-based self-assessment. Digital data collection is also much more convenient, as it makes data available for analysis by humans and computers immediately. Smartphones additionally provide a way to deliver immediate feedback to users based on their data whereas today data is only evaluated periodically in routine outpatient consultations with a risk of overlooking important patterns due to recall and observer biases. Whether the use of a smartphone-based system can improve treatment of bipolar disorder and depression in terms of primary disease outcomes, including mitigating the severity of affective episodes and preventing hospitalizations, is still unknown and answering this question is the main objective of the continued research in the RADMIS project.
Forecasting Mood in Bipolar Disorder from Smartphone Self-assessments with Hierarchical Bayesian Models

Original Paper

**Forecasting Mood in Bipolar Disorder from Smartphone Self-assessments with Hierarchical Bayesian Models**

Authors:

Jonas Busk, MSc1; Maria Faurholt-Jepsen, MD, DMSc2; Mads Frost, PhD3; Jakob E. Bardram, Professor, MSc, PhD4; Lars Vedel Kessing, Professor, DMSc2,5; Ole Winther, Professor, PhD1,6,7

1 Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

2 Copenhagen Affective Disorder Research Center (CADIC), Psychiatric Center Copenhagen, Rigshospitalet, Copenhagen, Denmark

3 Monsenso ApS, Copenhagen, Denmark

4 Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

5 Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

6 Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Denmark

7 Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark

* Corresponding author:

Jonas Busk

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark

Technical University of Denmark (DTU) B321
DK-2800 Lyngby

Email: jbusk@dtu.dk
Abstract

**Background:** Bipolar disorder is a prevalent mental disease imposing a high societal burden. Accurate forecasting of symptom scores can be used to improve disease monitoring, enable early intervention and eventually help prevent costly hospitalizations. While several studies have examined the use of smartphone data to detect mood, only few studies deal with forecasting mood one or more days ahead of time.

**Objective:** The objective of this work is to examine the feasibility of forecasting daily subjective mood based on daily self-assessments collected from bipolar disorder patients via a smartphone-based system in a randomized clinical trial.

**Methods:** We apply hierarchical Bayesian regression models, a multi-task learning method, to account for individual differences and forecast mood up to seven days ahead based on 15,975 smartphone self-assessments from 84 bipolar disorder patients participating in a randomized clinical trial. We report the results of two time-series cross-validation one day ahead prediction experiments corresponding to two different real-world scenarios and compare the outcomes to commonly used baselines methods. We then apply the best model to evaluate a seven-day forecast.

**Results:** The best performing model used a history of 4 days of self-assessments to predict future mood scores with historical mood being the most important predictor variable. The proposed hierarchical Bayesian regression model outperformed pooled and separate models in a one-day forecast time series cross-validation experiment and achieved predicted $R^2 = 0.51$ and $RMSE = 0.32$ for mood scores on a scale of -3 to 3. When increasing the forecast horizon, forecast errors also increase and the forecast regresses towards the mean of the data distribution.

**Conclusions:** We found that our proposed method can forecast mood several days ahead with low error compared to common baseline methods. The applicability of a mood forecast in clinical treatment of bipolar disorder is also discussed.

**Keywords:** bipolar disorder; mood; early medical intervention; machine learning; forecasting; Bayesian analysis; hierarchical model; multi-task learning.
**Introduction**

Bipolar disorder is estimated as one of the most important causes of disability worldwide [1, 2]. Bipolar disorder is characterized by recurrent episodes of depression, (hypo)mania and mixed episodes intervened by periods of euthymia [3] and with a high degree of comorbidity, functional impairment and increased risk of suicide [4]. The World Health Organization (WHO) estimates that about 60 million people are affected by bipolar disorder worldwide and that the burden of depression and other mental health conditions is on the rise globally [5]. Clinical treatment of bipolar disorder includes symptom monitoring, early identification of subsyndromal symptoms of depression and mania and intervention to prevent or reduce the severity of affective episodes.

In this paper, we analyze daily self-assessments including mood scores collected from bipolar disorder patients through a smartphone-based system. Ecological momentary assessments (EMA) reflect the methods used to collect assessments of individuals’ real-time states repeatedly over time during naturalistic settings, and may reduce recall bias [6]. Today a median of 76% of adults across 18 advanced economies reported having a smartphone [7], and many people use a smartphone daily [8]. The rapid evolution and ubiquity of mobile networks have resulted in increasing growth of e-mental health technologies, including electronic platforms offering tools for remote self-monitoring [9]. By using daily smartphone-based self-monitoring potential recall bias in self-reported patient data is minimized. Thus, smartphones extend the use of EMA beyond its classical use for self-reports and offer the opportunity to collect fine-grained data unobtrusively and outside the clinical settings [10]. By replacing paper-based self-assessments of traditional treatment methods with a smartphone-based system, users can ubiquitously enter and review their own data and share it with clinicians who can intervene if something appears alarming. Thus, smartphones provide a unique platform for monitoring and managing depression and mania [11, 12, 13]. Furthermore, modern smartphones provide the means for collecting rich sensor data, which is believed to capture valuable behavioral information that can be related to disease outcomes [14]. Digital self-reporting and data collection has the additional benefit of making data available for automatic analysis immediately, which can help support continuous disease monitoring.

We find it useful to distinguish between *mood detection*, i.e. predicting mood based on data from the same day, and *mood forecasting*, i.e. predicting mood one or more days ahead based on historical data. Smartphone-based mood detection, is well studied but remains a difficult problem. Several papers have examined the use of passive smartphone data, such as location, communication logs and device usage, to detect or classify daily, self-reported mood labels [15, 16, 17, 18, 19, 20, 21]. A few recent studies address mood forecasting, which is a more challenging task than mood detection, since the causal chain between cause and outcome is longer and because of the uncertainty inherently associated with future events. DeepMood is a solution for forecasting severely depressed mood from self-reported histories using
a Recurrent Neural Network [22]. They found that long-term historical information up to 14 days improves the accuracy of forecasting depressed mood classes and that mood the previous day is the most important predictor when forecasting severe depression one day ahead. A limitation of this method is that it needs labeled observations every day in a 14-day history in order to make predictions. Another study employs a selection of multi-task learning (MTL) techniques to train personalized models for predicting future mood, stress, and health one day ahead [23]. They found that using MTL to account for individual differences provides substantial performance improvements over traditional machine learning methods. By utilizing a clustering of users based on age, gender and personality, a new user needs only to be assigned to a cluster in order to enable prediction on new data, when labeled data from a population of similar users is available to fit the initial model.

A major challenge when reviewing work on mood prediction and behavior tracking is that researchers often have different data collection strategies and apply custom modelling and labeling approaches, consequently making results difficult to compare and sometimes contradicting [14]. Another problem is that most studies involve healthy subjects (i.e., not diagnosed patients), and it is therefore hard to generalize to patients suffering from affective disorders. Nonetheless, some common observations stand out. Several studies found that personalized models generally outperform generic models when sufficient data is available [16, 19, 23, 24, 25], demonstrating the importance of accounting for individual differences in the data. This can be accommodated by applying MTL, which provides a way of improving generalization by learning several related tasks simultaneously [26]. By considering each individual of a population its own task, MTL techniques can produce personalized predictions in a straightforward manner.

Our study differs from the prior work in a number of ways. Where many studies collect data from non-clinical subjects (like students and volunteers recruited online), our data is collected in a randomized clinical trial from patients who are diagnosed and treated for bipolar disorder. Moreover, to the best of our knowledge, the size of our patient population (N=84) is significantly larger than any prior clinical studies. We also found that even though most studies record subjective mood on a continuous or ordinal scale, the prediction task is often reduced to a classification problem by binning the values into two or more classes, such as neutral, depressed and manic. In this work, we treat mood prediction as a regression problem, which is more direct given the way data is collected and interpreted by users. Finally, rather than mood detection, we address the more challenging task of mood forecasting, and we apply a hierarchical Bayesian modelling approach, which is a popular method of MTL able to account for individual differences in the data.

The main objective of this work is to examine the feasibility and technical foundation of forecasting daily mood scores in bipolar disorder based on daily smartphone self-assessments. We hypothesize that utilizing these data to establish an accurate, online mood forecast solution can help improve disease monitoring by
providing additional insights that enable early intervention and thus eventually can prevent relapse of affective episodes and burdensome and costly hospitalizations.

**Methods**

**Data description**

Data used in the present study was collected between September 2014 and January 2018 during the MONARCA II randomized clinical trial [27] investigating the effect of smartphone-based monitoring. All patients with a diagnosis of bipolar disorder who had previously been treated at the Copenhagen Clinic for Affective Disorder, Copenhagen, Denmark in the period from 2004 to January 2016 and who at the time of recruitment were being treated at community psychiatric centers, private psychiatrists and general practitioners were invited to participate in the trial. Patients were included in the study for a nine-month follow-up period if they had a bipolar disorder diagnosis according to ICD-10 using the Schedules for Clinical Assessments in Neuropsychiatry (SCAN) [28] and previously were treated at the Copenhagen Clinic for Affective Disorder. Patients with schizophrenia, schizotypal or delusional disorders, previous use of the MONARCA system, pregnancy and lack of Danish language skills were excluded. Patients with other comorbid psychiatric disorders and substance use were eligible for the trial. As part of the MONARCA II trial, patients were randomized to either using a smartphone-based monitoring system (the Monsenso system) for daily self-monitoring (the intervention group) or to treatment as usual (the control group). Patients included in the intervention group collected daily smartphone-based self-monitoring data and were included in the analyses in the present report. Inclusion and exclusion criteria were investigated and assessed by two clinical researchers (MFJ).

Study participants were provided with an Android smartphone application configured for the study and instructed to complete daily self-assessments consisting of a questionnaire including the items listed in Table 1. Specifically, mood was scored on a scale of -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3, where negative values indicate various degrees of depression, positive values indicate mania and zero indicates neutral mood (euthymia).
Table 1. Items of the daily self-assessment questionnaire.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Level of physical activity.</td>
<td>-3 to 3</td>
</tr>
<tr>
<td>Alcohol</td>
<td>Alcoholic drinks consumed.</td>
<td>0 to 10+</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Level of anxiety.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Irritable</td>
<td>Level of irritability.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Cognitive difficulty</td>
<td>Level of cognitive discomfort.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Medicine</td>
<td>Medicine adherence.</td>
<td>0 to 2</td>
</tr>
<tr>
<td>Mixed mood</td>
<td>Experienced mixed mood.</td>
<td>0 to 1</td>
</tr>
<tr>
<td>Mood</td>
<td>Experienced mood.</td>
<td>-3 to 3</td>
</tr>
<tr>
<td>Sleep</td>
<td>Hours of sleep.</td>
<td>#hours</td>
</tr>
<tr>
<td>Stress</td>
<td>Level of stress.</td>
<td>0 to 2</td>
</tr>
</tbody>
</table>

Additionally, participants of the study were periodically evaluated by trained psychiatrists on clinical rating scales for depression and mania: the Hamilton Depression Rating Scale (HDRS) [29] and Young Mania Rating Scale (YMRS) [30]. Each rating scale consists of a series of questions that are scored and added up to summarize the current state of the patient with higher scores indicating more severe symptoms. All clinical assessments were conducted by clinical researchers, who were blinded to all smartphone-based data. Thus, data on severity of depressive and manic symptoms was collected rater-blinded. Both rating scales are clinically validated and generally accepted as accurate measures of illness severity in bipolar disorder.

Data preprocessing
Two of the self-assessment items were preprocessed prior to the analysis. Because the answer to the medicine item is categorical by design (medicine not taken, medicine taken, medicine taken with changes), we encoded it as two exclusive binary variables indicating if medicine was not taken (medicine omitted) or if medicine was taken with changes (medicine changed). Additionally, we did not expect sleep duration to have linear effect on mood, why the sleep variable was replaced with two new features by subtracting the individual mean and splitting the result into a negative and positive component (sleep negative and sleep positive) indicating decreased or increased sleep relative to the mean. Finally, we have normalized all self-assessment variables by their allowed minimum and maximum value. We also experimented with forward filling missing data from the previous day, but found that very little additional data was gained, so we left this step out of the final analysis presented in this work.

Forecasting
Forecasting is the task of predicting the future given available information from the past and the present [31, 32]. For forecasting to be feasible, it should be reasonable to assume that the history of recorded information somehow relates to the future events of interest. A typical forecasting task is illustrated in Figure 1. We denote by $w$ the size of the history used in the forecast and by $h$ the horizon of how far into the
future the target is predicted. In our particular case of using daily self-reports both 
w and h are measured in days.

Figure 1. Forecasting is the task of predicting the future given relevant information 
from the past and the present. The window size, w, is the size of the history defining 
the predictor variables and the horizon, h, is how far in the future the target variable 
is predicted.

Several methods for producing forecasts exist [32]. The simplest forecasting 
methods use only historic information about the target variable and do not take into 
account any other information but are designed to utilize time dependent patterns, 
such as seasonality and trend, to extrapolate observed data into the future. Another 
approach is to apply standard regression or classification models to predict the 
variable of interest based on relevant information, such as prior (lagged) 
observations of the target variable along with additional predictor variables. This 
approach has the benefit of allowing the use of a variety of different methods from 
the machine learning and statistical inference literature, but may not be as good at 
capturing long-term time dependent patterns. For short-term forecasts or data 
without long-term time dependence, however, this might not be a problem. For 
these reasons, we chose to apply the latter approach in this work.

Special care should be taken when evaluating the performance of a forecast. A 
genuine forecast uses only data available at the forecast time, and thus no future 
data, to estimate its parameters [32]. Consequently, the size of in sample residuals 
(training error) is not a reliable indicator of the true forecast error. The forecast 
performance can only be determined by fitting the model on training data observed 
prior to the test data. This needs to be considered when designing the experiment 
used to evaluate the forecast model, such as cross-validation. Time series cross-
validation addresses this by splitting the data into a sequence of consecutive test 
sets. The corresponding training sets consist only of data observed prior to each test 
set. Thus, no future information is included when constructing the forecasts. The 
cross-validation error is then computed across all the test sets. Because we consider 
data from multiple individuals, we apply two different time series cross-validation 
in our experiments:
• **Leave-all-out time series cross-validation**: Each individual’s data is partitioned into a sequence of $T$ consecutive similar sized test sets. Then the test sets are pooled across all individuals. The corresponding training sets consist of all data observed prior to each test set, resulting in $T - 1$ test/training set pairs (the first test set has no prior data).

• **Leave-one-out time series cross-validation**: Each individual's data is partitioned into a training set and a subsequent test set. The training set is then pooled with all data from all other individuals, resulting in a number of test/training set pairs equal to the number of individuals, $J$.

These two experiments correspond to two different scenarios: Leave-all-out time series cross-validation simulates a situation where a group of patients starts monitoring at the same time without any additional historical data, while leave-one-out time series cross-validation simulates a situation where each participant starts monitoring when data is already available from a population of similar individuals.

**Hierarchical Bayesian models**

When analyzing data consisting of multiple related sets of measurements, such as individuals in a population, a basic approach is to completely pool all the data into a common model, assuming all sets have similar properties. A drawback of this method is you risk losing important information at the individual level. To overcome this problem, an alternative approach is to model each set of data separately, assuming all sets are independent. However, you might miss information about how the individual sets relate to each other at the population level. Especially when each individual dataset is too small to construct a meaningful separate model, it is useful to include information from the population to make analysis feasible. A hierarchical (multi-level) Bayesian model is an intermediate solution allowing partial pooling of the data, thus providing a compromise between the completely pooled and separate models [33, 34]. The hierarchical approach captures the overall characteristics of the population while allowing individual differences and enables modeling of small related datasets, each getting a gradually more personalized model as more data is collected and included in the training set. Additionally, it allows for reasoning about previously unobserved individuals, assuming they come from the same population, which helps to overcome the cold start problem. Applying a Bayesian approach has the additional benefit of providing uncertainty in all model parameters and predictions, allowing for improved interpretability. Because of these desirable properties, we apply hierarchical models in our analysis. In particular, we explore the use of hierarchical implementations of both linear and ordinal regression models.
Figure 2. Bayesian network of a hierarchical linear regression model. Individual regression intercept $\alpha_j$ and weights $\beta_j$ are drawn from population distributions parameterized by $\mu_{\alpha}, \tau_{\alpha}$ and $\mu_{\beta}, \tau_{\beta}$. This allows the model to account for individual differences while constraining individual parameters to be similar across the population.

Ordinary linear regression is a method for predicting the outcome of a continuous variable, modeled as the linear combination of the model parameters and predictor variables. Hierarchical Bayesian linear regression can be expressed by assuming each individual set of parameters are drawn from a common population distribution (Figure 2). For individual $j = 1:J$, observation $i = 1:N$, target variable $y_{ji}$ and predictor variables $x_{ji}$:

$$ y_{ji} = \text{Normal}(\alpha_{ji} + \beta_j^T x_{ji}, \sigma) $$

where $\alpha_j$ and $\beta_j$ are sampled from population distributions:

$$ \alpha \sim \text{Normal}(\mu_{\alpha}, \tau_{\alpha}) $$
$$ \beta \sim \text{Normal}(\mu_{\beta}, \tau_{\beta}) $$

and the population means $\mu_{\alpha}, \mu_{\beta}$ and variances $\tau_{\alpha}, \tau_{\beta}$ as well as the standard error $\sigma$ have independent Normal priors.

Ordinal regression (sometimes referred to as ordinal classification) is a method for predicting a discrete variable that has a relative ordering of the possible outcomes. Thus, it can be thought of as an intermediate between regression and classification. An example of ordinal regression is ordered logistic regression. For an outcome belonging to one of $K$ categories, $y_{ji} \in 1:K$, ordered logistic regression is determined by a latent continuous variable, $z_{ji} = \beta_j^T x_{ji}$, along with a sequence of $K + 1$ ordered cutpoints $c_j$, such that $c_{k-1} < c_k$ and $c_0 = -\infty, c_K = \infty$ by definition. If $z_{ji}$ falls between two cutpoints, $c_{k-1}$ and $c_k$, the outcome is predicted to belong to
the corresponding category, \( y_{ji} = k \), with high probability. This type of model can be justified by assuming the category, \( y_{ji} \), is an incomplete measurement of the latent variable \( z_{ji} \):

\[
y_{ji} \sim \text{OrderedLogistic}(z_{ji}, c_j)
\]

Hierarchical Bayesian ordinal regression can be expressed by assuming each individual set of model parameters is drawn from a common population distribution:

\[
\beta_j \sim \text{Normal}(\mu_\beta, \tau_\beta) \\
c_j \sim \text{Normal}(\mu_c, \tau_c)
\]

with independent Normal priors on the population parameters \( \mu_\beta, \tau_\beta, \mu_c, \tau_c \) along with ordering constraints on \( \mu_c \) and \( c_j \). In practice, we reparameterize the hierarchical models to achieve more efficient sampling [35]. A practical difference of using ordinal regression over linear regression is that ordinal regression can never produce predictions (or uncertainties) outside the range of the training data. This can be an advantage when the target variable represents a bounded scale where values outside the scale does not have any meaning. Ordinary linear regression does not provide this guarantee, thus the ordinal model can lead to more interpretable outcomes.

We use the statistical modeling platform, Stan [36], to specify and perform inference in the hierarchical models. Generally, the models were fitted using 4 sampling chains and 5,000 iterations, where the first half was warm-up and parameter tuning, resulting in 10,000 posterior samples. Our prior believe was that self-reported mood would be the strongest predictor of future mood, hence the population parameters corresponding to mood were assigned less restrictive priors than the other population parameters, which were assigned more restrictive priors to introduce regularization. The Stan code of the hierarchical models and more details on the priors is included in Multimedia Appendix 1. To provide appropriate baseline results for comparison, a suite of naïve and standard machine learning regression models from the Scikit-learn machine learning library [37] and the popular XGBoost Python package [38] are also evaluated. These models are fitted both with pooled and separate data where applicable.

**Ethical considerations**

The trials were approved by the Regional Ethics Committee in the Capital Region of Denmark (H-2-2014-059) and the Danish Data protection agency (2013-41-1710). The law on handling of personal data was respected. Prior to commencement the trials were registered at ClinicalTrials.gov (NCT02221336). Electronic data collected from the smartphones were stored at a secure server at Concern IT, Capital Region, Denmark (I-suite number RHP-292 2011-03). The trial complied with the Helsinki Declaration of 1975, as revised in 2008.
Results

Descriptive statistics

The dataset consists of 15,975 daily self-assessments and 280 clinical evaluations from 84 participants. The population consisted of 52 females (62%) and ages 21 to 71 years (mean=43.1, SD=12.4). Figure 3 presents the distribution of self-reported mood scores across all individuals in the dataset (mean=-0.14, SD=0.48). The majority of observed mood scores, $y$, are centered around zero indicating euthymia ($-0.75 < y < 0.75 = 89.64\%$) with only few values indicating depression ($y < -0.75 = 8.68\%$) and even fewer values indicating mania ($y > 0.75 = 1.68\%$). As expected, the self-reported mood scores and HDRS are negatively correlated ($r = -0.40, P < .001$) and self-reported mood and YMRS are positively correlated ($r = 0.22, P < 0.001$).

Figure 3. Distribution of all self-reported mood scores (left) and of individual mean mood scores (right). The mood scores are generally close to zero indicating neutral mood with only a few exceptions indicating depressed or elevated mood.

Figure 4 shows correlations of self-assessment items with self-reported mood lagged for up to seven days. Self-reported mood has positive auto-correlation for the entire duration of 1-7 days. Additionally, activity has a positive correlation with mood for a few days, indicating high activity levels coincides with elevated mood, and anxiety has a small negative correlation with mood, indicating anxiety often coincides with negative mood scores. The remaining self-assessment items show small, diminishing correlations with lagged mood. A seasonality analysis of self-reported mood revealed no significant monthly or daily seasonality in the data and is left out for brevity.

Figure 4. Mean of individual correlations of self-assessment items and mood lagged up to seven days. Non-zero correlations indicate that items have some relation to mood on subsequent days that can be utilized for mood forecasting.
Window size selection
To find the optimal window size, \( w \), for forecasting mood, we evaluated a one day forecast with window sizes from one to seven days. Each window size was evaluated in a \( T = 24 \) leave-all-out time series cross-validation experiment with data partitions of size one week. The predicted coefficient of determination (\( R^2 \)) and root mean squared error (RMSE) were computed across all the test sets.

Figure 5 shows the \( \text{RMSE} \) of the cross-validation for \( w = 1 \) through 7 and \( h = 1 \). The errors of the naïve mean models are almost constant, varying only due to the difference in datasets available for different values of \( w \). The hierarchical Bayesian linear regression model achieved the lowest RMSE of all models for every window size with the best result at \( w = 4 \) days, which we then used in the following analysis.

![Figure 5. Window size selection results. The RMSE was evaluated in time series cross-validation experiments for \( w = 1 \) through 7 and \( h = 1 \). The lowest RMSE was achieved by the hierarchical linear model at \( w = 4 \).](image)

Model checks and feature importance
To evaluate how well the proposed hierarchical linear and ordinal models fit the data distribution, we trained them on the entire dataset of participants with at least two data points for \( w = 4 \) and \( h = 1 \) (\( N = 5881 \)). The hierarchical models achieved...
similar fit with in sample $R^2 = 0.56$ and in sample $RMSE = 0.29$. We then performed posterior predictive checks by testing the ability of the models to replicate (predict) the observed distribution of future mood from the observed history of predictor variables. In particular, we computed the ratio of observed mood values and replicated mood values less than -0.75 and greater than 0.75, respectively. The hierarchical linear model replicated 93% of the small values while the ordinal model replicated 65% of the small values. The hierarchical linear model replicated 73% of the large values while the ordinal model only replicated 24% of the large values. Thus, the hierarchical linear model is better at capturing the tails of the distribution whereas the ordinal model underestimates extreme values.

The importance of a predictor variable in a linear regression model can be measured as the absolute value of the $t$-statistic of its regression weight, $\beta$, computed as the mean weight scaled by its standard error: $t_\beta = \beta / SE(\beta)$ [39]. Table 2 presents the mean absolute $t$-statistic of the individual-level regression weights in the hierarchical Bayesian linear regression model for each of the predictor variables in a four-day history. This shows that self-reported mood is the most important variable for predicting mood the next day, which is not surprising considering mood has a strong auto-correlation (Figure 4).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$x_t$</th>
<th>$x_{t-1}$</th>
<th>$x_{t-2}$</th>
<th>$x_{t-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood</td>
<td>4.53 (3.35)</td>
<td>2.34 (0.55)</td>
<td>0.47 (0.28)</td>
<td>2.78 (0.18)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>2.78 (0.05)</td>
<td>0.71 (0.02)</td>
<td>1.29 (0.01)</td>
<td>0.76 (0.00)</td>
</tr>
<tr>
<td>Irritable</td>
<td>2.74 (0.11)</td>
<td>1.22 (0.01)</td>
<td>0.95 (0.01)</td>
<td>1.30 (0.00)</td>
</tr>
<tr>
<td>Mixed mood</td>
<td>2.09 (0.06)</td>
<td>2.51 (0.02)</td>
<td>1.96 (0.01)</td>
<td>0.52 (0.01)</td>
</tr>
<tr>
<td>Medicine changed</td>
<td>0.36 (0.10)</td>
<td>0.08 (0.01)</td>
<td>2.15 (0.01)</td>
<td>0.64 (0.00)</td>
</tr>
<tr>
<td>Sleep positive</td>
<td>1.65 (0.01)</td>
<td>0.72 (0.00)</td>
<td>0.37 (0.00)</td>
<td>0.16 (0.00)</td>
</tr>
<tr>
<td>Cognitive difficulty</td>
<td>1.48 (0.09)</td>
<td>0.58 (0.02)</td>
<td>0.19 (0.00)</td>
<td>1.57 (0.00)</td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.67 (0.02)</td>
<td>0.77 (0.01)</td>
<td>1.56 (0.01)</td>
<td>0.87 (0.00)</td>
</tr>
<tr>
<td>Medicine omitted</td>
<td>0.13 (0.01)</td>
<td>1.31 (0.00)</td>
<td>0.60 (0.00)</td>
<td>0.14 (0.00)</td>
</tr>
<tr>
<td>Stress</td>
<td>1.22 (0.12)</td>
<td>0.91 (0.02)</td>
<td>0.71 (0.01)</td>
<td>0.28 (0.01)</td>
</tr>
<tr>
<td>Activity</td>
<td>1.04 (0.03)</td>
<td>1.14 (0.02)</td>
<td>0.49 (0.01)</td>
<td>1.14 (0.01)</td>
</tr>
<tr>
<td>Sleep negative</td>
<td>0.41 (0.01)</td>
<td>0.52 (0.00)</td>
<td>0.48 (0.00)</td>
<td>0.52 (0.00)</td>
</tr>
</tbody>
</table>

**Time series cross-validation results**

The results of leave-all-out and leave-one-out time series cross-validation experiments for $w = 4$ and $h = 1$ are presented in Table 3. In both experiments the naïve pooled mean model scores a predicted $R^2$ close to zero because it does not explain any variance in the data. A predicted $R^2$ greater than zero indicates that
some variance is explained while a negative $R^2$ is worse than the pooled mean model. The last observed model simply repeats the last observed mood value, which performs considerably better than the mean model and represents a solid baseline.

The leave-all-out time series cross-validation experiment was evaluated with $T = 24$ and data partitions of size one week, resulting in $T - 1 = 23$ iterations of cross-validation. The hierarchical Bayesian linear model achieved the best result with predicted $R^2 = 0.511$ and predicted $RMSE = 0.324$, beating the naïve baselines and pooled and separate regression models. The hierarchical Bayesian ordinal model is a close second best. The leave-one-out time series cross-validation experiment was evaluated for each individual with the first two weeks of data pooled with data from the rest of the population in the training set and evaluated on the next 22 weeks of data from that individual, resulting in $J = 58$ iterations of cross-validation. The hierarchical Bayesian linear model achieved the best predicted $R^2 = 0.347$ and predicted $RMSE = 0.337$, but is very similar to the best pooled regression models, indicating the hierarchical model does a lot of pooling as well. The separate models fail to generalize to the held-out test data in this experiment, achieving negative $R^2$ scores, because the training sets contains only two weeks of data. Overall, the hierarchical and pooled models performed better than the separate models and all regression models generally outperformed the naïve baseline models when sufficient data was available.

Table 3. Results of leave-all-out time series cross-validation (left) and leave-one-out time series cross-validation (right). The hierarchical Bayesian linear regression model achieves the best results. The pooled models are better than the separate models overall.

<table>
<thead>
<tr>
<th>Model</th>
<th>Leave-all-out</th>
<th>Leave-one-out</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2 \uparrow^a$</td>
<td>$RMSE \downarrow^b$</td>
</tr>
<tr>
<td>Last observed</td>
<td>0.342</td>
<td>0.376</td>
</tr>
<tr>
<td>Pooled mean</td>
<td>-0.007</td>
<td>0.465</td>
</tr>
<tr>
<td>Pooled ridge</td>
<td>0.450</td>
<td>0.344</td>
</tr>
<tr>
<td>Pooled XGBoost</td>
<td>0.455</td>
<td>0.342</td>
</tr>
<tr>
<td>Separate mean</td>
<td>0.213</td>
<td>0.412</td>
</tr>
<tr>
<td>Separate ridge</td>
<td>0.345</td>
<td>0.375</td>
</tr>
<tr>
<td>Separate XGBoost</td>
<td>0.302</td>
<td>0.388</td>
</tr>
<tr>
<td>Hierarchical Bayesian linear</td>
<td><strong>0.511</strong></td>
<td><strong>0.324</strong></td>
</tr>
<tr>
<td>Hierarchical Bayesian ordinal</td>
<td>0.495</td>
<td>0.330</td>
</tr>
</tbody>
</table>

$^a$ Coefficient of determination ($R^2$): higher is better.

$^b$ Root mean squared error ($RMSE$): lower is better.

Seven-day forecast

So far we have focused on evaluating a one day forecast, but it is also interesting to forecast mood on a more distant horizon. Figure 6 shows the mean $RMSE$ of cross-validation for $w = 4$ and $h = 1$ through 7. The hierarchical Bayesian linear regression model achieves the lowest $RMSE$ of all models for every value of $h$. As might be expected, the error generally grows with the size of the horizon. The errors
of the naïve mean models are almost constant, varying only due to the difference in datasets available for different values of $h$. But even at $h = 7$ the best regression models are able to outperform the mean models, meaning they are able to capture useful information from prior self-assessments.

Figure 6. Results of forecasting mood up to seven days ahead. The RMSE was evaluated in time series cross-validation experiments for $w = 4$ and $h = 1$ through 7. As expected, the RMSE increases when forecasting further ahead. The proposed hierarchical models achieved consistently lower RMSEs than the baseline models.

Discussion

Principal Results
In this study, we have analyzed smartphone-based self-assessment data from a large population of 84 bipolar disorder patients with the purpose of forecasting subjective mood. The initial data analysis showed that the majority of observed mood scores are close to zero, indicating weak or no symptoms among the population for most of the study period. Still we found a significant negative correlation between self-reported mood and HDRS ($r = -0.40, P < .001$) and a significant positive correlation between self-reported mood and YMRS ($r = 0.22, P < .001$). This confirms prior findings [40, 41, 42], suggesting subjective mood is a valid indicator of mental state in bipolar disorder and thereby also a clinically relevant feature for daily monitoring and forecasting. We did not observe any substantial seasonality or long-term trend of subjective mood, indicating that time-series models designed to utilize such time-dependent patterns [32] are not appropriate for mood forecasting. However, the recorded mood scores do show auto-correlation several days ahead. Thus, we employed a multiple regression approach based on a history of predictor variables to forecast future mood scores. In particular, we proposed using a hierarchical Bayesian model to perform MTL, enabling personalized predictions while considering common characteristics of the population. The hierarchical approach additionally makes it possible to reason about individuals for whom we have observed little data, thus overcoming the cold start problem.
Employing a regression model approach to produce a forecast required us to find an appropriate window size defining the predictor variables included in the model. With perfect data and a model robust to overfitting, increasing the window size should never result in a worse model, since any added non-informative variables could simply be ignored. In a real-world application, however, increasing the window size often results in fewer training examples because of missing data and similarly requires more data to enable prediction on new instances. Thus, finding the optimal window size is a trade-off that depends on data quality and model robustness. In our experiment, we found that including a history of up to 4 days improved the prediction error, but with more complete data there is no reason the window size could not be increased even further. For instance, Surhara et al. found that their model for classifying depression benefited from long data histories up to 14 days [22], although it is our experience that collecting complete self-assessment histories over an extended period is very difficult.

By inspecting the inferred regression parameters of the hierarchical Bayesian model, we found historical mood to be the most important predictor of future mood. This result is not surprising since substantial changes in mood often occur over several days and thus future mood is likely to be similar to mood in the immediate past. Consequently, the forecast is inclined to extrapolate mood from previous days and gradually regress towards the mean of the data as uncertainty grows when forecasting further ahead. While this forecast behavior succeeds at achieving low error, its utility in a practical monitoring setting must be studied further. We see this as an interesting topic for future research.

The proposed hierarchical linear and ordinal models achieved the best predictive performance in time series cross-validation experiments. In leave-all-out cross-validation the hierarchical Bayesian linear regression model achieved the best result ($R^2 = 0.511, RMSE = 0.324$) with the hierarchical Bayesian ordinal model being a close second. In leave-one-out cross-validation the hierarchical Bayesian linear regression model also achieved the best result ($R^2 = 0.547, RMSE = 0.337$), but was much closer to the performance of the best pooled models. These results show how the hierarchical approach solves the cold start problem by including information from the population when little individual data is observed and gradually becoming more personalized as more data becomes available. In contrast to previous work we found that pooled models outperformed separate models indicating the individual datasets did not contain sufficient information to produce accurate forecasts. Thus, the separate models were biased and consequently it proved more useful to disregard individual differences and include data from the population in a general model. The hierarchical models succeeded in finding a compromise between the pooled and separate approach by regularizing the personalized models with data from the population.

In forecasting mood several days ahead, the hierarchical models likewise achieved the best results. As expected, the forecast error increased when forecasting further ahead, however, we observed that the best regression models performed better
than the naïve mean models for up to seven days ahead. It is a remarkable result that a short self-assessment history of just a few days can forecast mood several days ahead of time, the most important reason being that substantial mood changes often happen gradually over a horizon longer than seven days.

Limitations
We observed low prevalence of severe symptoms in our data sample leading to some limitations. Because the mood values have low variance, regression models will tend to regress towards the mean of the data, and naïve mean models are able to achieve low errors relative to the full range of the mood scale. It prevented us from assessing how well the proposed method performs in a population with more severe symptoms and neither can we assess how well the forecast is at anticipating severe cases.

A major motivation of our research and the MONARCA II study is to establish an online mood forecasting solution to improve monitoring and enable early intervention in bipolar disorder [27]. However, it is still not clear how an online forecast system is affected by interventions, since the intervention can change the future outcome and thus the future training data, which could lead to a biased model that underestimates future mood scores. Thus, it would be crucial to monitor the performance of an online system continuously using held out, unbiased data for validation.

Perspectives
The mood forecast presented in this paper have used a history of self-reported features as input. However, several research projects have been investigating the use of sensor-based and automatically collected data as input for mood prediction. Sensor technology in modern smartphones enables tracking of variety of behavioral features such as physical activity, location and sleep along with logs of communication and device usage. Even more detailed data can be captured with wearables such as wristbands and fitness trackers. Such sensor-based features could be used to augment or even reduce self-assessment in mood prediction tasks and thus reduce the need to prompt users for daily self-assessments. There is great potential in utilizing objectively collected sensor data in semi-automatic mood detection and forecasting.

Mood prediction and forecasting can be used as early warning signs in clinical treatment. This could be useful in e.g. a telemedicine setup in which trained nurses or other clinical personnel supervise patients in outpatient treatment. This could help catch early onset of major depressive or manic phases which can be addressed and handled early, which again could reduce the severity of symptoms and the degree of treatment. Hence, the need for re-admission could be reduced. The authors are currently working on implementing an online forecasting system evaluated as part of the RADMIS trials [43] in order to study its practical application, including investigating if such a system could potentially reduce readmission and hospitalization.
In this paper, we have examined the technical foundation of mood forecasting aimed at improving continuous disease monitoring. However, for a patient the prospect of experiencing depressed or elevated mood in the future might lead to changes in behavior and state of mind and in the worst case become a self-fulfilling prophecy. Therefore, online mood forecasting should be used with care and applied exclusively as a monitoring and early intervention tool for professionals rather than being presented directly to users.

**Conclusions**
Continuous symptoms monitoring and early detection are important components in the treatment of patients with bipolar disorder. Smartphones provide a unique platform for self-assessment and management of depression and mania and has the additional benefit of making data available for immediate analysis. In this work, we have examined the feasibility of establishing a mood forecast system based on self-assessments to provide additional insights and enable early intervention. We found that our proposed method of applying hierarchical Bayesian regression models is able to consistently outperform commonly used machine learning methods and forecast subjective mood up to seven days ahead.

**Acknowledgements**
This study was funded by the Innovation Fund Denmark through the RADMIS project and the Copenhagen Center for Health Technology (CACHET). We would like to thank everyone who participated in the MONARCA II trial as well as the clinical staff at the Psychiatric Center Copenhagen who helped facilitate the trial and collect the dataset used in this work.

**Conflicts of Interest**
JB, MFJ, and OW have no conflicts of interest. MF and JEB are founders and shareholders of Monsenso. LVK has during recent three years been a consultant for Sunovion and Lundbeck.

**Abbreviations**
EMA: ecological momentary assessment
HDRS: Hamilton Depression Rating Scale
MTL: multi-task learning
RMSE: root mean squared error
SD: standard deviation
YMRS: Young Mania Rating Scale

**Multimedia Appendix**
1. Stan code of the hierarchical linear regression and ordinal regression models and details on choice of model priors.
References


Multimedia Appendix 1: Stan code

This multimedia appendix includes the Stan code of the hierarchical linear regression and ordinal regression models (see below). In practice, the Stan code was wrapped in Python classes and used a string Template to customize the code. In particular, the numerical parameters of the hyperpriors were adjustable and the generated quantities section could be left out for increased sampling speed when it was not required. The Python classes additionally provided a scikit-learn style interface for fitting the model and making predictions as well as automatic caching of the compiled Stan model.

Priors

Because of our prior believe that current and previous observations of mood would be the strongest predictor of future mood, the population parameters of the hierarchical models corresponding to mood were assigned wider priors and the other population parameters were assigned narrower, more restrictive priors. In the hierarchical linear regression model, mean parameters, $\mu$, corresponding to mood were assigned Normal(0, 1) priors and population variance parameters, $\tau$, were assigned Normal(0, 0.1) priors. Mean parameters, $\mu$, not corresponding to mood were assigned Normal(0, 0.1) priors and population variance parameters, $\tau$, were assigned Normal(0, 0.01) priors. In the hierarchical ordinal regression model, mean parameters, $\mu$, corresponding to mood were assigned Normal(0, 10) priors and population variance parameters, $\tau$, were assigned Normal(0, 1) priors. Mean parameters, $\mu$, not corresponding to mood were assigned Normal(0, 1) priors and population variance parameters, $\tau$, were assigned Normal(0, 0.1) priors.

Additionally, our prior believe was that observed variables closer in time to the predicted future mood were more important, thus each prior was scaled with $1/\text{lag}$ of the corresponding predictor variable, i.e. for window size 4 ($w = 4$), population priors corresponding to predictor variables from the current day were scaled with 1/1, priors corresponding to the previous day were scaled with 1/2, priors corresponding to two days ago were scaled with 1/3 and priors corresponding to three days ago were scaled with 1/4. This helped to further regularize the model.
// Hierarchical linear regression model

data {
  int<lower=1> N; // number of examples
  int<lower=1> D; // number of predictors (dimensions)
  int<lower=1> J; // number of groups
  int<lower=1,upper=J> tid[N]; // group identifier index vector
  row_vector[D] X[N]; // example vectors
  real y[N]; // target vector
}

parameters {
  real mu_a; // intercept mean
  real<lower=0> tau_a; // intercept deviation
  vector[D] mu_b; // coefficient means
  vector<lower=0>[D] tau_b; // coefficient deviations
  real alpha_raw[J]; // intercepts
  vector[D] beta_raw[J]; // coefficients
  real<lower=0> sigma; // noise scale
}

transformed parameters {
  real alpha[J];
  vector[D] beta[J];
  for (j in 1:J) {
    alpha[j] = mu_a + tau_a * alpha_raw[j];
    beta[j] = mu_b + tau_b .* beta_raw[j];
  }
}

model {
  mu_a ~ normal(0, 1);
  tau_a ~ normal(0, 1);
  mu_b ~ normal(0, 1);
  tau_b ~ normal(0, 1);
  for (j in 1:J) {
    alpha_raw[j] ~ normal(0, 1); // normal(mu_a, tau_a)
    beta_raw[j] ~ normal(0, 1); // normal(mu_b, tau_b)
  }
  sigma ~ normal(0, 1);
  vector[N] z;
  for (i in 1:N)
    z[i] = alpha[tid[i]] + X[i] * beta[tid[i]];
  y ~ normal(z, sigma);
}

generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = normal_lpdf(y[i] | alpha[tid[i]] + X[i] * beta[tid[i]], sigma);
}
// Hierarchical ordinal regression model

data {
  int<lower=1> N; // number of examples
  int<lower=1> D; // number of predictors
  int<lower=2> K; // number of classes
  int<lower=1> J; // number of groups
  int<lower=1,upper=J> tid[N]; // group identifier index vector
  row_vector[D] X[N]; // examples
  int<lower=1, upper=K> y[N]; // targets
}

parameters {
  ordered[K-1] mu_c; // cutpoint means
  real<lower=0> tau_c; // cutpoint deviations
  ordered[K-1] c_raw[J]; // cutpoints
  vector[D] mu_b; // coefficient means
  vector<lower=0>[D] tau_b; // coefficient deviations
  vector[D] beta_raw[J]; // coefficients
}

transformed parameters {
  ordered[K-1] c[J];
  vector[D] beta[J];
  for (j in 1:J) {
    c[j][1] = mu_c[1] + tau_c * c_raw[j][1];
    for (k in 2:(K-1))
      c[j][k] = fmax(c[j][k-1] + 1.0e-5, mu_c[k] + tau_c * c_raw[j][k]);
    beta[j] = mu_b + tau_b .* beta_raw[j];
  }
}

model {
  mu_c ~ normal(0, 1);
  tau_c ~ normal(0, 1);
  mu_b ~ normal(0, 1);
  tau_b ~ normal(0, 1);
  for(j in 1:J) {
    c_raw[j] ~ normal(0, 1);
    beta_raw[j] ~ normal(0, 1); // normal(mu_b, tau_b);
  }
  for (i in 1:N)
    y[i] ~ ordered_logistic(X[i] * beta[tid[i]], c[tid[i]]);
}

generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = ordered_logistic_lpmf(y[i] | X[i] * beta[tid[i]], c[tid[i]]);
}
APPENDIX B

Daily Estimates of Clinical Severity in Bipolar Disorder from Smartphone Self-Assessments

Daily Estimates of Clinical Severity in Bipolar Disorder from Smartphone Self-Assessments

Jonas Busk (jbusk@dtu.dk), Maria Faurholt-Jepsen, Mads Frost, Jakob E. Bardram, Lars Vedel Kessing and Ole Winther

Abstract

The current golden standard for assessing the severity of symptoms of bipolar disorder is periodic evaluations using validated clinical rating scales. Frequent automatic estimation of clinical symptom severity can help support disease monitoring and treatment tasks between outpatient visits. The present study aims to study the feasibility of producing daily estimates of clinical severity based on smartphone self-assessments from a group of patients with bipolar disorder. We demonstrate how the estimates can be utilized to compute individual daily risk of relapse scores. Using 280 clinical severity ratings from 84 patients along with self-assessments we applied a hierarchical Bayesian modelling approach capable of providing individual estimates while learning characteristics of the population. The proposed method was evaluated in cross-validation experiments and compared to naive baselines and common machine learning methods. The depression severity model achieved a mean predicted $R^2$ of 0.57 (SD=0.10) and RMSE of 3.85 (SD=0.47) on the Hamilton Depression Rating Scale while the mania severity model achieved a mean predicted $R^2$ of 0.12 (SD=0.31) and RMSE of 3.74 (SD=0.46) on the Young Mania Rating Scale. In both cases, self-reported mood was the most important predictor variable of all the self-assessment items. Smartphone self-assessments can be used to automatically estimate clinical severity of depression and mania in patients with bipolar disorder and identify individuals with high risk of relapse.

Introduction

Bipolar disorder is a common and complex illness with an estimated prevalence of 1-2%, and is regarded as one of the most important causes of disability worldwide [1, 2]. It is characterized by recurrent episodes of depression, (hypo)mania and mixed episodes intervened by periods of euthymia [3] and with a high degree of comorbidity and functional impairment [4]. Bipolar disorder is associated with an elevated risk of mortality due to suicide and medical comorbidities such as cardiovascular disease and diabetes [5, 6, 7], and among people with bipolar disorder, life expectancy is decreased 8 to 12 years [8, 9]. In clinical practice,
there are major challenges in diagnosing and treating bipolar disorder [10]. Patients with bipolar disorder are often misdiagnosed, and the correct diagnosis can be delayed for several years after illness onset [11, 12, 13]. Currently, due to the lack of objective tests, the diagnostic process and the clinical assessment of the severity of depressive and manic symptoms relies on subjective information, clinical evaluation and rating scales [14]. Periodic clinical evaluations using clinical rating scales such as the Hamilton Depression Rating Scale (HDRS) [15] and the Young Mania Rating Scale (YMRS) [16] are currently used as the golden standard for assessing the severity of depressive and manic symptoms in patients with bipolar disorder. Each of these rating scale consists of a series of items that are scored and finally added up to produce a total score summarizing the current severity of affective state of the patient. This subjective evaluation involves a risk of patient recall bias, other recall distortions, decreased illness insight (mainly during affective episodes) and individual observer bias [17, 18, 19, 20, 21]. Furthermore, clinical evaluations are time consuming and require a specialist who is trained in using the rating scales to produce consistent and reliable results.

Additionally, patients may be asked to perform daily self-assessments to track day-to-day behavior in between clinical evaluations. Modern smartphones provide a unique platform for fine-grained real-time disease monitoring and management and a convenient means of self-assessment that have traditionally been carried out on paper [22, 23, 24]. A smartphone-based monitoring system enables users to ubiquitously record and review their own data, receive reminders and even share data with carers and clinicians. From the perspective of health care providers, it provides efficient, online monitoring of a group of patients and enables intervention in case any alarming behavior is observed. Electronic self-monitoring has the additional benefit of making data available for immediate and automatic analysis that can help support monitoring and treatment tasks between outpatient visits.

Correlations between self-reported mood scores and clinical severity ratings in patients with bipolar disorder have already been demonstrated by previous work [25, 26, 27]. To our knowledge this is the first study that aims to predict scores of clinical severity directly from combinations of smartphone-based self-assessed data in patients with bipolar disorder. In related work, detection of daily self-reported mood from smartphone sensor and usage data is well studied [28, 23, 29, 30] but remains a difficult problem due to noisy data. In [31], Grunerbl et al. classify affective states and change labels derived from clinical ratings and phone interviews of patients with bipolar disorder from a combination of smartphone sensor modalities and argue that detecting deviations from the normal state is more important than the recognition of a particular state in practical applications. Several studies in the field of affective computing have highlighted the need for personalized models to account for individual differences in order to achieve good predictive performance [29, 30, 32, 33]. However, a separate analysis is not feasible until sufficient data about each individual is available. Hierarchical Bayesian modelling is popular approach for providing individual models while burrowing statistical power from the population and
is especially useful when the individual datasets are too small to be analyzed separately [34].

The main objective of the study is to examine the feasibility of producing daily estimates of clinical severity scores of depression and mania based on smartphone self-assessments collected from a group of patients with bipolar disorder participating in a randomized controlled trial (RCT). Additionally, we demonstrate how uncertainty in the estimated quantities can be used to compute individual, daily risk of relapse, useful for identifying high risk individuals who need urgent assistance. We hypothesize that daily estimates of clinical severity scores extended with computation of individual relapse risk scores are more interpretable and actionable results than summarizing the self-assessments directly and can be a valuable tool in continuous disease monitoring and treatment of patients with bipolar disorder.

Materials and Methods

Patients and study design

Data used for this study was collected between September 2014 and January 2018 during the MONARCA II RCT investigating the effect of smartphone-based monitoring in patients with bipolar disorder [35]. All patients with a diagnosis of bipolar disorder who had previously been treated at the Copenhagen Clinic for Affective Disorder, Copenhagen, Denmark in the period from 2004 to January 2016 and who at the time of recruitment were being treated at community psychiatric centers, private psychiatrists and general practitioners were invited to participate in the trial. The clinic is a specialized outpatient clinic with a catchment area consisting of the Capital Region in Denmark corresponding to 1.4 million people. Patients with a newly diagnosis of bipolar disorder or with treatment-resistant bipolar disorder were referred to the clinic. The staff consists of specialists in psychiatry, psychologists, nurses, and a social worker, all with specific experience and knowledge regarding bipolar disorder. Treatment at the clinic comprises a two-year program including combined evidence-based psychopharmacological treatment and supporting therapy, including group psychoeducation [36]. Patients were included in the study for a nine-month follow-up period if they had a bipolar disorder diagnosis according to ICD-10 using the Schedules for Clinical Assessments in Neuropsychiatry (SCAN) [37] and previously were treated at the Copenhagen Clinic for Affective Disorder. Patients with schizophrenia, schizotypal or delusional disorders, previous use of the MONARCA system, pregnancy and lack of Danish language skills were excluded. Patients with other comorbid psychiatric disorders and substance use were eligible for the trial. As part of the MONARCA II trial, patients were randomized to either using a smartphone-based monitoring system (the Monsenso system) for daily self-monitoring (the intervention group) or to treatment as usual (the control group). Patients from the intervention group who successfully provided smartphone-based self-monitoring data were included in the analyses in the present report.
Data description

The dataset consists of 280 clinical evaluations from 84 patients. Each clinical evaluation includes ratings for severity of depression and mania using the Hamilton Depression Rating Scale (HDRS) [15] and the Young Mania Rating Scale (YMRS) [16], respectively. All clinical assessments were conducted by researchers (MFJ), who were blinded to all smartphone-based data. Thus, data on severity of depressive and manic symptoms were collected rater-blinded. On both rating scales, low severity ratings indicate weak symptoms while high severity ratings indicate severe symptoms. A score of 13 or more on either scale is classified as a depressive or manic episode, respectively, while a high score on both scales at the same time constitutes a mixed episode. Clinical ratings with the HAMD and the YMRS are considered to be valid on the day of the evaluation as well as the 3 previous days, thus each rating is attributed a total of 4 days in the dataset.

In addition to periodic clinical evaluations, participants were instructed to carry out daily self-assessments via a smartphone application configured for the study [23]. The smartphone application enabled users to set and receive reminders and users were allowed to provide self-assessments retrospectively for up to 2 days in case they forgot do so daily. The self-assessment questionnaire included the following items: activity level (scored from -3 to +3); alcohol consumption (number of units from 0 to 10+); anxiety level (scored from 0 to 2); irritability level (scored from 0 to 2); cognitive problems (scored from 0 to 2); medicine adherence (not taken/taken/taken with changes); mixed mood (yes/na); mood (scored from -3 to +3 including -0.5 and +0.5); sleep duration (in hours); and stress level (scored from 0 to 2).

Data preprocessing

Three of the self-assessment variables required preprocessing prior to analysis. We split the mood variable into a negative and positive component, mood negative and mood positive, allowing for non-linear relationships with the clinical severity ratings as we expected negative mood to be associated mainly with severity of depression (HDRS) and positive mood to be associated mainly with severity of mania (YMRS). Additionally, we hypothesized that the relationship between symptom severity and sleep duration is linear as sleeping a lot or very little can both represent warning signs. To encode this, we subtracted the individual mean of the sleep variable and split the result into a positive and a negative component, sleep negative and sleep positive. The Medicine variable is categorical by design with categories: medicine not taken, medicine taken, medicine taken with changes. Therefore the answers were encoded with two exclusive binary variables indicating if medicine was not taken, medicine omitted, or if medicine was taken with changes, medicine changed. Finally, we normalized all variables by their allowed minimum and maximum values to allow for easier selection of model hyperparameters and interpretation of the inferred model weights.
It is a common problem for users to occasionally forget to fill in their daily self-assessment, resulting in missing values in the dataset. In most cases, self-assessments are either missing or complete, but in a few instances, they are only partially missing. To avoid discarding rows with only a few missing values, we experimented with filling in values from the previous day, which is a popular method for dealing with missing values in time series data [38]. However, it resulted in very few additional complete observations and we therefore decided to leave this step out for simplicity.

Statistical methods

When analyzing several related sets of measurements, such as data from individuals in a population, the two extreme approaches are to either pool the datasets in a one-size-fits-all solution or analyze them separately, which is only possible when sufficient data is available (also known as the cold start problem). A hierarchical Bayesian approach provides an intermediate solution that enables personalized models while learning the characteristics of the population [39]. In a hierarchical Bayesian regression model, individuals have their own set of regression intercept and weights, \( \alpha_j, \beta_j \), sampled from a common population distribution parameterized by population-level means \( \mu \) and variances \( \tau \) determining the amount of pooling:

\[
\alpha_j, \beta_j \sim \text{Normal}(\mu, \tau) \quad (1),
\]

\[
y_{ji} \sim \text{Normal}(\alpha_j + \beta_j^T x_{ji}, \sigma) \quad (2),
\]

where \( y_{ji} \) is the \( i \)th observation of the target variable for individual \( j \), \( x_{ji} \) are the corresponding predictor variables and \( \sigma \) is the standard error. This hierarchical tying together of parameters means that data from the population helps regularize the individual-level weights. An additional benefit of the Bayesian approach is that it expresses uncertainty in all the model parameters and predictions through their posterior distributions, which is important for interpretability of the model. For more details, a complete description of the hierarchical Bayesian model is provided in the SI.

We use Stan [40] to specify and perform inference in the hierarchical Bayesian model and compare the predictive results with pooled and separate naïve baselines and common machine learning methods: Ridge Regression from the scikit-learn machine learning library [41] and XGBoost regression from the XGBoost Python package [42]. Details of the Stan setup is also included in SI. To estimate the predictive performance of the models we designed a cross-validation experiment where in each iteration we hold out one randomly sampled clinical evaluation (consisting of up to 4 days of data) from each individual and use the remaining data to fit the models. This procedure is repeated \( K \) times and the predicted coefficient of determination (\( R^2 \)) and root mean squared error (RMSE) is computed on the held-out data in each iteration.
Computing risk of relapse

Applying a Bayesian approach provides a probability distribution of unobserved outcomes given previously observed data (the posterior predictive distribution), which can be utilized to reason about uncertainty in the predictions. Specifically, samples from the posterior predictive distribution can be used to compute the probability that an unobserved outcome, $\tilde{y}_{ji}$, exceeds a predefined threshold, $T$:

$$\Pr(\tilde{y}_{ji} \geq T).$$

(3)

When estimating scores of clinical severity, by applying a threshold $T = 13$ we can interpret this probability as the risk that an individual is experiencing severe symptoms and utilize it as a personal score indicating the risk of relapse. In some practical applications, it may be more relevant to accurately identify high risk individuals than it is to estimate the exact value of the severity score.

Ethical considerations

The MONARCA II RCT was approved by the Regional Ethics Committee in the Capital Region of Denmark (H-2-2014-059) and the Danish Data protection agency (2013-41-1710). The law on handling of personal data was respected. Prior to commencement the trial was registered at ClinicalTrials.gov (NCT02221336). Electronic data collected from the smartphones were stored at a secure server at Concern IT, Capital Region, Denmark (I-suite number RHP-292 2011-03). The trial complied with the Helsinki Declaration of 1975, as revised in 2008.

Results

Descriptive statistics

The MONARCA II dataset consists of 280 clinical evaluations along with 15,975 daily self-evaluations from 84 patients assigned to the intervention group. Ages ranged from 21 to 71 years (mean=43.1, SD=12.4) and the population had 61.9% females. Most patients presented with rather low severity of depressive and manic symptoms during the study period resulting in low HDRS and YMRS ratings. The mean HDRS rating was 7.56 (SD=6.29) and 20.4% of ratings were greater than or equal to 13. The mean YMRS rating was 2.85 (SD=4.17) and 5.0% of ratings were greater than or equal to 13. Similarly, the majority of the self-reported mood scores were close to zero (mean=-0.14, SD=0.48), indicating neutral mood (euthymia).
Figure 1: Distributions of clinical ratings of symptom severity and self-reported mood scores. A negative mood score is expected to indicate a high HDRS rating and a positive mood score is expected to indicate a high YMRS rating. HDRS and YMRS ratings are rarely high at the same time (indicating mixed mood). Consequently, data is expected to primarily occupy the white background areas of the scatter plots.

After filling back the clinical severity ratings 4 days there were 764 observations with associated self-assessments. Figure 1 shows the relationships between the clinical severity and self-reported mood scores. We see that generally a high HDRS rating corresponds to an experience of neutral or depressed mood ($r = -0.40, P < 0.01$) while a high YMRS rating corresponds to neutral or elevated mood ($r = 0.22, P < 0.001$). Only in a few instances were HDRS and YMRS rated high at the same time, indicating a mixed episode ($r = 0.13, P = 0.02$).

**Model estimates**

The fit of the hierarchical Bayesian regression model was evaluated on the entire dataset of clinical ratings combined with completed self-assessments for all participants with at least two data points ($N=433$). The HDRS model achieved an $R^2$ of 0.84, meaning that the model accounted for 84% of the variance in the data, and a residual root mean squared error (RMSE) of 2.41. The YMRS model achieved an $R^2$ of 0.81 and a residual RMSE of 2.07.

The distributions of inferred population-level mean, $\mu$, and variance, $\tau$, parameters in the hierarchical Bayesian regression HDRS and YMRS models are summarized in Table 1. The absolute t-statistic of the mean parameters, computed as the mean scaled by the standard error of the parameter: $t_\mu = \bar{\mu}/SE(\mu)$, is included as a measure of variable importance, following the intuition that larger absolute weights and lower variance implies importance [43]. This shows that negative mood was the most important predictor variable in the HDRS model while positive mood was the most important predictor and in the YMRS model. A visual representation of the population-level parameters is presented in Figure 2.

Figure 5 in the SI presents the effect size of each self-assessment variable
<table>
<thead>
<tr>
<th>HDRS</th>
<th>µ</th>
<th>τ</th>
<th>Predictor</th>
<th>Mean (SD)</th>
<th>95% CI</th>
<th>t</th>
<th>Mean (SD)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.43 (0.66)</td>
<td>5.13 - 7.73</td>
<td>9.67</td>
<td>4.10 (0.50)</td>
<td>3.23 - 5.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood negative</td>
<td>-9.11 (1.40)</td>
<td>-11.94 - -6.43</td>
<td>6.51</td>
<td>0.56 (0.40)</td>
<td>0.02 - 1.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep negative</td>
<td>-6.48 (1.66)</td>
<td>-9.72 - -3.19</td>
<td>3.89</td>
<td>0.42 (0.31)</td>
<td>0.02 - 1.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed mood</td>
<td>2.11 (0.67)</td>
<td>0.79 - 3.42</td>
<td>3.15</td>
<td>0.44 (0.32)</td>
<td>0.02 - 1.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>2.26 (0.86)</td>
<td>0.58 - 3.96</td>
<td>2.63</td>
<td>0.38 (0.28)</td>
<td>0.02 - 1.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine changed</td>
<td>-1.81 (0.71)</td>
<td>-3.19 - -0.40</td>
<td>2.55</td>
<td>0.35 (0.27)</td>
<td>0.01 - 0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive difficulty</td>
<td>1.09 (0.73)</td>
<td>-0.35 - 2.48</td>
<td>1.50</td>
<td>0.43 (0.32)</td>
<td>0.02 - 1.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood positive</td>
<td>-2.80 (1.94)</td>
<td>-6.59 - 0.94</td>
<td>1.44</td>
<td>0.42 (0.32)</td>
<td>0.02 - 1.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep positive</td>
<td>2.83 (2.05)</td>
<td>-1.09 - 6.90</td>
<td>1.38</td>
<td>0.41 (0.31)</td>
<td>0.02 - 1.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>0.53 (0.61)</td>
<td>-0.66 - 1.71</td>
<td>0.88</td>
<td>0.50 (0.35)</td>
<td>0.02 - 1.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress</td>
<td>0.56 (0.73)</td>
<td>-0.86 - 1.99</td>
<td>0.76</td>
<td>0.50 (0.36)</td>
<td>0.02 - 1.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.59 (1.01)</td>
<td>-1.39 - 2.54</td>
<td>0.59</td>
<td>0.41 (0.31)</td>
<td>0.02 - 1.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine omitted</td>
<td>0.52 (0.97)</td>
<td>-1.38 - 2.42</td>
<td>0.54</td>
<td>0.37 (0.28)</td>
<td>0.01 - 1.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irritable</td>
<td>0.05 (0.74)</td>
<td>-1.41 - 1.49</td>
<td>0.06</td>
<td>0.59 (0.42)</td>
<td>0.02 - 1.57</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>YMRS</th>
<th>µ</th>
<th>τ</th>
<th>Predictor</th>
<th>Mean (SD)</th>
<th>95% CI</th>
<th>t</th>
<th>Mean (SD)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.10 (0.68)</td>
<td>1.79 - 4.46</td>
<td>4.59</td>
<td>4.35 (0.50)</td>
<td>3.48 - 5.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood positive</td>
<td>12.83 (1.90)</td>
<td>9.09 - 16.53</td>
<td>6.75</td>
<td>0.57 (0.42)</td>
<td>0.02 - 1.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood negative</td>
<td>3.42 (1.30)</td>
<td>0.87 - 5.99</td>
<td>2.63</td>
<td>0.66 (0.46)</td>
<td>0.03 - 1.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irritable</td>
<td>1.31 (0.69)</td>
<td>-0.05 - 2.68</td>
<td>1.90</td>
<td>0.71 (0.47)</td>
<td>0.04 - 1.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed mood</td>
<td>1.02 (0.62)</td>
<td>-0.20 - 2.21</td>
<td>1.65</td>
<td>0.54 (0.36)</td>
<td>0.03 - 1.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress</td>
<td>1.15 (0.70)</td>
<td>-0.21 - 2.54</td>
<td>1.65</td>
<td>1.24 (0.53)</td>
<td>0.12 - 2.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep positive</td>
<td>-2.69 (1.84)</td>
<td>-6.30 - 0.84</td>
<td>1.46</td>
<td>0.40 (0.30)</td>
<td>0.01 - 1.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>-0.78 (0.56)</td>
<td>-1.88 - 0.30</td>
<td>1.39</td>
<td>0.63 (0.41)</td>
<td>0.03 - 1.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine changed</td>
<td>0.46 (0.71)</td>
<td>-0.99 - 1.81</td>
<td>0.64</td>
<td>0.80 (0.48)</td>
<td>0.05 - 1.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive difficulty</td>
<td>0.41 (0.69)</td>
<td>-0.92 - 1.78</td>
<td>0.59</td>
<td>0.94 (0.54)</td>
<td>0.05 - 1.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.18 (0.80)</td>
<td>-1.40 - 1.73</td>
<td>0.23</td>
<td>0.69 (0.48)</td>
<td>0.03 - 1.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep negative</td>
<td>0.30 (1.50)</td>
<td>-2.63 - 3.26</td>
<td>0.20</td>
<td>0.43 (0.32)</td>
<td>0.02 - 1.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol</td>
<td>0.05 (0.93)</td>
<td>-1.77 - 1.86</td>
<td>0.06</td>
<td>0.39 (0.30)</td>
<td>0.02 - 1.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine omitted</td>
<td>-0.02 (0.90)</td>
<td>-1.78 - 1.75</td>
<td>0.02</td>
<td>0.41 (0.31)</td>
<td>0.01 - 1.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Weight table showing the mean, standard deviation (SD) and 95% credible interval (CI) of the population-level parameters in the HDRS model (top) and YMRS model (bottom). The population-level regression weight means, µ, are summarized in the leftmost columns and sorted by variable importance computed as the absolute t-statistic of the mean parameter. The corresponding variances, τ, are summarized in the columns to the right and can be interpreted as the amount of pooling of the given variable in the hierarchical model.
Figure 2: Distributions of the population-level regression parameters, $\mu$, and variances, $\tau$, in the hierarchical HDRS model (top) and YMRS model (bottom). In both models, self-reported mood has the largest absolute weight and group variance.

Cross-validation results

The predictive performance of the hierarchical Bayesian model was evaluated in a $K=100$ cross-validation experiment on all data where participants had complete observations of clinical ratings and self-assessments from at least three different clinical evaluations ($N=329$). In each iteration, data from one randomly sampled clinical evaluation from each user was held out and the remaining data was used to fit the models. Table 2 presents the cross-validation results where the hierarchical Bayesian model is compared to naïve pooled and separate mean models along with pooled and separate ridge regression and XGBoost regression models. Because of low variance in the data, the naïve mean models performed relatively well. Still the hierarchical Bayesian regression model achieved the best performance and is significantly better than the separate mean model in both the HDRS and YMRS case according to independent $t$-tests ($P < 0.001$). Overall, the separate models performed better than their pooled counterparts.
Table 2: Results of $K=100$ cross-validation experiments showing predicted coefficient of determination ($R^2$) and root mean square error (RMSE) of the HDRS (left) and YMRS (right) models, respectively. The hierarchical Bayesian model achieves the best overall performance in both cases and can predict the clinical severity ratings within 4 points of RMSE on the original scale.

Predicted risk of relapse scores

The prediction results from the cross-validation experiment presented in the previous section was used to compute risk of relapse scores $\Pr(\tilde{y}_{ji} \geq T = 13)$. The ability of the model to correctly assign high risk to instances with high ratings can be evaluated as a binary classification problem with severity ratings equal to or greater than the threshold $T$ constituting the positive class. Figure 3 presents receiver operating characteristic (ROC) curves of the HDRS and YMRS models illustrating the trade-off between true positive rate (TPR) and false positive rate (FPR), comparing the hierarchical Bayesian regression model to the naïve pooled and separate mean models. The pooled mean model corresponds to a model that either classifies all instances as low risk or high risk, achieving an area under the curve (AUC) of 0.50 in both the HDRS and YMRS case. The separate mean model independently classifies each individual as either high or low risk based on observed values of the ratings and achieved an AUC of 0.67 in the HDRS case and AUC of 0.49 in the YMRS case. The hierarchical Bayesian regression model was able to account for information in the self-assessments as well as individual differences and achieved the highest AUC of 0.89 in the HDRS case and 0.84 in the YMRS case.

Discussion

In the present study, we have analyzed clinical severity ratings of depression (HDRS) and mania (YMRS) along with daily smartphone self-assessments including self-reported mood in a population of 84 patients with bipolar disorder participating in the MONARCA II RCT. As hypothesized, our data analysis revealed a negative correlation between HDRS and self-reported Mood ($r = -0.40, P < 0.001$) and a positive correlation between YMRS and Mood ($r = 0.22, p < 0.01$). This confirms previous work [25, 26, 27], and suggests that self-reported mood is a valid indicator of symptom severity in patients with bipolar
disorder and thereby a clinically relevant feature for monitoring and analysis.

Next, we found that the proposed approach of applying hierarchical Bayesian regression models was able to fit the data distributions well and accounted for more than 80% of the variance in the data (HDRS: $R^2 = 0.84$, YMRS: $R^2 = 0.81$). Using the absolute $t$-statistic of the population-level regression weights as a measure of variable importance, decreased and increased self-reported mood were the most important variables for predicting the severity of depression and mania. This is not surprising since sampling of self-reported mood from the patients was designed to collect indicators on the patient’s affective state and thus should reflect the clinically rated symptoms. Other important variables in the HDRS model were decreased sleep and feelings of mixed mood and anxiety while in the YMRS model only mood ranked important (see Table 1).

To assess the predictive performance of the hierarchical Bayesian model compared to pooled and separate baseline models, we performed cross-validation experiments of HDRS and YMRS, respectively. The results presented in Table 2 showed that we were able to estimate clinical severity ratings using regression models based on smartphone self-assessments. The hierarchical Bayesian model achieved the best performance in predicting HDRS ($R^2 = 0.57 (0.10)$, RMSE = 3.85 (0.47)) and was significantly better than the separate mean model ($P < 0.001$). Similarly, the hierarchical Bayesian model was best at predicting YMRS ($R^2 = 0.12 (0.31)$, RMSE = 3.74 (0.46)) and was significantly better than the separate mean model ($P < 0.001$). Overall the YMRS models did not account for much of the variance in the data, indicated by the low $R^2$ scores. We believe this is mainly because of low variation in the observed YMRS data and that YMRS ratings are not more difficult to recognize automatically than HDRS ratings when sufficient data is available.

In practical settings of disease monitoring, detecting individuals with a high
risk of relapse is important in order to enable intervention. Therefore, a sensitive indication if a symptom severity rating is above a critical threshold might be more useful than estimating the exact value of the severity rating itself. Thus, we demonstrated how uncertainty in the estimated severity ratings can be utilized to compute individual daily risk of relapse scores by considering samples from the posterior predictive distribution of the hierarchical Bayesian model. In both the HDRS and YMRS case, using hierarchical Bayesian approach achieved substantial improvements over naïve pooled and separate mean models. Figure 3 shows the trade-off between true positive rate and false positive rate in the HDRS and YMRS models for identifying high risk individuals. This shows that including self-assessments in a regression model provided additional useful information for estimating the level of the clinical severity ratings and hence the relapse risk scores, which is a promising result.

The findings that a combination of fine-grained daily self-assessment items can be used to estimate clinical severity ratings are interesting and innovative. Daily longitudinal self-monitoring of mood symptoms gives valuable information of mood fluctuation experienced by patients with bipolar disorder. Long-term monitoring of symptoms has been an essential part of the monitoring and treatment of bipolar disorder for decades [44] and rapidly evolving smartphone technologies have made it possible to monitor symptoms more continuously. This can be clinically relevant for detection of prodromal symptoms before the first or recurrent depressive or manic episodes [45]. In the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), increased activity level or energy is acknowledged as one of the core features of hypomania and mania together with mood changes [46]. Several studies using factor analysis have described activation and not mood state as the primary symptom in manic episodes [47, 48]. However, in the present study we found mood to be the most important predictor variable for estimating HDRS and YMRS severity ratings while activity presented with low importance in both models. Furthermore, sleep disturbances and anxiety has been identified as an early symptom of depression and mania [49, 50] which is in line with our findings in the HDRS model while sleep and anxiety are less important in the YMRS model. Nonetheless, variation in mood, sleep, anxiety and activity levels are central symptoms in bipolar disorder and in the prodromal period before onset of bipolar disorder [51]. The four symptoms are possible to monitor by daily self-evaluation and subjective experiences of symptoms are very valuable information and necessary for detecting prodromal symptoms in individuals with high risk of developing bipolar disorder and for monitoring symptom and preventing new episodes in patients with bipolar disorder.

**Advantages**

The patients included from the RCT in the present study were clinically well characterized and were receiving treatment or had received treatment at the Copenhagen Clinic for Affective Disorders, Denmark. The clinical evaluations were conducted multiple times during follow-up by experienced researchers with
a specific knowledge within bipolar disorder. The smartphone-based system used in the present studies (the Monsenso system) was developed by the authors and has been shown easy to use with a high usability, usefulness, ease of learning to use and interface quality – also when compared with other smartphone-based systems [22, 52]. The use of smartphones for real-time fine-grained monitoring reduced the risk of recall bias. The proposed hierarchical Bayesian modeling approach is well suited for analysis of small related datasets, especially when the individual datasets are too small to analyze separately. Additionally, the linear regression method and ability to express uncertainty in all estimated quantities makes the model easy to interpret, which is essential in a clinical setting. Overall, the findings from the present study are found to be generalizable to patients with bipolar disorder not presenting with affective episodes who are willing to use a monitoring tool during prolonged time periods.

Limitations

The dataset used in this study primarily contained low ratings of clinical severity indicating most participants did not experience severe symptoms of depression or mania during the study period. Similarly, most of the self-reported mood scores were close to zero (indicating euthymia) and had low variance. Consequently, the naïve mean models fit the data well and achieved good performance in the prediction task. However, the best regression model was still significantly better than the mean models showing that it is possible utilize self-reported data to produce more accurate estimates of the clinical ratings of symptom severity. Furthermore, the absence of high ratings makes it difficult to reason about how good the models are at detecting extreme cases, which are naturally the most critical in a monitoring and intervention application.

Perspectives and future implications

Smartphones have become an ubiquitous technology in modern society and can be utilized to provide improved and personalized disease management and monitoring. Smartphone-based self-assessment makes data available for immediate analysis and can enable new tools for improved disease monitoring. In particular, accurate, daily estimates of symptom severity could help identify critical cases and enable timely intervention. Additionally, advances in sensor technology and algorithms is making it possible to extract a growing range of increasingly accurate behavioral features directly from sensor data. Utilizing these automatically generated features to infer symptom severity scores could be used to eliminate the need for frequent, intrusive self-assessments and improve the user experience of disease monitoring systems going forward.
Conclusions

In the present study, clinical ratings of the severity of depression and mania were estimated from smartphone-based self-assessments collected from patients with bipolar disorder participating in an RCT. We found that our approach of applying a hierarchical Bayesian model was able to estimate severity of depression and mania with low error compared to commonly used baseline methods and within 4 points of RMSE on the original HDRS and YMRS rating scales. Furthermore, we showed how uncertainty in the estimates can be utilized to compute personal relapse risk scores suited for identifying critical cases of patients experiencing severe symptoms of bipolar disorder and that our approach achieved substantial improvements over pooled and separate mean models. The results presented in this work show that it is feasible to produce daily estimates of clinical severity ratings of depression and mania from smartphone-based self-assessments, which can be used to improve and automate continuous disease monitoring and treatment of bipolar disorder.

Acknowledgments and funding

We would like to thank the participants of the MONARCA II trial as well as the clinical staff at the Psychiatric Center Copenhagen who helped facilitate the trial and collect the dataset. The study was funded by the Innovation Fund Denmark through the RADMIS project and the Copenhagen Center for Health Technology (CACHET).

Conflicts of interest

JB, MFJ, and OW have no conflicts of interest. MF and JEB are founders and shareholders of Monsenso. LVK has during recent three years been a consultant for Sunovion and Lundbeck.
References


Supplementary Information

Complete hierarchical Bayesian regression model

For individual $j = 1 : J$, observation $i = 1 : N$, target variable $y_{ji}$ and predictor variables $x_{ji}$:

$$y_{ji} \sim \text{Normal}(\alpha_j + \beta_j^T x_{ji}, \sigma)$$

where $\alpha_j$ and $\beta_j$ are sampled from population distributions:

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \tau_\alpha)$$
$$\beta_j \sim \text{Normal}(\mu_\beta, \tau_\beta)$$

with independent normal priors on $\mu$, $\tau$ and $\sigma$. A Bayesian network of the hierarchical Bayesian regression model is presented in Figure 4.

![Bayesian network of a hierarchical linear regression model](image)

Figure 4: Bayesian network of a hierarchical linear regression model. Individual regression intercept $\alpha_j$ and weights $\beta_j$ are drawn from population distributions parameterized by $\mu_\alpha$, $\tau_\alpha$ and $\mu_\beta$, $\tau_\beta$. This allows the model to account for individual differences while constraining individual parameters to be similar across the population.

Stan details

The pooled and hierarchical Bayesian models were fitted using the statistical modeling platform Stan [40]. Stan provides tools for specifying statistical models in a probabilistic programming language and performing inference in the models. The outputs of the inference are posterior samples of model parameters along with an inferential summary. The models in this work were fitted using 4 sampling chains and 5,000 iterations where the first half was warm-up.
Figure 5: Effect plots of the HDRS model (left) and YMRS model (right) computed on the entire dataset. The effect is computed as the product between individual-level weights of the hierarchical Bayesian HDRS and YMRS models and the observed data. Because the dataset is sparse, i.e. many values are zero, the effects are often small and, consequently, substantial effects represents outliers.
Appendix C

Objective smartphone data as a potential diagnostic marker of bipolar disorder

Objective smartphone data as a potential diagnostic marker of bipolar disorder

Maria Faurholt-Jepsen, Jonas Busk, Helga Þórarinsdóttir, Mads Frost, Jakob Eyvind Bardram, Maj Vinberg and Lars Vedel Kessing

Abstract

Objective: Currently, the diagnosis in bipolar disorder relies on patient information and careful clinical evaluations and judgements with a lack of objective tests. Core clinical features of bipolar disorder include changes in behaviour. We aimed to investigate objective smartphone data reflecting behavioural activities to classify patients with bipolar disorder compared with healthy individuals.

Methods: Objective smartphone data were automatically collected from 29 patients with bipolar disorder and 37 healthy individuals. Repeated measurements of objective smartphone data were performed during different affective states in patients with bipolar disorder over 12 weeks and compared with healthy individuals.

Results: Overall, the sensitivity of objective smartphone data in patients with bipolar disorder versus healthy individuals was 0.92, specificity 0.39, positive predictive value 0.88 and negative predictive value 0.52. In euthymic patients versus healthy individuals, the sensitivity was 0.90, specificity 0.56, positive predictive value 0.85 and negative predictive value 0.67. In mixed models, automatically generated objective smartphone data (the number of text messages/day, the duration of phone calls/day) were increased in patients with bipolar disorder (during euthymia, depressive and manic or mixed states, and overall) compared with healthy individuals. The amount of time the smartphone screen was ‘on’ per day was decreased in patients with bipolar disorder (during euthymia, depressive state and overall) compared with healthy individuals.

Conclusion: Objective smartphone data may represent a potential diagnostic behavioural marker in bipolar disorder and may be a candidate supplementary method to the diagnostic process in the future. Further studies including larger samples, first-degree relatives and patients with other psychiatric disorders are needed.

Keywords
Bipolar disorder; healthy control individuals, smartphone, diagnostic behavioural marker

Introduction

Bipolar disorder (BD) is characterized by changes in mood with episodes of depression, (hypo)mania and mixed episodes with intervening periods of euthymia (Phillips and Kupfer, 2013). Core clinical features of BD include changes in psychomotor activity and behavioural activities, and episodic shifts in energy, activity, sleep and other behavioural aspects that may be quantified objectively (Beigel and Murphy, 1971; Faurholt-Jepsen et al., 2012; Kuhs and Reschke, 1992; Kupfer et al., 1974; Mitchell et al., 2008; Popescu et al., 1991; Sobin and Sackeim, 1997). More specifically, changes in social activity (Weinstock and Miller, 2008), i.e., engaging in social relations, as well as physical activity (Faurholt-Jepsen et al., 2012; Kuhs and Reschke, 1992; Kupfer et al., 1974) represent the central aspects of BD that may be possible to measure objectively.

1Department O, Psychiatric Center Copenhagen, Rigshospitalet, Copenhagen Affective Disorder Research Center, Copenhagen, Denmark
2DTU Compute, Danish Technical University, Lyngby, Denmark
3The Copenhagen Center for Health Technology, Danish Technical University, Lyngby, Denmark
4The Pervasive Interaction Laboratory (PIT Lab), IT University of Copenhagen, Copenhagen, Denmark

Corresponding author:
Maria Faurholt-Jepsen, Psychiatric Center Copenhagen, Rigshospitalet, Blegdamsvej 9, Copenhagen 2100, Denmark. Email: maria@faurholt-jepsen.dk
Currently, due to the lack of objective tests, the diagnostic process as well as the clinical assessment of the severity of depressive and manic symptoms in BD relies on patient information and information from relatives, clinical observations and evaluations, and clinical rating scales (Phillips and Kupfer, 2013). Thus, the reliance on patient information and clinical evaluations raises issues including recall bias, decreased illness insight and differences in assessment experience (Cassidy, 2010). Therefore, objective methods for diagnosis and illness activity monitoring would be a tremendous clinical advantage.

Mobile health (mHealth) refers to health services delivered by mobile devices, such as mobile phones, mobile monitoring devices, personal digital assistants (PDAs) and other wireless electronic devices (World Health Organization (WHO), 2011). mHealth is a relatively new area within health care, and the use of sensors embedded within mobile monitoring devices could provide opportunities for new areas of research, development and treatment. A report by the WHO in 2011 stated ‘the use of mobile and wireless technologies to support the achievement of health objectives (mHealth) has the potential to transform the face of health service delivery across the globe’ (WHO, 2011). Currently, approximately one-third of the world’s adult population owns and uses a smartphone, and it has been estimated that by the year 2020 this proportion will increase to 80% (ChargeItSpot, 2016). Data suggest that more than half of smartphone users seek health-related information on their phone, and more recently the use of sensors embedded within mobile devices to monitor behavioural aspects has provided new areas of research. A digital marker has been defined as consumer-generated physiological and behavioural measures collected from digital tools that can be used to explain, influence and/or predict health-related outcomes (The Medical Futurist, 2018). Smartphones have been suggested as an easy and inexpensive way to monitor daily illness activity in BD including daily data on social and physical activity (Bardram et al., 2013; Faurholt-Jepsen et al., 2015). The use of smartphones to monitor BD provides unique opportunities to collect large amounts of fine-grained data in an unobtrusive, passive and continuous way in the long term and outside of clinical settings and could lead to the identification of new behavioural digital markers and digital phenotyping of BD (Glenn and Monteith, 2014; Hidalgo-Mazzei et al., 2016b; Insel, 2017; Monteith et al., 2015).

Within major depressive disorder and schizophrenia, few preliminary pilot studies, feasibility studies and case reports including automatically generated objective smartphone data, i.e., information on phone usage, mobility and voice features, have been published (Burns et al., 2011; Dang et al., 2016; Doryab, 2014; Wahle et al., 2016; Zhang et al., 2016). One of these studies investigated differences in voice features between patients with schizophrenia and healthy control individuals (HC) and suggested that several voice features may be able to discriminate between patients with schizophrenia and HC. Findings from studies including patients with major depressive disorder and collecting automatically generated objective smartphone data suggest that the overall concept of digital diagnosis is accepted positively by the patients (Burns et al., 2011; Dang et al., 2016; Doryab, 2014). Only one of the pilot studies investigated the use of data on acceleration, WiFi and GPS in combination with self-reported depression survey data for the delivery of a tailored cognitive behavioural therapy intervention (Wahle et al., 2016). However, the use of automatically generated smartphone data as an additional diagnostic behavioural marker in these populations has not been investigated. Recently, there has been an increased interest in the use of both self-monitored and automatically generated objective smartphone data within BD. Few observational studies (Faurholt-Jepsen et al., 2014, 2015, 2016a, 2016c; Gideon et al., 2016), pilot studies (Abdullah et al., 2016; Alvarez-Lozano et al., 2014; Beiwinkel et al., 2016; Grünber et al., 2015; Karam et al., 2014; Muaremi et al., 2014; Palmius et al., 2016), feasibility studies, case reports (Guidi et al., 2015; Hidalgo-Mazzei et al., 2016a; Saunders et al., 2017; Vanello et al., 2012) and study protocols (Faurholt-Jepsen et al., 2016b, 2017; Hidalgo-Mazzei et al., 2015; Kessing et al., 2017; Ritter et al., 2016) using automatically generated objective smartphone data have been published. Studies by Our previous studies (Faurholt-Jepsen et al., 2014, 2015, 2016a, 2016c) and those of others (Abdullah et al., 2016; Beiwinkel et al., 2016; Guidi et al., 2015) suggest that the automatically generated objective smartphone data reflected by communication logs, screen activation, location and voice features may reflect illness activity in BD and discriminate between affective states. However, these studies included rather small samples and did not include HC, and thus the use of automatically generated objective smartphone data as a potential diagnostic behaviour digital marker discriminating between BD and HC has not been investigated and is unknown.

The aim of this pilot study was to investigate whether objective smartphone data could discriminate between patients with BD and HC including analyses of sensitivity and specificity, and thus potentially represent a potential diagnostic behavioural digital marker in BD. We hypothesized that automatically generated objective smartphone data would be able to discriminate (1) between patients with BD during euthymia compared with HC, (2) between patients with BD during depressive state and manic or mixed state, respectively, compared with HC, and (3) between patients with BD overall compared with HC.

We have previously reported on the association between smartphone data and illness activity among patients with BD included in this report, i.e., not including HC or comparisons between patients with BD and HC. In that study,
we found that objective smartphone data reflect illness severity in BD and differ between affective states (Faurholt-Jepsen et al., 2016c).

**Materials and methods**

This pilot case–control study investigated the use of automatically generated objective smartphone data as an electronic diagnostic behavioural marker in patients with BD compared with HC.

The study was approved by the Committee on Health Research Ethics of the Capital Region of Denmark (H-7-2014-007 & H-2-2011-056) and the Danish Data Protection Agency (2013-41-1710). Smartphone data were stored at a secure server at Concern IT, Capital Region, Denmark (1-suite no. RHP-2011-03). The participants were offered to loan a smartphone free of charge by the study. Written informed consent was obtained from all participating subjects. The study complied with the Declaration of Helsinki.

We developed a software (‘MONARCA’) for smartphones to monitor self-assessed items and objective activities prior to this study (Bardram et al., 2013).

**Study participants**

**Patients with BD.** The patients were recruited from The Clinic for Affective Disorder, Psychiatric Centre Copenhagen, Denmark from October 2013 to December 2014. The inclusion criteria was BD diagnosis according to ICD-10 using Schedules for Clinical assessment of Neuropsychiatry (SCAN; Wing et al., 1990). The exclusion criteria were the lack of Danish language skills and pregnancy. In order to collect data during different affective states, the patients participated for a 12-week study period during the very early phase of their course of treatment at the clinic and received various types, doses and combinations of psycho-pharmacological treatment during the study. The participants were invited to participate in the study following referral to the clinic. Clinical and socio-demographic data were collected at inclusion. Analyses on the smartphone data and the association with illness activity in BD collected as part of this study have been published elsewhere (Faurholt-Jepsen et al., 2016c).

**HC.** As part of the study investigating stress in healthy individuals, a group of HC were recruited consecutively from the Blood Bank at Rigshospitalet, Copenhagen University Hospital, Denmark, by approaching blood donors in the waiting room on random occasions from September 2015 to August 2016. The inclusion criteria were as follows: women and men over the age of 18 years, no history of psychiatric illness and no first-generation family history of psychiatric illness and use of an Android smartphone as the regular mobile phone. The exclusion criteria were as follows: lack of Danish language skills and pregnancy. The HC participated in the study as part of a larger cohort study (Kessing et al., 2017). In this study, baseline data from HC were included in the analyses.

Since it was not possible to collect automatically generated objective smartphone data from iPhones at the time of the study, patients with BD and HC not willing to use Android smartphones during the study were excluded.

**Settings and assessments**

The study was conducted at The Clinic for Affective Disorder, Psychiatric Centre Copenhagen, Denmark.

**Clinical assessments. Patients with BD.** The BD diagnosis according to ICD-10 was confirmed using SCAN (Wing et al., 1990). The severity of depressive and manic symptoms was clinically rated fortnightly using the Hamilton Depression Rating Scale-17 items (HDRS-17; Hamilton, 1967) and the Young Mania Rating Scale (YMRS; Young et al., 1978) for the 12-week study period.

**HC.** The absence of any psychiatric diagnoses according to ICD-10 was confirmed using SCAN (Wing et al., 1990). The severity of depressive and manic symptoms was clinically rated at inclusion using the HDRS-17 (Hamilton, 1967) and the YMRS (Young et al., 1978).

All participants were instructed to carry their smartphones with them during the day and to use it for usual communicative purposes. Participants did not receive economic compensation for participating in this study.

**Smartphone data.** The MONARCA software used to monitor subjective and objective activities of BD was developed in our previous study (Bardram et al., 2013). After inclusion, the participants were instructed to use the MONARCA software for smartphones for self-evaluations on a daily basis during the study period. At the time of the study, the following objective smartphone data were available and automatically collected around the clock: the number of outgoing and incoming calls and text messages/day, the duration of phone calls (min/day), the number of times the smartphones’ screen was turned ‘on/off’ per day (reflecting the number of times the participants interacted with the smartphones), the duration the smartphone screen was ‘on’ per day. The particular smartphone data included in this report were due to technical aspects available in both cohorts (patients with BD and HC) at the time of the study. In addition, data on voice features during phone calls were collected, but due to the need for more advanced and technical statistical modelling data on the use of voice features for discriminating between patients with BD and HC and discriminating between affective states will be presented in future reports.

The researchers conducting the clinical assessments did not have access to the objective smartphone data and were therefore blinded to these data at the time of the clinical assessments. An overview of data collection during the study is provided in Table 1.
Table 1. Data collection during the study including patients with bipolar disorder (BD; n = 29) and healthy control individuals (HC; n = 37).

<table>
<thead>
<tr>
<th></th>
<th>BD</th>
<th>HC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euthymia</td>
<td>Depression</td>
</tr>
<tr>
<td>Hamilton Depression Rating Scale</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Young Mania Rating Scale</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Smartphone data</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Statistical analysis

The statistical analyses were defined a priori. To investigate differences in smartphone data between BD and HC, two-level linear mixed-effects regression models, which allow for variation of the outcome variables both within subjects (intra-individual variation) and between subjects (inter-individual variation), were employed since patients were investigated multiple times during different affective states. Level one represented repeated measures of objective smartphone data within each group (BD and HC) and level two represented between-subject variations (variations between BD and HC). Analyses comparing objective smartphone data between patients with BD (including different affective states) and HC were conducted. The linear mixed-effects regression models included a random intercept and the specification of individual participants as a random effect to accommodate correlations within each individual in the outcome variables over time. Covariates including age and gender were specified as the fixed effects. First, unadjusted linear mixed-effects regression models with levels of objective smartphone data as the dependent variables were conducted. Second, linear mixed-effects regression models with levels of smartphone data as the dependent variables adjusted for age and gender as the possible confounding factors were conducted. There is no consensus on how to report the performance on hierarchical linear models. In this paper, we used the Snijders and Bosker’s method (Snijders and Bosker, 1994).

In patients with BD and HC, the averages of the automatically generated objective smartphone data were taken for the days the outcome measures (HDRS-17 and YMRS) were referring to (the current and past 3 days). Automatically generated objective smartphone data from patients as well as HC were represented by 4 days from each rating and with the number of ratings ranging between 1 (HC) and 7 (BD). A depressive state was defined as an HDRS-17 score ≥ 13 and a YMRS score < 13. A manic or mixed state was defined as a YMRS score ≥ 13. A euthymic state was defined as an HDRS-17 score < 13 and a YMRS score < 13.

To calculate the classification accuracy of the objective smartphone data, machine learning techniques (scikit-learn gradient boosting classifier) were used. In many cases, we observed class imbalance; one class was represented by a large amount of examples (euthymia and depression), while the other was represented by a few examples (mania). To mitigate this problem, random oversampling, sampling the minority class with replacement, was used to create a balanced training set before learning the classifier. The gradient boosting classifier combines an additive sequence of simple decision tree classifiers into a single stronger classifier. At each iteration, a tree is generated from a subsample of the training data and using a random subset of features to ensure maximal degree of independence among the trees and prevents overfitting (Breiman, 2001). Model evaluation was done by 10-fold cross-validation. The sensitivity, estimating the probability that a test will identify ‘disease’ among those with ‘disease’, was calculated as true positive/positive, and the specificity, estimating the fraction of those without ‘disease’ among those with a negative test result, was calculated as true negative/negative. Area under the curve (AUC) was used as a metric to assess the performance of a model and the tradeoff between the sensitivity and specificity. The proportions of positive and negative results that were true positive and true negative were reflected by the positive predictive value (PPV), calculated as true positive/true positive + false positive, and the negative predictive value (NPV), calculated as true negative/true negative + false negative, and these metrics were used to describe the performance of the automatically generated objective smartphone data as a diagnostic test.

As no prior study has compared automatically generated objective smartphone data among patients with BD and HC, we were not able to make statistical power analyses prior to the study since potential effects are unknown. The hypotheses tested in this pilot study were made initially before data analyses and based on our prior findings within BD patients (Faurholt-Jepsen et al., 2016c). Although outcomes and covariates were specified prior to the analyses, the machine learning analyses were conducted post hoc. Consequently, we find it most correct to account for multiple testing in the statistical models (Bonferroni correction conducting 10 comparisons); p-values below 0.005 in individual models were considered statistically significant. Data were entered using Excel and EpiData®, Stata version 12.1 (StataCorp LP, College Station, TX, USA) was used for statistical analyses, and Python with the scikit-learn library was used for classification analysis.

Results

Background characteristics

A total of 51 patients with a diagnosis of BD and 72 HC were invited to participate in the study, of which 32 (62.7%)
patients and 37 (53.4%) healthy control individuals agreed to participate. The main reasons for declining to participate were as follows: (1) it would be too time consuming and (2) a preference of iPhone. Three patients dropped out immediately after inclusion (changed their minds regarding participation), and thus a total of 29 patients participated in the study. None of the participants dropped out during the follow-up period. A total of 10.3% of the patients’ visits with the researcher were missing, thus leaving a total of 182 clinical ratings available. The present results are based on 29 patients with BD clinically evaluated fortnightly during the 12 weeks and different affective states and a group of 37 HC and represent a total of 219 clinical ratings (182 from BD and 37 from HC). Apart from three participants, all the included participants wished to use their own smartphone during the study. Furthermore, socio-demographic characteristics including the severity of depressive and manic symptoms according to affective states represented by raw and unadjusted mean scores on the HDRS-17 and YMRS are presented in Table 2.

Differences in automatically generated objective smartphone data in BD overall and during affective states compared with HC

Table 3 presents the results of linear mixed-effects regression models for differences in objective smartphone data between affective states (euthymia, depressive state, manic or mixed state) in patients with BD and HC. Results from the unadjusted and adjusted analyses were similar, and thus results from the adjusted analyses (adjusted for age and gender) are presented.
In the adjusted models, the duration of phone calls was increased during a euthymic state, a depressive state, a manic or mixed state and overall ($B = 53.33$, 95% confidence interval (CI): [45.60, 61.05], $p < 0.001$, Snijders and Bosker’s estimate: 0.10), compared with HC.

The number of incoming text messages/day was increased during a manic or mixed state compared with HC. The duration the smartphone screen was on/day was decreased during a euthymic state ($B = –72.44$, 95% CI: [–112.70, –32.16], $p < 0.001$) and overall ($B = –50.00$, 95% CI: [–86.82, 13.19], $p = 0.008$, Sniders and Bosker’s estimate: 0.060) compared with HC.

Analyses including models adjusted for employment status were omitted, due to high collinearity between the groups (BD or HC).

**Sensitivity, specificity, PPV and NPV of automatically generated objective smartphone data**

Table 4 presents the results of gradient boosting classifier models for sensitivity, specificity, PPV and NPV of objective smartphone data. In models classifying BD overall versus HC, there were a sensitivity of 0.92, a specificity of...
0.39, a PPV of 0.88 and an NPV of 0.52. In models classifying patients with BD during euthymia versus HC, there were a sensitivity of 0.90, a specificity of 0.56, a PPV of 0.85 and an NPV of 0.67.

In addition, exploratory analyses on the classification of affective states within BD based on combined objective smartphone data were conducted. A depressive state versus a euthymic state was classified with a sensitivity of 0.36, a specificity of 0.68, a PPV of 0.33 and an NPV of 0.65. A manic state versus a euthymic state was classified with a sensitivity of 0.37, a specificity of 0.92, a PPV of 0.47 and an NPV of 0.88.

### Discussion

In this pilot study, we investigated differences in automatically generated objective smartphone data as well as the sensitivity, specificity, PPV and NPV between patients with BD and HC. The results indicated that automatically generated objective smartphone data differed between BD and HC, and that these had a rather high sensitivity and PPV but a low specificity and NPV.

Interestingly and in accordance with our a priori hypotheses, we found that the sensitivity and PPV of objective smartphone data were rather high in classifying BD overall and during euthymia versus HC. Furthermore, automatically generated objective smartphone data on changes in communicative activities and the use of smartphone differed between patients with BD during euthymia, depressive states, manic or mixed states and overall, respectively, and HC. However, not all of the automatically generated objective smartphone data differed between patients with BD and HC, and thus the findings in regard to the hypotheses were both supportive and some rejected.

The most intriguing and novel results from this pilot study were that (1) the sensitivity and PPV in classifying BD overall and during euthymia compared with HC were relatively high; (2) several of the automatically generated objective smartphone data differed between patients with BD during a euthymic state and HC; and (3) several of the automatically generated objective smartphone data differed between patients with BD, regardless the affective state, and HC.

A diagnosis of BD needs careful clinical evaluation and judgement. This is the first report on the use of automatically generated objective smartphone data as a potential diagnostic behavioural digital marker discriminating between BD and HC. The results from this study indicate that alterations in automatically generated objective smartphone data reflecting behavioural activities may represent a diagnostic behavioural digital marker of BD and could potentially represent a supplementary and assisting tool that could facilitate the clinical diagnostic process that currently, due to the lack of objective markers, relies on patient information and information from relatives, clinical observations and evaluations. However, in this study, the specificity and NPV were quite low. Thus, the tradeoff between the sensitivity and specificity reflected by the AUC should be considered in future studies and considerations on whether a high sensitivity could be important even though it could be at a cost of lower specificity and thereby the risk of false-positive classifications of patients with BD versus HC or false-positive classifications of affective states. Since still few studies have investigated the use of smartphones for monitoring in BD, more studies investigating this area more in depth both clinically and methodologically are needed. Large long-term cohort studies investigating the use of automatically generated objective smartphone data as a diagnostic marker, differentiating BD from HC and relatives at risk of BD and as a state marker, and differentiating among euthymia, depression and mania are ongoing (Kessing et al., 2017).

Surprisingly, as can be seen from the results of this pilot study, regardless of the affective state, patients with BD received and sent more text messages than HC. This may reflect the fact that the social network including relatives and people caring for the patients may be activated by the patients’ condition. Patients may thus contact others and may be contacted more due to worrying, help and care. During manic or mixed states, patients with BD did not

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity (SD)</th>
<th>Specificity (SD)</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD, overall versus HC</td>
<td>0.92</td>
<td>0.39</td>
<td>0.88</td>
<td>0.52</td>
<td>0.66</td>
</tr>
<tr>
<td>Euthymic state versus HC</td>
<td>0.90</td>
<td>0.56</td>
<td>0.85</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>Depressive state versus HC</td>
<td>0.79</td>
<td>0.50</td>
<td>0.71</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>Manic state versus HC</td>
<td>0.47</td>
<td>0.89</td>
<td>0.69</td>
<td>0.76</td>
<td>0.75</td>
</tr>
</tbody>
</table>

BD: bipolar disorder; HC: healthy control individuals; SD: standard deviation; PPV: positive predictive value; NPV: negative predictive value; AUC: area under the curve.


*Sensitivity = true positive/positive.

*Specificity = true negative/negative.
send more text messages than HC. However, during manic or mixed states, the duration of phone calls was increased so it may be that patients and their social network (including relatives) experienced increased need for verbal communication by phone calls and not by writing text messages. Also, during a manic or mixed state, the smartphones’ screen was turned ‘on’ (e.g., interacting with the smartphone by turning on the smartphones’ screen) more frequently compared with HC. However, the duration of time when the screen was turned ‘on’ was not increased. It may be that patients during a manic or mixed state had increased activity levels and restlessness reflected by turning the screen ‘on/off’ an increased number of times but not letting the screen ‘on’ for a prolonged period of time. Future long-term and large cohort studies investigating the use of composite measures based on automatically generated objective smartphone data reflecting behavioural activities to discriminate between patients with BD and HC and to discriminate between affective states will hopefully provide more insight to the area, and investigate which combinations of smartphone data that will give the highest sensitivity, specificity, PPV and NPV so that the use of automatically generated objective smartphone data potentially can be clinically relevant for monitoring and treating patients with BD.

Currently, approximately one-third of the world’s adult population owns and uses a smartphone, and it has been estimated that by the year 2020 this proportion will increase to 80%.

Overall, since most people carry their phone with them during most of the day and that it has been estimated that by the year 2020 the proportion of the world’s adult population that owns and uses a smartphone will increase to 80% (ChargeItSpot, 2016), smartphones can allow for collection of data on behavioural aspect that otherwise would be difficult to access. Furthermore, fine-grained data can be collected during real-time and naturalistic settings with a low level of intrusiveness without the need for people to interact with a software programme, minimizing the risk of missing data and fatigue during long-term monitoring. Accordingly, this type of automatically generated objective smartphone data has clear advantages in the monitoring of BD.

Limitations

Several clinical as well as methodological limitations to this study should be mentioned. Case–control studies carry an inherent risk of bias at different levels, such as selection bias, information bias and confounding, necessitating strict methodological requirements and thorough considerations of the study design and analyses. The patients included in the study consisted of those with bipolar I and bipolar II disorders treated in a highly specialized mood disorder clinic, during the very early phase of their course of treatment. This may have contributed to the relatively low number of manic/mixed episodes as well as the relatively low symptom level during affective episodes. However, the patients were recruited during the beginning of their course of illness presenting with rather high levels of depressive and manic symptoms assessed using standardized clinical rating scales during the study period. In addition, patients were newly diagnosed with BD and this could have had an impact on the use of technology regarding support and information seeking. Thus, findings may be more specific to this group of patients, and future studies including patients during different stages of illness could provide more generalizable findings.

Furthermore, a potential confounding effect of psychopharmacological treatment cannot be ruled out. Defining and recruiting a proper control group in case–control studies is always difficult. The HC included in this study were recruited from the Blood Bank at Rigshospitalet, Copenhagen University Hospital, Denmark, and thus may represent a ‘super-healthy’ comparison group.

Along this line, a potential confounding effect of employment status cannot be excluded. However, analyses including models adjusted for employment status were omitted, due to high collinearity between the groups (BD or HC). Future observational studies could consider investigating this aspect further and perhaps consider matching of groups on employment status. Many factors that are not related to the mental health status of the participants could also influence the results. People may vary considerably in how they interact with their phones and many people use other platforms to communicate with others including WhatsApp and Facebook messenger. To account for some of these differences, the statistical analyses were adjusted for age and gender. However, the unknown variability in phone usage should be addressed in future studies by including larger samples and adjusting the statistical analyses for additional confounding factors.

The study included a rather small sample of patients and HC and did not include a power analysis prior to the study. However, each patient with BD was assessed several times during the follow-up, thereby increasing the statistical power. Future studies should include a priori power analysis estimated when designing the study. Regarding the statistical analyses, although outcomes and covariates were specified prior to the analyses, the machine learning analyses were conducted post hoc and consequently multiple testing was accounted for in the statistical analyses.

None of the included patients dropped out during the study, but patients who were unwilling to use an Android smartphone were excluded, since it was not possible to collect automatically generated objective smartphone data from iPhones. Thus, the participating patients could represent a sample of particularly motivated patients not having problems with interacting with Android Smartphones and could introduce a potential bias as described by others (Montes et al., 2012; Spaniel et al., 2008).
Using automatically generated objective smartphone data as diagnostic digital markers highlights a number of challenges with digital/mHealth, and this sort of smartphone system in particular. When using this kind of platform, there are challenges regarding regular smartphone/iOS update and the need for constant updating of software. There will be challenges regarding whether the validity evaluations of older versions can apply to newer versions. This is one of the key problems in digital health, and considerations regarding this matter should be one of the key points to address in future studies. Furthermore, future studies including a priori power calculations investigating the sensitivity, specificity, PPV and NPV of automatically generated objective smartphone data in larger samples in BD comparing affective states, healthy relatives at risk of BD and HC are needed to evaluate the clinical utility. Including healthy first-degree relatives at risk of developing BD, as we are currently doing (Kessing et al., 2017), could provide further information on the use of objective smartphone data as an early predictive marker of later onset of BD and could provide important knowledge regarding the causality of changes in objective smartphone data in BD. Finally, few studies on the use of smartphone data as a marker of BD have been published and the thus the findings from this study are hypothesis generating and should be investigated further in future studies.

Conclusion
This pilot study demonstrated rather high sensitivity and PPV of objective smartphone data between BD and HC, but with low specificity and NPV. Furthermore, there were differences in levels of automatically generated objective smartphone data reflecting behavioural activities in patients with BD (during euthymia, depressive and manic or mixed states and overall) compared with HC. Objective smartphone data may represent a potential diagnostic behavioural digital marker and could potentially supplement, assist and facilitate the diagnostic process within BD, but further studies including larger samples are needed.

Acknowledgements
The authors would like to thank the patients and HC for participating in this study and statistician Christian Ritz, PhD, for consulting on the statistical analyses. M.F.-J., H.B. and L.V.K. authored the protocol, conceived the study and performed the statistical analyses. All authors have contributed to and approved the manuscript.

Declaration of Conflicting Interests
M.F.-J., J.B. and H.B. have no conflicts of interest. M.F. and J.E.B. are shareholders and co-founders of Monsenso providing the MONARCA system. M.V. has within the recent 3 years received speaker fees from Lundbeck. L.V.K. has within recent 3 years been a consultant for Lundbeck, Servier and Astra Zeneca.

Funding
This study was funded by Lundbeckfonden (R167-2013-16138), the European Union (EU) 7th Frame Programme, the Mental Health Services, Copenhagen, Denmark, the Danish Foundation Trygforden (ID 109766), the Gert Einar Joergensens foundation, and the A.P. Moeller and the Hustru Chastine Mc-Kinney Moellers foundation for general purposes. The funders had no role in the study design, data collection, analyses and preparation of the manuscript.

ORCID iD
Maria Faurholt-Jepsen https://orcid.org/0000-0002-0462-6444

References


Palmius N, Tsanas A, Saunders KEA, et al. (2016) Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*. Available at: https://doi.org/10.1109/TBME.2016.2611862


Wing JK, Babor T, Brugha T, et al. (1990) SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* 47: 589–593.


This appendix includes the Stan code of the Hierarchical linear regression model and the Hierarchical ordinal regression model, respectively. In practice, the Stan code is wrapped in Python classes that use a string Template to customize the code. For example, the numerical parameters of the hyperpriors can be adjusted and the generated quantities section can be left out when it is not required for increased sampling speed. The Python classes additionally provide methods for getting and setting model parameters, a scikit-learn style interface for fitting the model and making predictions as well as automatic caching of the compiled Stan model.
// Hierarchical linear regression model

data {
  int<lower=1> N; // number of examples
  int<lower=1> D; // number of predictors (dimensions)
  int<lower=1> J; // number of groups
  int<lower=1,upper=J> tid[N]; // group identifier index vector
  row_vector[D] X[N]; // example vectors
  real y[N]; // target vector
}

parameters {
  real mu_a; // intercept mean
  real<lower=0> tau_a; // intercept deviation
  vector[D] mu_b; // coefficient means
  vector<lower=0>[D] tau_b; // coefficient deviations
  real alpha_raw[J]; // intercepts
  vector[D] beta_raw[J]; // coefficients
  real<lower=0> sigma; // noise scale
}

transformed parameters {
  real alpha[J];
  vector[D] beta[J];
  for (j in 1:J) {
    alpha[j] = mu_a + tau_a * alpha_raw[j];
    beta[j] = mu_b + tau_b .* beta_raw[j];
  }
}

model {
  mu_a ~ normal(0, 1);
  tau_a ~ normal(0, 1);
  mu_b ~ normal(0, 1);
  tau_b ~ normal(0, 1);
  for (j in 1:J) {
    alpha_raw[j] - normal(0, 1); // normal(mu_a, tau_a)
    beta_raw[j] - normal(0, 1); // normal (mu_b, tau_b)
  }
  sigma ~ normal(0, 1);

  vector[N] z;
  for (i in 1:N)
    z[i] = alpha[tid[i]] + X[i] * beta[tid[i]];
  y ~ normal(z, sigma);
}

generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = normal_lpdf(y[i] | alpha[tid[i]] + X[i] * beta[tid[i]],
                             sigma);
}
// Hierarchical ordinal regression model
data {
  int<lower=1> N; // number of examples
  int<lower=1> D; // number of predictors
  int<lower=2> K; // number of classes
  int<lower=1> J; // number of groups
  int<lower=1,upper=J> tid[N]; // group identifier index vector
  row_vector[D] X[N]; // examples
  int<lower=1, upper=K> y[N]; // targets
}
parameters {
  ordered[K-1] mu_c; // cutpoint means
  real<lower=0> tau_c; // cutpoint deviations
  ordered[K-1] c_raw[J]; // cutpoints
  vector[D] mu_b; // coefficient means
  vector<lower=0>[D] tau_b; // coefficient deviations
  vector[D] beta_raw[J]; // coefficients
}
transformed parameters {
  ordered[K-1] c[J];
  vector[D] beta[J];
  for (j in 1:J) {
    c[J][1] = mu_c[1] + tau_c * c_raw[1][1];
    for (k in 2:(K-1))
      c[J][k] = fmax(c[J][k-1] + 1.0e-5, mu_c[k] + tau_c * c_raw[k][k]);
    beta[J] = mu_b + tau_b .* beta_raw[J];
  }
}
model {
  mu_c ~ normal(0, 1);
  tau_c ~ normal(0, 1);
  mu_b ~ normal(0, 1);
  tau_b ~ normal(0, 1);
  for(j in 1:J) {
    c_raw[j] ~ normal(0, 1);
    beta_raw[j] ~ normal(0, 1); // normal(mu_b, tau_b);
  }
  for (i in 1:N)
    y[i] ~ ordered_logistic(X[i] * beta[tid[i]], c[tid[i]]);
}
generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = ordered_logistic_lpmf(y[i] | X[i] * beta[tid[i]], c[tid[i]]);
Bibliography


