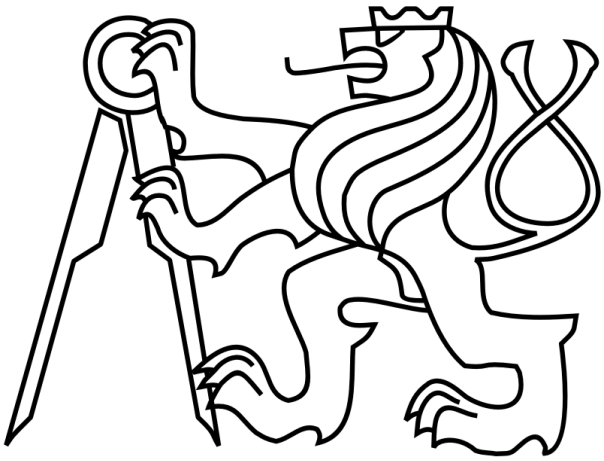
CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF ELECTRICAL ENGINEERING

DEPARTMENT OF CYBERNETICS



Long-Term Actigraphy in Bipolar Disorder: Processing, Analysis, and Applications in Diagnostics

DOCTORAL THESIS

Ing. Jakub Schneider

Prague, April 2021

Ph.D. Study Programme: P2612 – Electrotechnics and Informatics
Branch of Study: 3902V035 – Artificial Intelligence and Biocybernetics

Supervisor: **doc. Ing. Daniel Novák, Ph.D.**
Supervisor-Specialist: **Ing. Eduard Bakštein, Ph.D.**

Abstract

Bipolar affective disorder (BD) is a severe mental illness burdening 2 % of the global population, considerably shortening their lives by 15-20 years. The traditional treatment involves permanent medication and several examinations in a year. Thus, many clinical episodes are overlooked, which may lead to hospitalisation or even suicide. The links between changes in circadian rhythm and progression of BD are studied for years. But only the recent novel possibilities of continuous data sharing allow monitoring of circadian characteristics in the long-term by actigraphy wearables. In this thesis, statistical analysis and advanced machine learning concepts are applied to these data to deepen the knowledge about BD and its episodes and explore the feasibility of automatic episode detection.

The thesis contributes in three areas:

First, the adjustment of the actigraphic features for long-term monitoring. The traditional (non-parametric circadian rhythm analysis) features were updated to overcome the limits of long-term monitoring. Their robustness to missing data was evaluated. Moreover, the features set was extended to assess circadian rhythm changes typically connected with BD symptoms. Particular focus was given to descriptors of circadian phenotype preferences (chronotypes), where we offered clear guidelines for the use of actigraphy for chronotyping purposes.

Second, the diagnostic BD recognition. The differences between BD patients and healthy people have been explored, focusing on long-term variability that has not been studied to this extent before. Using machine learning methods, we have shown that distinguishing between non-symptomatic BD patients and healthy people is possible based on actigraphy alone. The proposed model achieved an accuracy of 88 %.

Third, detection of BD patient's state via machine learning techniques. The circadian rhythm changes in the patients' natural environment associated with BD symptomatic episodes were explored. These associations are vital for better understanding the undergoing processes in bipolar depression and mania, and they may support individual treatment.

This thesis shows that actigraphy presents a great opportunity in treatment objectivization in psychiatry. Hopefully, the use of objective biomarkers will facilitate evidence-based and efficient clinical decision-making to prevent severe BD conditions in the future.

Keywords: Actigraphy, Bipolar disorder, Circadian rhythms, Chronotype, Statistical analysis, Machine learning

Abstrakt

Bipolární afektivní porucha (BAP) je závažné mentální onemocnění, které postihuje 2 % světové populace a zkracuje život o 15 až 20 let. Tradiční léčba sestává z neustálé preventivní medikace a několika lékařských vyšetření ročně. To může vést k přehlédnutí mnoha klinických epizod, což může vyústit v nutnost hospitalizace a v extrémních případech i k sebevraždě pacienta. Propojení mezi cirkadiálními rytmy a průběhem BAP je studováno už léta. Nicméně až nové možnosti sdílení dat online umožnují dlouhodobé sledování pomocí autografu. Tato doktorská práce zpracovává tyto dlouhodobé záznamy, pomocí metod strojového učení a statistických analýz, za účelem rozšíření znalostí o BAP a jejích klinických epizodách s cílem ověřit možnosti jejich automatické detekce.

Tato práce rozšiřuje znalosti ve třech oblastech:

Za prvé aktualizuje tradiční aktigrafické příznaky tak, aby mohly být využity pro dlouhodobé sledováni stavu pacientů, včetně ověření jejich odolnosti vůči chybějícím datům. Navíc jsou přidány další příznaky, u nichž je předpokládané propojení s BAP. Speciální pozornost je věnována využití aktigrafie pro určování chronotypů, kde je poskytnut přehledný návod, jak dosáhnout co největší shody s klasickými dotazníky chronotypů.

Za druhé se zabývá možností podpory diagnostiky. Zde jsou zkoumány rozdíly v pohybové aktivitě během dne, a zvláště v jejich dlouhodobých změnách v rozsahu, který doposud nebyl studován. Pomocí metod strojového učení ukazuje, že aktigrafie je schopná odlišit bezpříznakové pacienty a zdravé lidi. Námi navržený model je byl schopen odlišit s 88 % přesností.

Za třetí se zabývá možností automatického rozpoznávání stavu pacientů. Zde jsou analyzovány souvislosti mezi změnami v cirkadiálním rytmu a neodhalenými či ambulantně léčenými klinickými epizodami. Tyto souvislosti jsou důležité pro odhalení vnitřních procesů během bipolárních epizod deprese a mánie, navíc mohou být použity jako podpora pro individuální nastavení léčby.

Tato práce ukazuje, že aktigrafie je velmi přínosnou metodou pro posuzování průběhu léčby pacientů s BAP. Jsem přesvědčen, že již brzy bude díky takto objektivizované a cílené péči snazší předcházet závažným stavům u pacientů s BAP.

Klíčová slova: aktigrafie, bipolární porucha, cirkadiální rytmy, chronotyp, statistická analýza, strojové učení

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published work of others has been acknowledged in the text, and a list of references is given.

Prague, Czech Republic — Jakub Schneider

April 2021

Acknowledgement

I would like to express my sincere gratitude and many thanks to people who helped me or supported me during the preparation of this thesis.

- To my supervisors Daniel Novák and Eduard Bakštein, for their guidance and support throughout my work on the thesis and my whole doctoral studies.
- To my fellow students, namely Pavel Vostatek, Jindřich Prokop, Václav Burda, Jiří Anýž, and Tomáš Sieger, for many fruitful research consultations, help, inspiration, as well as for being great people, with whom it was a pleasure to work.
- To my colleagues from the National Institute of Mental Health in Klecany (NIMH), mainly Filip Španiel, the father of the thesis topic, who shared with me his vast clinical insight with Bipolar Disorder and his enthusiasm for this topic. To Eva Fárková for providing me with data from her research, on which I could test many of my methods. To Marian Kolenič, Martina Ungrmarová, and all the other people from the NIMH Bipolar Disorder Clinic who helped me with collecting the data used for my research.
- To all people from the Mindpax.me company. It would not be possible to record such long actigraphy records from so many patients without their enthusiasm and outstanding work.
- And last but not least, I want to thank my family and friends for keeping me sane while writing the thesis, especially in the crazy times of the COVID-19 pandemic.

Jakub Schneider, April 2021

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADA – Average Daily Activity

ASERT – Aktibipo Self-rating EMA

AUC – Area Under the Receiver Operating Characteristic Curve

BD – Bipolar Affective Disorder

- BD-I – Bipolar Disorder Type 1
- BD-II – Bipolar Disorder Type 2

BMI – Body Mass Index

BT – Bluetooth

CBT – Cognitive Behavioural Therapy

CNS – Central Nervous System

CQ – Circadian Quotient

CTS – Circadian Timing System

ECG – Electro Cardio Graph

EMA – Ecological Momentary Assessment

GOF – Goodness of Fit

GPS – Global Positioning System

HC – Healthy Controls

IQR – Interquartile Range

IS – Interdaily Stability

IV – Intradaily Variability

LOSO – Leave One Subject Out (Cross-validation)

LTTV –Long-term Temporal Variability

MADRS – Montgomery-Åsberg Depression Rating Scale

MAE – Mean Absolute Error

MCTQ – Munich Chronotype Questionnaire

- MSFsc – Mid-Sleep Time on Free-regime Days
- SJL – Social Jetlag

MEMS – Microelectromechanical System

MEQ – Morningness-Eveningness Questionnaire (Chronotype)

MESOR - Midline Estimating Statistics of Rhythm

NPCRA – Non-parametric Circadian Rhythm Analysis

MSE – Mean Square Error

NIMH – National Institute of Mental Health in Klecany, Czech Republic

PSG – Polysomnography

RA – Relative Amplitude

RF – Random Forest

RMSSD – Root Mean Squares of Successive Differences

RSS – Sum of Residuals Squares (Error)

SD – Standard Deviation

SleDur – Sleep Duration

APSO – Activity Prior Sleep Onset, AASO – Activity After Sleep Onset

APWU – Activity Prior Wakeup, AAWU – Activity After Wakeup

SMD – Standardised Mean Difference

WASO – Wake After Sleep Onset

YMRS – Young Mania Rating Scale

1. Introduction

The treatment of patients suffering from a mental disorder is a complicated process. Diagnosing these diseases is quite different and more complicated than in other fields of medicine. Many diagnoses in psychiatry may have unprecedented physiological causes and effects. Therefore, they commonly cannot be obtained by mere physiological measurement, genetic tests, or medical imaging techniques. From this point of view, psychiatry differs from other fields of medical care. Despite modern technical advances, the diagnoses are commonly obtained using a structured interview (with patients, relatives, etc.). Such an approach is highly time-consuming and partly subjective. It requires an excessive level of training and experience to suppress the subjectivity. Especially as interrater reliability is not lower than in other medical fields, Cohen's $\kappa \sim 0.7$ (Pies, 2007). Nonetheless, it may take years before the patient is correctly diagnosed and receives the optimal treatment (Baldessarini *et al.*, 2007; Kessing *et al.*, 2015). While many diagnoses (such as BD, major depressive disorder, or schizophrenia) are not fully curable, the patients may still live a full life if they receive the correct treatment.

In bipolar disorder, which is the objective of this thesis, the patients suffer from irregularly recurring episodes of either elevated or depressed mood. In between the episodes (inter-episode time), they may live a valuable life with normal work and family life. Many famous and highly successful people such as Carrie Fisher, Francis Ford Coppola, Miloš Kopecký, Sting, or Winston Churchill[1] are known to have been diagnosed with BD. It is also suspected that some other historical celebrities, such as Isaac Newton, Abraham Lincoln, Vincent Van Gogh, Ludwig Von Beethoven, etc.[2] may have suffered from BD as well. On the other hand, symptomatic episodes (relapses) of BD substantially reduce the quality of life, affecting both personal and professional life, with a deteriorating tendency (adding comorbidities). Timely detection of relapses could significantly increase the quality of patients' lives and reduce the treatment expenses, as early detected episodes can usually be managed in an ambulatory setting and not by hospitalisation.

[1] https://olympiahouserehab.com/celebrities-with-bipolar (2021-Jan)
[2] https://www.butler.org/blog/famous-people-and-depression (2021-Jan)

During classical treatment, the patients visit their doctor only a few times a year, which increases the risk, that onset of the episode is not detected before the point when hospitalisation is required. Moreover, the disease's long-term state development is usually based on the patient recalling mood fluctuations between visits, which is highly obscured by recall bias. Fortunately, there is a revolution, called digital phenotyping, starting in psychiatry, which may transfer it into a data-driven medicine in a similar way as genetic testing transferred oncology (Hsin *et al.*, 2018). Digital phenotyping uses devices, such as wearables and smartphones, to evaluate behaviour changes and circadian rhythmicity, and use them as warning signs.

This thesis aims to extract relevant clinical information from the long-term actigraphy recordings to be used as a supportive tool for diagnostics and BD treatment by objectively evaluating BD patient's state. The thesis contributes to two fields. First, it explores and updates the actigraphic features, commonly used to describe the circadian rhythm. Second, it uses the updated features to explore the changes in the circadian rhythm connected with BD diagnosis.

In order to achieve that, the longest (to our knowledge) continuous actigraphy data were recorded in a large group of patients. The recording was done in cooperation with the National Institute of Mental Health (NIMH) and Mindpax Co. Ltd. (a Czech company developing a digital tool for people with severe mental illnesses).

1.1. Goals of the Thesis

The methodological goals targeting data processing are:

- to design a set of traditional and novel circadian features and enhance the features, where necessary, in order to comply with the requirements of long-term actigraphy monitoring.
- to provide an explainable physical activity descriptor that may be used in the physician-patient communication (and evaluation) and in enhancing patients' self-awareness.
- to examine limitations in actigraphic features used for long-term monitoring

The clinically relevant goals include analyses:

- to objectify the estimation of chronotype using actigraphy in comparison to clinically used questionnaires.
- to evaluate differences between BD patients and healthy controls, and use machine learning technics to evaluate the utility as a diagnostic tool.
- to identify features that may be used for automatic detection of patient state and perform patient state estimation based on these features.

1.2. The Structure of the Thesis

The thesis is structured as follows:

Chapter 2 gives a brief introduction to the epidemiology of bipolar disorder and its treatment. Section 2.1 provides information about bipolar disorder, its prevalence, and its symptoms. Sections 2.2 and 2.3 describe a standard treatment procedure with an overview of commonly used clinical scales. Section 2.4 provides information about patients' momentary self-assessment and emerging methods of digital phenotyping.

Chapter 3 provides basic information about actigraphy and its derived features describing sleep and circadian rhythmicity. Sections 3.1 and 3.2 give a brief history and describe the principles of function of actigraphy wearables. Section 3.3 presents procedures of commonly used pre-processing techniques and their limitations. In section 3.4, we introduce parameters

of commonly used actigraphy wearables, from both – scientific and commercial spheres. And finally, section 3.5 presents the used actigraphic features with additional updates and extensions to be used in long-term recordings.

Chapter 4 contains information about all of the datasets used, including the onboarding procedures, recording methodologies, and basic health and demography summaries of volunteers included in the studies.

Chapter 5 focuses on variability in circadian features during long-term monitoring and the reliability of these features when they are estimated over samples, including missing values, which are the major problem of long-term actigraphy.

Chapter 6 evaluates the possibility and benefits of objectification of the chronotype estimation using actigraphy. It focuses on the accuracy and stability of actigraphy based estimation of chronotype (chronotyping). The results are validated by comparison to clinical questionnaires.

Chapter 7 explores the differences in circadian and sleep features developed in Chapter 4 between BD patients in remission and healthy controls, focusing on variation obtained from long-term monitoring. Within, we evaluate the usability of the actigraphy recordings in clinical practices. An example of a supportive diagnostic tool is provided using a machine learning task of classification BD patients and healthy controls.

Chapter 8 provides a preliminary exploration of circadian rhythm changes during symptomatic periods. Several actigraphic features (section 3.5) were identified as the most promising in detecting the relapse. The feasibility of such detection is tested using two machine learning approaches.

Chapter 9 concludes the thesis while highlighting the achievements and contributions.

2. Background

2.1. Bipolar Affective Disorder

Bipolar disorder (BD), previously known as manic depression, is a summary name for a complex group of severe chronic mood disorders that are defined as the repetitive occurrence of relapses, episodes of depression, mania, hypomania, or their mixture, with non-symptomatic euthymic periods (remissions) in between. The BD group contains, according to DSM-5 (APA, 2013), three conditions: Bipolar 1 disorder (BD-I), Bipolar 2 disorder (BD-II), and Cyclothymic disorder (BD-III). These are sometimes accompanied by other disorders of the bipolar spectrum 'not otherwise specified', where episodes are too short or too few, so they don't meet definitions of mania or hypomania (Towbin *et al.*, 2013). The difference between BD-I and BD-II is the severity of manic episodes – mania and hypomania. Cyclothymic disorder (BD-III) describes a condition of frequently cycling brief episodes of hypomania and depression.

The global prevalence of BD (BD-I and BD-II) is expected to be between 1-2 % worldwide (Merikangas *et al.*, 2011), though it is reported even higher in specific localities, e.g. 3-4 % in South Africa (Steel *et al.*, 2014). The WHO signified BD as the 6th leading source of disability affecting about 5 % of the global population (BD-I, II, III, and spectrum) (Colombo, Fossati and Colom, 2012). BD is typical by its early onset. 70 % of BD individuals manifest clinical symptoms before the age of 25 years (Nowrouzi *et al.*, 2016). Individuals with this disorder are symptomatic about half of their lives (Judd and Akiskal, 2003; Judd *et al.*, 2003). The consequences of the disease are quite severe. The mortality studies associate it with loss of approximately 10-20 potential years of life (McIntyre *et al.*, 2020). The reported suicide rate is 20-30 times higher in BD patients compared to the general population (Dome, Rihmer and Gonda, 2019; Dong *et al.*, 2019). Additionally, compared to the general population, adults with BD experience elevated rates of obesity, diabetes, cardiovascular disease, and metabolic syndrome (Fagiolini *et al.*, 2003; McIntyre *et al.*, 2020). Total estimated annual treatment costs are over 202 billion US$ in the USA (McIntyre *et al.*, 2020) and 113 billion € in the EU (Gustavsson *et al.*, 2011). The diagnosis of BD is often delayed because some of the symptoms, such as impulsivity, affective instability, anxiety, cognitive disorganisation, depression, and psychosis, are shared with many other mood and mental disorders – for

example, major depressive disorder (MDD), schizophrenia, attention-deficit hyperactivity, borderline personality disorder, etc. Moreover, BD is commonly accompanied by a plethora of comorbidities, such as sleep disorders and alcohol or substance abuse. Therefore, it takes approximately 6-10 years from the first occurrence of symptoms to obtain an accurate diagnosis (Baldessarini *et al.*, 2007; Kessing *et al.*, 2015). The length of the diagnostic process is also given by the predominance of depressive episodes (Akiskal *et al.*, 2000) and overlooked hypomania episodes, which are usually not considered pathological by patients, and therefore not reported (Angst, 1998).

The pathogenesis of BD is poorly understood. Recent findings (Andreazza, Duong and Young, 2018) associate it with disturbances in mitochondrial function. The genetic origin is well documented. Inheritability of BD is about 70 %, and common genetic variants were already detected (Stahl *et al.*, 2019).

The mood changes associated with relapses are accompanied by extreme shifts in energy, activity, sleep, and behaviour. The mania and hypomania are manifested by increased activity, energy, or agitation, euphorically exaggerated senses of self-confidence, abnormal cheerfulness, decreased need for sleep, over-talkativeness, racing thoughts, high distractibility, and poor decision making. The depressive episodes, which are typically both more prolonged and frequent, are manifested by depressed mood (feelings of sadness, emptiness, hopelessness, sometimes accompanied with higher irritability), loss of interest in most activities, changes in appetite connected with weight changes, changes in sleep insomnia or hypersomnia, loss of energy, feeling worthless and guilty, decreased concentration, and/or suicidal thoughts. When a patient develops several symptoms from both depression and mania simultaneously, we talk about a mixed state (APA, 2013). All of these relapses are life-threatening. In depression, the suicide risk is higher, especially in a depression with mixed symptoms (Dome, Rihmer and Gonda, 2019). In mania, the risk of poor decision making is combined with a reduced need for sleep, which (when untreated) may cause life-threatening exhaustion (Plante and Winkelman, 2008).

The factors contributing to relapse in BD are also not clearly understood. Still, it has been suggested that there could be an association with dysregulation of circadian (*circa* = about, *dies* = day) rhythm (Murray and Harvey, 2010; Alloy *et al.*, 2017) and disturbed sleep (Millar, Espie and Scott, 2004; St-Amand *et al.*, 2013; Geoffroy, Boudebesse, *et al.*, 2014; Bellivier *et al.*, 2015; Gold and Sylvia, 2016).

The circadian rhythm dysregulation appears in acute episodes as well as in inter-episode periods (see Chapter 7). Therefore, measurements of circadian rhythm via motor activity profiles may provide a valid trait marker of BD (Milhiet *et al.*, 2011), and a deeper understanding of this dysregulation may contribute to improved management of the disease (Scott, Vaaler, *et al.*, 2017; Merikangas *et al.*, 2019). For example, depression induces a lack of physical activity, which is associated with many comorbidities in adults with BD, and which may become one of the future clinical treatment targets (Fagiolini *et al.*, 2003; Janney *et al.*, 2014; Vancampfort *et al.*, 2017).

2.2. Clinical Practice - Standard Treatment

The clinical treatment consists of pharmacotherapy, including mood stabilisers, antidepressants, and antipsychotics, psychological interventions, and electroconvulsive therapy (McIntyre *et al.*, 2020). The first line of pharmacological therapy is monotherapy by mood stabilisers or antipsychotics. The oldest mood stabiliser used in BD is Lithium, which is also highly disease-specific. It is efficient in approximately one-third of patients (Hui *et al.*, 2019), even in monotherapy, in treatment of both types of acute relapses, as well as in relapse and suicidality prevention. The main disadvantage is that the effective dosage is only slightly lower than the toxic levels, and therefore it should be periodically updated/tested, also with respect to the renal function (every 2-3 months). It also requires a salt-restricted diet and avoidance of certain medications. Other monotherapies include valproate or antipsychotics (olanzapine, quetiapine, aripiprazole, etc.). During acute episodes, the medications are commonly combined with other mood stabilisers (lamotrigine, carbamazepine, etc.), antipsychotics, and possibly antidepressants. The use of antidepressants is not generally advisable as it may cause rapid cycling or manic shift (Látalová, 2010). Due to additional comorbidities, BD patients are often prescribed multiple medications. For these medication mixtures, it usually takes a longer time to adjust the optimum dosage.

Psychological interventions are in most cases focused on the education of the patient on how to cope with his illness. One of these methods is cognitive behaviour therapy (CBT), which is non-pharmacological psychotherapy, focusing on teaching patients how to become aware of, and examine their distorted thinking, and cognitively test it against reality judgments. CBT is often combined with psychoeducation, which focuses on the education of patients, and possibly their relatives, in better understanding of the mental illness, in order to better cope with it (Bäuml *et al.*, 2006; Miziou *et al.*, 2015). Interpersonal and social therapy is another type of psychological intervention. During this therapy, the patients are educated on possible changes in social rhythm, which pose a risk of relapse onset, as well as risks posed by low medication adherence. Patients learn about the need for a regular daily routine and how to avoid or cope with daily stressful events (Frank, 2007) as these may cause disruptions of the circadian rhythm, which are reported as a possible relapse trigger (Scott, Vaaler, *et al.*, 2017; Merikangas *et al.*, 2019).

2.3. Clinical Practice - State Assessment

Periodic ambulatory examinations commonly evaluate the course of the patient's state. The re-evaluation period varies around 3-4 months (Wang *et al.*, 2005). The most objective evaluation of patient state is possible through clinical-administered scales, which are recommended for the treatment (Tohen *et al.*, 2009). Most of the clinical scales evaluate separately manic and depressive symptoms. Manic symptoms may be assessed by **Young Mania Rating Scale** (Young *et al.*, 1978) (YMRS), Bech-Rafaelsen Mania Rating Scale (Bech, P., Rafaelsen, O. J., Kramp, P., & Bolwig, 1974), Clinical-Administered Rating Scale for Mania (Altman *et al.*, 1994), and Observer-Rated Scale for Mania (Krüger *et al.*, 2010). Depressive symptoms may be assessed by **Montgomery-Åsberg Depression Rating Scale** (Montgomery and Åsberg, 1979) (MADRS), Quick Inventory of Depressive Symptomatology (Trivedi *et al.*, 2004), the five-item Hamilton Depression Rating Scale (González-Pinto *et al.*, 2009), Inventory of Depressive Symptomatology (Trivedi *et al.*, 2004), or Bipolar Depression Rating Scale (Berk *et al.*, 2007). Several scales evaluate both manic and depressive symptoms together. These are, for example, the National Institute of Mental Health's Prospective Life Chart Methodology - Clinician (Denicoff *et al.*, 2000), Clinician Monitoring Form (Sachs, Guille and McMurrich, 2002), Brief Bipolar Disorder Symptom Scale (Dennehy *et al.*, 2004), and Bipolar Inventory of Symptoms Scale (Gonzalez *et al.*, 2008).

The administration of clinical scales is time-consuming; therefore, their use is optional in most clinical practices. The clinical scales, if used, are usually utilized to monitor the state only during acute episodes. Out of these episodes, the patients are usually evaluated by a semi-structured personal interview. The long re-evaluation period may cause missing an episode's onset, which is the best moment for intervention. The long re-evaluation period also causes that most of the minor subclinical episodes are unnoticed. In order to cope with this issue, there are emerging long-term monitoring systems.

2.4. Emerging Approaches for Long-term Monitoring

The need for a finer sampling of patient illness state progression leads to the development of long-term monitoring systems, which may be divided into three categories:

- Patient's self-assessment questionnaires are an illness progression monitoring approach, where the patient himself evaluate his state/mood. They are subjective but highly focused. This approach is already used in practice. Patient-filled 'diaries' may help to follow the course of illness between medical check-ups. When completed online, the self-assessment could additionally be used for a timely warning.
- Behavioural analyses explore the development of illnesses development/state based on objectively measured changes in smartphone usage. These are mostly in the research stage, and wide usage would be substantially limited by regulatory restrictions and the patient's willingness to share sensitive data.
- Physical activity, measured using an actigraph or a smartphone, which monitors changes in circadian rhythm may be used to assess a patient's state. This approach is presently also in the research/development stage.

The optimal system would probably combine all three approaches, or at least two: the focused self-evaluations and one of the other two objective measures.

2.4.1. Self-assessment (Ecological Momentary Assessment - EMA)

The self-assessment mood reports usage in clinical practice and research is gaining importance in the last years (Barrigón *et al.*, 2017; Cerimele *et al.*, 2019). There are obvious advantages of their use over the clinical-administered scales. First of all, the reporting may be much more frequent, which is extremely important. It has been reported that physicians may miss up to half of the patients' symptoms and underestimate the severity of symptoms (Cerimele *et al.*, 2019). Another advantage is the reduction of measurement cost, as no clinical personnel is needed. Moreover, the administration in patients' natural environment may increase the acquired data's accuracy (introducing the so-called 'ecological validity'), as it reduces the degree to which the examination by a physician and clinical environment affects the results (the so-called 'white cloak syndrome'). Some even argue that the accuracy may be increased by avoiding the clinician interpretation (FDA, 2006). And finally, it may be used as a part of CBT, as the patient cognitively contemplates his/her state.

On the other hand, the need for increased patients' adherence appears to be the main disadvantage of this method, as the adherence in BD patients' has generally been reported low (Chakrabarti, 2016). Among others, the patients may experience fear of possible interventions based on the reports, and there may be a loss of insight during more severe symptomatic periods.

The value of self-assessed reports, commonly referred to as ecological momentary assessment (EMA), can be seen in the Cerimele's meta-analysis summarising existing studies using patient-observed and clinician-observed symptoms. These studies indicate that patients from psychiatric clinics who use self-assessment reports have a better outcome than those who don't (Cerimele *et al.*, 2019).

As in clinical scales (section 2.3), the EMAs may be divided into those assessing only depression, those assessing only mania, and those assessing both polarities at once. Cerimele et al (2019) evaluated EMAs considering selected parameters: briefness, possible public use, the inclusion of remission indicator and suicidal ideation indicator, test-retest repeatability, sensitivity to change, etc. In case of manic symptoms, the following best EMAs achieved comparable or better performance than clinician-administered: Altman Self-Rating Mania Scale (Altman *et al.*, 1997), Self-Report Manic Inventory (Shugar *et al.*, 1992), and Computerized Adaptive Testing-Mania (Achtyes *et al.*, 2015). Concerning EMAs focused on depressive symptoms, the best (and only comparable to the clinician-administered) was the Quick Inventory of Depressive Symptomatology (Bernstein *et al.*, 2010). The best self-assessed questionnaires targeting both polarities together were the Internal State Scale (Huang *et al.*, 2003), Affective Self-Rating Scale (Adler *et al.*, 2008), and National Institute of Mental Health's Prospective Life Chart Methodology - Self (Born *et al.*, 2014).

The EMAs have been used already for more than two decades now, as indicated by the introduction times of individual scales stated above. Nowadays, when e-Health is on the rise, the inclusion of smartphones may upgrade the field of psychiatry to a new level. Though there is some concern about the deterioration of the patients' state by focusing them on studying their symptoms, some studies (Faurholt-Jepsen, Geddes, *et al.*, 2019) suggest that it doesn't have to be that case. Smartphones are becoming a more and more common part of life. Out of 92 % of people who own a mobile phone in the USA, 77 % have smartphones (Orsolini, Fiorani and Volpe, 2020). The ownership of mobile phones by patients suffering from mood disorders is similar to the general population (86 % in 2013), and the expected rate of

smartphones is also similar (Matthews *et al.*, 2017). Therefore, the incorporation of EMAs may increase patients' adherence, even using some gamification techniques. Moreover, mobiles may be used even beyond the collection of EMAs, as it is documented in the next section.

2.4.2. Behavioural Analysis

The inclusion of smartphones into psychiatric care may represent the dawn of long-term monitoring and, therefore, precise and early identification of many health conditions, which allows for timely interventions. There is a plethora of evidence that human behaviour may be monitored using smartphones and personal wearable sensors. Such an approach is called digital phenotyping (Orsolini, Fiorani and Volpe, 2020). Concerning patient involvement, there are two types of measured data:

1) Actively acquired data, usually obtained through a survey, which requires the participation of the patient.
2) Passively acquired data, which are usually recorded using smartphone statistics and sensors readings. These data, which may be collected without patients active participation, include: information about movement (accelerometers, GPS readings, mobile towers connections, etc.), social interactions (number and duration of calls, number of messages, number of running apps, Wi-Fi, and Bluetooth readings as number of available devices, screen time, number of unlocks, etc.), and other physiological variables (speech parameters, typing dynamics, and possibly heart rate, weight, etc.). The list of possible collected data streams may be extended by the usage of other personal sensors.

There is evidence suggesting that passive smartphone data may be used to detect relapses in depressive disorders (Onnela and Rauch, 2016), schizophrenia (Barnett *et al.*, 2018), symptom severity in anxiety (Jacobson, Summers and Wilhelm, 2020). Considering BD, promising results are obtained for speech (Karam *et al.*, 2014; Muaremi *et al.*, 2014; Gideon, Provost and McInnis, 2016; McInnis, Gideon and Mower Provost, 2017). Because of privacy issues, the recorded data are focused only on speech characteristics, such as pitch frequency, number of utterances, etc. Other possible biomarkers include typing dynamics (Cao *et al.*, 2017), movement (GPS and accelerometers) (Grünerbl *et al.*, 2015; Palmius *et al.*, 2017), and general mobile usage statistics (Faurholt-Jepsen, Busk, *et al.*, 2019). Palmius *et al.* reported 85 % accuracy in depressive episode detection based on geographic location recording. The findings

suggest that generated objective smartphone data (the number of text messages/day, the duration of phone calls/day) were increased in BD patients compared to the control group (Faurholt-Jepsen, Busk, *et al.*, 2019). Increased physical activity may present a warning signal for BD phase transition (Beiwinkel *et al.*, 2016).

A machine learning model using smartphone-collected visual analogue scales for mood, energy, and anxiety finds the self-assessed energy to be an important BD state predictor, even better than mood (Ortiz, Bradler and Hintze, 2018).

Using a combination of activity, sleep, light exposure, heart rate, clinical scales, and EMA, Cho et al. (2019) train a model with an AUC of around 0.9 and an accuracy of about 80 % in predicting remissions, depressions, and hypomanias.

Additionally, these applications may provide a utility for patients with BD to manage their activity levels and exposure to light to coordinate with their circadian rhythm to maintain a stable mood state (Perna *et al.*, 2018).

Although the results seem extremely promising, the studies provided so far are based on relatively small samples of people. Also, many measures recorded during the studies could pose legal issues in privacy, security, and responsibility for technical errors. Moreover, the publicly available applications, such as Beiwee[3] and MindLamp[4] (which may be obtained for free on google and apple app-stores), do not work on all smartphones and do not support all the features mentioned before - mainly the speech characteristics are missing. Also, the patient's physical activity is not measured when he/she does not have a smartphone with him/her. Therefore, the use of a readily accessible device - an actigraph - may present a plausible starting point for broader clinical use.

[3] https://www.hsph.harvard.edu/onnela-lab/beiwe-research-platform/ (2020-Dec)
[4] https://www.digitalpsych.org/lamp.html (2020-Dec)

2.4.3. Actigraphy

Actigraphy (Chapter 3) is a non-invasive method of measuring sleep and circadian rhythm in the natural environment. Its use does not require additional patients' participation, as the data are collected mostly passively. Thanks to the increased use of different types of sport testers and activity monitors, it also poses a low risk of stigmatisation. As there is ample evidence of a connection between BD and changes in sleep and circadian rhythm (Section 2.1), both in euthymic state and relapse episodes (Tazawa *et al.*, 2019). Actigraphy is a highly promising tool for long-term monitoring of the course of the illness. More details on actigraphy measured differences between a healthy population and BD patients could be found in section 7.1.2.

3. Actigraphy

3.1. History of Actigraphy

The first wrist-worn actigraph, a device that records body movement, was developed in the 1970s (McPartland, Kupfer and Gordon Foster, 1976). Its usage was limited at the beginning, but from the 1980s, actigraphy started to be used for sleep research, mainly to analyse sleep-wake patterns. When compared to polysomnography (PSG), one of the advantages is that the sleep may be continuously measured for 24-hours a day, including daily naps. Unlike PSG, it can easily be measured for several consecutive days or weeks. Since that time, actigraphy is still a largely expanding field, additionally including monitoring of circadian rhythm (Sadeh *et al.*, 1995; Ancoli-Israel *et al.*, 2003). The development in microelectromechanical systems (MEMS), battery power, and memory media have given rise to modern digital lightweight wearables that allow recording physical activity data for weeks, with high sampling frequencies. Such capability hugely enhanced the possibilities of research in circadian rhythms and sleep disorders. In order to include such wearables in psychiatric care, it is necessary to monitor, collect and evaluate the data online to provide real-time feedback. This is achievable by mobile network devices, e.g. smartphones which additionally allow for the acquisition of EMAs at the same time.

3.2. Actigraph - Operating Principle

The key component of an actigraph is a three-axis accelerometer. The accelerometer is usually a MEMS sensor that consists of a fixed part and a mass attached to the fixed part by springs allowing movement in one direction. According to Newton's first law of motion, when the sensor accelerates (changes movement velocity), the weight tends to stay at the original position. This leads to the displacement of the mass relative to the fixed part leading to a change in the electrical characteristic of the sensor (for example, capacity), as is shown in Figure 3.1. The three-axis accelerometer consists of three such sensors oriented perpendicular to each other and therefore allows for measuring of acceleration in 3D. Such a sensor has a limited oscillation frequency range, which is still much higher than the frequency range common for biological movements.

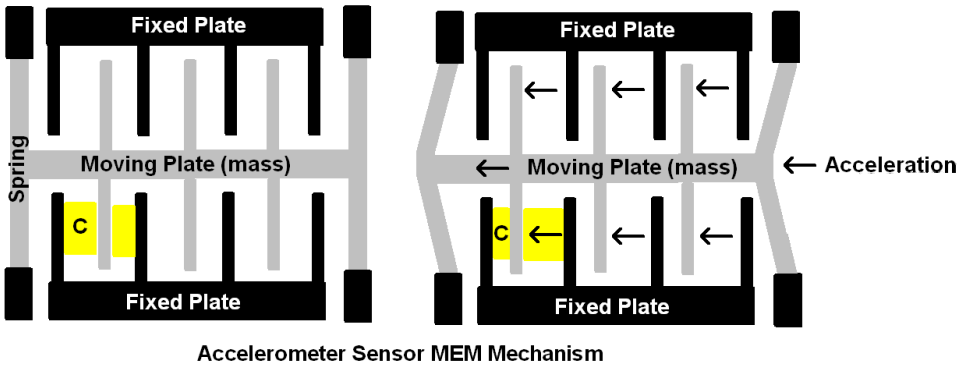


Figure 3.1 - Working principle of MEMS acceleration mechanism[5]

Generally, the frequency of voluntary physiological movements rarely exceeds the frequency of 3-4 Hz. Involuntary tremors can exceed 5 Hz frequency (Redmond and Hegge, 1985). The Nyquist-Shannon sampling theorem indicates that the sampling frequency should be at least two times higher than the highest recorded frequency. Hence the lowest sampling frequency required for physiological activities is about 8 Hz, while the recommended sampling frequency should be higher than that in order to cover the involuntary tremors.

3.3. Actigraphy - Data Pre-processing

The signal recorded from the accelerometer has to be digitalised and pre-processed by filtering out low and high frequencies to remove gravitational acceleration and high-frequency artefacts (such as using a drill, driving a car, etc.). The typical filtered frequency range for records with a sampling frequency of tens of Hz is 0.25 – 4 Hz using a bandpass filter. (Redmond and Hegge, 1985) Some approaches use a higher upper frequency (about 10 Hz) to include faster movements that may occur in younger people. (Ancoli-Israel *et al.*, 2003)

There are wearables that allow for the storage of raw (unfiltered) values even for a relatively long time, depending on sampling frequency and battery capacity. Most wearables (and all used for long-term monitoring and online processing) aggregate the raw activity data into so-called epochs. The duration of these epochs is arbitrary, but most are ranging from seconds to a few minutes.

[5] https://www.electronicwings.com/sensors-modules/adxl335-accelerometer-module (2020-Nov)

As of today, there are no strict standards in the method used to aggregate the raw data into epochs. Therefore, there is also no specific physical unit assigned to the measured epoch score. Scientific wearables usually represent the data as activity counts. This goes back to the origin of chronobiology, where the activity of an animal was measured by counting events as a movement of a wheel, or passages of an animal through an infrared light beam, or similar measuring methods (Sokolove *et al.*, 1977; Matikainen-Ankney *et al.*, 2019). In the case of the wrist-worn wearable, this method doesn't hold anymore, but the unit stays the same. The most common approaches to aggregate the raw data into epoch activity counts according to the paper (Ancoli-Israel *et al.*, 2003) are presented here:

- **A.** Time above threshold: In this strategy, the amount of time where the activity is above a selected acceleration threshold (usually 0.1-0.2 G after low pass filtering) is cumulatively counted per selected epoch
- **B.** Zero-crossing: In this approach, the number of times when the acceleration passes a value close to zero is counted for a selected epoch.
- **C.** Digital integration: This method is used with high sampling rate accelerometers, where the output is an integration of acceleration in a given epoch (after filtering).
- **D.** Maximum acceleration: In this approach, only the highest acceleration (after filtering) is saved for a given epoch.[6]

The drawbacks of these approaches are that in **A.** and **B.**, the acceleration level of the movements is not reflected. The **B.** (Zero-crossing) approach is additionally vulnerable to high-frequency artefacts. The **C.** (Digital integration) needs a high sampling frequency that requires more battery power, and therefore does not allow for long actigraphy recording. The **D.** (Maximum acceleration), compared to the **C.**, does not reflect the duration of the activities in each epoch. And finally, both **C.** and **D.** fail to represent the frequency of the movements.

The aggregated epoch activity counts are then used to estimate features describing circadian rhythms, sleep, etc. – see description in section 3.5. While there are differences between the wearables in the units measuring activity, the patterns of activity onset, offset, and peaks are usually similar across the wearables (Bellone *et al.*, 2016).

[6] Method used in Mindpax MindG and MGK wearables - see section 3.4

The commercial sport (fitness) trackers use similar accelerometers sensors like the scientific wearables. Although the data pre-processing is mostly not public, it also has to include bandpass filtering and aggregation of data into epochs. In sport-trackers, the epochs are not represented by activity counts but by higher-level aggregations, which are step counts per epoch during the active part of a day, and sleep phases (states) during automatically detected sleep. The steps are typically obtained using frequency analysis of the raw activity data. Some more advanced sport-tracker devices also detect different types of activities like walking, running, biking, swimming, etc. These outputs have only limited use in actigraphy analysis, but most of the devices would be able to measure the activity counts (as presented previously in this section) if used with different firmware.

Many modern devices, especially commercial sport-trackers, are also equipped with other sensors for measuring light, temperature, heart rate, pressure, GPS, ECG, etc.

3.4. Common Actigraphy Wearables

Actigraphy is becoming a standard measurement for sleep and circadian research as well as fitness tracking. Nowadays, there are over 200 different portable activity trackers made by various companies. Widely used in sleep and circadian research are, for example, wearables from ActiGraph corporation (Florida, USA) and CamNtech Ltd. (Cambridge, UK). Other wearables, more specifically oriented, but approved by published research, are developed by Condor (San Paulo, Brazil) (Bellone *et al.*, 2016), Vivago (Helsinki, Finland) (Lötjönen *et al.*, 2003), and Mindpax (Prague, Czech Republic) (Fárková *et al.*, 2019; Cuesta-Frau *et al.*, 2020; Schneider *et al.*, 2020).

The ActiGraph wearables allow raw recording storage in the high-frequency range of 30-256 Hz. This allows for a calculation of activity counts using any pre-processing method mentioned in section 3.3. It makes it also possible to evaluate body position, especially when the wearable is attached to the waist or thigh. The trade-off are higher requirements for storage capacity and for higher sample rates, which limit battery life. The old ActiGraph design is of bulky construction and might cause some discomfort to wear. Most of the ActiGraph wearables have to be read out manually. Nowadays, ActiGraph provides some improved models, where CentrePoint Insight can share data online through a reading station or a

smartphone. For models with Bluetooth technology (BT), there is a possibility to accompany the wearable with Polar heart rate monitors.

The main CamNtech actigraph model is called MotionWatch. It provides data from an accelerometric sensor and a light sensor. Data are pre-processed on the wearable and stored as epoch aggregates only. The epoch length may be set from 1-60 sec, where the settings affect the recording's maximal duration. The wearable is equipped with an event button, which may be used for different purposes based on study design. Due to lower sampling frequency, the requirements for storage and battery are also lowered. Therefore, the CamNTech wearable is significantly smaller than the ActiGraph. Similarly to the ActiGraph, the CanNTech cannot share data during recording, limiting its clinical use possibilities and the control over the data acquisition process. Other CamNTech products include a MotionWatch with increased mechanical endurance and a model and tiny actigraph similar to an NFC chip with one axis accelerometer designed for animal studies.

Figure 3.2 - Example of activity monitoring devices. In the upper row are presented the research wearables (MindG, ActiGraph, MotionWatch), which provide raw data for future analyses. In the bottom row are presented commercial smartwatches (Garmin, Apple, Withings).[7]

[7] Images were obtained from the respective producers' webpages

The Condor's ActTrust wearables are primarily oriented on sleep measurements, adding light and temperature measurement. Vivago's WristCare wearables are used in elderly care, focusing on automatic alarms. Mindpax's MingG is focused on psychiatric care, using a system for sharing and preparing data for physicians. Technical parameters and features of individual wearables shown in Figure 3.2 are presented in the following Table 3-1.

Table 3-1: The technical parameters of selected actigraphs

Manufacturer	*Model*	*Sampling frequency Data type*	*Storage capacity*	*Battery life*	*Online reading*	*Additional features*
ActiGraph	wGT3X-BT	30-100 Hz Raw data	4 GB (180 days at 30 Hz)	25 days (without BT at 30 Hz)	No	BT, Water resistance (WR) 1m 30 min
	GT9X	30-100 Hz Raw data	4 GB (180 days at 30 Hz)	14 days (sleep mode, at 30 Hz)	No	Display, WR, BT, Event button, Gyroscope, magnetometer, additional 16G 3-axis accelerometer
	CentrePoint Insight	32-256 Hz Raw data	512 MB (30 days at 32 Hz)	30 days at 32 Hz	**Yes**	Display, WR, BT, possible mobile app
CamNtech	MotionWatch 8	50 Hz 1-60 sec epochs	4Mbit (1,5 days for 1-sec epochs to 91 days 60-sec epochs)	91 days	No	Event marker, light sensor, WR
Condor	ActTrust 2	25 Hz 1-86400 sec epochs	8MB (90 days for 60-sec epochs)	90 days for 60-sec epochs	No	Display, shower resistant, event marker, temperature sensors, light sensors (all, colours and UVA/UVB)
Mindpax	MindG[+]	6.5 Hz 30-sec epochs	256 kB (27 days - but is periodically read, so till the end of the battery)	Over 7 months	**Yes**	WR, accompanied mobile app, or reading station
Vivago	WristCare	Aggregated sleep, activity, and circadian rhythm data[*]	Unknown, but as it is read over FM it is till the end of the battery	2-4 months	Over FM	Alarm button, WR, display, mobile app for user and family

[*] it is possible to record epochs, but the detailed specification is not provided publicly by the manufacturer are not clear

[+] this wearable is used in studies presented in this thesis

The online access to the data during recording represents a great advantage, as it allows for monitoring of patient compliance and wearable errors. Inclusion of the display also helps, as the wearable may be presented as a watch, and it shall not cause stigmatisation feelings to patients. Battery capacity is the main factor limiting the possible study duration in most wearables. Water resistance is another important feature as it is not uncommon that patients forget to take the wearable back on after it is removed, e.g. for hygiene. A display and mobile application may also increase patient compliance, as it provides feedback information (if it is allowed by the research design).

The market for commercial sport-trackers and smartwatches is huge. The higher-end consumer devices are commonly equipped with GPS, accelerometer, optical heart rate sensor, ambient light and infrared sensors, temperature sensor, altitude sensor, gyroscope, magnetometer, and some recent models also with electrical ECG monitor (FitBit, Apple, Withings, Samsung), and therefore may provide many interesting bio-measurements. While most of the devices are medically not validated, some of the flagship devices from well-established companies (Apple, Fitbit, Coros, Garmin, Polar, Suunto, Withings, etc.) were tested for accuracy in many domains. The accuracy of step count is generally quite high. The mean absolute percentage error (MAPE) is usually below 1 %, though it varies between devices. Estimated distance is much less accurate (MAPE > 10 %) unless using a GPS. In that case, it is more accurate, though usually slightly underestimated (MAPE 3-6 %) (Wahl *et al.*, 2017; Gilgen-Ammann, Schweizer and Wyss, 2020). Heart rate measurements are relatively accurate in a resting state, while during higher intensity activities, the accuracy drops. The error interval[8] obtained from the Blan-Altman plot (Altman and Bland, 1983) is approximately +/- 25 bpm (Claes *et al.*, 2017; Wang *et al.*, 2017). In comparison, the chest strap (which may be added to ActiGraph) has an error interval only up to +/- 10 bpm. The heart rate and activity measurements show valid proportional changes, and therefore they may be used to assess state alternation rather than absolute measurements (Hernando *et al.*, 2018; Henriksen *et al.*, 2020).

The accuracy of electrical ECG achieves a reasonable level under rest conditions (Saghir *et al.*, 2020). Concerning sleep, the accuracy of detected sleep duration was acceptable for all actigraphs (Ancoli-Israel *et al.*, 2003). The sleep onset and offset time are less accurate, as actigraphs typically detect sleep more likely than wake periods (Sadeh, 2011). Therefore, sleep

[8] Measured as limits of agreement, the interval that contains 95% errors

is generally well detected in a healthy population, but the accuracy drops for people with lower sleep efficiency (as detected by PSG).

All smartwatches can transfer data through BT to a smartphone, and therefore, they may be shared online with the caretaker/researcher. This is an important advantage because many possibly valuable measurements can be obtained this way to support treatment decisions. The automatic detection of selected activities (such as running, walking, biking, elliptical, swimming, etc.), blood oxygen saturation (SpO2), and heart rate is yet another advantage. Though some devices are relatively accurate, extreme caution is needed while using the outputs for any kind of medical consideration. For example, during the 2020 COVID-19 crisis, the SpO2 measurements could be great for homecare monitoring, but unfortunately, the reported accuracy is not sufficient for clinical use (Tomlinson *et al.*, 2018; Tarassenko and Greenhalgh, 2020). In spite of that, Mishra *et al* (2020) show that smartwatch measurements may present a timely warning sign of respiratory infection.

The unreliability of smartwatch measurements, together with the fact that circadian rhythmicity is not monitored by smartwatches or fitness trackers by default, still limits the selection of the wearables for actigraphy studies to those mentioned in Table 3-1.

3.5. Actigraphic Features

The most commonly used approaches assessing the circadian rhythmicity may be divided into parametric (usually cosinor) and non-parametric. Additionally, sleep is one of the main state- and trait-markers in BD, while another serious candidate for a state-marker is rhythm instability.

3.5.1. Cosinor Analysis

Cosinor analysis (Minors and Waterhouse, 1988; Cornelissen, 2014; Gonzalez *et al.*, 2018) is the most commonly used parametric approach to describe the circadian rhythm. The regular activity patterns are estimated by fitting a cosine function (see Eq. 3.1) with a fixed period, typically set to 24-hour. The resulting features are named the **Acrophase** – the time shift of the fitted function – the time of the activity peak, the **MESOR** (Midline Estimating Statistics Of Rhythm) – the offset of the fitted cosine function – the overall average activity, and the **Amplitude** – the difference between active and resting activity. The **Circadian Quotient** (CQ), computed as the ratio of the Amplitude and MESOR, represents an estimation of how well-circumscribed periods of activity are during a day - a proxy of rhythm robustness. (Gonzalez *et al.*, 2018)

The formula for cosinor:

$$Y(t) = M + A \cdot \cos(\omega \cdot t + \varphi) + e(t), \tag{3.1}$$

where $Y(t)$ – is the data-point measured at time t, M – MESOR, A – Amplitude, ω - angular frequency of the curve, φ – phase angle of the maximum value of the fitted curve (or acrophase, commonly represented in a daytime hour $= -24\frac{\varphi}{2\pi}$), and $e(t)$ – residual error at time t.

The formula 3.1 can be rewritten according to trigonometric angle sum identity as:

$$Y(t) = M + \beta x + \gamma z + e(t), \tag{3.2}$$

where $\beta = A \cdot \cos(\varphi)$, $\gamma = -A \cdot \sin(\varphi)$, $x = \cos(\omega \cdot t)$, and $z = \sin(\omega \cdot t)$.

The parameters of the fitted cosinor may be obtained, using the least-squares method, by minimising the sum of residual squares (RSS):

$$\mathrm{RSS} = \sum_{i=1}^{N} \left[Y_i - \left(\widehat{M} + \hat{\beta}\, x_i + \hat{\gamma}\, z_i \right) \right]^2$$

3.3

In the equation above, N is the total number of valid samples, Y_i is the measured value at the time i, and $\widehat{M}, \hat{\beta}$, and $\hat{\gamma}$ are the parameter estimations. The minimising triple is given by:

$$\begin{pmatrix} \widehat{M} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} N & \sum_i x_i & \sum_i z_i \\ \sum_i x_i & \sum_i {x_i}^2 & \sum_i x_i z_i \\ \sum_i z_i & \sum_i x_i z_i & \sum_i z_i \end{pmatrix}^{-1} \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i x_i \\ \sum_i Y_i z_i \end{pmatrix}$$

3.4

Amplitude is then obtained as $A = \sqrt{\beta^2 + \gamma^2}$ and phase shift as:

$$\varphi = \begin{cases} -\tan^{-1}\left|\frac{\gamma}{\beta}\right| & \text{for } \gamma \geq 0,\ \beta > 0 \\ -\pi + \tan^{-1}\left|\frac{\gamma}{\beta}\right| & \text{for } \gamma < 0,\ \beta > 0 \\ -\pi - \tan^{-1}\left|\frac{\gamma}{\beta}\right| & \text{for } \gamma \leq 0,\ \beta < 0 \\ \tan^{-1}\left|\frac{\gamma}{\beta}\right| & \text{for } \gamma > 0,\ \beta < 0 \end{cases}$$

3.5

Additionally, the quality of data approximation may be shown by the two measures:

1. The mean square error (MSE) of fit is:

$$\mathrm{MSE} = \frac{1}{N} \sum_t e(t)^2$$

3.6

2. The percentage of data explained by the cosinor model is Goodness of Fit (GOF), calculated based on the MSE (3.6) of cosinor model, and MSE of a constant model – (total mean square error - *TMSE*):

$$\mathrm{GOF} = 100 \cdot \frac{TMSE - \mathrm{MSE}}{TMSE} \qquad 3.7$$

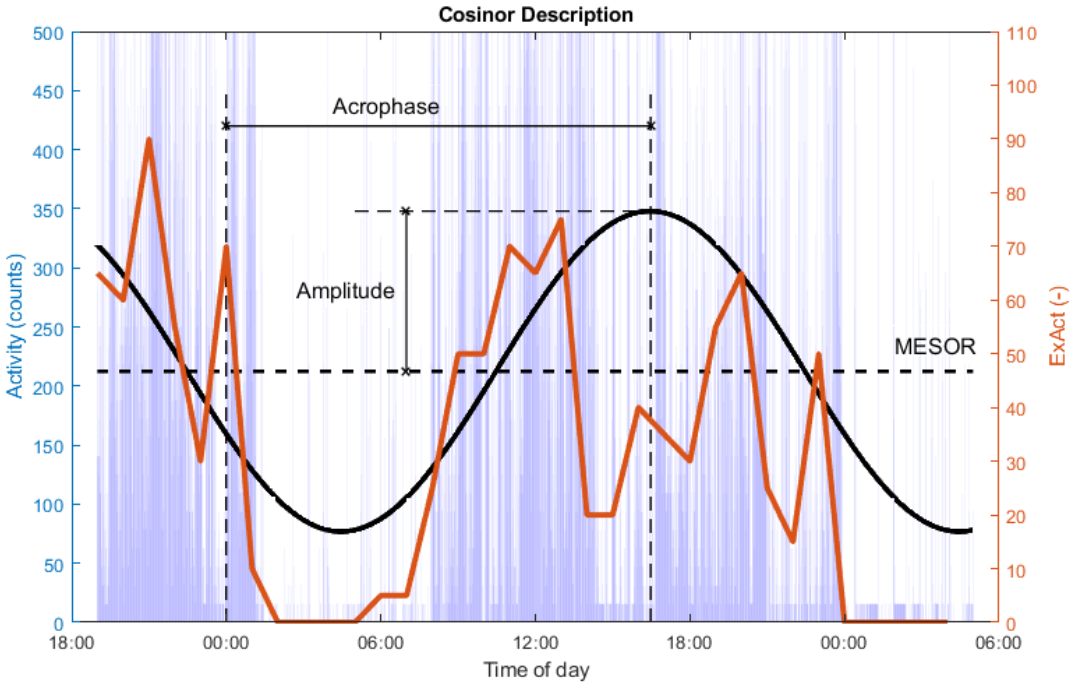


Figure 3.3 - Description of Cosinor rhythm characteristics. The figure shows a cosine function (black) fitted on the epoch actigraphy data (blue) with marked cosinor features Amplitude, Acrophase, and MESOR. The 24-hour period is fixed. The shown actigraphy data have low midday activity, which is increasing residuals error MSE and reducing GOF (12.9 %). For comparison, the figure also shows the hourly ExAct score (described in section 3.5.5).

Physiologically, the MESOR represents the mean overall level of activity. Increased MESOR values may be seen due to high levels of daily activity or due to shortened, disturbed, or fragmented sleep. The Amplitude feature represents the distinction between the rest and active part of days. It is increased in cases of high and stable daily activity with sound sleep. It may be reduced by disturbed sleep or fragmented daytime activity. The Acrophase represents the actual 'morningness' or 'eveningness' of the subject (an approximation of chronotype, see Chapter 6). During low activity days, the Acrophase is less stable. Additionally, it shows shifts in activity timing, e.g. due to the start of the daylight-saving time or travelling across time zones. The GOF represents the stability of the rhythm as well as the fragmentation of activity during both day and night.

The main advantage of the cosinor approach is its robustness. The method may be used for non-equidistantly sampled data or records with missing values. On the other hand, five assumptions (described below) are required for its use (Cornelissen, 2014), most of which are difficult to meet in actigraphy recordings.

1. The model should fit the data well: this requirement is commonly not met. This is the first challenge: Although the circadian activity pattern may be divided into the low activity part of sleep and high activity part of the awake state, the activity levels seldom follow the increasing and lowering trend of the cosine function. Quite commonly, there is a visible drop in activity in midday (see Figure 3.3). The data may also be skewed by either morning or evening activity peak, which is typical for extreme chronotypes (see Chapter 6).
2. Normal distribution of residuals: This is also not always met as the values are lower bounded (when there is no activity, the minimum actogram value is zero) and not from above, as the sensor saturation is not commonly met.
3. Homogeneous variance: This collides with much higher data variance during the waking hours than during sleep.
4. Independence of residuals: This is also not always met, as continuous actigraphy data measured for a given type of activity tends to stay similar over the whole course of the activity.
5. Stability of the features over time: This collides with the fact that the workdays and weekend (free) days are usually different. Moreover, variations of feature values are expected to change with the BD patient state, which is connected to the instability in the circadian rhythms (see Chapters 7 & 8).

In order to follow these assumptions, the cosinor features should be obtained from a shorter estimation window. This is contrary to obtaining a valid rhythm description, where the estimation window should be long to eliminate socially induced noise. Considering both of these contradictory requirements, we have chosen rather a shorter window, which is better determined in time (see Chapter 6 for the impact of window length in the task of chronotype estimation from actigraphy). In our research, we use features estimated from one-week and two-week-long windows. In this way, the distribution of working and free days should be similar in all windows.

An additional way to fulfil the assumptions, which may help with the exploration of the evening and morning activity peaks, is the use of a multicomponent cosinor (Cornelissen, 2014). In this way, the data are fitted by a mixture of two or more cosine functions of given periods. Similarly, the use of the wavelet function could increase the percentage of explained data. Unfortunately, such approaches obscure interpretability and comparability, as the classical cosinor is adopted by most of the actigraphy studies even despite the issues mentioned above (Jones, Hare and Evershed, 2005; Salvatore *et al.*, 2008; Faedda *et al.*, 2016; Krane-Gartiser *et al.*, 2019).

Throughout this work, the cosinor parameters are obtained from the fit of the cosine function with a 24-hour long period on one or two weeks of data (except for chronotype evaluation – Chapter 6). In order to maintain causality, the value for each day is based on data from the previous week (or two).

3.5.2. Non-parametric Circadian Rhythm Analysis

The non-parametric circadian rhythm analysis (NPCRA) (Witting *et al.*, 1990; van Someren *et al.*, 1996; Jones, Hare and Evershed, 2005) is a summary name for a set of features that describe activity patterns for each day, or few consecutive days, without assuming a particular underlying analytical function. Such an approach may be beneficial, as many assumptions of cosinor analysis may be violated when used on actigraphic data, especially in BD patients, where the rhythm is expected to be more disturbed. Such a less restricted approach to estimate features is then more prone to noise and errors based on missing data (see Chapter 5).

The traditional actigraphic features estimated using the NPCRA are the average activity during the most active ten hours (M10) and its mid-time (M10-time), the average activity during the least active five hours (L5), and its mid-time (L5-time). Moreover, L5 and M10 are combined in Relative Amplitude (RA) (Eq. 3.8), which is closely related to the CQ in cosinor analysis.

$$\mathrm{RA} = \frac{\mathrm{M10} - \mathrm{L5}}{\mathrm{M10} + \mathrm{L5}} \tag{3.8}$$

The M10 and L5 features, together with their timing, may be estimated for each day (Figure 3.4), or from a window of consecutive days (Figure 3.5). In this thesis, 7- and 14-day sliding windows were used in order to maintain the same ratio of free and working days.

As the way of estimation of daily L5 and M10 is not strictly standardised in the actigraphy studies (Gonçalves *et al.*, 2015), we have suggested a long-term monitoring approach, shown

in Figure 3.4. In this way, we expect to reduce the overlapping detections for consecutive days while allowing a maximum range of L5-time and M10-time in each of the days. In our approach, we add 5 hours from the previous and the following days to the evaluated daily window. In this way, the M10-time and L5-time can be detected for each hour of the evaluated day. Still, the possible overlapping M10 or L5 segments may cause a problem. To reduce this risk while keeping a reasonable range of possible L5-time and M10-time, the estimation times for L5-time are limited to 21:30 from the previous day (-2:30) to 21:30 of the current day. In the case of M10-time, the limits were set to 2:30 to 24:00 of the current day. Such an approach reduces unjustified jumps (crossing midnight) in L5-time and M10-time while keeping L5 as a sleep descriptor, where it is common to add sleep to the day following the wakeup. The daily L5 and M10 features are calculated for the respected detected epochs.

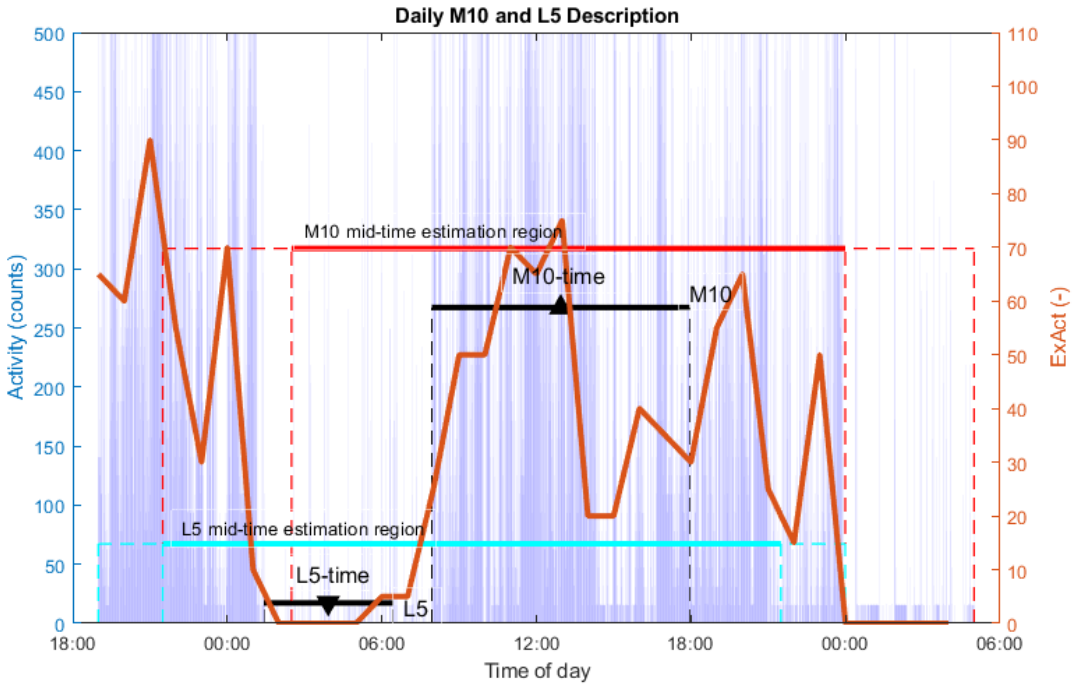


Figure 3.4 - Estimation of daily M10 and L5 values. The figure shows estimated L5 and M10 values for actigraphic data (blue) and regions where the L5 and M10 features are searched. The bold red section is where the M10 mid-time is searched. Similarly, the bold cyan region shows where L5 mid-time is searched. The dashed extensions represent the maximal position where the whole M10 or L5 region may expand. The regions do not correspond precisely with a calendar day to eliminate the possibility of the same region being detected twice on consecutive days. RA is, in this case, 0.88. For comparison, the figure also shows the hourly ExAct score (described in 3.5.5).

The daily values L5 and M10 are features that change quickly from day to day, compared to cosinor and other NPCRA features. This allows comparison between working and free days, etc., but on the other hand, it makes them much more prone to noise and missing data. Therefore, there is a modified version of both features, based on a week (or 2 weeks) average day. In this approach, the epoch data are aggregated into 5-minute segments. The main daily activity profile is computed by averaging daily values through the week (weeks). The mean profile is then, again, expanded by 5 hours before and after the 24-hour cycle in order to widen the possible mid-times ranges. The expansion is done by copying the first and last 5 hours, as shown in Figure 3.5.

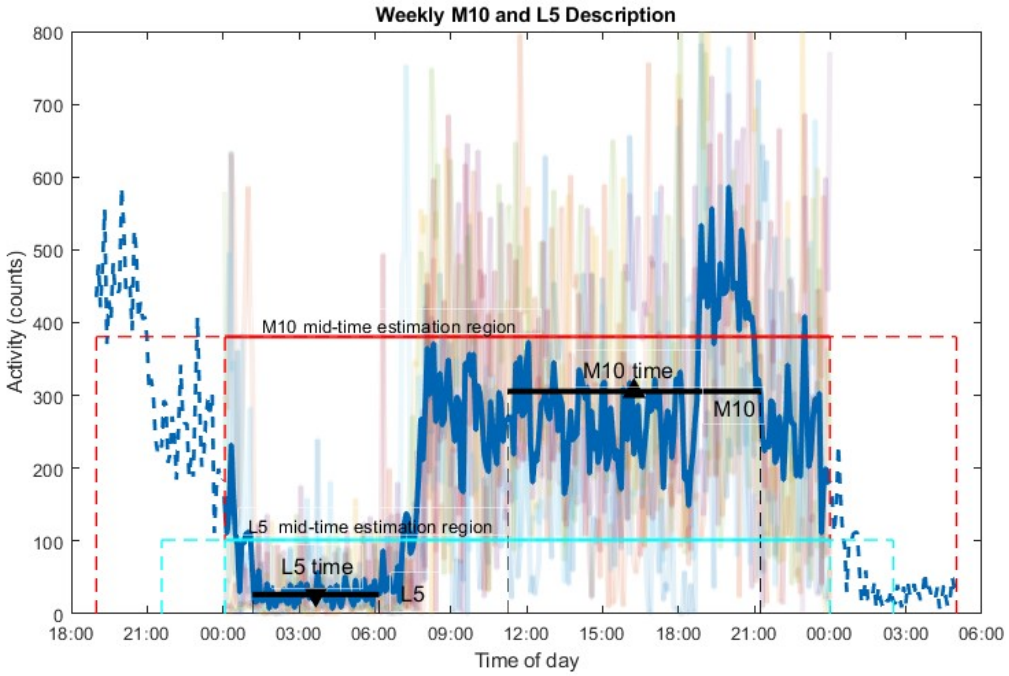


Figure 3.5 - Estimation of weekly M10 and L5 values. Daily data aggregated into 5-minute segments (coloured semi-transparent lines) are averaged into an average day (blue) and expanded by 5 hours before and after the averaged day (blue dashed). The M10 and L5 values are then obtained from the marked estimation regions. The data used for estimation were the same as for cosinor (Figure 3.3), and therefore may be compared. The RA for this data epoch is 0.84.

Concerning the physiological interpretation, the M10 represents the amplitude of rest-activity rhythm, which is connected to the motor capability and function of the circadian timing system (CTS) and cortical function, especially the frontal cortex's integrity (Gonçalves *et al.*, 2015).

The CTS functionality depends on the integrity of neurons in the suprachiasmatic nucleus (SCN) located near the optic nerve in the frontal part of the hypothalamus. Through SCN-Pineal complex, SCN is responsible for the secretion of the hormone melatonin. It has

been demonstrated that lowering the number of vasoactive intestinal polypeptide expressing neurons (stimulated by the light-dark cycle) causes a deficit in the circadian process (Hofman, 1950; Gonçalves *et al.*, 2015).

The reduction in M10 is associated with motor difficulty, exercise reduction, or CTS degradation. The reduction is typical for later stages of Alzheimer's disease (Gonçalves *et al.*, 2015). The M10-time is also connected with CTS, and it also represents the actual morningness or eveningness. Therefore, it is a promising feature for objective estimation of chronotype (see Chapter 6), such as cosinor Acrophase. Later activity midpoint was also observed in BD patients (Kaufmann *et al.*, 2018). BD patients are associated with evening chronotype (Gershon *et al.*, 2018). Daily-based M10-time may be used for social jetlag estimation as it may be computed separately for free and working days. In our published article (Fárková *et al.*, 2019), we observed an earlier M10-time peak in participants who were successfully undergoing a weight-reducing program compared to the unsuccessful group.

The L5 feature, as a measure of the rest phase, is also connected with the function of the CTS. The value of daily L5 is low when sleep is more efficient (few arousals) and increases with CTS degradation, commonly connected with ageing, and neurodegenerative diseases, such as Alzheimer's and Parkinson's (Gonçalves *et al.*, 2015). A higher value of daily L5 is expected in BD patients, who have typically low sleep efficiency (Harvey *et al.*, 2005; Gershon *et al.*, 2012; Geoffroy, Boudebesse, *et al.*, 2014). The weekly L5 may be used to represent the regularity of the sleep-wake regime. Higher values are obtained for the irregular regime, i.e. shift-work. The L5-time feature may be used similarly as the M10-time for chronotype estimation, but as it is correlated with mid-sleep time, it should be closer to the mid-sleep-based definition of chronotype (Juda, Vetter and Roenneberg, 2013b) – see Chapter 6. The variation of daily L5-time may be used to monitor and evaluate sleep hygiene, which is highly important in BD patients.

The derived RA feature is associated with the maturation of the central nervous system (CNS). It drops at a later age as the locomotor activity reduces and sleep efficiency decreases (Gonçalves *et al.*, 2015).

Other NPCRA features that focus on the daily rhythm are the Intradaily Variability (IV), which describes the fragmentation of daily rhythm, and the Interdaily Stability (IS), which evaluates the similarity between days. These features do not have a matching counterpart in the cosinor analysis. Unlike M10 and L5, these features may not be estimated for each day separately. The

IV and IS originate from the Chi-square periodogram (Sokolove and Bushell, 1978). And as such, they have to be estimated from a longer time window (beneficially 10+ days). We have used 7-days and 14-days windows. IV and IS are calculated from equations derived by Witting *et al.* (1990). These equations are:

$$\mathrm{IS} = \frac{N \sum_{h=1}^{p} (\overline{X_h} - \bar{X})^2}{p \sum_{i=1}^{N} (X_i - \bar{X})^2} \tag{3.9}$$

$$\mathrm{IV} = \frac{N \sum_{i=2}^{N} (\mathrm{X}_i - \mathrm{X}_{i-1})^2}{(N-1) \sum_{i=1}^{N} (\bar{X} - \mathrm{X}_i)^2}, \tag{3.10}$$

where N is the number of samples in a window, p is the number of samples per day (24-hour cycle), $\bar{X}_h$ are the hourly means, $\bar{X}$ is the average of the data, X_i are the individual data-points. While Witting *et al.* used X_i original samples, modern devices allow much finer sampling, which affects the results of IS and IV. Gonçalves et al. found that resampling the data into specific segment length (approximately 15-30 minutes segments) has favourable discriminative properties in both simulated and experimental data. Our findings (Hlaváč, 2020) support this suggestion as well: the best results were achieved by 10- to 20-minute segments for IV and 20-minutes and longer for IS). Consequently, we have resampled the data for the calculation of IS and IV into 20-minute segments.

The IV feature represents a rest-activity fragmentation. Therefore, it is associated with the maturation of CTS and diseases of the sleep-wake cycle. Fragmentation includes daytime sleepiness and nocturnal arousals. Additionally, a link between IV and sleep quality, poorer cognitive and motor performance, and reduced social interactions has been reported (Gonçalves *et al.*, 2015). Moreover, a lower IV value, i.e. smaller rest-activity fragmentation, is associated with better sleep consolidation. Physical exercise and bright-light therapy, which is expected to improve sleep structure, has been shown to improve the IV parameter in elderly people and patients suffering from dementia and Alzheimer's disease (Van Someren *et al.*, 1997, 1999).

The IS feature represents the similarity in day-to-day activity profiles. It is used as a marker of synchronisation with the light-dark cycle (24-hour zeitgeber) and stability of daily rhythm

(Alloy *et al.*, 2017). Therefore, it is associated with CNS function, mainly its photic and non-photic synchronisation, where higher IS represents a better state. IS has also been associated with CNS development (in new-borns) and mental disorders, where higher IS is associated with better cognitive functions. IS is highly affected by social factors and lifestyle. The typical value for a daytime worker is 0.66, while for a shift-worker it is typically 0.25. (Gonçalves *et al.*, 2015)

In addition to these widely used non-parametric features, we have included in our dataset also **Average Daily Activity** (ADA – Eq. 3.11) and average activity in four quarters of a day ($AQA_{1\text{-}4}$, where 1-4 represents the quarters of the day).

$$\mathrm{ADA} = \frac{1}{N_d}\sum_{i=1}^{N_d} X_i \qquad 3.11$$

In the above equation, N_d is the number of valid samples in each day (midnight to midnight) and X_i are the recorded activity data-points. $AQA_{1\text{-}4}$ were obtained in a similar manner over six-hour-windows (0:00-6:00-12:00-18:00-24:00).

Another set of added features is designed to directly compare changes in behaviour while the effects of different lifestyles are reduced. The data are divided into four categories (low activity, sedentary activity, moderate activity, and high activity), based on individual thresholds. The thresholds T_{1-3} are estimated as 1-3 quartiles (quantiles 0.25, 0.5, 0.75), and T_4 as the maximal value of each patient's activity distribution (using the whole set of recorded patient's data). The features, named DA_{1-4} (in the further text, the subscript is replaced by the category low/sedentary/moderate/high) represent a percentage of each day's activity, which belongs to the categories estimated by the following equation:

$$\mathrm{DA}_k = \frac{\sum_1^{N_d} X_i \le T_k}{N_d} - \mathrm{DA}_{k-1}, \qquad 3.12$$

where and $\mathrm{DA}_0 = 0$.

And finally, as a description of fragmentation activity in the active part of a day, the Root Mean Squares of Successive Differences (RMSSD – Eq. 3.13) was calculated for the daily M10 segments ($RMSSD_{M10}$). The advantage of RMSSD over standard deviation (SD) is that

it measures variability reflecting both temporal order and amplitude of the data (McGowan *et al.*, 2020).

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=2}^{N}(X_i - X_{i-1})^2}\,, \tag{3.13}$$

where X_i are individual data-points and N is their count.

3.5.3. Sleep Detection and Sleep Derived Features

As previously stated, actigraphy is often used for sleep/wake monitoring (section 3.4). Changes in sleep are connected to BD and both types of relapses, shortened or missing sleep in mania and prolonged sleep in depression (Plante and Winkelman, 2008). While the distinction between sleep and wake time using actigraphy is generally considered reliable, the detection of sleep phases as obtained from polysomnography, was not successful when using actigraphy recordings (Kaplan *et al.*, 2012; Kosmadopoulos *et al.*, 2014). Most sleep detectors for actigraphy data are based on detecting epochs of low activity. However, the algorithms are different between wearables, as their sensitivity for low activity and noise is system-dependent (Meltzer *et al.*, 2012; Cellini *et al.*, 2013; Smith *et al.*, 2020). There are no standards, how sleep (and especially the main daily sleep) should be detected, and some research-oriented systems (such as CamNtech – MotionWatch – section 3.4) work based on patient-marked laydown time or in a semi-supervised way (which requires manual correction from a researcher). Such an approach is not feasible for large long-term studies.

In our research, we have used sleep detection using the Mindpax algorithm. Firstly, sleep/wear-offs epochs are detected using logistic regression model odds calculated on the basis of features using 10- and 30-minute long widows. These features include average value in a centralised window, standard deviation, and percentages of increasing and decreasing values in windows preceding and following the classified sample. Secondly, the epochs of sleep are distinguished from wear-offs using thresholding of activity within detected epochs, excluding its edges. And finally, short gaps between the detected periods of sleep are filled, and very short periods of sleep are removed. (Vostatek, 2018)

Based on the detected sleep segments, a main sleep of the day - the night sleep - is detected as the longest sleep within 24-hour starting from 15 o'clock on (Vostatek, 2018):

1) filling up to 64-minute long gaps between sleep epochs

2) removal of sleep epochs shorter than 200 minutes

3) filling up to 240-minute-long gaps between sleep epochs

The primary sleep parameters are obtained using the detected night sleeps. These include night sleep duration (SleDur), the sleep onset time (SleON), sleep offset time (SleOFF), and mid-sleep for each day. Overall daily amount of sleep ($SleDur_{18}$) is obtained as a sum of durations of all detected sleep epochs within 24-hour, starting at 18 o'clock or at midnight ($SleDur_{daily}$). This way, the main night sleep is usually not divided between two days.

Sleep quality is assessed by several features, among them are:

- Wake After Sleep Onset (WASO). This is the duration of all segments during the main night sleep, which the sleep detector didn't identify as sleep.
- Variability in data during the main daily sleep $RMSSD_{sleep}$ assessed using Eq. 3.13.
- Restless sleep (RSL) and immobile sleep (ISL)

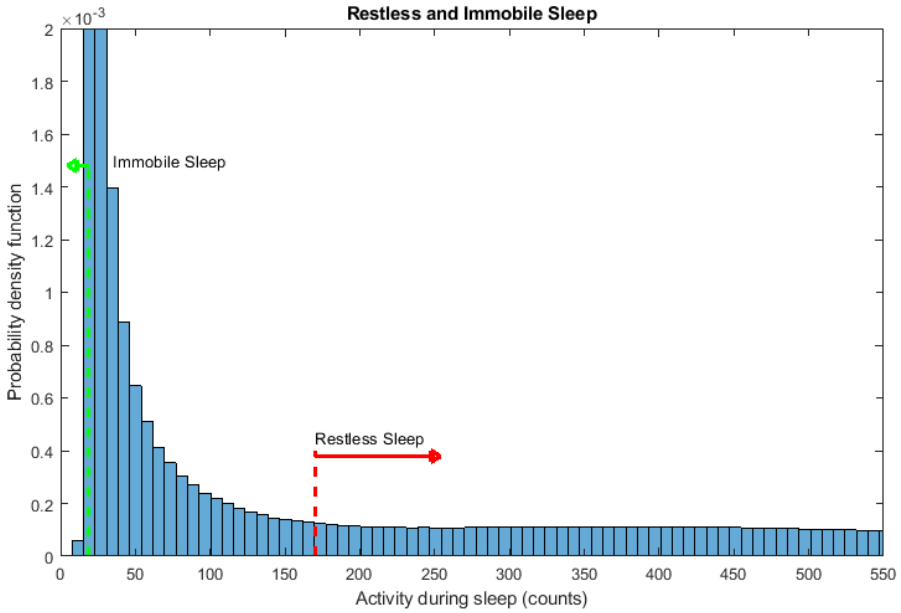


Figure 3.6 - Thresholds for Restless Sleep (RSL) and Immobile Sleep (ISL) (ACTIBIPO 2 dataset). The figure shows the probability density function estimate for activity counts during night sleep. The ISL threshold is chosen, so it includes the two least bits of recorded activity counts (includes approximately 85 % of night sleep activity). The RLS threshold is based on the distribution of activities during night sleep (includes approximately 5 % of night sleep activity).

Restless sleep (RSL) is a percentage of samples surpassing high levels of activity during sleep, and ISL is a percentage of samples under low sleep activity thresholds. These thresholds were obtained from the main daily periods of sleep (ACTIBIPO 1 and 2 datasets – Chapter 4). The ISL threshold is set as the two lowest epoch activity counts (based on the used analogue-digital converter used). The RLS is set based on the distribution activities during night sleeps[9] (see Figure 3.6).

The commonly used diary-based (other markings of bedtime) parameter, the sleep onset latency, was substituted by automatic approximation features. These features are average activity during 2 hours prior to sleep onset (APSO) and after sleep onset (AASO). Similarly, waking up has been described by average activity during 2 hours prior to wake-up (APWU) and activity after wake-up (AAWU). Then combined into the ratio of APSO/AASO and AAWU/APWU based on a shorter (30 minutes) period were used to evaluate the steepness of falling asleep and waking-up processes.

Based on the main night sleep offsets and onsets, the variability in data during the active part of a day was assessed as $RMSSD_{actday}$ (Eq. 3.13) from data resampled into 5-minute segments and autocorrelation lag (ACL) (correlation of signal moved by 5 minutes with the original) of the resampled active day data.

[9] The position of Restless Sleep threshold is based on minimal point in sleep activity from Mindpax GMK (older version of MindG) see supplement (Schneider *et al.*, 2020) and transformed to MindG (Table 3-1) activity counts.

3.5.4. Chronotype Measures

Chronotype related features can also be calculated based on actigraphy. The detailed connection between chronotype and chronotype-related features (such as L5-time, M10-time, mid-sleep, Acrophase, etc.) is described in Chapter 6.

The Munich Chronotype Questionnaire (MCTQ) (Roenneberg, Wirz-Justice and Merrow, 2003; Juda, Vetter and Roenneberg, 2013a) is based on patient-filled typical sleep times during working and free days. The two main features, which are estimated based on the MCTQ, are a corrected mid-sleep time on free days (MSFsc) and social jetlag (SJL), the difference between sleep habits in freely running regime (free days) and working days. Using the detected night sleep epochs in combination with reported free days or calendar-based free days, the values of MSFsc and SJL may be obtained from actigraphy as:

$$\mathrm{MSFcs} \;=\; \mathrm{MS}_{free} - \frac{\mathrm{SD}_{free} - \left(p_{free} \cdot \mathrm{SD}_{free} + p_{work} \cdot \mathrm{SD}_{work}\right)}{2} \tag{3.14}$$

Here the MS_{free} is the average mid-sleep on free days (waking up on a free day), SD_{free} is the average sleep duration on free days, SD_{work} is the average sleep duration on work days and p_{free} is the probability (or ratio from all days) of a free-day and p_{work} is the probability of a working day. For a normal full-time worker, this may be $p_{free} = {}^{2}/_{7}$ and $p_{work} = {}^{5}/_{7}$, based on a five-day workweek.

$$\mathrm{SJL} \;=\; \mathrm{MS}_{free} - \mathrm{MS}_{work} \tag{3.15}$$

Where MS_{work} is the average mid-sleep on working days, and MS_{free} is the average mid-sleep on free days.

3.5.5. Explainable Activity Measure (ExAct)

Unlike the steps recorded by fitness tracking devices, the activity measured in counts based on the acceleration filtered acceleration measurements (section 3.3) has no straightforward interpretation. Therefore, we have engineered a feature that could be presented and explained to patients as a part of motivation and for better self-understanding. (Schneider, 2021)

This explainable activity (ExAct) assigns a score to each 5-min segment based on the expected activity on such activity level. The 5-min segments are used to filter short bursts of physical activity, which may be artefacts caused by the environment (as transport, mechanical drill, etc.) and therefore hard to interpret. The activities were divided into 4 levels (see Table 3-2).

Table 3-2: Explainable activity levels description

Level	Description	score	Activities
1	Sleep	0	sleep, passive TV/movie watching, deep meditation, etc.
2	Low activity	5 (1/min)	computer work, driving, leisure walking, and low demanding housework, etc.
3	Medium activity	10 (2/min)	walking, swimming, active housework etc.
4	High activity	20 (4/min)	sports as running, biking, floorball, etc.

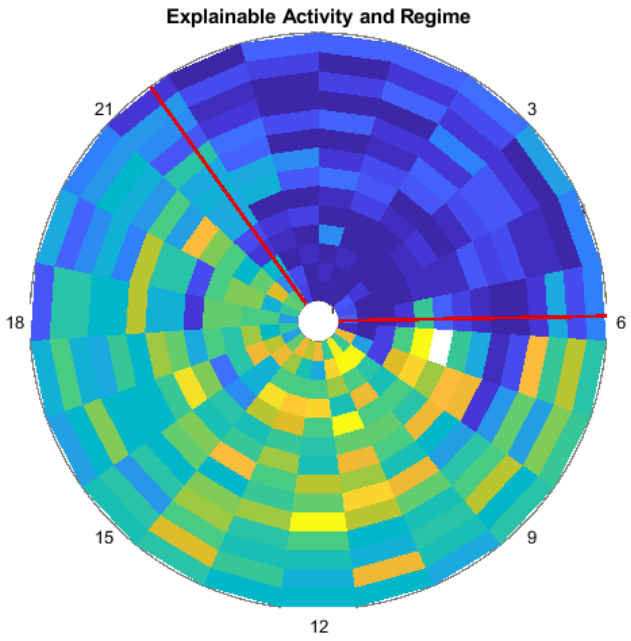


Figure 3.7 - Explainable activity and regime visualisation. The figure presents visualisation of hourly explainable activity (ExAct) scores for 14 consecutive days. The presented data show a regular regime. The red lines present average sleep onset and offset times. The days are organised so that the most outer circle is the actual day and the most inner circle the first (oldest) day of the defined window

The activity levels' thresholds were trained in a microstudy where 6 young, healthy individuals recorded their activities (145) while writing an activity logbook. The ExAct score can be added to hourly scores. These scores may then be presented to patients to visualise their rhythm regularity, as in Figure 3.7. Daily values represent the overall activity level during the day and could be interpreted as the number of active minutes during a day. The daily ExAct may be used as an additional measurement of daily activity or, when it is divided by the interval between wakeup and sleep onset (hours), as the active-day-part activity measurement ($ExAct_{active}$ per active hour). (Schneider, 2021)

4. Datasets

This chapter presents details about recorded and used datasets, including individual studies procedures and recorded data summaries. In the presented thesis, three datasets are used:

- ACTIBIPO 1 – a dataset for comparison between healthy controls (HC) and BD patients (Chapter 7)
- ACTIBIPO 2 – a dataset for exploration and classification of symptomatic episodes in the course of BD (Chapters 5 & 8)
- CHRONOBIO – a dataset combined from two related datasets, including a set of women undergoing a weight reduction program and a set of healthy women. This dataset is used to assess the relationship between selected actigraphic features and results of chronotype questionnaires (Chapter 6)

4.1. ACTIBIPO 1 Dataset

4.1.1. Participants and Procedure

Actigraphy data were recorded for more than 90 days in 35 BD patients mainly with BD-I diagnosis, recruited from the outpatient BD clinic at the National Institute of Mental Health (NIMH), in Klecany, Czech Republic, and in 26 HC, matched for age and sex, who were recruited by advertisement in the community. All BD patients underwent a baseline psychiatric examination by a trained institutional psychiatrist, confirming euthymic state or low levels of depressive/manic symptoms, using the Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979) and the Young Mania Rating Scale (YMRS) (Young *et al.*, 1978).

Inclusion criteria: all BD patients were diagnosed according to DSM-5 criteria (APA, 2013). At the study entry, all patients had to be euthymic or in a remitted state (i.e., YMRS $\leq$ 12 and MADRS $\leq$ 9, see Table 4-1) with no reported mood episodes for $\geq$ 60 days prior to study entry (Tohen *et al.*, 2009).

Exclusion criteria for BD patients were the presence of an acute depressive episode, dysthymia, suicidal thoughts, a (hypo)manic episode, or diagnosis of schizo-affective disorder at enrolment.

HC exclusion criteria were: past or acute presence of a moderate depressive or (hypo)manic episode or suicidal thoughts, diagnosed neurological, sleep, or mood disorders, or a family history of mood or psychotic disorder among their first-order relatives.

All BD patients and HC who fulfilled the inclusion criteria were equipped with an actigraphy wearable and instructed to wear it preferably on their non-dominant hands.

On enrolment into the study, all participants answered a demographic questionnaire. The HC pool was contacted through an emailed screening questionnaire which asked for some basic information (age, sex, and employment status) and family disease history (neurological: epilepsy, Parkinson's disease, etc., sleep disorders: insomnia, sleep apnea, narcolepsy, etc.). Subjects who fulfilled the screening criteria were further evaluated using the M.I.N.I. structured questionnaire (Lecrubier *et al.*, 1997) for neuropsychiatric disorders. All participants were equipped with a wrist-worn actigraphic monitoring wearable (Mindpax – GMK[10]) and were instructed to remove it only when necessary.

During follow-up (i.e., the period when the data were recorded), BD participants were assessed monthly by their treating physician via in-person visits or a telephone interview to identify their current psychiatric state. We allowed for some minor increase of symptoms during follow-up (i.e., YMRS $<$ 15 and MADRS $<$ 15). The clinical episodes criteria included psychiatric hospitalizations, work incapacity, MADRS $\geq$ 15, YMRS $\geq$ 15 (Macfadden *et al.*, 2009), suicidal ideation, or substantial deterioration of the patient's clinical state.

The ethical committee of the NIMH in Klecany, Czech Republic, has approved the study and all BD patients and HC signed written informed consent.

[10]A device with similar parameters as MindG (Table 3-1), but due to different hardware the counts scores are not directly comparable

Table 4-1: Demographic, health and activity characteristics in ACTIBIPO 1 patients and controls groups

Metadata	BD	HC	p-value
Participants count final (original) [BD-I/BD-II]	25 (35) [16 BD-I, 9 BD-II]	25 (26)	-
Age	39.72 (SD 12.85, range 22-63)	39.68 (SD 11.19, range 25-63)	$p^{\ddagger} = 0.6549$
Sex	60 % female (N=15)	68 % female (N=17)	$p^{\dagger} = 0.7688$
Days in study (recorded / valid)	134 (SD 39) [range 61 - 179] / 86 (SD 21) [range 50 - 124]	97 (SD 12) [range 72 - 126] / 86 (SD 13) [range 62 - 108]	(for recorded) $p^{\ddagger} = 0.0018$* (for valid) $p^{\ddagger} = 0.2977$
Working status (full-time/part-time/no work)	6 / 12 / 7	22 / 3 / 0	$p^{\S} = 0.0125$*
Total days in the study (recorded/valid)	3341 / 2158	2426 / 2113	-
MADRS	At admission to the study 2.3 (SD 3.9) Through the study - with relapses 8.8 (SD 7.1) - without relapses 6.6 (SD 3.8)	-	-
YMRS	At admission to the study 0.4 (SD 0.81) Through the study - with relapses 1.8 (SD 3.3) - without relapses 2.1 (SD 3.5)	-	-
Mood episodes	In the study: 2 patients with episode/s (2 dep., 1 hypomania) Excluded: 5 patients with episode/s (7 dep., 1 mania - psychosis)	-	-
Lithium	***7 (28 %)***	-	
Antipsychotics used:	**12 (48 %)**	-	
Quetiapine	10 (40 %)	-	
Olanzapine	4 (16 %)	-	
Antidepressants used:	**6 (24 %)**	-	
Bupropion	5 (20 %)	-	
Sertraline	2 (8 %)	-	
Antiepileptic used:	**9 (36 %)**	-	
Lamotrigine	6 (24 %)	-	
Valproate	3 (12 %)	-	
Carbamazepine	1 (4 %)	-	
Mixed/Single treatment	19 patients/6 patients	-	
Average features††			
ADA (average daily activity)	605 (SD 110)	778 (SD 92)	The statistical comparison is presented in Table 7-2
Sleep duration	8.98 (SD 1.22) hours	7.40 (SD 0.51) hours	
Circadian quotient	0.78 (SD 0.12)	0.66 (0.07)	
M10-time	14.7 (SD 1.3) o'clock	14.9 (SD 1.3) o'clock	
RSL (restless sleep)	2.6 (SD 0.9) %	2.1 (SD 0.6) %	
MSFsc (chronotype)	3.7 (SD 1.0) o'clock	3.6 (SD 1.1) o'clock	
LTTV in features††			
ADA	103 (SD 32)	94 (SD 25)	The statistical comparison is presented in Table 7-2
Sleep duration	1.7 (SD 0.6) hours	1.3 (SD 0.3) hours	
IV (Intradaily variability)	0.07 (SD 0.02)	0.05 (SD 0.01)	
IS (Interdaily stability)	0.06 (SD 0.02)	0.06 (SD 0.02)	
M10	166 (SD 50)	148 (SD 42)	
M10-time	2.1 (SD 0.8) hours	2.6 (SD 0.5) hours	
L5	31 (SD 23)	38 (SD 23)	
L5-time	1.8 (SD 0.6) hours	1.4 (SD 0.5) hours	
RSL	1.8 (SD 0.8) %	1.5 (SD 1.0) %	

Significance * < 0.05 ** <0.01 *** <0.001

‡ Mann-Whitney test; † Fisher exact test; § χ^2 test (chi2 = 8.77)

††Selected features, for all, see Table 7-1 and the supplement of Schneider *et al.*, 2020 article

LTTV stands for long-term temporal variability.

4.1.2. Subjects Characteristics

The BD patients and HC group characteristics, after exclusions, are shown in Table 4-1. Ten (29 %) of the 35 BD patients enrolled in the study were excluded: 5 subjects were excluded due to insufficient length of recorded inter-episode data (≤ 50 days - 4 for depression episodes and 1 for psychosis after childbirth), 4 for an excessive amount of missing actigraphy data (due to wearable removal or its malfunction), and 1 resigned from the study upon personal request, resulting in 25 BD patients in the final set. All of the subjects were attending a standard BD treatment program and were using clinicians' choice medication. Among 26 HC, 1 subject was excluded due to an excessive amount of missing data, resulting in 25 HC in the final set. A lower dropout rate in HC vs BD patients was expected.

4.2. AKTIBIPO 2 Dataset

4.2.1. Procedure

The actigraphy data were recorded for 18 months and longer in 369 BD patients. The patients were contacted through their treating physicians or an online campaign. The onboarding was done in two phases. In the first phase, patients were instructed to fill in an online structured pre-screening questionnaire to confirm their BD diagnoses. Those whose diagnosis was confirmed were instructed to read information about the study, confirm the informed consent, and fill in their contact information and their treating physician contact information.

In the second phase, the patients have been divided into groups:

1. CORE group patients have been evaluated by the institutional psychiatrist at NIMH in Klecany, who confirmed their diagnosis.
2. PERIFERY1 group containing those who could not be personally evaluated at NIMH or for whom the institutional psychiatrist at NIMH could not confirm their diagnosis beyond any doubt. The information about the patients, which was needed during the study, was provided by their treating physician.
3. PERIFERY2 group is similar to the group PERIFERY1. Only the treating physicians of the individual patients did not provide the needed additional information.

All patients were contacted, either during the personal examination (CORE) or by phone (PERIFERY1 and PERIFERY2). Additional personal information was collected, and all patients were evaluated using MADRS and YMRS clinical scales. The MADRS and YMRS

were chosen for their applicability over the phone. All patients, who fulfil the inclusion criteria, were instructed to wear an actigraphy wearable (Mindpax – MindG – see section 3.4), preferably on their non-dominant hands. Furthermore, they were asked to fill in a weekly Aktibipo SElf-RaTing questionnaire (ASERT – introduced in section 4.2.2) using a provided Mindpax mobile application.

The inclusion criteria were the diagnosis of BD disorder according to DSM-5 criteria (APA, 2013) with 2 or more affective episodes in the anamnesis and actual remitted state (MADRS $\leq$ 12 and YMRS $\leq$ 9).

Exclusion criteria were the same as in the ACTIBIPO 1 dataset - the presence of an acute depressive episode, dysthymia, suicidal thoughts, a (hypo)manic episode, or diagnosis of schizo-affective disorder at enrolment.

Additionally, patients in the CORE group agreed to be evaluated using MADRS and YMRS rating scales by phone each month. For CORE and PERIFERY1 groups, the health record (hospitalisations with BD diagnosis, work insufficiencies caused by worsening of the BD state, and suicidal attempts) was collected at the end of the patient recorded period.

Relapses according to study protocol were all states with high rating scales scores (YMRS $\geq$ 15, or MADRS $\geq$ 22), or hospitalisation for BD diagnosis, or work insufficiency caused by BD, or suicidal attempt. This study was organised as a non-intervention study. Therefore, there were no alternations of treatment based on the periodical scaling. Only when the patient had a high score in MADRS suicidal thoughts question, this information was shared with his treating physician.

The ethical committee of the NIMH in Klecany, Czech Republic, has approved the study and all BD patients and HC signed written informed consent.

4.2.2. Self-rating Questionnaire (ASERT)

The Aktibipo SElf-RaTing (ASERT) is a questionnaire for ecological momentary assessment (EMA) of mood in bipolar disorder invented at NIMH (Anýž *et al.*, 2021 preprint). The questionnaire (Table 4-2) contains 10 items, mapping depressive (4 items), manic (4 items), and non-specific (2 items) symptoms, with 5 possible response levels for each symptom.

Table 4-2: ASERT description, questions, questions grouping and possible answers in Czech and English

No	Group	Questions English version	Questions Czech version
1	**depressive**	I feel sad, downhearted	Cítím se smutně, sklesle
2		I do not enjoy anything, and nothing pleases me	Nic mě nebaví, netěší
3		I have no energy	Nemám energii
4		I feel gloomy and pessimistic about the future	Budoucnost vidím černě, pesimisticky
5	**manic**	I feel unusually great, optimistic	Cítím se neobvykle skvěle, optimisticky
6		I have excess energy	Mám nadmíru energie
7		My thinking is very fast, others cannot keep up with me	Myslí mi to hodně rychle, ostatní mě nestíhají
8		I need to sleep less than usual	Potřebuji spát méně, než obvykle
9	**non-specific**	I feel restless, tense	Cítím neklid, napětí
10		I cannot focus	Nemohu se soustředit
		Reply options	**Možné odpovědi**
		0: I do not agree	0: nesouhlasím
		1: more likely, I do not agree	1: spíše nesouhlasím
		2: I probably agree	2: asi souhlasím
		3: I agree	3: souhlasím
		4: I completely agree	4: naprosto souhlasím

As stated in the previous section, ASERT was administered on a weekly basis through a smartphone application developed by Mindpax. The ASERT was provided in Czech, and English versions, based on the settings of the mobile phone.

The questionnaire has been validated against MADRS and YMRS symptomatic episodes (MADRS ≥ 15, or YMRS ≥ 15). In the case of symptomatic episodes, the average episode thresholds obtained from logistic models (Anýž *et al.*, 2021 preprint) were either the sum of 5 in the manic ASERT group for a manic episode or the sum of 15 in ASERT depression and nonspecific group for a depressive episode. The depressive and nonspecific groups were combined because MADRS also includes both of these symptom types.

4.2.3. Subjects Characteristics

From 369 BD patients, who were included in the study, 88 were excluded for a short duration of recorded data - less than 6 months of valid days data (80 % of data-points per day), 5 were excluded for too few recorded ASERTs (at least 12 ASERTs – approx. 3 months were required), and 1 was excluded based of missing gender and age information, leaving 275 BD patients for analyses. The differences between groups are presented in Table 4-3.

Table 4-3: Demographics, activity and health characteristics in ACTIBIPO 2 patients

Group/parameter	CORE	PERIPHERY1	PERIPHERY2
Participants count	98	122	55
Sex (female)	58	72.5[+]	45
Birth year	1980 (SD 11)	1978 (SD 11)	1975[***] (SD 12)
BMI	27.85 (SD 5.72)	28.83 (SD 6.48)	27.44 (SD 5.93)
Height	173.7 (SD 9.2)	173.2 (SD 8.7)	171.7 (SD 8.6)
Weight	84.2 (SD 19.2)	86.4 (SD 19.6)	81.0 (SD 18.7)
Recorded days per patient	796 (SD 267)	601[***] (SD 213)	568[***] (SD 245)
Valid days (%)	85.7 (SD 14.1)	85.1 (SD 16.9)	91.2[**/#] (SD 11.9)
ASERTs (count)	90.4 (SD 35.3)	67.5[***] (SD 30.6)	66.6[***] (SD 33.1)
ASERTs per week (count)	0.80 (SD 0.19)	0.80 (SD 0.24)	0.82 (SD 0.18)
ASERTs average dep. score	3.72 (SD 2.93)	5.24[***] (SD 3.41)	5.11[*] (SD 3.28)
ASERTs average man. score	1.42 (SD 1.39)	2.28[***] (SD 1.88)	2.28[*] (SD 2.11)
ASERTs average all score	7.10 (SD 4.83)	10.45[***] (SD 6.15)	10.31[***] (SD 5.54)
Relapses[‡] (scales and hospitalisation)	365[x]	66	24
Relapses per participant[‡] (min-max)	3.7 (0-20)	0.54[***] (0-7)	0.43[***] (0-7)

[+] one participant underwent a change of gender during the study
[‡] in the PERIPHERY groups, the scales were usually not provided. Therefore, scales relapses may not occur. Additionally, PERIPHERY2 have quite limited information about relapses. Therefore, the number of relapses may be quite underestimated
[x] all individual scales (MADRS ≥ 15, or YMRS ≥ 15) are considered as separate relapses
[*] < 0.05, [**]< 0.01, [***]< 0.001 for a CORE – PERIPHERY difference
[#] < 0.05 for a difference between PERIPHERIES

4.2.4. Expert Labels

As can be seen from above, the clinical patient state has been assessed only by health records, plus in CORE group by monthly scaling. Such data (the ground truth) are too scarce to be compared to the daily actigraphic features. Therefore, an expert labelling of patients' states has been introduced.

Using the MADRS and YMRS scales, ASERTs, medical records, and notes from monthly scaling (only CORE group), a team of experts annotated time epochs whenever there was sufficient information to conclude a patient state. The team consisted of 3 psychologists, who were providing the monthly scaling, and analysts (2-3) who were processing the data. No actigraphic features were visualised during annotation, except for regions where the actigraphy data were missing. Based on the general agreement of this team of experts, there were marked epochs of remission, depression-onset, depression, depression-offset, mania-onset, mania,

mania-offset. For cases of mixed symptoms, multiple labels were used for the same epoch. These cases were extremely rare. Generally speaking, the scale-based or medical record-based relapse (symptomatic) periods were expanded using ASERTs and internal notes from monthly scaling. Additionally, the regions around, in which the onset (offset) of symptoms most likely occurred were marked as (-onset, or -offset).

Similarly, regions with elevated symptoms that didn't reach the threshold for relapse (MADRS ≥ 15, or YMRS ≥ 15) were marked as onset or offset. Hospitalisations periods haven't been annotated with expert labels, because the activity during hospitalisation is considerably restricted. A summary of the expert annotations is presented in the following Table 4-4.

Table 4-4: Expert labels summary information

Parameter/state	**Remission**	**Mania onset**	**Mania**	**Mania offset**	**Depression onset**	**Depression**	**Depression offset**
Labels (count)	218	94	48	78	156	109	150
In patients (count)	87	46	27	42	64	50	67
Labelled days (count)	21424	1960	1189	1554	4893	4232	4154
Valid actigraphy for labelled days (%)	88.76	89.59	85.95	90.48	88.39	88.40	87.53
Labelled segments duration (days) Median (Q1-Q3)	55 (26-101)	16.5 (9-29)	19 (10.5-34)	14 (8-26)	22 (11.5-36)	27 (8.75-47.25)	18 (10-31)
Valid actigraphy during segments (%) Median (Q1-Q3)	98.32 (85.5-100)	100 (94.74-100)	100 (80.03-100)	100 (93.24-100)	100 (90-100)	100 (82.77-100)	100 (92.59-100)
Labelled segments per patient Median (Q1-Q3)	2 (1-3.75)	1.5 (1-3)	1 (1-2)	1 (1-2)	2 (2-3)	2 (1-2)	2 (1-3)

4.3. CHRONOBIO Dataset

4.3.1. Procedure

The CHRONOBIO dataset includes a group of women undergoing weight-loss treatment (obese and overweight – obesitology subset), and a control group of healthy women (normal weight – control subset). Part of the sample (the obesitology subset) overlaps with the sample we have examined previously in a study by Fárková, Schneider *et al.* (2019); however, the data were subjected to different analyses and have now been restudied in the context of chronotyping methodology. Both groups underwent the same procedure.

Inclusion criteria were as follows: age 18+, but not of menopausal and post-menopausal age, BMI 18.5-55 kg/m^2 during onboarding, no shift work, no pharmacologically treated psychiatric illness and written consent to participate in the study. Exclusion criteria were having less than four weeks (28 days) of valid recorded days of actigraphy data, time zone travel, and BMI ≥ 55.0 kg/m^2 at any time during follow-up.

In this study, we have equipped a group of women participants with an actigraphy wristband (GMK[11] actigraph by Mindpax Ltd.) and instructed them to wear it at all times for up to three months. All participants filled the Morningness-Eveningness Questionnaire (MEQ) (Horne and Ostberg, 1976) and Munich Chronotype Questionnaire (MCTQ) (Roenneberg, Wirz-Justice and Merrow, 2003) chronotyping questionnaires at study admission. For details on MCTQ features estimation, see section 3.5.4. All participants were also surveyed for work status, dates of between-time-zone (long-distance) travel, and free days other than weekends and public holidays (the free days' survey).

Additionally, after a follow-up period of approximately 18 months, a subgroup of participants from the control group re-filled the MCTQ and MEQ questionnaires.

4.3.2. Subjects Characteristics

From the total of 122 women recruited to the study, 2 women were excluded due to BMI > 55 or more; 1 woman was excluded due to travelling between time zones in the observed period, and 19 women were excluded due to less than 28 valid consecutive days. The resulting set contained actigraphy recordings from 100 participants available for analyses. Descriptive statistics of the complete set, as well as of the two original subsamples, can be found in

[11]A device with similar parameters as MindG (Table 3-1), but due to different hardware the counts scores are not directly comparable

Table 4-5. The most extreme chronotypes (min-max) in the set were MEQ 27-72, MCTQ-MSFsc 1.7 – 6.7 o'clock, and the most extreme relative social jetlag values (MCTQ-SJLrel) were 0 – 3.5 hours. The key differences between the obesitology subset and control subset were in BMI and actigraphy duration by design. No significant differences were observed between the two subsets in terms of chronotype, gender or age. Moreover, while there are differences in BMI, their combined range is similar to that in the Czech population.

Table 4-5: Demography, health, chronotype and actigraphy data characteristic for the CHRONOBIO data set and its subsets

Group/parameter	**Whole dataset median, (Q_1 - Q_3)$^+$**	**Obesitology subset median, (Q_1 - Q_3)$^+$**	**Control subset median, (Q_1 - Q_3)$^+$**	**p-value$^\#$**
Participants (females)	100 (100)	61 (61)	39 (39)	-
Age (years)	37 (30 - 43)	39 (30 - 44)	34 (29 - 41)	0.124
Height (cm)	168 (164 - 173)	167 (164 - 173)	170 (164 - 174)	0.240
Weight (kg)	82.0 (64.0- 99.0)	95.0 (84.2 – 104.8)	63.0 (59.0 – 69.5)	<0.001*
BMI (at onboarding)	30.4 (23.1 – 34.7)	33.4 (30.7 – 37.3)	22.1 (20.5 – 24.0)	<0.001*
Chronotype characteristics				
MEQ	53 (47 - 60)	53 (48 - 60)	53 (46 - 59)	0.544
MCTQ-MSFsc	3.40 (2.80 - 4.08)	3.38 (2.75 – 3.97)	3.44 (2.82 – 4.40)	0.321
MCTQ-SJLrel	1.22 (0.55 – 1.76)	1.13 (0.67 – 1.66)	1.25 (0.50 – 1.98)	0.813
Actigraphy recordings				
Actigraphy (follow up days)	83 (61 - 90)	89 (84 - 92)	56 (47 -69)	<0.001*
Actigraphy (valid days)	69 (46 – 88)	84 (68 - 90)	47 (43 - 62)	<0.001*

$^\#$ obtained using Wilcoxon rank-sum test between the obesitology subset and the control subset.
*denotes significant differences between the obesitology and control set at $p < 0.05$
$^+$ Q1 and Q3 represent the first and third quartile

5. Robustness of Actigraphic Features to Missing Data

5.1. Introduction

Alteration of circadian rhythm is commonly reported as a risk factor and symptom of several psychiatric disorders; for example, BD is directly connected with vulnerable and unstable circadian rhythm. Still, it is uncertain whether dysregulation is the cause that provides vulnerability for the development of BD or whether rhythm dysregulation is one of the symptoms (Alloy *et al.*, 2017; Walker *et al.*, 2020). Other mental disorders which are associated with vulnerability to (or considered a cause of) disruption of circadian rhythm are anxiety disorders, major depressive disorder, and schizophrenia. The disruptions of sleep and circadian rhythm are commonly associated with the severity of disease (major depressive disorder and schizophrenia) or as a risk factor (especially jetlag) (Walker *et al.*, 2020).

Additionally, sleep disturbances, measured as the variation of sleep duration and timing, have also been associated with physical health (i.e., heart condition, diabetes, obesity, and sleep disorders) and stress (Bei *et al.*, 2016). While the state-of-the-art methods for circadian rhythm measurement assess melatonin levels and body core temperature, actigraphy is also widely accepted as a less invasive method (Reid, 2019). Therefore, the use of long-term actigraphy accompanied by online data collection as described in section 3.4 may present a plausible method for monitoring risk factors. It may prove to be a supportive approach for evaluating symptoms severity.

One of the long-term monitoring system downsides is the need for constant checking the data collection to achieve a reasonable ratio of missing data-points. The data are usually missing either due to wearable removal (selected sports, inconvenience of use – typical for psychiatric patients, etc.) or technical difficulties (as may be erroneous data transfer from the wearable to the server). As the actigraphic features were not originally developed with consideration of missing values, we have analysed these basic actigraphic features (sections 3.5.1 and 3.5.2) to see how robust they were against missing data-points in the recording.

The goals of this chapter are:

1. To enumerate the natural fluctuation in individual actigraphic features over a long time evaluated as the long-term temporal variability (LTTV) (**Exp. 1**)
2. To evaluate and explore the errors caused by missing data in actigraphic features estimation - both the estimation error offset (average estimation error) and its variation. (**Exp. 2**)
3. To evaluate the severity (impact) of the estimation errors and the possibility of their correction. (**Exp. 1** & **Exp. 2**)

5.2. Methods

5.2.1. Natural Long-term Variation in Features (Exp. 1)

In order to interpret the meaning of measurement error, its size has to be compared to the average value or range of possible values. Since many actigraphic features don't have a defined range of values, it is necessary to define a meaningful range to which the error can be compared. In order to meaningfully evaluate the estimation error of actigraphic features caused by missing data-points, we compare it to the long-term natural fluctuation of features. Therefore, our first step is to obtain the long-term natural fluctuation of each feature (assessed by long-term temporal variability – LTTV). These fluctuations are helpful not only for the interpretation of estimation errors, but also for interpreting other changes happening in the actigraphic features (see Chapters 7 & 8).

A subset of patients was selected from the ACTIBIPO 2 dataset (section 4.2). The inclusion criterion was to have at least one year (365 days) of valid days. Valid in this context means that there are not more than 10 minutes of missing data-points in one day, or not more than 1 hour of missing data-points in a 7-day window, or not more than 2 hours of missing datapoints in 14-day window.

In these patients, the cosinor and NPCRA features (see sections 3.5.1 & 3.5.2) were calculated for all of the valid days (and windows). Afterwards, temporal variability of each feature for each patient was calculated in the form of (1) standard deviation (SD), (2) interquartile range (IQR) (as some features – mainly the timing features [L5-time, M10-time, Acrophase] - do not follow a normal distribution), and (3) coefficient of variation (CV) (the ratio of SD and MEAN, which is easier to be compared to actograms recorded using a different actigraphy

wearable, see Chapter 7). These values are then referred to as long-term temporal variability ($LTTV_{SD}$, $LTTV_{IQR}$, and $LTTV_{CV}$).

The LTTV may differ considerably between individuals, based on their different lifestyles, movement possibilities, etc. Therefore, for evaluation/interpretation purposes, we have defined three hypothetical patients: 1. the minimal LTTV of each feature is referred to as to the most stable patient, 2. the median LTTV of each feature is referred to as typical patient and 3. the maximal LTTV of each feature is referred to as the most unstable patient.

5.2.2. Estimation Error in Features Based on Missing Data (Exp. 2)

This experiment again uses the ACTIBIPO 2 dataset (see section 4.2). In this case, a patient was included based on the completeness of his/her data. The inclusion criterion was to have a 14-day-long data segment without any missing values. Using these complete segments, we have evaluated the impact of a percentage of missing values. A set of cosinor and NPCRA features (the same as in Exp. 1) was estimated, as described in section 3.5 (3.5.1 and 3.5.2).

In addition to features calculated from the segments – samples ($samp$) – without missing data $F_i(samp, 0)$, where F_i represents individual features (see Table 5-1), and $samp$ represents data from one patient (14-day segments), the features were also calculated from segments with randomly distributed dropped values $F_i(samp, miss)$, where the $miss$ represents percentage of missing values in the feature-specific estimation window (1-day, 7-day, 14-day). The 1-day window features were calculated from the 2nd day of the segment. The 7-day window features were calculated from the 1st week of the segment. The 14-day window features were estimated from the whole segment. The range of missing values $miss$ was 2-60 % (with a 2 % step up to 20 % and with a 5 % step onwards). The estimation error (EE) was obtained as

$$E_{F_\mathrm{i}}(samp, miss) = F_\mathrm{i}(samp, miss) - F_\mathrm{i}(samp, 0) \quad (5.1)$$

The estimation error offset for an individual feature is defined as an average of $E_{F_\mathrm{i}}(samp, miss)$ over all of the samples with fixed $miss$:

$$\bar{E}_{F_\mathrm{i}}(miss) = \frac{1}{N_{samp}} \sum_{samp} E_{F_\mathrm{i}}(samp, miss) \quad (5.2)$$

And the estimation error variability for an individual feature with fixed $miss$ is defined as SD:

$$\tilde{E}_{F_i}(miss) = \sqrt{\frac{1}{N_{samp} - 1} \sum_{samp} \left(E_{F_i}(samp, miss) - \bar{E}_{F_i}(miss) \right)^2}$$

5.3

The data removal procedure was repeated 100 times to reduce the effect of random distribution of missing values.

5.2.3. The Nature of Missing Data-based Features Errors (Exp. 2)

Linear regression models with zero intercepts (Eq. 5.45.2) were trained to represent the average ($\bar{E}_{F_i}(miss)$) and $\tilde{E}_{F_i}(miss)$ for each feature. Further, the model predicted $\widehat{\bar{E}_{F_i}}(miss)$ represents a common (predictable) offset, which may be corrected. The model predicted $\widehat{\tilde{E}_{F_i}}(miss)$ represents a common uncertainty of the feature.

$$\bar{E}_{F_i}(miss) = 0 + \beta_1 \cdot miss \, [+\beta_2 \cdot miss^2] + \varepsilon_{F_i}(miss)$$

5.4

Where $\varepsilon_{F_i}(miss)$ represents residual errors, and the part in the square brackets is optional, based on the type of tested dependency. The model formula for $\tilde{E}_{F_i}(miss)$ is the same as for $\bar{E}_{F_i}(miss)$.

Two dependencies were tested, a linear (Eq. 5.4) and a quadratic (Eq. 5.4, including the square bracket). For significant models, a better dependency was chosen by comparison of adjusted R^2 (Eq. 5.5), where the residual-based R^2 (a coefficient of determination – Eq. 5.6) is adjusted by the number of predictors. The quadratic model is used when it explains data better for more than 5 % using the $R^2_{adjusted}$.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

5.5

In the formula above, N is the total sample size, and p is the number of predictors.

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{miss} \varepsilon_{F_\mathrm{i}}(miss)^2}{\sum_{miss}\left(\bar{E}_{F_\mathrm{i}}(miss) - \overline{\bar{E}_{F_\mathrm{i}}}\right)^2}$$

5.6

In the formula above, SST is the sum of squares total, SSE is the sum of squares error, and $\overline{\bar{E}_{F_\mathrm{i}}}$ stands for the average of estimation errors offsets (Eq. 5.2).

5.2.4. Effect of Aggregation of Missing Values into Blocks (Exp. 2)

The random distribution of missing values is not common in the actigraphy recordings. The data are usually missing in blocks, which correspond most commonly to wearable removal or malfunction.

The individual strength of dependencies was tested by a model combining, changing the percentage of missing values ($miss$), and aggregating these missing values into non-overlapping blocks, as it would be more probable in a real actigraphy recording. For simplification, all blocks were of the same size, and only linear dependency was tested for both number of blocks (n_{block}) and $miss$. In this case, the features were estimated for missing data $miss$ (0 - 60 %). for each $miss$ (except $miss = 0$ %), the missing data-points were arranged into a different number of blocks n_{block} (1 – 50). The procedure was repeated 100 times for each sample to reduce the effect of random distribution of missing values' blocks.

$$\bar{E}_{F_\mathrm{i}}(miss, n_{block}) = 0 + \beta_1 \cdot miss + \beta_2 \cdot n_{block} + \varepsilon_{F_\mathrm{i}}(n_{block}, miss)$$

5.7

The $\bar{E}_{F_\mathrm{i}}(miss, n_{block})$ is obtained similarly as in Eq. 5.4, as average over samples, the model for uncertainty $\tilde{E}_{F_\mathrm{i}}(miss, n_{block})$ was trained in similar manner.

5.3. Results

5.3.1. Natural Long-term Variation in Features (Exp. 1)

Out of 275 patients in the ACTIBIPO 2 dataset, 27 fulfilled the inclusion criteria – i.e. enough data-points to be included in features' natural variability estimation (section 5.2.1). The results are presented in Table 5-1. The typical patient (median) $LTTV_{CV}$ of most features was in the range of 5-20 % (see the fifth column of the table). The largest $LTTV_{CV}$ were observed for L5-time features. The comparison of $LTTV_{SD}$ and $LTTV_{IQR}$ shows that most of the features LTTV follow approximately Gaussian distribution.

Table 5-1: Natural variability of selected features in BD patients

Feature	Estimation window size	$LTTV_{SD}$ median (min, max)	$LTTV_{IQR}$ median (min, max)	$LTTV_{CV}$ (%) median (min, max)
M10	One day	73.12 (47.67, 160.43)	96.39 (57.11, 274,09)	16 (9,29) %
L5		11.32 (3.96, 112.39)	10.03 (4.09, 147.17)	28 (13, 101) %
M10-time (hours)		2.26 (1.06, 4.99)	3.06 (0.57, 6.28)	16 (8, 28) %
L5-time (hour)		1.58 (0.77, 5.33)	2.24 (0.83, 7.93)	60 (28, 872) %
RA		0.055 (0.024, 0.233)	0.055 (0.019, 0.401)	6 (3, 34) %
M10 RMSSD		241.04 (179.85, 366.83)	315.84 (215.45, 461.71)	13 (10, 18) %
$MESOR_7$	Seven days	28.95 (15.36, 57.50)	37.14 (20.06, 110.39)	10 (6, 23) %
$Amplitude_7$		26.88 (18.99, 79.25)	34.25 (21.37, 99.18)	14 (7, 42) %
$Acrophase_7$ (hour)		0.57 (0.28, 5.00)	0.69 (0.35, 4.44)	4 (2, 30) %
CQ_7		0.076 (0.036, 0.182)	0.095 (0.051, 0.275)	10 (4, 45) %
GOF_7		4.08 (2.69, 9.12)	5.17 (3.41, 11.61)	20 (9, 77) %
$M10_7$		43.36 (27.31, 82.19)	55.93 (24.84, 133.43)	11 (6, 18) %
$L5_7$		10.64 (3.89, 79.90)	9.76 (2.75, 122)	22 (10, 45) %
M10-$time_7$ (hour)		1.44 (0.37, 4.85)	1.58 (0.08, 4.42)	10 (3, 29) %
L5-$time_7$ (hour)		0.90 (0.32, 4.51)	1.17 (0.25, 5.25)	34 (13, 133) %
RA_7		0.045 (0.019, 0.176)	0.050 (0.012, 0.279)	5 (2, 41) %
IV_7		0.069 (0.043, 0.111)	0.081 (0.049, 0.164)	14 (11, 22) %
IS_7		0.066 (0.043, 0.146)	0.087 (0.054, 0.210)	12 (6, 50) %
$MESOR_{14}$	Fourteen days	28.03 (13.29, 57.25)	34.70 (15.63, 113.15)	9 (4, 23) %
$Amplitud_{14}e$		24.05 (16.13, 75.01)	29.64 (17.36, 91.92)	12 (7, 46) %
$Acrophase_{14}$ (hour)		0.48 (0.25, 4.51)	0.62 (0.30, 3.91)	3 (2, 28) %
CQ_{14}		0.064 (0.027, 0.171)	0.089 (0.039, 0.317)	9 (3, 44) %
GOF_{14}		3.43 (2.17, 7.59)	4.37 (2.86, 10.30)	17 (7, 78) %
$M10_{14}$		43.12 (21.66, 77.79)	48.74 (24.14, 138.74)	10 (5, 18) %
$L5_{14}$		9.14 (3.83, 70.54)	9.67 (2.60, 125.71)	19 (9, 39) %
M10-$time_{14}$ (hour)		1.10 (0.27, 3.78)	1.33 (0.08, 3.75)	7 (3, 22) %
L5-$time_{14}$ (hour)		0.74 (0.26, 3.93)	1.00 (0.25, 5.17)	26 (12, 90) %
RA_{14}		0.038 (0.014, 0.162)	0.044 (0.011, 0.313)	5 (2, 40) %
IV_{14}		0.059 (0.031, 0.099)	0.071 (0.039, 0.140)	12 (8, 18) %
IS_{14}		0.060 (0.037, 0.125)	0.074 (0.047, 0.202)	11 (5, 56) %

SD- standard deviation, IQR- inter-quintile range, CV- coefficient of variance

5.3.2. Features Estimation Error and its Variation (Exp. 2)

Out of 275 patients, 112 patients (segments) were included in the set, estimating the effect of missing data on the estimation error for individual features. The rest 163 patients were excluded for not having a 14-day long data segment without missing values.

The Offsets of Estimation Error Based on Missing Data-points

The dependency of estimation error (EE) offsets (Eq. 5.2) on the amount of missing data-points may be divided into three types (see examples in Figure 5.1):

1) Zero (no common change with missing values): cosinor **Acrophase** and **MESOR**.
2) Linear (usually with a tiny slope): non-parametric **M10-time**, **L5-time**, **M10$_{7,14}$**, and cosinor **CQ$_{7,14}$**, and **GOF$_{7,14}$**
3) Nonlinear: non-parametric **L5**, **RMSSD$_{M10}$**, **RA**, **IV$_{7,14}$**, **IS$_{7,14}$** and **M10**

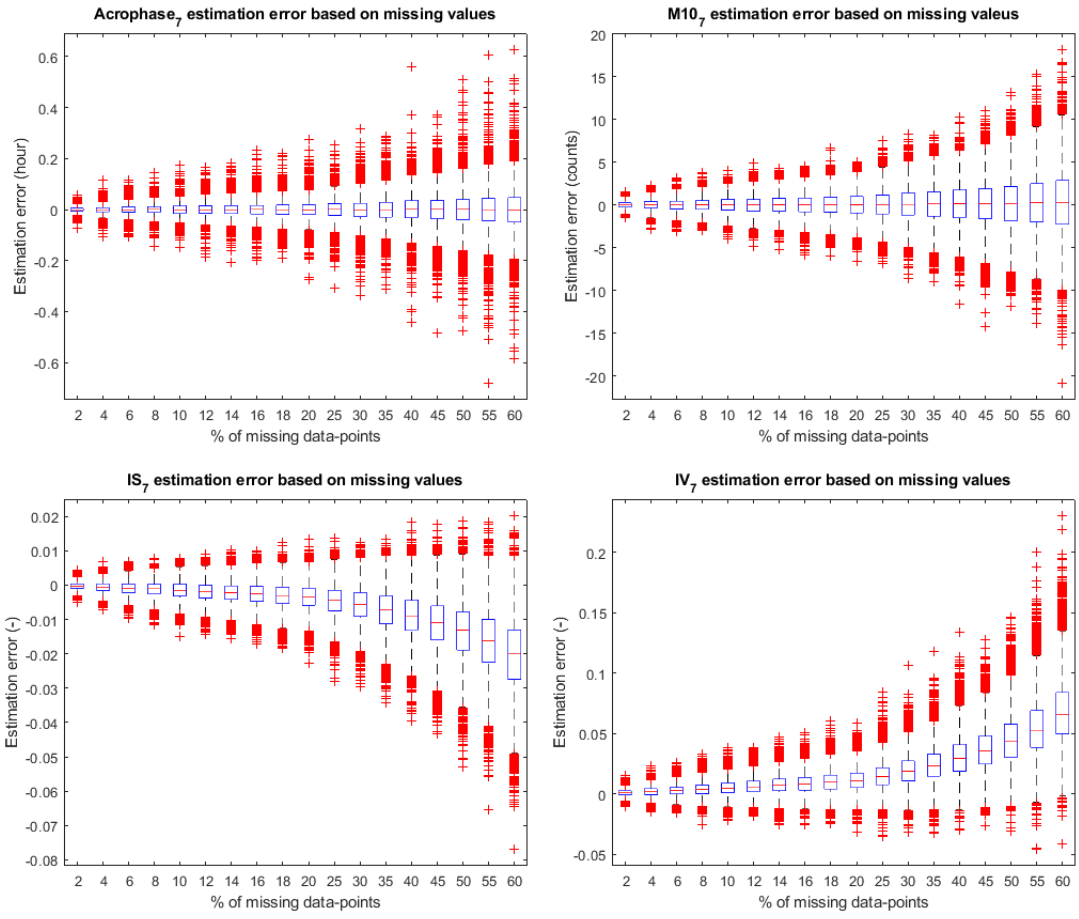


Figure 5.1 - Typical distributions of estimation errors based on samples with a specific amount of missing values. On the left top, a distribution of errors in Acrophase$_7$ a feature, which doesn't systematically change with an increasing number of missing values and only the estimation uncertainty (variation) grows. On the top right is an M10$_7$ feature, where there is a small offset of EE, which is negligible compared to the error variation. At the bottom, there are distributions of errors in IS and IV. There is a considerable offset of EE in these features, which is nonlinearly dependant on the amount of missing values.

For features where a linear common trend of estimation error offset was observed (significant β coefficient; Eq. 5.4– see Table 5-2), its slope was small compared to $LTTV_{SD}$ (as in $M10_7$). The detailed results for the fitted models - coefficients, their significance, and effect sizes - are presented in Table 5-2. The table also presents EE offsets estimated (by models) and measured (from data) for 20 % of missing data-points, both randomly distributed and aggregated into 4 blocks.

5.3.3. Features Estimation Error - Blocks of Missing Values

The arrangement of the missing values into blocks affected the EE of the features. In most features, the aggregation of missing values into blocks affected the variation of EE. In some features (see Figure 5.2.), it also affected the EE offsets.

The way how the aggregation of missing values into blocks affects the EE variation may be divided into three categories:

1. The variation of EE increases with the amount of missing data-points and lowers with the number of blocks into which are the missing data-points divided. Most of the features belong to this category, namely the one-day-based non-parametric features (**M10**, **L5**, **M10-time**, **L5-time**, **RA**), the cosinor **MESOR$_{7,14}$,** and the multiple-days-based non-parametric features (**M10$_{7,14}$**, **L5$_{7,14}$**, **RA$_{7,14}$**).
2. The variation of EE changes in the same way as in the first category, only in this case, the variation is low when the missing data-points are in 1 or 2 blocks. The cosinor **Amplitude$_{7,14}$**, **CQ$_{7,14}$**, **GOF$_{7,14}$**, and **Acrophase$_{7,14}$** belong to this category, as well as the multiple-day-based nonparametric **M10-time$_{7,14}$**, **L5-time$_{7,14}$**, and **IS$_{7,14}$**. The change is small, compared to $LTTV_{SD}$, in the timing features, especially in cosinor **Acrophase$_{7,14}$**, **L5-time$_{14}$**, and **M10-time$_{14}$**.
3. Both variation and offset of the EE change nonlinearly with the number of blocks. The features that describe the instability in the daily activities **RMSSD$_{M10}$** and **IV$_{7,14}$** belong to this category. The EE variation in the **RMSSD$_{M10}$** is lowest when the data are missing in 2-6 blocks, for other numbers of blocks it is higher. In the **IV$_{7,14}$** features, the variation is low for 1 block, it increases slightly for 2-6 blocks, and decreases for 10-20 blocks, and eventually increases rapidly for a large number (30+) of blocks. In both **RMSSD$_{M10}$** and **IV$_{7,14}$** features, the highest EE variation is reached for a large percentage of missing data-points.

The variation of EE observed for a different amount of missing values divided into blocks, and its comparison to natural long-term variability in the features are presented in Table 5-3. The predictions of EE variation $\widehat{\tilde{E}_{F_1}}$ estimated from models Eq. 5.7 are presented in the supplement Table S-1.

Although the EE offsets (Eq. 5.4) are affected by the number of blocks in some features, these offsets are substantially (by orders of magnitude) smaller than the $LTTV_{SD}$ for these features. The only exception is the **$RMSSD_{M10}$**. Concerning the EE offsets, the effects of the aggregation of missing values into blocks (Figure 5.2) may be divided into four categories:

1. Offsets are neither affected by the amount of missing data-points, nor by the number of blocks. These include the daily activity peaks and troughs timings features (cosinor **$Acrophase_{7,14}$**, non-parametric **$M10\text{-}time_{1,7,14}$, $L5\text{-}time_{1,7,14}$** both one- and multiple-day-based) and the cosinor **$MESOR_{7,14}$**.
2. Offsets are not affected by the amount of missing data-points, but they change with the number of blocks. The **L5** feature and **$IV_{7,14}$** are the representatives of this category. In **L5**, the result is not affected by a small number of blocks, but it is slightly affected if there are more blocks (8+). In **$IV_{7,14}$**, the effect is nonlinear, but the EE offset decreases with the number of blocks in general.
3. Offsets are affected by the amount of missing data-points, but they are not affected by their arrangement into blocks. The **$M10_{7,14}$** and **$L5_{7,14}$** features belong to this category. EE offset increases with the amount of missing data-points for **$M10_{7,14}$** and decrease for **$L5_{7,14}$**.
4. Offsets are affected by both the amount of missing data-points, and by the number of blocks. They grow with the increasing amount of missing data-points and decrease with the number of blocks. The non-parametric **M10** and **RA, $IS_{7,14}$,** and cosinor **$Amplitude_{7,14}$**, **$CQ_{7,14}$** and **$GOF_{7,14}$** belong to this category. In **$IS_{7,14}$**, the change based on the number of blocks is so pronounced, that it leads to decreased instead of increased values for a high number of blocks. For **$RMSSD_{M10}$**, the value decreases for a few blocks and increases for many blocks. The relation is not linear, as shown in Figure 5.2.

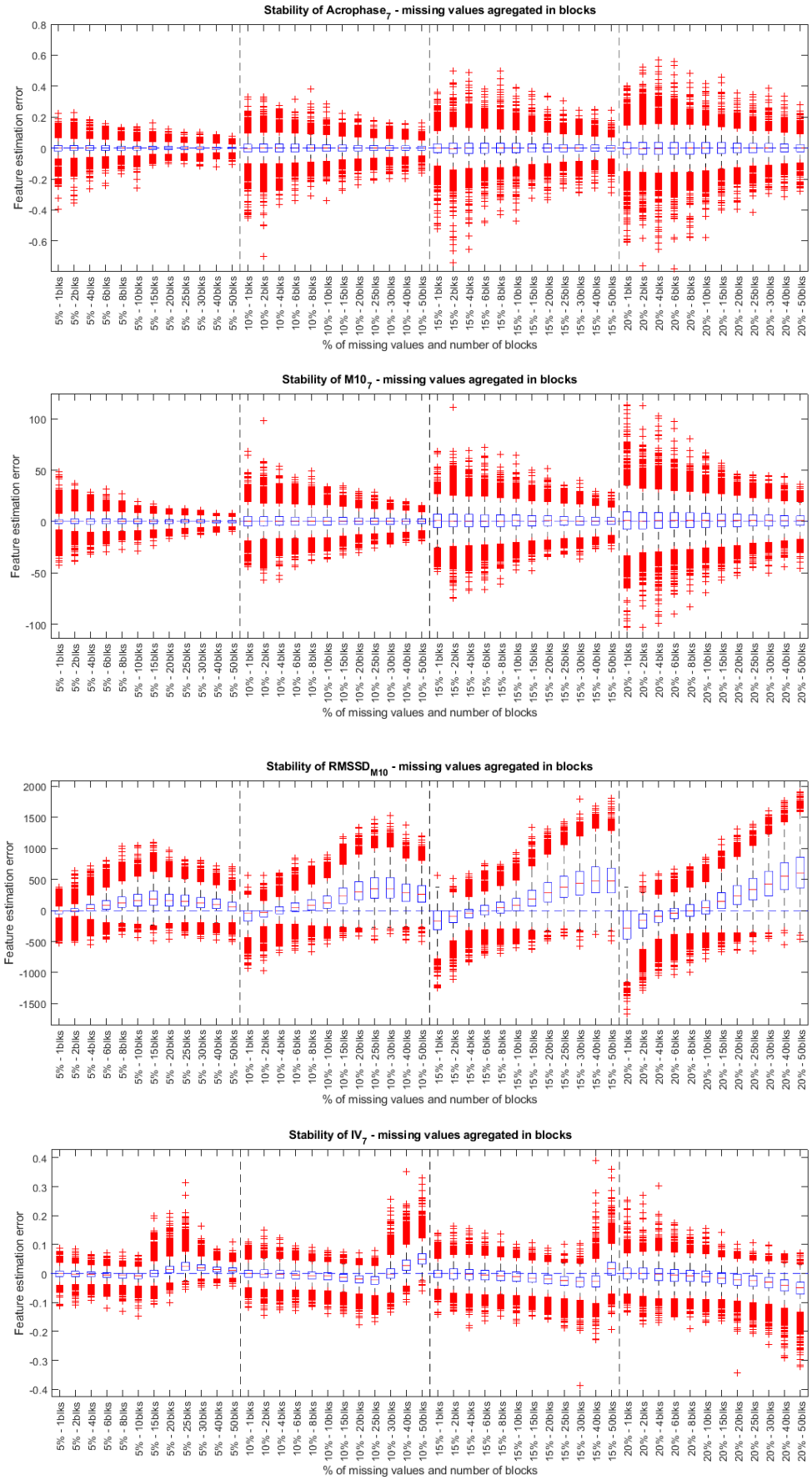


Figure 5.2 - Estimation error in features based on the amount of missing data-points distributed into blocks. The first two graphs show Acrophase and M10 features, where the estimation error variation drops with an increasing number of blocks. The last two graphs show $RMSSD_{M10}$ and IV_7 features, that have non-linear relation between EE variation and number of blocks.

Table 5-2: Features stability modelled for data with missing values (estimation error offset)

Feature	Window size	Miss model				Miss and Block model			Effect of *miss* = 20 % compared to natural $LTTV_{SD}$	
		Quad-lin	Coeff β_1	Coeff β_2	R^2 adjusted	Coeff miss β_1	Coeff block β_2	R^2 adjusted	Effect Miss model⁺ model $\widehat{\bar{E}_{F_i}}$ (measured $\bar{E}_{F_i}$)	Effect Miss and Block model‡ for n_{block} = 4 model $\widehat{\bar{E}_{F_i}}$ (measured $\bar{E}_{F_i}$)
M10	One day	Quadratic	4.52E-4*	4.31E-4***	0.9859	0.0281	0.0291	0.0064	0.2630 (0.3066)	0.6790 (3.7056)
L5		Quadratic	-0.0139***	-0.0004***	0.9992	-0.0076	-0.0218	0.0296	-0.4469 (-0.5021)	-0.2393 **(-1.5025)**
M10-time		Linear	1.25E-3***	-	0.8884	-9.97E-4	5.97E-4	0.0163	0.0251 (0.0303)	-0.0175 (-0.1000)
L5-time		Linear	1.27E-3***	-	0.6902	0.0046***	-7.88E-4**	0.4931	0.0255 (0.0511)	0.0884 (0.1177)
RA		Quadratic	5.93E-5***	1.9E-6***	0.9989	5.21E-5	9.51E-5*	0.0214	1.94 E-3 (2.18E-3)	1.42E-03 (**7.41E-3**)
RMSSD$_{M10}$		Quadratic	-3.591***	-0.077***	0.9998	-6.1632***	11.7286***	0.5989	**-102.58 (-103.92)**	**-76.3493 (-108.81)**
MESOR$_7$	Seven days	Linear	-7.52E-5	-	0.0369	-0.0055***	0.0018***	0.3927	-1.50 E-3 (2.88E-3)	-0.1023 (-0.0549)
Amplitude$_7$		Linear	3.97E-4**	-	0.2704	0.0468***	-0.0152***	0.6968	7.94E-3 (-1.52E-3)	0.8745 (0.7163)
Acrophase$_7$		Linear	-1.71E-6	-	0.0104	1.01E-5*	2.55E-6	0.0245	-3.40E-05 (3.57E-4)	2.13E-04 (-1.38E-3)
CQ$_7$		Linear	1.88E-6**	-	0.3096	1.91E-4***	-5.99E-5***	0.7255	3.76E-05 (-4.91E-6)	3.58E-03 (2.93E-3)
GOF$_7$		Linear	1.75E-4***	-	0.5625	0.0085***	-0.0024***	0.8469	3.50 E-3 (1.11E-4)	0.1614 (0.0964)
M10$_7$		Linear	4.26E-3***	-	0.8395	0.0996***	-0.0184***	0.8534	0.0852 (0.0498)	1.9190 (1.2200)
L5$_7$		Quadratic	-5.4E-3*	-9.2E-5***	0.9852	-0.0737***	0.0174***	0.8827	-0.0475 (-0.0661)	**-1.4054 (-0.9700)**
M10-time$_7$		Linear	8.49E-4***	-	0.9269	0.0017***	8.96E-4***	0.4558	0.0170 (0.0167)	0.0382 (**0.0775**)
L5-time$_7$		Linear	-1.3E-4**	-	0.3761	-0.0014***	1.29E-4*	0.6596	-2.57 E-3 (-9.46E-3)	-0.0270 (-0.0282)
RA$_7$		Quadratic	2.94E-6*	4.09E-7***	0.9910	2.99E-4***	-6.55E-5***	0.8908	2.22 E-4 (2.85E-4)	**5.73E-03 (4.06E-3)**
IV$_7$		Quadratic	2.58E-4***	1.37E-5***	0.9949	-5.29E-4***	-5.80E-5	0.1568	**0.010628 (0.0115)**	**-0.0108** (-3.11E-3)
IS$_7$		Quadratic	-7.5E-5***	-4.1E-6***	0.9947	0.0012***	-6.74E-4***	0.7660	-3.14 E-3 (-3.40E-3)	**0.0222 (0.0173)**
MESOR$_{14}$	Fourteen days	Linear	5.5E-5	-	0.0186	-0.0054***	0.0022***	0.4338	1.10 E-3 (-8.03E-3)	-0.0995 (-0.1029)
Amplitude$_{14}$		Linear	3.68E-4***	-	0.3733	0.0256***	-0.0058***	0.7217	7.35 E-3 (9.80E-3)	0.4882 (0.4112)
Acrophase$_{14}$		Linear	-1.97E-6	-	0.0080	1.81E-4***	-3.83E-5***	0.6764	-3.90E-05 (-9.55E-6)	3.47E-03 (7.51E-3)
CQ$_{14}$		Linear	1.26E-6***	-	0.4040	1.18E-4***	-2.86E-5***	0.8023	2.51E-05 (6.37E-5)	2.24E-03 (1.95E-3)
GOF$_{14}$		Linear	9.41E-5***	-	0.7429	0.0052***	-0.0012***	0.8683	1.88 E-3 (1.93E-3)	0.0996 (0.0880)
M10$_{14}$		Linear	2.524E-3***	-	0.8016	0.0491***	-0.0050***	0.8234	0.050481 (0.0324)	0.9628 (0.6981)
L5$_{14}$		Quadratic	-7.5E-4*	-5.24E-5***	0.9794	-0.0454***	0.0087***	0.8881	-0.0359 (-0.0362)	**-0.8741** (-0.5635)
M10-time$_{14}$		Linear	-1.6E-4*	-	0.3257	-0.0026***	-485E-4***	0.5879	-3.23 E-3 (5.53E-3)	-0.0531 (**0.0910**)
L5-time$_{14}$		Linear	-6.662E-5***	-	0.4053	-8.01E-4***	5.44E-5*	0.7675	-1.32 E-3 (-3.68E-3)	-0.0158 (0.0242)
RA$_{14}$		Quadratic	2.71E-6*	2.43E-7***	0.9794	2.0E-4***	-354E-5***	0.9065	1.51E-4 (1.54E-04)	**3.87E-03** (2.70E-3)
IV$_{14}$		Quadratic	2.55E-4***	1.37E-5***	0.9954	-1.69E-4***	-2.97E-4***	0.2907	**0.0106 (0.0115)**	-4.57E-03 (-1.68E-3)
IS$_{14}$		Quadratic	-7.3E-5***	-4.0E-6***	0.9953	5.87E-4***	-2.93E-4***	0.8433	-3.04E-3 (-3.27E-3)	**0.0106 (7.42E-3)**

Coefficient significance * < 0.05, ** < 0.001, *** < 0.0001 (*t*-test) for mean estimation error (EE) offset in the 100 repetitions;

⁺Miss model (Eq. 5.4) and ‡Miss and Block model (Eq. 5.7) – model prediction (measured from data)

The **bold** effect value indicates where the EE offset reached 20+ % of natural $LTTV_{SD}$ of the most stable patient (Table 5-1); the **bold** shaded effect values indicate where the EE offset reached 40+ % of natural $LTTV_{SD}$.

Table 5-3: Reliability of feature estimation – The variation of estimation error calculated for data missing in blocks

Feature	Window size	Variation of estimation error $\tilde{E}_{F_i}$ for selected percentages of missing values aggregated into a specific number of blocks											
		5 % 4 blocks	5 % 6 blocks	5 % 10blocks	10 % 4 blocks	10 % 6 blocks	10 % 10blocks	15 % 4 blocks	15 % 6 blocks	15 % 10blocks	20 % 4 blocks	20 % 6 blocks	20 % 10blocks
M10	One day	6.4451	5.6544	4.6821	10.4582	9.2703	7.6894	**16.3404**	14.2394	12.0758	**20.8263**	**18.2279**	**15.1478**
L5		**2.3553**	1.9477	1.8018	**3.3251**	**3.0734**	**2.8249**	**4.9638**	**4.5527**	**4.0800**	**5.8316**	**5.5461**	**5.1461**
M10-time		**0.6152**	**0.5748**	**0.5226**	**0.8320**	**0.7822**	**0.7237**	**1.1348**	**1.0238**	**0.9029**	**1.4102**	**1.2750**	**1.1082**
L5-time		**0.7972**	**0.8056**	**0.7709**	**1.0218**	**1.0290**	**1.0162**	**1.3241**	**1.2464**	**1.1842**	**1.4795**	**1.3875**	**1.3571**
RA		0.0095	0.0081	0.0071	**0.0136**	**0.0127**	**0.0114**	**0.0198**	**0.0185**	**0.0163**	**0.0243**	**0.0227**	**0.0206**
RMSSD$_{M10}$		**98.7593**	**118.6427**	**148.1665**	**108.6813**	**122.6964**	**154.9828**	**133.6587**	**140.2980**	**162.9907**	**154.1249**	**154.0750**	**174.6695**
MESOR$_7$	Seven days	3.2306	2.8427	2.3790	5.0467	4.5527	3.9527	**7.3471**	**6.8417**	**6.0142**	**8.8041**	**8.6030**	**7.5478**
Amplitude$_7$		4.0984	3.6207	3.0366	**6.4764**	**5.8572**	5.0503	**9.2443**	**8.7148**	**7.7017**	**11.6401**	**10.8146**	**9.5127**
Acrophase$_7$		0.0276	0.0247	0.0210	0.0440	0.0391	0.0342	0.0626	0.0574	0.0522	0.0922	0.0726	0.0657
CQ$_7$		0.0140	0.0125	0.0105	**0.0222**	**0.0204**	**0.0175**	**0.0316**	**0.0300**	**0.0269**	**0.0392**	**0.0370**	**0.0333**
GOF$_7$		0.7612	0.6673	0.5501	**1.2140**	**1.0921**	**0.9312**	**1.7340**	**1.6310**	**1.4519**	**2.1058**	**2.0272**	**1.7839**
M10$_7$		5.0353	4.4857	3.7890	7.7898	7.1615	6.2032	**11.3615**	**10.5216**	**9.3287**	**14.2277**	**13.1622**	**11.7217**
L5$_7$		**3.0739**	**2.9253**	**2.3321**	**5.0621**	**4.6107**	**3.8743**	**7.4872**	**7.0952**	**6.2581**	**8.9633**	**8.4383**	**7.6274**
M10-time$_7$		**0.6242**	**0.6185**	**0.5623**	**0.8086**	**0.7559**	**0.7418**	**1.0160**	**0.9268**	**0.8837**	**1.1176**	**1.0428**	**1.0107**
L5-time$_7$		**0.3050**	**0.3002**	**0.2705**	**0.4108**	**0.4033**	**0.3495**	**0.5339**	**0.5267**	**0.4948**	**0.6105**	**0.5907**	**0.5965**
RA$_7$		**9.21E-03**	8.73E-03	7.35E-03	**0.0150**	**0.0136**	**0.0117**	**0.0216**	**0.0205**	**0.0182**	**0.0265**	**0.0245**	**0.0225**
IV$_7$		**0.0139**	0.0136	**0.0143**	**0.0199**	**0.0191**	**0.0187**	**0.0279**	**0.0266**	**0.0255**	**0.0332**	**0.0323**	**0.0307**
IS$_7$		0.0125	0.0124	**0.0109**	**0.0192**	**0.0178**	**0.0169**	**0.0285**	**0.0266**	**0.0238**	**0.0356**	**0.0332**	**0.0302**
MESOR$_{14}$	Fourteen days	2.6763	2.4748	2.1048	3.9480	3.7324	3.3970	5.4080	5.2772	5.0410	**6.6548**	**6.2923**	**6.1983**
Amplitude$_{14}$		3.4327	3.1835	2.7002	**5.0967**	4.8520	**4.4257**	**7.0343**	**6.9386**	**6.3817**	**8.0654**	**8.3904**	**7.9514**
Acrophase$_{14}$		0.0320	0.0297	0.0276	**0.1091**	0.1109	**0.0399**	**0.1135**	**0.1271**	**0.1400**	**0.1955**	**0.1709**	**0.1755**
CQ$_{14}$		0.0120	0.0110	9.39E-03	**0.0175**	**0.0169**	**0.0153**	**0.0234**	**0.0238**	**0.0221**	**0.0269**	**0.0283**	**0.0275**
GOF$_{14}$		0.6454	0.5869	0.4978	**0.9254**	**0.8982**	**0.8124**	**1.1988**	**1.2417**	**1.1906**	**1.3620**	**1.4534**	**1.4533**
M10$_{14}$		4.0821	3.7806	3.2505	6.1736	5.7225	5.2236	**8.7670**	8.4095	7.6363	**10.4815**	**10.1856**	**9.4741**
L5$_{14}$		**2.5916**	**2.3622**	**1.9939**	**3.9294**	**3.5413**	**3.2838**	**5.4023**	**4.9758**	**4.7303**	**6.2848**	**6.2820**	**6.0266**
M10-time$_{14}$		**1.0136**	**1.0096**	**0.9996**	**1.1867**	**1.2089**	**1.1173**	**1.3039**	**1.2180**	**1.2211**	**1.3441**	**1.2837**	**1.3041**
L5-time$_{14}$		**0.2748**	**0.2756**	**0.2580**	**0.3853**	**0.3697**	**0.3411**	**0.4662**	**0.4188**	**0.4408**	**0.6213**	**0.6107**	**0.5730**
RA$_{14}$		**8.38E-03**	**7.64E-03**	6.59E-03	**0.0122**	**0.0116**	**0.0105**	**0.0174**	**0.0166**	**0.0156**	**0.0198**	**0.0202**	**0.0193**
IV$_{14}$		0.0105	0.0101	9.80E-03	**0.0150**	**0.0144**	**0.0139**	**0.0195**	**0.0196**	**0.0193**	**0.0225**	**0.0229**	**0.0231**
IS$_{14}$		9.44E-03	8.85E-03	8.38E-03	**0.0141**	**0.0131**	**0.0123**	**0.0195**	**0.0193**	**0.0177**	**0.0230**	**0.0234**	**0.0224**

Bold red shaded text indicates the missing data-points setting, where the standard deviation (variation) of estimation error (EE) reaches 40+ % of the natural $LTTV_{SD}$ of a feature for the typical patient (Table 5-1).
Bold text indicates the setting where the EE variation reaches 20+ % of the natural LTTV.

5.4. Discussion

The natural $LTTV_{CV}$ in features presented as percentage (see Table 5-1) is relatively low, about 5-10 % in most of the feature, in spite of the fact that these results are obtained from BD patients where the overall variability in circadian rhythm (including episodes) is expected higher (Alloy *et al.*, 2017). The variability is higher for the L5-time feature describing the daily sleep, where the average value is small, and therefore even small changes represent larger percentual change.

The long-term natural variability of circadian rhythm is associated with photoperiodic regulation and circannual clock. In a healthy population, it is known that phase advances in response to bright light in the morning and that sleep timing changes (even with suppressed the photoperiodic time clues) throughout year seasons (phase delay in spring and summer, and phase advance in winter and autumn) (Honma *et al.*, 1992). In BD patients, the phase changes (M10-time, Acrophase, and L5-time) are expected higher because clinical episodes and their severity are known to affect circadian rhythm and sleep stability (Alloy *et al.*, 2017; Schneider *et al.*, 2020; Walker *et al.*, 2020).

The increased instability of circadian rhythm in BD patients may be the reason why we observe large intra-individual differences in our sample (Table 5-1). The natural LTTV in features, which describe physical activity levels (i.e. M10, MESOR, Amplitude), was approximately 5-6 times higher in the most unstable patient compared to the most stable patient. Concerning the **daily regime stability** (timings of sleep and activity throughout days - i.e. M10-time, L5-time and Acrophase), the most stable patient had LTTV of M10-time and L5-time around 20 minutes for both 7-day and 14-day based features (for 1-day based features, the LTTV was about 1 hour). The most unstable patient had the LTTV in these features about 4-5 hours (therefore 5-15 times higher). The typical patient had LTTV of physical activity features only 2 times higher than the most stable patient. His daily regime stability was about 1 hour for 7-day and 14-day based features (and approximately 2 hours for 1-day based features). Honma *et al.* observed a sleep shift between summer and winter of about 2.5 hours. This shift corresponds to the LTTV in L5-time for the typical patient (slightly less than 1 hour when assessed by the SD metric and slightly more when assessed by the IQR metric).

In the features where there were common offsets of estimation errors (EE) (significant model coefficients) based on the percentage of missing data, these offsets were mostly very small. When there were linear relations between missing data-points and EE offsets, these offsets for 20 % of missing data-points didn't reach 1 % of the features' natural LTTV for the typical patient. The only features where the EE offset was comparable to their natural LTTV were the $RMSSD_{M10}$, IV and IS features (43 %, 19 %, and 5 % of their natural LTTV). The correction of the EE offset in these features would be complicated because: (1) the relation between missing data and offset of EE is not linear (in both global model and individual models), and (2) it is strongly affected by the arrangement of missing data-points into blocks. Additionally, for two of them ($RMSSD_{M10}$, and IV), the random distribution of samples represents the most extreme case. The nonlinearity of the relationship is probably based on the resampling of the data for estimation of these features (20-minute segments for IV and IS, and 5-minute for $RMSSD_{M10}$), where the missing data blocks of a certain length are more harmful for the feature estimation. Similar behaviour was not observed in $M10_{7,14}$ and $L5_{7,14}$ features, where the data were also resampled into 5-minute segments. Unlike in $RMSSD_{M10}$, the data in $M10_{7,14}$ and $L5_{7,14}$ are averages over several days, which probably reduces the effect caused by the length of missing data blocks. The correction may be possible when both effects of the amount of missing data-points and size of blocks are considered, but this was beyond the scope of this analysis, as only blocks of the same length were tested.

On the other hand, the variation of EE changes considerably with the increasing amount of missing data-points. As was expected, the parametric (cosinor) analysis was more resistant to missing data-points, especially the Acrophase feature (where even with 20 % of missing data-points, the EE variation did not reach 20 % of its natural LTTV). From the non-parametric features, the most stable was the M10 feature. The high stability of M10 could be caused by high levels of its LTTV in BD patients, where higher fluctuation of activity is a symptom of the disease (Schneider *et al.*, 2020). In other non-parametric features, many were considerably (20 % of their typical patient's LTTV) affected already at 5 % level of missing values, especially when the data were missing in a single or few blocks. Unfortunately, as wearable removal is the most common cause of missing data-points (for example, for hygiene or sport), the data are typically missing in one or a few blocks. While the wear-off for evening hygiene or even the whole night would affect the feature's EE less, as it is usually connected with low levels of physical activity, the removal for sports would considerably affect the features, mainly those based on 1-day data. The rhythm stability describing features IS and IV are more

robust. Still, the stability (measured as EE variation compared to typical patient's LTTV) drops already for 10 % of missing data-points in the sample. The EE variation could be accurately predicted using regression models (compare the EE variation - Table 5-3 and their estimations Table S-1 in the supplement). The comparison reveals that the model predictions of EE variation are more benevolent for smaller (5-10 %) amounts of missing data-points.

The correction of EE offset is complicated because, in features where EE caused by missing data-points are consistent and stable (predictable by the models), the EE offsets are typically quite small (around 1 % of their LTTV). And therefore, their correction would have a negligible effect on the results. On the other hand, in features where the EE offset caused by missing data-points is large (20 % of their LTTV), it is also largely variable and therefore not predictable by the used models. More complex models (which are, however, beyond the scope of this study) would have to be trained, using simulated data with variable length of missing data segments (blocks) in order to correct these EE offsets. A design of such models could be beneficial because these features ($RMSSD_{M10}$, and IV) represent descriptors of rhythm fragmentation, which is connected to cognitive and motor performance (Gonçalves *et al.*, 2015).

5.5. Limitations

The presented results need to be interpreted while considering some limitations:

Firstly, the features' LTTV may be possibly different in the healthy population than in the BD patients, whose data are presented here. Some aspects, such as medication used, or clinical episodes, may alter and possibly increase the features' LTTV (Schneider *et al.*, 2020). On the other hand, many BD patients have a free daily regime (due to disability pension), which may decrease the features' LTTV (Schneider *et al.*, 2020). Though the use of a representative patient should suppress these extremes, it still may cause some interpretation issues.

Secondly, the arrangement of missing data-points into blocks of the same size is not something that is happening in real-world scenarios. This may affect the features that use resampled data, as the resampling may be susceptible to the specific block length. This would not happen in real-world scenarios, as the length of segments would be variable.

Thirdly, the random distribution of these blocks is also a considerable simplification. In real-world settings, the wearable is seldom removed during sleep, unless it was removed

before bedtime. Therefore, the effects on night-time based features (L5 and L5-time) are probably exaggerated in our results. In contrast, the effect on daily activity features (Amplitude, M10, etc.) may be underrated, as the probability of wearable removal is higher during the daytime.

5.6. Conclusion

There are large interpatient differences in LTTV of actigraphic features, even when they are obtained from highly reliable (complete) recorded data. The uncertainty introduced into a sample by segments of missing values reaches relatively high levels (20+ % LTTV) already for small amounts of missing data-points (5-10 %). The largest estimation error variation (uncertainty) was usually reached when the missing data were in a single or only a few blocks. The only exceptions were features describing the variability in daily activities patterns ($RMSSD_{M10}$, and IV). In these features, the variation of estimation error behaved highly irregular and reached a maximum for high number of blocks. The cosinor-based features are more robust to missing data (especially Acrophase) than the non-parametric features. The difference is not as large (approx. 5-10 %), because for more than 15 % of missing data-points, the estimation error variation is considerably large (20+ % LTTV) for most of the features. A typical consistent offset of the estimation errors, observed in some features, is negligible (< 1 %), and therefore it doesn't need to be corrected. In features where the estimation error offset reaches high values, comparable to its LTTV, it is dependent nonlinearly on both the percentage of missing data and the number of blocks (potentially on the block size). Therefore, it cannot be corrected be by the methods proposed in this chapter.

6. Objectivisation of Chronotype Estimation Through Actigraphy

In this chapter, we are re-using data from our impacted journal article Fárková, E., **Schneider, J.** *et al.* (2019) 'Weight loss in conservative treatment of obesity in women is associated with physical activity and circadian phenotype: A longitudinal observational study', *BioPsychoSocial Medicine*, 13(1), pp. 1–10. doi: 10.1186/s13030-019-0163-2. In this paper, we have used the selected actigraphic features to determine how the circadian rhythm is associated with weight reduction or gain during a weight reduction programme.

The results presented in this chapter are about to be submitted as a journal article **Schneider, J.,** Fárková, E., Bakštein, E. (2021) 'Chronotyping Objectivisation Through Wrist-Worn Actigraphy'.

6.1. Introduction

Since the 1980s, physical activity has been a standard and established marker of circadian rhythms, which is also helpful for exploring rhythm disturbances. More precisely, human physical activity (measured using wrist-worn actigraphy) shows an individual circadian pattern in the individual's private environment, which may be beneficial for clinical practice or long-term studies (Portaluppi, Smolensky and Touitou, 2010; Smith *et al.*, 2018). Moreover, actigraphy was reported to be an accurate estimation of sleep onset and offset times (Kaplan *et al.*, 2012; Kosmadopoulos *et al.*, 2014).

The long-duration recordings allowed us to produce plots to see changes in circadian rhythms over time. From this perspective, actigraphy is a very convenient method for studying circadian rhythms in humans (Ancoli-Israel *et al.*, 2003; Portaluppi, Smolensky and Touitou, 2010; Calogiuri, Weydahl and Roveda, 2011).

The popularity of actigraphy in chronobiology stems from its apparent benefits: it is cheap, easy to use, reliable, and provides objective data about individuals' daily routines (Portaluppi, Smolensky and Touitou, 2010). Over the past years, the circadian system came into the research spotlight for its wide implications for overall health (Abbott, Malkani and Zee, 2020).

Although actigraphy has been an established means to measure physical activity and sleep, there is no consensus on the use of actigraphy for determining the circadian phenotypes: a set of subjectively defined variables, such as a rate of social jetlag or a chronotype (individual circadian preference). A variety of methods for analysing circadian aspects of activity data show promising results (Gupta and Pati, 1994; Gonçalves *et al.*, 2014, 2015). However, the quality and range of evidence-based methodology do not seem to be fully in line with the level of popularity of actigraphy as a research method. We thus believe further detailed methodology-oriented research is needed.

6.1.1. Actigraphy-based Circadian Parameters

With respect to chronobiology, existing studies provide evidence that wrist activity covaries with the phase of melatonin secretion (Ancoli-Israel *et al.*, 2003), core body temperature (Ancoli-Israel *et al.*, 2003), oral temperature (Gupta and Pati, 1994), and heart rate (Gupta and Pati, 1994). Midpoints of sleep are significantly correlated with the dim light melatonin onset phase in adolescents, indicating that the middle of sleep may be a useful circadian phase marker (Crowley *et al.*, 2006, 2016).

6.1.2. Subjective Chronotype and Actigraphy

Several studies have already explored the proper functioning of the subjective tool, the chronotype estimating (chronotyping) questionnaires (i.e. Munich Chronotype Questionnaire (MCTQ) and Morningness-Eveningness Questionnaire (MEQ)), employing selected actigraphic parameters and vice versa.

Gershon and her team (2018) replaced the subjective determination of chronotype with objective actigraphy-based determination. They found that subjective and objective chronotypes correlate significantly with each other (Gershon *et al.*, 2018; Kaufmann *et al.*, 2018).

Additionally, a significant negative association was observed between Korean MEQ score and activity Acrophase timing in the Korean study's validation subset. The lower the MEQ score was (towards eveningness), the more delayed was the activity Acrophase. Furthermore, Lee *et al.* (2014) found that the mean activity Acrophase of the Evening types (E-type) group was nearly two hours later than that of the Morning types (M-type) group. This difference was even more significant on free days than on workdays. (Lee *et al.*, 2014)

Vitale *et al.* (2015) also observed a significant difference in the Acrophase: The M-types have shown an early Acrophase (14:32 h) compared with both other types - the Neither types (N-types) (15:42 h) and the E-types (16:53 h).

Roenneberg, Wirz-Justice and Merrow (2003) described a phase shift of sleep to later phases during free days using the MCTQ questionnaires. Actigraphy shows a strong correlation with the MCTQ (Santisteban, Brown and Gruber, 2018). The corrected sleep mid-time (MSFsc) on free days as measured by the MCTQ was not significantly different from the actigraphy based sleep mid-time (mid-sleep) (Santisteban, Brown and Gruber, 2018). Furthermore, an actigraphy study by Lehnkering *et al.* (2006) supported the Roennebergs' findings concerning the sleep phase.

Lee and his team (2014) found a significant negative correlation between Korean MEQ score and bedtime or wake time, with a stronger correlation in both cases, when weekend (free day) versus weekday wake time was used. Such correlation is caused by typical patterns, when weekday sleep-wake times are shortened because of the need to get up to go to work, while weekend sleep-wake times may better reflect the underlying chronotype. This concept has been described as a "social jetlag" (SJL) (Wittmann *et al.*, 2006).

Overall, actigraphy is an elegant and non-invasive method based on the continuous measurement of physical activity. Actigraphy is widely used in sleep and circadian rhythms studies. It, however, lacks coherence in its use. This chapter provides insight into how well actigraphy can replace questionnaires in chronotyping and how long records should be.

The aims of this study are:

1. To evaluate the connection between the questionnaire-based chronotype estimates obtained through chronotyping questionnaires (MEQ and MCTQ), and objectively measured through actigraphy parameters connected to earliness or lateness of physical activity (and sleep).
2. To assess the accuracy of approximation of chronotyping questionnaire scores from actigraphy and select actigraphy measures that best resemble the questionnaire-based chronotypes.
3. To evaluate how the length of the actigraphy observation period affects the accuracy of chronotype estimation.
4. To evaluate the stability (test-retest) of the actigraphy based chronotypes.

6.2. Methods

In this study, we have equipped a group of women participants (see the CHRONOBIO dataset details in section 4.3) with an actigraphy wristband and instructed them to wear it at all times for up to three months. All participants filled the screening chronotyping questionnaires at study admission, together with the MEQ and MCTQ (for MSFsc and SJLrel). A subgroup of patients also filled the MCTQ and MEQ questionnaires after a follow-up period of 18 months. The participants also completed a survey where they indicated work status, long-distance travel dates, and free and vacation days other than weekends and public holidays.

6.2.1. Subjective Methods – Chronotype Questionnaires

Czech versions of the MCTQ and MEQ questionnaires, validated by a double-reverse translation from the originals, were used. Previously, the Czech translations of both questionnaires were validated, and their relationship investigated (Fárková *et al.*, 2020).

Morningness-Eveningness Questionnaire (MEQ)

The gold standard in chronotype detection is the self-assessment inventory developed by Horne and Ostberg (1976), the MEQ (Di Milia *et al.*, 2013). The MEQ consists of 14 multiple choice questions, and 5 open questions, inquiring about individual preferred times for different activities. The MEQ score ranges from 16–86, with lower values indicating a more evening chronotype (Horne and Ostberg, 1976). It can be categorized into the E-type (16–41), the N-type (42–58), and the M-type (59–86) (Roenneberg, 2015; Ryu *et al.*, 2018).

Munich Chronotype Questionnaire (MCTQ)

The Munich Chronotype Questionnaire (MCTQ), designed by Roenneberg and his team (2003), quantifies the chronotype according to the phase of entrainment based on the reported mid-sleep and takes into account its occurrence on free-regime and scheduled-regime, e.g. working days. The MCTQ parameters are mid-sleep on weekdays (MSW), mid-sleep on free days (MSF) and mid-sleep on free days corrected for sleep debt on weekdays (MSFsc). The latter represents a continuum of circadian preference (Levandovski *et al.*, 2011). MCTQ was primarily used for determining MSFsc and the rate of SJL - i.e. the difference between hours of sleep on free days and working days. The rate of SJL quantifies (in hours and minutes) the discrepancy between circadian and social clocks, which can lead to chronic sleep loss (Roenneberg *et al.*, 2012).

Questionnaire Chronotype Test-retest Stability

Based on the data from the 18 volunteers who re-filled the chronotyping questionnaires (see section 4.3.1), we evaluated the test-retest long-term stability for both MCTQ and MEQ chronotypes, using the Spearman correlation in a similar way as in the study of Lee *et al.* (2014).

6.2.2. Actigraphy

The actigraphy wearable (GMK[12] actigraph by Mindpax Ltd.) was worn on the non-dominant arm's wrist and was set to collect activity counts in 30s epochs. The data were wirelessly transferred to a server, using base stations at participants' homes, and stored for offline processing.

First, an exploratory analysis was performed: actograms were studied macroscopically to verify the absence of artefacts or abnormalities. The sleep and wake periods, and periods when the actigraph wasn't worn were detected automatically (see section 3.5.3). The wear-off periods (when the wearable was removed) were excluded from subsequent analyses.

The mid-sleep feature was obtained from the detected main daily sleep. Additional features included in the analyses were the cosinor Acrophase (3.5.1), the chronotyping $MSFsc_{acti}$, and $SJLrel_{acti}$ (3.5.4), and the NPCRA M10-time and L5-time (3.5.2). The features were estimated using 1-6 weeks' time window segments of actigraphy recordings. Except for the Acrophase, all features were calculated from individual days and then averaged for the required segment length. Acrophase was calculated from the whole segments. The features obtained from different length segments were used to evaluate the effect of recording length on 1) the feature stability in time and 2) the chronotype estimation accuracy.

Moreover, for SJL evaluation, mid-sleep, M10-time and L5-time daily values were averaged for free-regime and working days of the segments. Sleep, and therefore mid-sleep, is always assigned to the day of wake-up. The absolute differences between working and free days are referred to as mid-sleep_{diff}, M10-time_{diff}, and L5-time_{diff}.

All feature extraction, data processing and statistical analyses were with Matlab software (MATLAB 2018b, The MathWorks, Inc., Natick, Massachusetts, United States.).

[12] A device with similar parameters as MindG (Table 3-1), but due to different hardware the counts scores are not comparable

6.2.3. Chronotype Estimation from Actigraphy

A set of univariate linear models was used to evaluate chronotype estimation accuracy for each of the selected actigraphic features. We trained linear regression models with the scale-based circadian phenotype score CPSc (chronotypes MSFsc, MEQ, and social jetlag SJLrel) as the outcome variable and actigraphic features obtained from the first valid actigraphy window as predictors.

$$\mathrm{CPSc} = \beta_0 + \beta_1 F_i + \varepsilon_i, \tag{6.1}$$

where F_i represents the *i*-th selected feature and ε_i stands for residual error. To evaluate model performance on unseen data, we used a 5-fold cross-validation procedure: the study participants were divided into 5 folds (groups), where 4 folds were used for training (training group) and one for testing (testing group). The performance was evaluated using the test group. This procedure was repeated five times. Therefore, each participant had been one time in a testing group.

We chose the mean absolute error (MAE) as a primary accuracy measure, which may be interpreted in the response variable's original units.

$$\mathrm{MAE} = \frac{1}{n_{test}} \cdot \sum_{t}^{n_{tests}} \left|\mathrm{CPSc}_t - \widehat{\mathrm{CPSc}_t}\right|, \tag{6.2}$$

where n_{tests} is the number of values in each test set, CPSc_t is the chronotype or SJL recorded score and $\widehat{\mathrm{CPSc}_t}$ is the model estimation of the chronotype or SJL score.

For each of the selected actigraphic features, and each window length, three linear regression models were fitted predicting one of the three circadian phenotype scores:

a) the MEQ score

b) MCTQ-MSFsc

c) MCTQ-SJLrel

To predict the MEQ score or the MCTQ-MSFsc time, we have used MSFsc_{acti}, Acrophase, M10-time, L5-time, and mid-sleep as predictors. To predict MCTQ-SJLrel we have used SJLrel_{acti}, and differences between working days and free days in $\mathrm{M10\text{-}time}_{diff}$, $\mathrm{L5\text{-}time}_{diff}$, and

mid-sleep$_{diff}$ as predictors. For the sake of the analysis, the working days and free days were based on the Czech calendar weekends and holidays and corrected when the participant filled a free day survey.

To evaluate feature strength and necessary actigraphy recording duration for chronotyping, the features were estimated from the 1–6 week-long segments. The usefulness of each actigraphic feature was assessed for each questionnaire-based response: (1) by the models' prediction accuracy (MAE) and (2) the Spearman's correlation between the questionnaire-based response and actigraphic feature.

To illustrate the extent to which the actigraphic features improve chronotype estimation for each individual, we also computed the performance scores of an intercept-only, null model, defined as:

$$\mathrm{CPSc} = \beta_0 + \varepsilon_i, \tag{6.3}$$

The null model is equivalent to using the sample mean to estimate individual chronotype. The same evaluation procedure as for the standard models was used for the null model.

Sensitivity Analysis for the Impact of Confounders: Age and BMI

Both age and BMI have been reported previously to be connected with chronotype and physical activity (Mecacci *et al.*, 1986; Roenneberg *et al.*, 2007; Bass, 2012; Sridhar and Sanjana, 2016). As the dataset used in this study is relatively heterogeneous, we exploratively evaluated the relationship between individual actigraphic features, age, and BMI using univariate linear regression. The evaluation was done by cross-validation of an additional set of linear models, using a procedure identical to the main analysis. The linear models were modified by adding two regressors for age and BMI. The resulting form of the models was:

$$\mathrm{CPSc} = \beta_0 + \beta_1 \cdot F_i + \beta_2 \cdot BMI + \beta_2 \cdot AGE + \varepsilon_i, \tag{6.4}$$

Gender was not used as cofounder because all included volunteers were women.

Test-retest Stability of Actigraphy-based Chronotype

To evaluate the stability of actigraphy chronotypes' over time, we have used the test-retest procedure, similar to the one used to evaluate the questionnaire-based chronotype stability. The stability was estimated using Pearson's correlation coefficient between actigraphic feature

values calculated from two time-windows, separated by a gap. The tested window lengths were set to 1-3 weeks. The time gap between the windows was set in the range of 0-3 weeks. A scheme of examination procedure is shown in Figure 6.1.

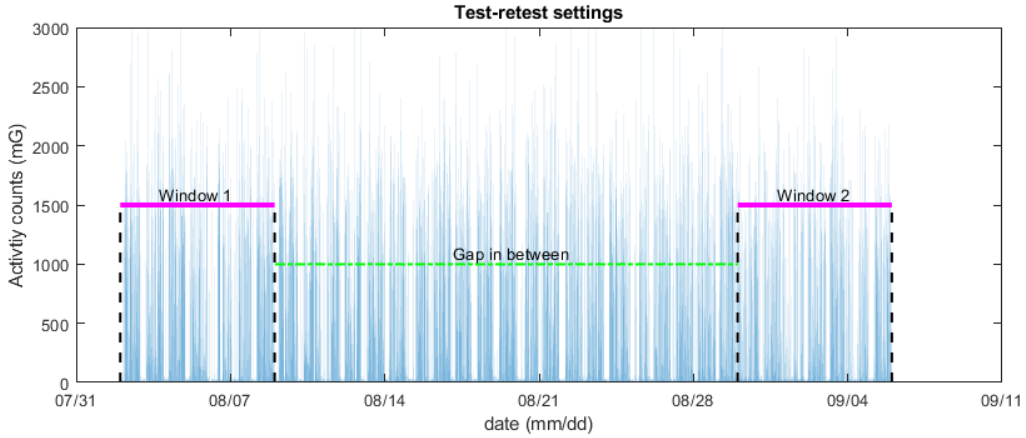


Figure 6.1 - Test-retest evaluation settings. The features are estimated from selected segments (1-week purple marked windows at the beginning and the end) with a set gap in between (3 weeks – green dashed line). Features are estimated for all patients with sufficient valid data in both windows (maximum 20% missing data-points). The test-retest is obtained as a Pearson's correlation coefficient between Window 1 and Window 2 feature estimates.

6.3. Results

The details about participants are presented in the Datasets Chapter 4 - section 4.3, the CHRONOBIO dataset.

6.3.1. Chronotype Estimation from Actigraphy

All of the selected actigraphic features have shown a significant connection to their respective questionnaire counterparts. Table 6-2 summarises the cross-validated linear models' results for features calculated from windows, for which the best prediction accuracy was achieved. Analogous results for models using different all used window lengths to compute the actigraphic features can be found in Table S-2 in the supplement.

For **MEQ**, all of the five features selected for evaluation were significant. The best predictor based on low train MAE was the Acrophase (MAE = 5.6 points, R-squared = 0.37), followed by the daily M10-time (MAE = 5.9 points, R-squared = 0.29), and mid-sleep (MAE = 6.0 points, R-squared = 0.36). AGE was a significant confounder for MEQ score prediction. The BMI was not. Using a linear model with confounders brought a little MAE improvement for all mentioned strong predictors (< 0.1 MEQ points). The window length with minimum error varied from 3-6 weeks, depending on the feature. If an MCTQ-MSFsc

questionnaire-based score was used for MEQ prediction, it achieved test MAE = 5.4 and train R-squared = 0.47.

Similarly, for the **MCTQ-MSFsc**, all features selected for comparison have shown a significant connection. The best predictor, based on low MAE, was the $MSFsc_{acti}$ (MAE = 0.57 hours, R-squared = 0.47), followed by Acrophase (MAE = 0.61 hours, R-squared = 0.40), mid-sleep (MAE = 0.62 hours, R-squared = 0.46), and L5-time (MAE = 0.65 hours, R-squared = 0.32). AGE was a significant confounder, while BMI score was not. The improvement of MAE using the model with confounders was < 2 minutes for all mentioned predictors. As the MCTQ MSFsc and $MSFsc_{acti}$ have a theoretical 1:1 dependency, the fitted and theoretical models are shown in Figure 6.2. The window length achieving minimum error varied between 4-6 weeks, depending on the feature. When the questionnaire-based MEQ score was used for MCTQ-MSFsc estimation, it achieved test MAE = 0,61 hours and train R-squared = 0.47.

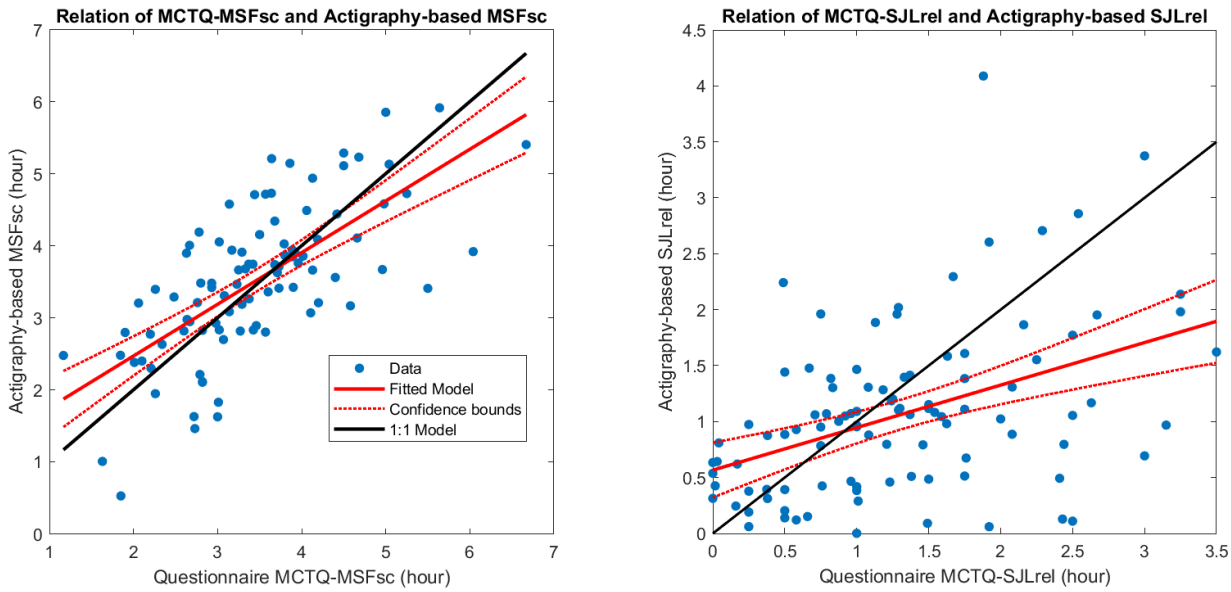


Figure 6.2 - MCTQ and actigraphy circadian phenotypes dependency. Showing patients data and the fitted and theoretical models for MSFsc (left side) and SJLrel (right side) The fitted model (solid red) is shown with the 95% confidence interval (dashed red). The solid black line represents the theoretical 1:1 model. The difference between SJLrel models suggests that many participants tended to overestimate their SJL.

Unlike in the previous two cases, only two of the selected features were significant predictors of the **MCTQ-SJLrel**. The best predictor, based on low MAE was the $SJLrel_{acti}$ (MAE = 0.62 hours) followed by and $mid\text{-}sleep_{diff}$ (MAE = 0.65 hours). The other non-significant features, the $M10\text{-}time_{diff}$ and $L5\text{-}timed_{iff}$, showed slightly inferior results (MAE > 0.67 hours, where 0.69 was the result of a null model). As in previous cases, AGE was a significant confounder, while the BMI score was not. The incorporation of confounders into the model almost didn't improve the model. The MAE improvement for the models with confounders was less than 1 minute for any of the actigraphy predictors. The window length achieving minimum error was varying from 3-4 weeks, depending on the feature. As the MCTQ-SJLrel and SJLrelacti have an explicit theoretical 1:1 dependency, the data and the fitted theoretical and actual model are shown in Figure 6.2. Additionally, we have observed high $SJLrel_{acti}$ variability over time (the median SD of 4-week overlapping windows was 15 minutes, the first quartile Q1 = 10 minutes, and the third quartile Q3 = 23 minutes).

Moreover, the correlation of the individual actigraphic features with AGE and BMI can be found in Table 6-1. The results of models including confounders can be found in Table S-3 in the supplement.

Table 6-1: Pearson's correlation coefficients between individual actigraphic features, age and BMI

Feature	Correlation with AGE	Correlation with BMI
$MSFsc_{acti}$	-0.421^{***}	-0.223^{*}
Acrophase	-0.528^{***}	-0.350^{***}
M10-time	-0.426^{***}	-0.399^{***}
L5-time	-0.302^{*}	-0.240^{*}
Mid-sleep	-0.413^{***}	-0.232^{*}
$SJLrel_{acti}$	0.088	-0.016
$M10\text{-}time_{diff}$	-0.130	0.107
$L5\text{-}time_{diff}$	0.025	0.020
$Mid\text{-}sleep_{diff}$	-0.008	-0.180

$^{***} < 0.001$, $^{**} < 0.01$, $^{*} < 0.05$

Table 6-2: Actigraphy vs questionnaire-based chronotype: linear model results. For each actigraphic feature, the optimal window length providing minimum test-set MAE is shown.

1) MEQ		**TRAIN: β1**		**TRAIN: R-squared**		**TEST: MAE**	
Feature	***Window (weeks)***	***mean***	***SD***	***mean***	***SD***	***mean***	***SD***
Acrophase	3	-5.884	0.377	**0.372**	0.042	5.608	0.453
$MSFsc_{acti}$	4	-4.249	0.450	0.260	0.043	6.287	0.514
M10-time	4	-3.871	0.267	0.289	0.038	5.869	0.498
L5-time	6	-5.994	0.337	0.313	0.021	6.125	0.781
Mid-sleep	6	-6.400	0.705	**0.359**	0.042	5.995	1.148
Where MEQ based on questionnaire MSFsc has an MAE of 5.446 MEQ points and R-squared 0.47. The null model MAE is 7.216 points.							

2) MCTQ-MSFsc		**TRAIN: β1**		**TRAIN: R-squared**		**TEST: MAE**	
Feature	***Window***	***mean***	***SD***	***mean***	***SD***	***mean***	***SD***
Acrophase	6	0.679	0.041	**0.402**	0.037	0.605	0.113
$MSFsc_{acti}$	6	0.658	0.035	**0.471**	0.034	0.569	0.089
M10-time	6	0.437	0.061	0.287	0.046	0.691	0.086
L5-time	6	0.729	0.062	0.356	0.022	0.665	0.155
Mid-sleep	6	0.830	0.083	**0.455**	0.061	0.615	0.107
Where MSFsc based on questionnaire MEQ has an MAE of 0.612 hours and R-squared 0.47. The null model MAE is 0.804 hour.							

3) MCTQ-SJLrel		**TRAIN: β1**		**TRAIN: R-squared**		**TEST: MAE**	
Feature	***Window***	***mean***	***SD***	***mean***	***SD***	***mean***	***SD***
$SJLrel_{acti}$	4	0.497	0.059	**0.188**	0.024	0.622	0.116
$Mid\text{-}sleep_{diff}$	*3*	*0.223*	*0.064*	*0.086*	*0.047*	*0.647*	*0.160*
$M10\text{-}time_{diff}$	*6*	*0.142*	*0.019*	*0.038*	*0.015*	*0.687*	*0.141*
$L5\text{-}time_{diff}$	*6*	*0.214*	*0.069*	*0.069*	*0.016*	*0.674*	*0.196*
The SJLrel may be predicted from questionnaire MSFsc with MAE of 0.640 hours and R-squared 0.19. The null model has MAE 0.691 hour.							

Bold names mark features, which is significant and explains considerable variation in the data (R-squared > 0.35) (for SJL the one with significant coefficient), *italic* names mark features with non-significant coefficients. For comparison, the questionnaire-based prediction accuracies and results of null models are provided.

6.3.2. Impact of the observation period

In the previous section, we mentioned that the best results were obtained for specific estimation window lengths. The results presented in the Table 6-2 suggest that the longer time window usually provided the lowest chronotyping error for both the MEQ and MSFsc chronotypes (especially for MSFsc). On the other hand, shorter windows seem better suited for the estimation of SJL, where the maximum is reached for an approximately 3-4 weeks long window. Figure 6.3 presents the impact of estimation window length on the degree of correlation with the respective questionnaire-based chronotypes.

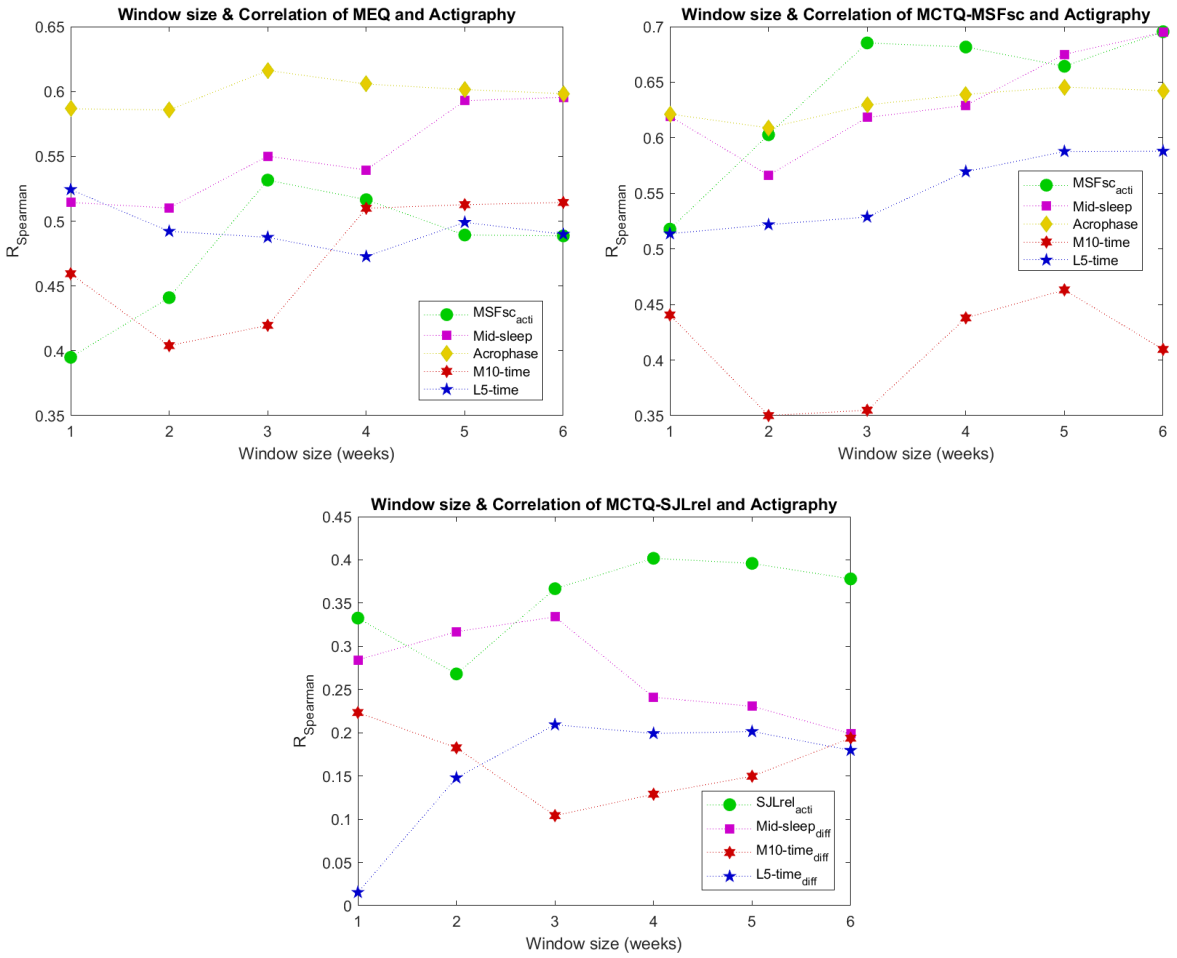


Figure 6.3 - Impact of actigraphy estimation window length on the level of association (Spearman's correlation coefficient) with each questionnaire-based chronotype measurement.

According to these results, Acrophase is a stable and valuable predictor of chronotype. It had a stable correlation across different observation window lengths ($R_{Spearman}$ ~ 0.6) with both questionnaire-based chronotype measures. For MEQ, the Acrophase reached the best score.

For MCTQ-MSFsc, there were better-correlated features: the $MSFsc_{acti}$ ($R_{Spearman} \sim 0.7$) and mid-sleep. However, for very short recording lengths (one or two weeks), the Acrophase achieved the best results also for MCTQ-MSFsc.

For the MCTQ-SJLrel the closest actigraphic feature is $SJLrel_{acti}$ with a moderate level of correlation ($R_{Spearman} \sim 0.4$), providing consistent results for windows of three weeks and longer. The $M10\text{-}time_{diff}$ and $L5\text{-}time_{diff}$ features showed a low level of correlation ($R_{Spearman} < 0.3$).

6.3.3. Test-retest Results for Actigraphic Features and Chronotype

We examined the test-retest stability of selected features by computing each actigraphic feature using two estimation windows of the same length, separated by a gap – see Figure 6.1. A Pearson's correlation was calculated for each of the features between the sets of values obtained from the two estimation windows. The results are shown in Table 6-3.

The most stable chronotyping feature was the Acrophase, where the test-retest score reached high values (~ 0.7) already for short estimation windows (1 week) with a long gap (3 weeks) in between. For longer estimation windows (3 weeks), it reached a very high (~ 0.8) test-retest score. The second most stable feature was mid-sleep. Compared to the Acrophase, it had lower stability (~ 0.3) for a short estimation window (1 week) and a long gap in between (3 weeks). For a long observation period windows (3 weeks), the results were similar to Acrophase (~ 0.8).

In general, the actigraphic features selected for comparison with MCTQ-based phenotypes had much lower levels of long-term stability. Their respective test-retest scores were dependent mainly on the length of the estimation window. For the $MSFsc_{acti}$, the maximum test-retest score (~ 0.7) was reached for two consecutive 3 week-long windows. For the long-gap scenario (3 weeks), the test-retest score was only moderate (~ 0.5). As the best result for the $SJLrel_{acti}$ feature, a moderate test-retest score (~ 0.5) was observed for 3-week estimation windows without any gap in between. For shorter estimation windows, the test-retest score was even lower. Again, the test-retest score is highly dependent on the estimation window length. The length of the gap does not seem to have much effect on these results. All the test-retest scores were significant ($\alpha < 0.05$), except for the $MSFsc_{acti}$ and $SJLrel_{acti}$ from the one-week estimation windows with a one-week gap in between.

Table 6-3: Stability (test-retest) of chronotype predicting features (Pearson's correlation coefficients)

Acrophase

Win length/gap	0 weeks	1 week	2 weeks	3 weeks
1 week	0.801	0.742	0.683	0.688
2 weeks	0.817	0.756	0.791	0.547
3 weeks	0.851	0.838	0.592	0.807

Mid-sleep

Win length/gap	0 weeks	1 week	2 weeks	3 weeks
1 week	0.789	0.745	0.655	0.283
2 weeks	0.784	0.473	0.720	0.740
3 weeks	0.823	0.825	0.849	0.782

$MSFsc_{acti}$

Win length/gap	0 weeks	1 week	2 weeks	3 weeks
1 week	0.293	*0.196*	0.451	0.456
2 weeks	0.410	0.597	0.598	0.544
3 weeks	0.735	0.669	0.549	0.419

$SJLrel_{acti}$

Win length/gap	0 weeks	1 week	2 weeks	3 weeks
1 week	0.222	*0.120*	0.257	0.281
2 weeks	0.303	0.421	0.392	0.470
3 weeks	0.559	0.553	0.409	0.477

Italic values were **not** significant ($p > 0.05$)

Questionnaire Chronotype Test-retest Stability

A subset of 19 women re-filled the MEQ and 17 women re-filled the MCTQ. The median time between filling the first and second set of questionnaires was 81 weeks with a minimum of 72 weeks and a maximum of 107 weeks. The Pearson's correlations between the first and second set of questionnaires were for MEQ $R_{MEQ} = 0.956$ ($p < 0.001$), for MCTQ-MSFsc $R_{MSFsc} = 0.634$ ($p = 0.004$), and for MCTQ-SJL $R_{SJLrel} = 0.718$ ($p < 0.001$).

6.4. Discussion

Both subjective (questionnaires) and objective (actigraphy) methods have been used for chronotype estimation across the world (Ancoli-Israel *et al.*, 2003; Zavada *et al.*, 2005; Thun *et al.*, 2012; Di Milia *et al.*, 2013; Roenneberg *et al.*, 2019). Still, to our knowledge, ours is the first study, which determines the quality of their substitutability and defines clear guidance for further clinical and research usage. In particular, we focused on the overlap between the subjective and objective measures of the chronotype and social jetlag, impact of the observation period and repeatability of the measurements (test-retest).

6.4.1. The Connection Between Questionnaire Chronotypes and Actigraphy

While multiple actigraphic features were significantly correlated with the questionnaire-based chronotypes (Figure 6.3), the overlap between the two chronotyping methods, and therefore, their accurate substitutability is limited.

According to our first hypothesis, we found that the chronotype may be objectively measured by actigraphy. Multiple actigraphic features were significantly correlated with the questionnaire-based chronotypes. The highest correlation was achieved between the **MCTQ-MSFsc** and its actigraphy counterpart, the $MSFsc_{acti}$. This correlation was observed for a six-week window and was $R_{Spearman} = 0.70$, which is comparable to correlation $R_{Pearson} = 0.73$ previously observed by Santisteban and his team (2018). A similar level of correlation with MCTQ-MSFsc was also achieved by the time of mid-sleep averaged from a six-week window ($R = 0.69$). These high correlations could be expected because MSFsc is defined as a personally perceived average mid-sleep time on free days corrected for socially induced jetlag. While being questionnaire-based, the focus of the questions is the regular time of sleep onsets and offsets on working and free days (Roenneberg, Wirz-Justice and Merrow, 2003), the data which could be obtained from the actigraphy-based sleep detection (Kaplan *et al.*, 2012; Kosmadopoulos *et al.*, 2014; Bellone *et al.*, 2016). In terms of substitutability, the $MSFsc_{acti}$ achieved a prediction with a mean average error (MAE) of approximately 34 minutes while explaining 47 % of the interindividual variability. Compared to the typical range of MSFsc of approximately 2-9 hours (Wittmann *et al.*, 2006) (the min-max range in our study was 1.7-6.7 hours), this does not appear very high, and is also substantially more accurate than the accuracy of the null model (MAE of 48 minutes). The largest residual errors were observed on the extreme chronotypes (see Figure 6.2). If we consider actigraphy to be

precise in sleep detection, we may be suggesting that people may overestimate their perceived inclination toward morningness or eveningness. Additionally, the majority of differences were positive (actigraphy later than MCTQ), which suggests that people are prone to report the waking time earlier than it actually is. Generally, the actigraphy-based sleep is reported to overestimate the sleep duration, and not otherwise (Sadeh, 2011; Smith *et al.*, 2020).

The highest correlation for **MEQ**, the R = 0.62 (Acrophase three-week window), was slightly lower than the maximum correlation between actigraphy and the MCTQ-MSFsc. This could be caused by the MEQ being more an idealised notion of a daily regime. The questions in the questionnaire are mainly focused on activity and its ideal timing during the day (Horne and Ostberg, 1976). That could also be the reason why the most correlated actigraphic feature is the Acrophase. This measure assesses at the same time both the daily activity peak and the time of night sleep trough. As in MSFsc, the second most correlated feature (with R = 0.60 for a six-week window) is mid-sleep time. This, combined with a lower correlation between MEQ and M10-time (R = 0.51), signifies the importance of sleep time for chronotyping. When it comes to substitutability, the Acrophase archived MAE of 5.6 points, while explaining 37 % of the interindividual variability. Considering this in relation to the range of MEQ scores, it should be sufficient to distinguish between the standards chronotype categories (M-type, N-type, and E-type people). A pending limitation is that, similarly to the MSFsc, the highest residual errors are observed at the extreme chronotype values. This suggests that the optimal regime, expressed by the MEQ score, is not reachable for these extreme types.

Additionally, to put the results into context, we tested possible substitutability of the questionnaire chronotyping methods between themselves. The MAE for MEQ prediction from MCTQ-MSFsc was 5.4 points, which is slightly better than the MAE for actigraphy (5.6 points). The MAE of MSFsc predicted from MEQ was 36 minutes, which is slightly worse than the MAE from actigraphy (34 minutes). Moreover, this corresponds to the explained interindividual variability, which was 47 % between questionnaires. Therefore, it is better than the actigraphy – MEQ relation and the same as in the actigraphy - MSFsc. These differences clearly show the need to choose the chronotyping method for planned research correctly. This also shows that the actigraphy is more closely related to the MSFsc type of chronotype.

While the correlation between the questionnaires based chronotypes and selected actigraphic features is high, the maximal correlation reached for **MCTQ-SJLrel** and its actigraphy counterpart $SJLrel_{acti}$ was only R = 0.40 for a four-week window. The difference in mid-sleep

time between free and working days reached the second-highest correlation (R = 0.33 for a three-week window), which is relatively low. Similarly, the free vs working days differences for M10-time$_{diff}$ and L5-time$_{diff}$ showed even a lower level of agreement. The prediction MAE of about 37 minutes is also relatively high compared to the typical range of SJL of approximately 0-4 hours (Wittmann *et al.*, 2006), and the model explained only 19 % of the intraindividual variance. Such bad results are most likely connected with the high variability observed on the SJL value +/- 15 min based on a 28 days long estimation window with 27 days overlap.

A possible cause could be that the regime on free days may still be affected by other social factors other than work. Many authors agree that the difference between the regime on free and working days is enormous for some individuals, especially in terms of length, timing and mid-sleep time and amount of physical activity (Monk *et al.*, 2000; Roenneberg *et al.*, 2012). In general, there is no agreement on the clear distinction between daily regimes (free, working, working with a flexible schedule, free with a fixed schedule, etc.). In spite of that, the SJL concept is based on these differences (Wittmann *et al.*, 2006). While we considered both Saturday and Sunday as free days in this study, the sleep on these days is not the same. The Friday-to-Saturday night may still be affected by the work on Friday. Similarly, the Sunday-to-Monday night may differ from other working days based on free Sunday evening.

6.4.2. The Actigraphy Period Length for Chronotyping

Due to the fact that the length of actigraphic records in published studies varies considerably (from records of several days to longitudinal studies measuring continuously for several weeks or months), we consider it necessary to identify the sufficient length of actigraphic record for circadian phenotyping. The ideal record length for the study of sleep parameters has already been investigated by several studies (Acebo *et al.*, 1999; Aili *et al.*, 2017). Aili and her team (2017) found that more than 7 nights are needed for an accurate actigraphy-measured total sleep time. Another study examined record lengths to monitor major sleep attributes in children and adolescents. Its authors claim that five or more nights are enough for an objective assessment of sleep quality (Acebo *et al.*, 1999). However, these studies did not deal with the setting of circadian rhythms but studied only the necessary length of recording suitable for describing the main parameters of night sleep. According to our results, the most considerable improvement of chronotype prediction was observed for the first three weeks. Afterwards, the improvement is relatively small, as shown in Figure 6.3 (for SJL and MSFsc), by flattering the correlation score curve. However, the best results were obtained mainly for a six-week-long

window (which is the longest window we used). The SJL is hard to predict from actigraphy. As for the optimal window length, it behaves similarly to the chronotype prediction.

Although estimation windows longer than 4 weeks provided worse results in our sample, this was due to smaller sample sizes for the long windows. Furthermore, the second-best actigraphy predictor for SJLrel, the mid-sleep$_{diff}$, prediction power drops for 3 weeks and longer estimation windows. The cause for this may be additional inter-daily variability in sleep timings which is not evaluated in the simple MCTQ scenario.

6.4.3. The Test-retest Stability of Chronotypes

Chronotype is expected to be stable for most of adult life, with a slow move towards morningness in older age (Roenneberg *et al.*, 2004, 2007). Our results on the questionnaire support that theory. Again, there are differences between the questionnaires. While the MEQ is incredibly stable over time – test-retest score R = 0.96 after approximately one and half years, which is in agreement with Lee et al. 2014 finding R = 0.90, the MCTQ-MSFsc is much less stable, achieving only R = 0.63. These differences again signify the different focus of each respective chronotyping questionnaire.

The stability of actigraphic features varied across features. The Acrophase, which is among the top chronotyping features (the best for MEQ), is also the most stable R = 0.81 for a three-week estimation window and three-week gap. The MSFsc$_{acti}$, which is the actigraphy analogy to the MSFsc score, is much less stable, with R = 0.42 for the same settings. These differences are likely caused mainly by the necessity to define the free days to compute the MSFsc$_{acti}$, as the mid-sleep time by itself is almost as stable as the Acrophase, with R = 0.78.

6.5. Limitations

Results of this study need to be interpreted considering the following limitations:

Firstly, the sample consisted of 122 women. While such sample size, especially considering the study duration, is above the average for actigraphy studies (see Figure 7.1), it is not enough for global generalisations. Moreover, the dataset included only women. There is no consensus concerning gender differences in chronotypes or sleep. But many studies found morning chronotype to be more prevalent in women, while evening chronotype being more prevalent in men. On the other hand, in a recent Czech study, no significant sex chronotype differences were found (assessed by MEQ and MCTQ) (Fárková *et al.*, 2020). Considering actigraphy

circadian parameters and sleep, women and men exhibited a similar circadian activity profile; however, women exhibited better sleep-wake patterns (Jean-Louis *et al.*, 2000). Another study found no significant gender differences (Lehnkering *et al.*, 2006). Even with limitation, we consider our sample representative, as it includes a wide range of chronotypes.

Secondly, while sleep detection using actigraphy is generally considered reliable (see section 3.5.3), most actigraphy sleep detectors are based on detecting epochs of low activity, and therefore have a tendency to overestimate the sleep duration (Sadeh, 2011; Smith *et al.*, 2020). Additionally, the quality of sleep detection may be affected by the internal settings of actigraphy data pre-processing (within the wearable) (Meltzer *et al.*, 2012; Cellini *et al.*, 2013; Smith *et al.*, 2020) and other aspects, such as possible sleep disorders (Sadeh *et al.*, 1995), or mood disorders as bipolar disorder (Gruber *et al.*, 2009; Schneider *et al.*, 2020). In this study, these patients were excluded based on our exclusion criteria. Nonetheless, it is important to consider these limitations for any study, which would include actigraphy-based chronotyping.

Third, the stability of actigraphy chronotypes was evaluated for short actigraphy segments. The test-retest evaluation period for the actigraphic features stability evaluation was limited to a three-week maximum estimation window. As seen from our results, this is at the same time the minimum window length to consistently estimate the subjective chronotype, based on actigraphy. Moreover, the stability after a year and a half for the chronotyping questionnaire is not directly comparable to the stability of actigraphy chronotyping features.

6.6. Conclusions

Actigraphy is a popular method of estimating sleep and circadian rhythms patterns. As we have shown in this study, longer-term recordings of three and more weeks of duration may be used as an objective evaluation of the chronotype and show good agreement with the MSFsc and MEQ questionnaires, traditionally used to determine the chronotype. The Acrophase and $MSFsc_{acti}$ estimated from automatically detected sleep periods were the best parameters for chronotyping. In all cases, the actigraphy-derived chronotype showed a more conservative estimate (closer to the sample mean) than its questionnaire-based counterpart, suggesting the tendency of the participants to overestimate the extremity of their behaviour. Our study also highlighted the distinction between the idealized MEQ scores with high stability over an extended period of time and moderate predictability by

actigraphy, the MCTQ-MSFsc, which showed high predictability by actigraphy and lower stability over time, and the SJL, which was both highly variable over time and hard to predict from actigraphy.

The increasing availability of actigraphy wearables, allowing long-term monitoring of the sleep-wake patterns with relatively high accuracy, stimulates wide-ranging applications of monitoring the circadian rhythms on a large scale. We show, that however, the actigraphy-based chronotype may provide a slightly different view than the traditional questionnaires, it may be highly valuable, especially if an unbiased and momentary value is of interest.

7. Actigraphy-based Classification of BD Patients and HC

This chapter is based on the article published in CNS spectrums: **Schneider, J.** *et al.* (2020) 'Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls', *CNS Spectrums*, pp. 1–11. doi: 10.1017/S1092852920001777. In this chapter, I'm using whole text sections as they were published in the journal extended for some parts included in the article supplement.

7.1. Introduction

7.1.1. Actigraphy Studies in BD Patients

Actigraphy is a convenient way to study motor activity patterns. Existing findings from actigraphy studies suggest that circadian rhythm and sleep are disrupted in patients with BD, even in the remitted state. Current evidence, including reviews (meta-analyses) (Scott, 2011; Geoffroy *et al.*, 2015; Alloy *et al.*, 2017; Scott, Murray, *et al.*, 2017), documents lower overall activity (Harvey *et al.*, 2005; Jones, Hare and Evershed, 2005; Salvatore *et al.*, 2008; St-Amand *et al.*, 2013) and longer and more disrupted sleep in remitted BD patients than in healthy controls (HC) (Millar, Espie and Scott, 2004; Gershon *et al.*, 2012; Geoffroy, Boudebesse, *et al.*, 2014; McKenna, Drummond and Eyler, 2014). Similar observations have also been found in unaffected child and adolescent offspring of parents with BD (Sebela *et al.*, 2019). Although previous studies have improved the understanding of motor activity in BD patients, most existing studies are based on a limited period of actigraphy monitoring. They, therefore, miss the opportunity to assess and account for intra-individual temporal variations in actigraphy parameters. Variability in sleep and circadian parameters, obtained from actigraphy, suggests lower levels of synchronisation of BD patients with the day and night rhythm (Harvey *et al.*, 2005; Scott, 2011; Gershon *et al.*, 2012; Geoffroy, Etain, *et al.*, 2014; Bei *et al.*, 2016) and may be closely connected with the symptomatic periods (Krane-Gartiser *et al.*, 2014; Scott, Murray, *et al.*, 2017). The short duration of the studies (mostly < 14 days, the longest being 50 days - see Figure 7.1) is a limitation for variability assessment (Millar, Espie and Scott, 2004; Mullin, Harvey and Hinshaw, 2011; Gershon *et al.*, 2012). In order to overcome these issues, we increased the observation period in the actigraphy study presented

here to 90 days, aiming to focus on intra-individual long-term temporal variability (LTTV) in circadian rhythm and sleep parameters.

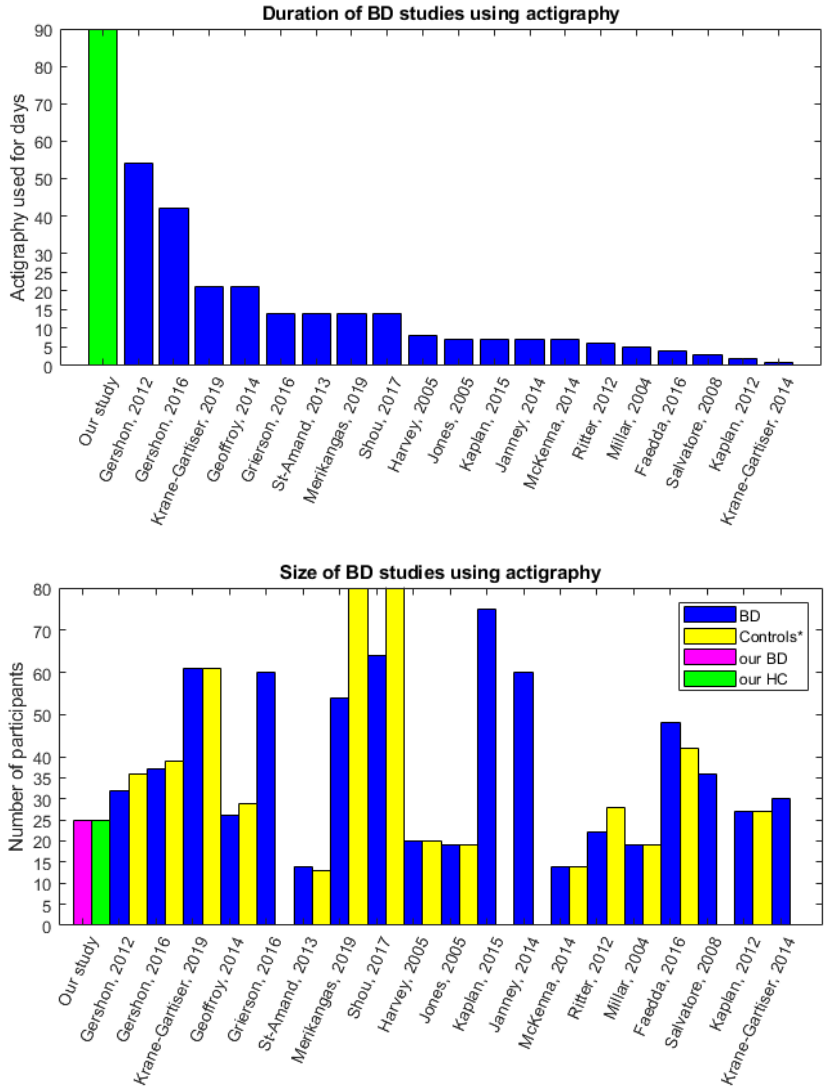


Figure 7.1 - Duration and sample size of BD actigraphic studies. The first bars (green/pink) represent our study, showing that the duration is one of the longest and the sample size is comparable to other studies. It may be seen that the duration is negatively associated with sample size and that newer studies tend to be longer or bigger.

Contrary to statistical evaluation, the machine learning techniques provide a means to quantify between-group differences by evaluating the classification power of a set of features (biomarkers), considering complex non-linear relationships among features. At least two recent actigraphy studies are employing this approach for actigraphy-based BD-HC classification. The first was done by Faedda *et al.*(2016), who reached 83 % accuracy with 64 % sensitivity, and 92 % specificity when using 3-5 days of actigraphy and diary data from children (5-18 years old). There was no medication used, and all data were recorded during a

similar regime (school days). The second recent study by Krane-Gartiser *et al.* (2019) applied classification algorithms to a set of 61 HC, and 61 remitted BD patients with stable medication, resulting in 78 % accuracy (75 % sensitivity and 80 % specificity) using selected actigraphic features and MADRS scores, resulting in 70 % accuracy using actigraphy alone. The main advantages were the use of matched groups (including employment status) and strict remission criteria (MADRS and YMRS ≤ 8 for ≥ 3 months).

7.1.2. Literature-based Differences Between BD Patients and HC

Following the available literature, we have expected a lower overall motor activity (Harvey *et al.*, 2005; Jones, Hare and Evershed, 2005; Salvatore *et al.*, 2008; St-Amand *et al.*, 2013; Janney *et al.*, 2014), and also lower peak activity (Gonçalves *et al.*, 2015) in BD patients versus HC. Based on diminished adaptability to changes in circadian rhythm, lower rhythm robustness was expected (Gonzalez *et al.*, 2018). Additionally, due to greater mood instability, higher fragmentation of activity profiles within a day, and instability between days were expected, including higher variability in most actigraphy parameters, both motor activity-based or time based (Kaufmann *et al.*, 2018).

Reduction in sleep quality has been reported in BD patients (Scott, 2011; De Crescenzo *et al.*, 2017); therefore, higher motor activity and longer awake or mobile periods were expected during night sleep. Further, since BD is associated with longer sleep (Millar, Espie and Scott, 2004; Ritter *et al.*, 2012; Geoffroy *et al.*, 2015; Ng *et al.*, 2015; Alloy *et al.*, 2017), though some reports did not confirm this finding (Jones, Hare and Evershed, 2005; Kaplan *et al.*, 2012; St-Amand *et al.*, 2013), we expected sleep time to be longer and more variable. Moreover, since longer sleep latency is associated with BD (Millar, Espie and Scott, 2004; Gershon *et al.*, 2012; Ritter *et al.*, 2012; Geoffroy, Boudebesse, *et al.*, 2014), we also expected lower activity before sleep onset and greater activity (restless sleep) after sleep onset, with higher variability in both sleep latency and restless sleep. Finally, BD is associated with later chronotype (Alloy *et al.*, 2017; Gershon *et al.*, 2018; Kaufmann *et al.*, 2018), represented as a later activity peak and a later sleep mid-time.

7.1.3. Variability Measurements and Primary Objectives

This chapter is focused on motor activity and intra-individual temporal changes in motor activity during waking hours and during sleep. Motor activity was measured using a wrist-worn actigraphy wearable, an instrument specifically tailored for use in psychiatry (Mindpax Ltd. – see section 3.4). Temporal variability is connected to changes in daily routine and in

circadian rhythm synchronization. Therefore, temporal variability may be a more straightforward way to measure the assumed triggers/predictors of BD symptoms (Milhiet *et al.*, 2011; Alloy *et al.*, 2017) than a standard comparison of average activity levels. We have also expected the variability measurements to be comparatively insensitive to basic differences in daily routine between BD patients and HC.

Aims of the study were:

1. To evaluate the motor activity profiles of inter-episode BD patients versus HC.
2. To use machine learning to distinguish between BD patients and HC using actigraphy-derived features focusing on variability measurements.
3. To evaluate the effect of employment status on the results (post hoc).

7.2. Methods

All methods and analyses described in this chapter are applied to the ACTIBIPO 1 dataset presented in Chapter 4 - section 4.1, and containing approximately three months' worth of data from 25 BD patients and 25 HC. All actigraphy features were calculated, as described in Chapter 3 - section 3.5. Only the one-day NPCRA and sleep features, and the seven-day cosinor features, were used. Chronotype and SJL were estimated using the whole actigraphy recordings.

7.2.1. Statistical Analysis

The LTTV and average values were calculated from all available daily values, using the standard deviation and the mean, respectively. Intergroup statistical comparison was performed on a preselected subset of features (Table 7-2), chosen based on the available literature.

The features were checked for normality using Q-Q plots, and they were normalized on the basis of skewness and kurtosis (for details, see section 7.3.2). When normality was not disproved in the transformed values (Jarque-Bera (1987) test, $\alpha = 5$ %), a student *t*-test was used; otherwise, the Wilcoxon rank-sum test was used for non-normally distributed data. One-sided tests were used, based on an apriori hypothesis from the existing literature (section 7.1.1). See Table 7-1 for details on feature normality. Additionally, see the data processing scheme in Figure 7.2.

The results were corrected for multiple comparisons using the Holm (1979) procedure (n = 25). The corrected results are marked as 'corr' after each result in the Results section. The effect size was calculated as the standardized mean difference (SMD). The area under the receiver operating characteristic (AUC) was computed to measure the classification power of individual actigraphic features.

All data processing and statistical analyses were conducted with Matlab 2015b, The MathWorks, Inc.

7.2.2. Classification

In order to illustrate the discriminatory power of the entire feature set combined (as opposed to the statistical analysis, which was aimed at individual features/biomarkers), we designed a set of classifiers discriminating between BD patients and HC. In total, we used three models differing by the features that were employed: **A**) a model with all the features presented above, **B**) a model based only on temporal variabilities, and **C**) a model using only features with low dependency on employment status (see section 7.3.4).

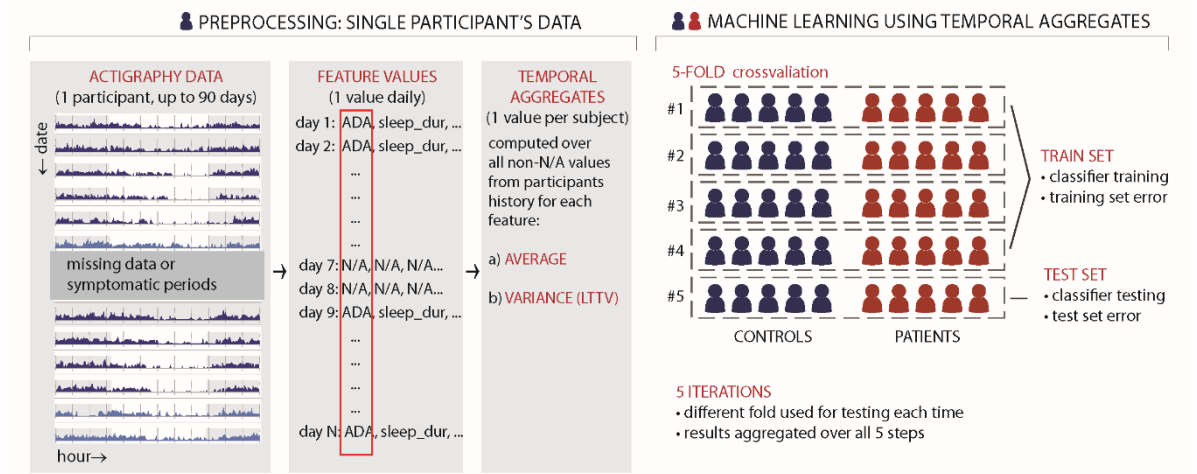


Figure 7.2 - Pre-processing and machine learning classification scheme; The left side shows the estimation of individual values (average and LTTV) from features, based on all valid days for each patient. In the right part, the machine learning cross-validation process

The models were trained using a **random forest (RF) classifier** (Breiman, 2001)**,** commonly used for heterogeneous biomedical data including actigraphy (Faedda *et al.*, 2016), and the out-of-sample performance was estimated using five-fold cross-validation. In each fold, data from 20 BD patients and 20 HC participants were used for training the classifier, the rest were used for evaluating the classification performance. In subsequent folds, the data from 5+5

different subjects were used for validation until all patients were iterated. The entire five-fold procedure was repeated 100 times, to estimate the uncertainty of the results, caused by the random division of the patients into folds and random feature selection in RF. See the data processing scheme in Figure 7.2.

7.2.3. Post hoc Analysis of Employment Status

An analysis of the classification results revealed a strong association between the misclassification of individual subjects and employment status. We, therefore, investigated the association between employment status, group membership (BD patients or HC) and individual actigraphic feature values. **A** set of linear models was built, so that the parameter value was a linear combination of BD patients/HC group status, employment status and intercept:

$$feature \sim 1 + BD/HC + employment\ status \quad 7.1$$

The model was fitted using a least-square means approach with robust bi-square weights, and the significance of the coefficient values was evaluated using a standard T-statistic. Based on the results, the identified features independent of employment status were used for training classification model **C**.

7.3. Results

7.3.1. Statistical Comparison

In terms of LTTV, compared with HC, BD patients showed significantly greater variation in the IV feature ($t(48) = -4.71$, $p_{corr} = 0.0005$, AUC = 0.85), greater variability in the activity-peak-time (M10-time; $z = 3.24$, $p_{corr} = 0.0107$, AUC = 0.77), and greater variability in the L5-time ($t(48) = -2.88$, $p_{corr} = 0.0500$, AUC = 0.75). In the IS feature, the variability had a higher predictive capacity than the mean value (both nonsignificant). For actual differences, see Table 4-1, and for effect sizes, see Table 7-2.

When evaluating individual averages (Table 4-1 and Table 7-2), compared to HC, BD was associated with lower ADA ($t(48) = 6.06$, $p_{corr} < 0.0001$, AUC = 0.90), longer sleep duration ($z = -4.35$, $p_{corr} = 0.0002$, AUC = 0.86), and lower CQ ($z = -4.25$, $p_{corr} = 0.0002$, AUC = 0.85). However, in some features (mainly in the overall averages) the observed differences were

highly associated with BD employment status. For more details and post hoc analysis on the effect of employment status, see section 7.3.4.

7.3.2. Features Normalisation

A transformation achieved the normality of distribution in most of the features except average CQ, LTTV in L5-time and M10-time and average sleep duration. Details of used transformations as well typical values for features averages and LTTV for the BD patients and HC are presented in following Table 7-1

Table 7-1: Features overview and normalisation

Feature	Variant	BD value‡	HC value‡	Transformation	Normality
Activity daily (ADA)	average	605 (SD 110)	778 (SD 92)	-	Yes
	LTTV	103 (SD 32)	94 (SD 25)	sqrt(x)	Yes
Circadian quotient (CQ)	*average*	0.78 (SD 0.12)	0.66 (SD 0.07)	*x^2*	**No**
	LTTV	0.07 (SD 0.02)	0.06 (SD 0.02)	sqrt(x)	Yes
IS	average	0.53 (SD 0.09)	0.53 (SD 0.06)	x^2	Yes
	LTTV	0.06 (SD 0.02)	0.06 (SD 0.02)	ln(x)	Yes
IV	average	0.49 (SD 0.11)	0.46 (SD 0.06)	-	Yes
	LTTV	0.07 (SD 0.02)	0.05 (SD 0.01)	-	Yes
Lest active 5 hours (L5)	average	73 (SD 24)	73 (SD 19)	sqrt(x)	Yes
	LTTV	31 (SD 23)	38 (SD 23)	ln(x)	Yes
Lest active 5 hours - time (L5-time)	average	2.85 (SD 0.94)	2.85 (SD 0.89)	ln(x)	Yes
	LTTV	1.80 (SD 0.55)	1.44 (SD 0.47)	*ln(x)*	**No**
Daily peak activity (M10)	average	994 (SD 179)	1179 (SD 137)	-	Yes
	LTTV	166 (SD 50)	148 (SD 42)	sqrt(x)	Yes
Time of daily peak activity (M10-time)	average	14.68 (SD 1.27)	14.89 (SD 1.34)	-	Yes
	LTTV	2.11 (SD 0.79)	2.61 (SD 0.51)	*ln(x)*	**No**
MSFsc	-	3.71 (SD 1.00)	3.62 (SD 1.05)	-	**No**
Activity after sleep onset	average	102 (SD 26)	80 (SD 17)	-	Yes
	LTTV	218 (SD 49)	232 (SD 50)	sqrt(x)	Yes
Activity prior sleep onset	average	758 (SD 184)	937 (SD 126)	-	Yes
	LTTV	71 (SD 32)	57 (SD 30)	sqrt(x)	Yes
Restless sleep (RSL)	average	2.6 (SD 0.9)	2.1 (SD 0.6)	sqrt(x)	Yes
	LTTV	1.78 (SD 0.79)	1.54 (SD 1.00)	ln(x)	Yes
Sleep duration	*average*	8.97 (SD 1.22)	7.40 (SD 0.51)	*x^2*	**No**
	LTTV	1.69 (SD 0.58)	1.32 (SD 0.33)	sqrt(x)	Yes

‡ Before normalisation, the *italic* text with orange shading significates that the normal distribution has not been achieved ($\alpha = 5\ \%$)

Table 7-2: Group differences between patients and controls

Temporal-variability				**Average values**				**Rationale**
Hypothesis - LTTV in feature is higher/lower in BD patients	**p-value**	**AUC**	**SMD (non-param.)**	**Hypothesis - Average value in feature is higher/lower in BD patients**	**p-value**	**AUC**	**SMD (non-param.)**	
Var. in IV is higher in BD patients	**< 0.001*****	0.8544	1.33 (0.97)	IV is higher in BD patients	0.131	0.5872	0.32 (0.26)	Fragmentation of activity within a 24-hour cycle. (Alloy *et al.*, 2017)
Var. of M10-time is higher in BD patients ‡	**< 0.001***	0.7680	-0.74 (-0.71)	M10-time is later in BD patients	0.707	0.5168	-0.15 (-0.24)	Finding daily activity extremes on a daily basis. (Alloy *et al.*, 2017)
Var. of L5-time is higher in BD patients	**0.003***	0.7456	0.71 (0.65)	L5-time is later in BD patients ‡	0.197	0.5712	0.04 (0.35)	See M10-time
Var. in sleep duration is higher in BD patients	**0.004**	0.7168	0.79 (0.56)	Sleep duration is higher in BD patients ‡	**< 0.001*****	0.8592	1.68 (2.03)	A basic sleep-describing feature.
Var. in AASO is higher in BD patients	**0.039**	0.6448	0.47 (0.40)	AASO is higher in BD patients	**< 0.001****	0.76	0.99 (0.88)	Used as an approximation of sleep latency.
Var. of CQ is higher in BD patients	**0.041**	0.6352	0.5 (0.27)	CQ is higher in BD patients ‡	**< 0.001*****	0.8512	1.23 (1.23)	An estimate of how well-circumscribed periods of activity are during a day; a proxy for rhythm robustness (Gonzalez *et al.*, 2018)
Var. of M10 is higher in BD patients	*0.074*	0.6112	0.41 (0.23)	M10 is lower in BD patients	**< 0.001****	0.792	-1.16 (-0.53)	Approximates the amplitude of peak daily activity (M10) and sleep quality (L5) for each day. It is related to motor capability (Gonçalves *et al.*, 2015)
Var. in RSL is higher in BD patients	*0.081*	0.6304	0.26 (0.41)	RSL is higher in BD patients	**0.012**	0.6624	0.67 (0.31)	Feature describing sleep quality. Represents sleep inefficiency based on actigraphy
Var. in IS is higher in BD	0.110	0.6272	0.29 (0.31)	IS is lower in BD	0.417	0.5328	-0.11 (-0.14)	Synchronization to the light-dark cycle and stability of daily rhythm.
Var. in ADA is higher in BD	0.151	0.5904	0.31 (0.34)	ADA is lower in BD	**< 0.001*****	0.8992	-1.71 (-1.05)	Describes how active a person is throughout the day.
Var. in APSO is lower in BD	0.160	0.5808	-0.27 (-0.24)	APSO is lower in BD	**< 0.001****	0.7712	-1.14 (-0.45)	See AASO
Var. in L5 is higher in BD	0.952	0.3424	-0.32 (-0.25)	L5 is higher in BD	0.932	0.3696	-0.38 (-0.38)	See M10
Significance after Holms correction (Holm, 1979) **(n = 25)** * **< 0.05** ** **<0.01** *** **<0.001** ‡**tested using Wilcoxon rank-sum test (non-normally distributed data)**				BD patients are generally a later chronotype	0.249	0.5568	0.08 (0.15)	Objectively assess the values from MCTQ questionnaire. (section 3.5.4)

7.3.3. Classification of BD and HC

The full actigraphy-based model (model **A**) successfully distinguished people with inter-episode BD and HC. Accuracy was around 88 % with a specificity of 91 % - see Table 7-3

When only time-variability of the actigraphic features was used (model **B**), the classification accuracy dropped, mainly due to a higher HC misclassification rate (i.e., a drop of specificity). The accuracy drop in the **B** model was apparently also due to the removal of the strongest feature, which was the average sleep duration (BD 8.97±1.22 hours vs HC 7.40±0.51[13]). Most of the misclassifications were in full-time/part-time working BD patients (see the last column in Table 7-3). For model **A**, in the working patients, 1.7 out of 6 were on average misclassified; in the part-time working patients, 2.1 out of 12 were misclassified, while there were no misclassifications in the unemployed/pensioned patients. For model **B**, 0.2 out of 7 unemployed/pensioned patients were misclassified. (For model **C**, which uses features that do not show dependency on employment status, see section 7.3.4)

Based on the out-of-bag estimation (Hastie, Tibshirani and Friedman, 2017), we assessed the importance of each feature in the classification task. Figure 7.3 shows features ordered by their average classification strength, depicting their approximate effect sizes based on model **A**. Models **B** and **C** differ by not including the unused features (the order of classification strength does not change).

Table 7-3: Random forest classifier results in participants whose data were not used during model training

Model	Accuracy mean (SD)	Sensitivity mean (SD)	Specificity mean (SD)	Misclassification in BD patients based on employment‡ Full-time/Part-time/Pensioned
A. All features	87.8 (2.6) %	84.8 (3.5) %	91.0 (4.0) %	29 % / 17.2 % / 0 %
B. Time variations	78.5 (4.2) %	77.7 (5.2) %	79.3 (5.8) %	36 % / 26.7 % / 3 %
C. employment status independent features	78.7 (3.4) %	76.2 (5.3) %	81.2 (4.2) %	33 % / 29 % / 7 %

‡The number of BD patients working full-time was 6 (therefore 1 patient corresponds to 16.7 %) part-time working n = 12 (1 patient ~ 8.3 %) and unemployed n = 7 (1 patient ~ 14.3 %)

[13] This corresponds to adult sleep duration (Roenneberg *et al.*, 2007)

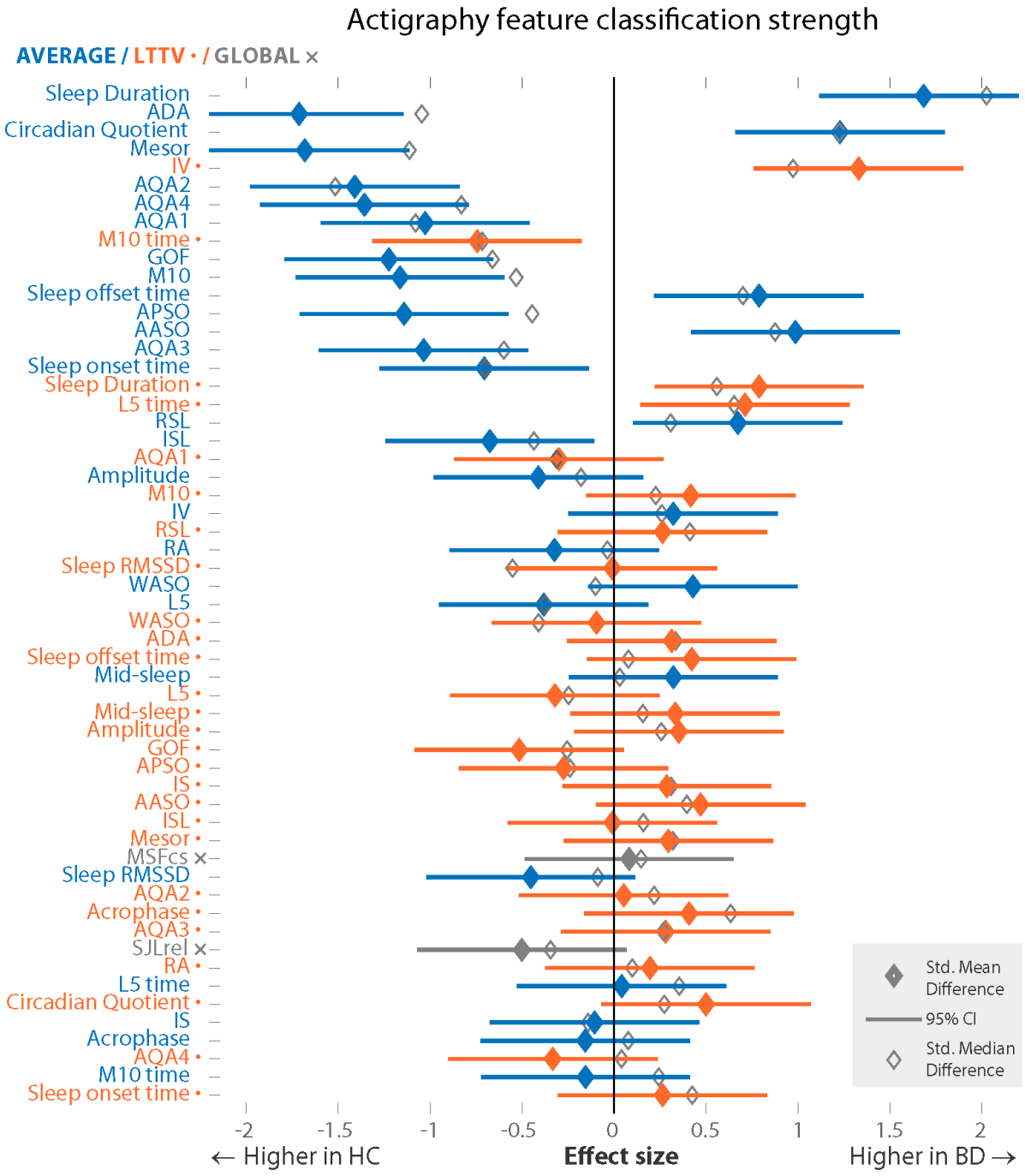


Figure 7.3 - Features used in classification the (Model A) ordered by their classification strength, showing the effect size for each feature (non-normalized data), LTTVs in blue, average values in orange and global features in grey. The effect size (with a 95 % confidence interval) is an approximation because the distribution was not always Gaussian. The grey diamond shows the effect size estimated by the median values and shows how precise the blue approximation is.

7.3.4. Effect of Employment Status

Using linear models, we identified four types of features based on their association with employment status (see Table 7-4). We trained a new random forest distinguishing BD patients and HC, using exclusively the variables that were most affected by the BD patients/HC group difference, and not by employment status. Model **C**, which used only type 1 features (LTTV in M10-time, IV, and SleDur and averages of M10, and APSO and AASO) reached an accuracy of 78.7 %; for details, see Table 7-3 - model **C**.

Table 7-4: Categories of features based on employment status effects

Category	Category description	Time-variability of the feature	Average of the feature
1	Features affected exclusively by BD/HC difference. (used for model C)	**M10-time, IV, SleDur**	**M10, APSO, AASO**
2	Features affected exclusively by employment status.	L5-time	L5-time
3	Features affected by both BD group and employment status. The ↑↑ reduce the BD effect (are in the same direction, and combine into a bigger difference) and the ↑↓ feature support the BD effect due to employment status (is in the opposite direction, and the difference is greater).		↑↑ CQ, ↑↑ ADA, ↑↑ SleDur‡, ↑↓ RSL
4	Features not significantly affected either by BD or by working status.	CQ, M10, IS, ADA, L5, APSO, AASO, RSL	L5, M10-time, IS, IV, MSFcs

‡ for sleep duration, the effect of the disease is twice as strong, resulting in the finding that even working BD patients differed significantly from HC

7.4. Discussion

This study shows that a machine-learning model using only actigraphic recording was capable of distinguishing between inter-episode BD patients and HC with 88 % accuracy on the test data. In addition, when the effect of working status was suppressed by empirically derived feature selection, our results indicated that actigraphic data on motor activity patterns in BD might contain a clinically informative and scalable biosignal, which differentiates between BD patients and HC. In her article, Ortiz *et al.* (2018) used machine learning for forecasting a clinical episode based on patient-perceived energy during the evening. Motor activity is associated with future mood and energy (Merikangas *et al.*, 2019); therefore, long-term actigraphy may be promising for relapse forecasting.

When compared to existing actigraphy-based machine learning studies of Krane-Gartiser *et al.* (2019) and of Faedda *et al.* (2016), our model using all features is more accurate than both. When only features with low dependency on employment status are used, our results are slightly better than the results by Krane-Gartiser (acc. 79 % vs 78 %) and lower than Faedda's

model (acc. 83 %) with higher sensitivity (76 % vs 64 %) and lower specificity (81 % vs 92 %). These results have to be considered bearing in mind that our remission criteria were more lenient than those in the Krane-Gartiser paper, whose dataset was matched for employment status. Faedda used children with a similar daily regime (school) and without any medication treatment and cleaned noisy data based on additional information obtained from parents. In addition to actigraphy, the Krane-Gartiser *et al.* (2019) employed MADRS as an additional predictor variable as well. Post hoc analyses demonstrated that the inclusion of this psychopathology score contributed critically to the overall efficacy of the model. When MADRS was excluded, the accuracy based selectively on motor activity dropped to 70 %, which is lower than our results.

Matching the BD patients and HC groups based on employment status, Krane-Gartiser *et al.* (2019) reduced the confounding effect of differences in social engagement. This approach has been substantiated by the identification of employment status as a significant confounder. The fact, that HC are typically employed, while at the same time many BD patients are either unemployed or pensioned, may by itself introduce a significant bias, due to the systematic effect of the dissimilar social clock and demands in the two groups. To address this problem at least partially, some studies have used shift work as an exclusion criterion (Millar, Espie and Scott, 2004; St-Amand *et al.*, 2013; Bullock and Murray, 2014). Only a small number of actigraphy studies have attempted to match HC on employment status (Jones, Hare and Evershed, 2005; Gershon *et al.*, 2012, 2016; Krane-Gartiser *et al.*, 2019). Unfortunately, even using age-matched HC groups with a similar rate of unemployment may introduce a different type of bias (Millar, Espie and Scott, 2004) due to the reasons causing a healthy person of productive age to be unemployed.

To control specifically for these potential biases, we identified and modelled a set of actigraphic features with low dependency on employment status and possibly other aspects affecting motor activity during the day, such as family status and type of employment. The contribution of these different factors to the BD-specific characteristics of motor activity patterns is beyond the scope of the presented dataset, and has to be evaluated in a separate study. According to our analysis, LTTV in interdaily variability feature, LTTV in M10-time, LTTV in SleDur and average M10, and average activity before sleep onset (APSO) and after sleep onset (AASO) fulfil these requirements. In a post hoc analysis, a model incorporating exclusively the features with low dependency on employment status achieved predictive accuracy of 79 % in discriminating between BD patients and HC.

7.4.1. Long-term Temporal Variability

Recent evidence suggests that not only previously reported changes in sleep and activity of BD patients, but also, the temporal variability of these parameters may be a disease-specific trait marker (Shou *et al.*, 2017). Despite this promising report, only a few studies have been able to specifically address time variation features in actigraphy. The reasons for this situation are mainly technical, as the analysis typically requires long-term continuous actigraphy. The variability of circadian rhythm has been already discussed in section 5.1; therefore, we mention here only the studies evaluating the variability for BD:

1. Increased standard deviation and RMSSD in actigraphy (Krane-Gartiser *et al.*, 2014)
2. Positive correlation between mood variability and variability in activity (Carr *et al.*, 2018)
3. Greater variability in afternoon activity (BD-I) and in night-time activity (BD-II), without differences in peak time variation (Shou *et al.*, 2017)
4. Greater variability in peak activity time (Kaufmann *et al.*, 2018)

Consistently, our analysis of the long-term time variability of actigraphy and sleep features revealed a significantly higher variability in the IV feature and in day-peak and day-trough activity times (M10-time and L5-time) in BD patients versus HC. In sleep features, we observed a difference in SleDur time-variability. Although the feature achieves only $p_{corr} < 0.1$ after correcting for multiple comparisons, this result should not be disregarded due to the limited power of the statistical test. Using differences in the sleep features observed by Geoffroy *et al.* (2015), the power for medium effect size (0.5 SMD) is only 0.68 for $\alpha = 0.1$.

7.4.2. Average Actigraphy and Sleep

As in previous studies, a lower overall activity (ADA) and flattening in rhythmicity (CQ) was detected in BD patients vs HC. Lower activity is a widely reported trait-marker of BD, even in remitted cases (Harvey *et al.*, 2005; Jones, Hare and Evershed, 2005; Salvatore *et al.*, 2008; St-Amand *et al.*, 2013; Bullock and Murray, 2014; Janney *et al.*, 2014; McKenna, Drummond and Eyler, 2014; Grierson *et al.*, 2016). Unfortunately, ADA also showed significant dependency on employment status. A lower daily activity peak was observed by previous studies (McKenna, Drummond and Eyler, 2014; Grierson *et al.*, 2016), where it was connected with worsening of the disease.

In contrast to previous studies (Scott, 2011; Alloy *et al.*, 2017; Kaufmann *et al.*, 2018), we did not observe an intergroup difference in chronotype based on motor activity, although all

subjects were evaluated at approximately the same time of the year. We also did not observe any later activity onset in BD patients versus HC, as had been observed previously (Salvatore *et al.*, 2008; Gershon *et al.*, 2016; Grierson *et al.*, 2016; Shou *et al.*, 2017; Kaufmann *et al.*, 2018).

Prolonged SleDur (in our study for > 1 hour) has been observed by some (Millar, Espie and Scott, 2004; Ritter *et al.*, 2012; Geoffroy, Boudebesse, *et al.*, 2014), but not by other studies (Jones, Hare and Evershed, 2005; Gershon *et al.*, 2012; St-Amand *et al.*, 2013). It is possible that the observed difference may be caused (1) by persistent sub-depressive symptoms, because even inter-episode BD patients show more depression-related symptoms (Judd, 2002; Geoffroy, Boudebesse, *et al.*, 2014), (2) by medication, whereby especially atypical antipsychotics are related to hypersomnia (Ng *et al.*, 2015), and (3) by the difference in employment status, as already has been mentioned.

Other commonly observed differences in BD are lower sleep efficiency (Millar, Espie and Scott, 2004; Harvey *et al.*, 2005) and prolonged sleep latency (Millar, Espie and Scott, 2004; Gershon *et al.*, 2012; Ritter *et al.*, 2012; Geoffroy, Boudebesse, *et al.*, 2014). These values cannot be estimated without the use of sleep diaries or patient markings of sleep time, which were not collected in our study. Our fully automatic approximation of these features is RSL, for sleep efficiency, and decline in activity on sleep onset, measured by APSO and AASO, for sleep latency. The between-group difference in RSL was not significant after corrections for multiple comparisons. Further, a slower decline in activity in BD versus HC during sleep onset was observed (APSO was lower, and AASO was higher in BD patients vs HC).

7.5. Limitations

Results of this study need to be interpreted considering the following limitations:

First, the relatively small sample size can reduce the power of the statistical tests. Although the sample size was small, it is in line with many previous actigraphy studies, each of which had a much shorter follow-up duration than our 90-day period.

Second, we had a relatively high dropout/exclusion rate of about 29 % in BD patients, due to loss of interest in participating in the study, the occurrence of a relapse, or technical difficulties. However, a comparable dropout rate is not exceptional in this type of study. For

example, Krane-Gartiser *et al.* (2019) had a dropout rate of 54 % in the BD patients group, as a consequence of very strict exclusion criteria.

Third, BD patients and HC were not matched for employment status, as only a few reasons might cause unemployment in 'healthy' productive age individuals. Contaminating the HC group with people with a possible risk of different morbidities might cause a different type of bias. To address this issue, we did not select the sample on the basis of employment status. Instead, we conducted a sensitivity analysis, creating a model in which those actigraphic features, which were highly correlated with employment status, were removed, showing the robustness of our discrimination/prediction models.

Fourth, all BD patients have been using their prescribed medication. There are reported effects of medication on sleep (Monti, 2016), and effects on activity can also be expected. However, Jones *et al.* (2005) stated that '*no evidence was found for a significant association between medication use and any of the circadian activity measures*' and Shou *et al.* (2017) did not observe any association between psychotropic medication and levels of activity. It has been shown that mood stabilisers can affect several circadian parameters (Hwang *et al.*, 2017). The assumed major mechanism is through the regularisation (normalisation) of the sleep and circadian rhythm as it has been shown for lithium (seven patients in our study) and valproate (three patients in our study) (Geoffroy, Boudebesse, *et al.*, 2014). Considering the combination of medications, Gonzalez *et al.* (2018) observed that individual medication type (mood stabilisers, antidepressants, antipsychotics, etc.) had a higher association with motor activity changes than the number of medications from each type. The medications may nonetheless impact the results, and therefore present a limitation of the study. At the same time, withdrawal from medication during the follow-up period is unacceptable due to the risk of relapse and related ethical issues.

Fifth, the BD subjects were not fully euthymic, and residual symptoms may have affected the results. Our relapse threshold allowed the presence of subclinical symptoms in the examined sample, e.g., residual depression (Judd, 2002). Monthly clinical assessments may also miss or underestimate briefer but clinically relevant mood shifts.

Sixth, there are findings of a high prevalence of comorbidities in BD (Hossain *et al.*, 2019). Although many are hard to distinguish from symptoms of bipolar disorder itself (sleep disorders, anxiety disorders, borderline personality disorder), other diseases have a higher prevalence in the BD group, such as drug/alcohol abuse, asthma, hypothyroidism, migraine,

etc., which may also affect circadian rhythm and motor activities throughout the day. These were not matched with the HC group and thus present a possible confounder and a limitation of the study.

Finally, we did not include patients with psychiatric disorders other than BD in order to evaluate the degree to which the identified actigraphic biosignature is specific to BD, or whether it is a more global marker of a mental illness. To overcome this limitation, future studies should include psychiatric control groups to investigate this issue.

7.6. Conclusions

There are significant differences in activity patterns between BD patients and HC. A clinically applicable, cost-effective and scalable classifier-based approach was able to distinguish BD patients from HC with approximately 88 % accuracy, which is better than previous studies by a large margin. Some of the strongest discriminants, e.g., ADA and SleDur, could be closely associated with differences in employment status and also with differences in the use of medications. The time-variation in some features (IV, M10-time, SleDur) showed lower dependency on employment status and may therefore be a preferable actigraphy biomarker candidate. When only such features, which are less dependent on employment status, were used, the model was still able to distinguish between BD patients and HC with approximately 79 % accuracy, which is still comparable with the best results obtained by other groups (Faedda *et al.*, 2016; Krane-Gartiser *et al.*, 2019). Future studies are needed in order to identify actigraphic features which are global trait-markers of mental illness from those, which are more specific to BD, and eventually, to identify features (state-markers) that may be associated with an impending relapse.

8. Actigraphy-based Clinical State Estimation

This chapter builds on the feasibility study, which we have published Cuesta-Frau, D.; **Schneider, J**. *et al.* (2020) 'Classification of Actigraphy Records from Bipolar Disorder Patients Using Slope Entropy: A Feasibility Study', *Entropy*, 22(11), p. 1243. DOI: 10.3390/e22111243.

The study compares entropy estimation methods applied to actigraphy data based on their ability to distinguish among BD relapses (depression and mania) and remissions. In the journal article, we used the ACTIBIPO 2 dataset, from which we selected 14-days segments from episodes of remission, depression, or mania, as described in section 8.2.1. At the pre-processing step, periods of high activity with substantial duration (1000+ data-points) were extracted from the segments. Afterwards, the entropy for each of the periods was estimated using different entropy estimation approaches. We used the following entropy measures: sample entropy (Richman and Moorman, 2000), permutation entropy (Bandt and Pompe, 2002), weighted permutation entropy (Fadlallah *et al.*, 2013), bubble-entropy (Manis, Aktaruzzaman and Sassi, 2017), and slope entropy (Cuesta-Frau *et al.*, 2019). The classification was done by three binary classifiers (dep-man, dep-rem, man-rem) and validated by leave-one-out (LOO) cross-validation (i.e. leaving always one of the extracted periods for testing) on bootstrapped samples.

The slope entropy was the only entropy estimation method that succeeded in all three classification tasks. In contrast, sample entropy was slightly better for the dep-rem classification and permutation entropy for dep-man classification. After parameters tuning, the classification results were for the dep-man[14] binary classification task: accuracy 73 %, sensitivity 75 %, and specificity 69 %; for the dep-rem binary classification task: accuracy 67 %, sensitivity 68 %, and specificity 66 %; and for the man-rem binary classification task: accuracy 62 %, sensitivity 75 %, and specificity 61 %.

[14] positive class is the first in each pair

8.1. Introduction

The possibility of predicting or detecting clinical mood episodes using objective means is the ultimate goal of digital phenotyping. It would significantly enhance the possibilities of treatment and, therefore, the wellbeing of BD patients, as the adverse medication effects could be diminished by state-based dose adjustment. Moreover, most episodes could be prevented using clinical interventions. The means how to achieve the goal may include actigraphy, other physiological measures as heart rate (HR) variability and electrodermal activity (Khan and Anwar, 2019), and behaviour measures (like smartphone usage – section 2.4.2).

The association of mood and physical activity in BD was shown and discussed in Chapters 2 and 7. Recently Merikangas *et al.* (2019) showed in the short-term (quarters of a day) model that the activity significantly affects the future mood, but not the other way. She also observed a connection between sleep and activity on consecutive days. Moreover, the activity affects both the subjective mood and the energy in the following part of the day. These observations make actigraphy a promising approach for patient state prediction. Unfortunately, only a few studies are focusing on actigraphy-based differences between episodes of depression and mania (Krane-Gartiser *et al.*, 2014; Gershon *et al.*, 2016; Scott, Vaaler, *et al.*, 2017; Cho *et al.*, 2019), or changes associated with episodes (like impulsivity and mood instability (McGowan *et al.*, 2020)). Moreover, half of them (Krane-Gartiser *et al.*, 2014; Scott, Vaaler, *et al.*, 2017) were conducted on acutely hospitalised inpatients.

Krane-Gartiser *et al.* (2014) recorded 24-hour actigraphy from 30 acutely hospitalised bipolar patients (18 for mania and 12 for depression), compared to 28 HC. Her analysis found significant differences between manic and depressive patients and patients and HC. She observed physical activity differences, where patients with both depression and mania were less active than HC. The results were highly significant for the morning but not for the evening activity. Moreover, higher fragmentation of activity was reported for the relapsed patients, and expressed for mania in the morning (but not in the evening) by lower autocorrelation lag. The observed morning's differences correspond to Gershon *et al.* (2016) findings based on 37 BD outpatients with a 6-weeks actigraphy recording, comparing daily activity profiles during the depression and euthymia. Patients were less physically active during depression episodes, with later activity onset (low morning activity) and low evening activity, with steep midday activity peak.

In another relevant study, McGowan *et al.* (2020) used 18-day actigraphy recordings to study the association between NPCRA (3.5.2) features and impulsivity and mood instability in 31 euthymic BD patients, 21 borderline personality disorder patients, and 31 HC. Both mood instability and impulsivity were associated with BD and may represent episode prodromes. Found associations were variability of activity between days (low IS) and lower rhythm amplitude (RA) for higher mood instability or impulsivity. Moreover, mood instability was associated with higher daily activity fragmentation (IV), higher nocturnal activity (L5) and delayed daily activity peak (M10-time). All these associations were significant only for the borderline personality disorder group (neither for BD, nor for HC), showing that the association expected in BD patients is more pronounced in borderline personality disorder, where the changes in actigraphic features appeared larger.

In a classification scenario, Scott, Vaaler, *et al.* (2017) used 24-hour long actigraphy recordings in 34 acutely hospitalised BD patients (16 manias, 12 depressions, 6 mixed states) to evaluate the possibility of actigraphy to distinguish among severe mood episodes. A discriminant function analysis reached classification accuracy (depression, mania, mixed state) about 79% on training data. The most significant cause of error were depressions misclassified as manias (42%). When using cross-validation, the model accuracy was 55 %.

In another study classifying BD episodes based on physical activity, and in this case, other physiological and environmental modalities, Cho *et al.* (2019) collected activity (steps and sleep parameters) and HR using a fitness tracker (Fitbit 2). The data were also combined with EMA and light exposure data, which have been collected using a smartphone. The dataset consisted of 55 volunteers whose data were collected for 2 years. The data were used to predict EMA for three following days. The model based on 130 physiological features from the previous 18 days reached an accuracy of about 65 %. Another model that distinguished mood episodes, confirmed by clinical scales in the MDD, BD-I, and BD-II in outpatients, reached overall accuracy slightly above 80 % on the unbalanced training sets. The best sensitivity at detecting episodes was in BD-II outpatients 64 % for depression and 67 % for hypomania. In BD-I outpatients, the sensitivity for depressive episodes was 25 %, and for manic episodes, it was 20 %.

The exploration of statistical differences in actigraphic features observed during BD episodes is important for better understanding the BD progression. Still, only the possibility of recognising episodes could revolutionize the treatment of BD patients. Especially as

actigraphy is collected passively, and therefore it is not demanding high patients' adherence. Contribution of such a system to better clinical therapy would be even larger if it could work for previously unseen patients.

There already were some partial successes using a combination of multiple sensors recording data from different domains (see section 2.4.2). At the same time, many of those approaches could face privacy issues. The combination of multiple sensors seems to have a higher dropout (as far as 88 % of excluded days due to missing data Cho *et al.*, (2019)). Possible reasons for such large amounts of missing data may be the frequent need for battery recharging in devices recording data from multiple domains, or feelings of stigmatisation when using multiple or recognisable sensors.

The goals of this analysis are to:

1. Statistically evaluate inter-episode differences using actigraphy-based circadian features in BD outpatients.
2. Explore the feasibility of the machine learning approach (using linear and nonlinear methods) to distinguish between BD episodes in unseen outpatients.
3. Find features that are commonly associated with the self-perceived worsened state, separately for depression and mania.

8.2. Methods

The work presented in this chapter uses data of patients from the ACTIBIPO 2 dataset, which is described in detail in Chapter 4 - section 4.2. For the statistical comparison (Goal 1) of actigraphic features during episodes (depression, mania, remission) and models for episode recognition (Goal 2), only the CORE group patients were included, as they have annotated episodes (see expert annotations 4.2.4.). For evaluating individual features' possibilities for distinction among clinical states in individual patients and their self-perceived episodes (Goal 3), the data of all ACTIBIPO 2 patients were used.

All data processing and statistical analyses and evaluation of logistic regression models were conducted with Matlab software (Matlab 2018b, The MathWorks, Inc), the evaluation of random forest classification was conducted in Python environment (Python 3.7.4, Python Software Foundation)

8.2.1. Data Pre-processing

The patients' states were annotated as (1) symptomatic clinically relevant episodes (relapses), (2) mildly symptomatic sub-clinical episodes, and (3) non-symptomatic episodes (remission) by a team of experts (for details, see 4.2.4). All hospitalisation periods were excluded, as the activity is considerably restricted in hospitals.

Only the episodes of remission, depression and mania, were included in the analysis. A 14-day long segment was selected from each episode in such a way that the segments contained a minimum amount of missing data-points. The segments with more than 10 % of missing values (1.4 days of missing values) were excluded from further analyses.

Features, corresponding to days in selected episodes' segments, were calculated in the way described in section 3.5. In cases where features' estimation window preceded the selected episodes' segments, all the features from the windows, which contained 10+ % of missing data-points, were excluded from further analysis. Additionally, Slope Entropy (Cuesta-Frau, 2019) was estimated from the M10 window for each day, and two environmental features were included: daylight duration (duration from dusk till dawn calculated for Prague altitude - SUN) and moon-illumination (moon cycle - MOON). The list of 63 actigraphic features is presented in Table 8-1 in the results.

In each segment, the excluded daily feature values were replaced by imputed ones. The imputation was done based on the feature estimation window as follows:

1. For the 1-day-based features, the excluded daily features were imputed as a linear interpolation of surrounding values.
2. For the 7-day-based features, the excluded feature's values were imputed as the average of values from the second week of the segment (episode). (The values in the first week may not originate from the episode only).
3. For the 14-day features, all excluded feature's values were replaced by the feature value for the last day of the segment (episode).

8.2.2. Statistical Comparison

First, we evaluated how symptomatic episodes differ from non-symptomatic episodes in their alternation of physical activity profiles (Goal 1). The Wilcoxon rank-sum test was used for testing the pairwise differences in feature's distributions during remission, depression, and mania. As this analysis is of exploratory nature and many of the features are highly correlated (Figure S.1 in supplement), no multiple comparison corrections were applied. Additionally,

per-patient feature averages were subtracted from the features for this comparison (Diff dataset). In this way, we have partially suppressed the differences in patients' lifestyles, which are apparent from Chapter 5 - Table 5-1. The results for original values are shown in Table S-4 in the supplement.

8.2.3. Models and Feature Selection

The feasibility of the machine learning approach to distinguish between clinical episodes (Goal 2) was tested using two classifiers. First was the logistic regression model (LRM). The second was the random forest (RF) classifier. In order to obtain results that could be directly comparable to the paper of Cuesta-Frau *et al.* (2020), the models have been trained and fine-tuned on each of the binary classification tasks separately (dep-man, dep-rem, and man-rem).

Logistic Regression Model

To reduce overfitting the model by too many features, the LRM was trained using a forward stepwise feature selection (FFS) procedure to choose from the dataset of 65 features, 63 actigraphic and 2 environmental (MOON and SUN). The logistic regression equation (Hastie, Tibshirani and Friedman, 2017) follows:

$$\log\left(\frac{p(F)}{1-p(F)}\right) = \beta_0 + \beta_1 F_1 + \cdots + \beta_p F_p \qquad 8.1$$

where $F = (F_1, \ldots, F_p)$ are the p features selected by the FFS procedure and $(\beta_0, \ldots, \beta_p)$ are the coefficients of the fitted model.

The FFS used deviance as the optimisation criterion. The deviance (Hastie, Tibshirani and Friedman, 2017) represents a difference between the log-likelihood of the actual model and the saturated model (a model with the maximum number of parameters that can be estimated), as below:

$$D = -2\left(\log(\boldsymbol{L}(bp, y)) - \log(\boldsymbol{L}(bs, y))\right), \qquad 8.2$$

where bs are coefficients $(\beta_{0s}, \ldots, \beta_{ps})$ of the saturated model and bp are coefficients $(\beta_{0a}, \cdots, \beta_{pa})$ of the actual model. The deviance has a Chi-square distribution, which was used to stop the feature selection when the difference from the previous model was smaller

than the 95th percentile of chi-square with *n-p* degrees of freedom, where *n* is the number of parameters in the saturated model and *p* is the number of parameters in the model.

A final set of features for the model is obtained according to the following procedure:

1. A set of features is selected from all features by the FFS procedure in every iteration of a Leave-One- Subject-Out (LOSO) cross-validation process (see section 8.2.4).
2. An expected contribution of each feature is estimated using the count of 'how many times such feature was selected by the FFS procedure', 'how statistically significant it was in every model where it was selected' and 'how successful was that specific model'. The contribution score (ConSc) for a specific feature was obtained from a heuristic using the significance (*p*-value) of the feature's coefficient and a probability ($p_{surrAUC}$) that a resulting area under the receiver operating characteristic (AUC) of the original model being better than AUCs of the same model using surrogates (random mixing of labels in the test set – see section 8.2.4).

$$\mathrm{ConSc}_k = \sum_{i}^{n} 2 \cdot \mathbf{H}(p_{surrAUC} < \alpha) \cdot \left(1 - p\left(\beta_k(\boldsymbol{LR}_i)\right)\right), \tag{8.3}$$

   where **H** stands for Heaviside function, α is the significance threshold ($\alpha = 0.05$), and $p\left(\beta_k(\boldsymbol{LR}_i)\right)$ is the statistical significance (*p*-value) for the β coefficient of the *k*-th feature in *i*-th logistic regression $\boldsymbol{LR}_i$ model, and *n* is the number of models (patients). For the features, which have not been selected to be used in the model, the $p\left(\beta_k(\boldsymbol{LR}_i)\right) = 1$. In this way, the features, which were frequently selected and had a significant impact on the models' outputs, are evaluated as having a high contribution value.

3. Then *N* features were selected based on the highest ConSc_k . The *N* has been arbitrarily set based on the visualised drop in ConSc.
4. The features were then clustered based on their correlation, and from those highly correlated (R ~ 0.85+), only those with higher ConSc have been selected into the final features set.

These final features sets were then used for training 3 LRMs (dep-man, dep-rem, and man-rem) without the FFS procedure. The results are presented in the results section.

Random Forest Model

The random forest has been chosen because it is a robust method, which usually doesn't tend to overfit as much as a single decision tree, and it doesn't require extensive hyper-parameter tuning, as in, e.g. gradient boosting machines.

First, the hyper-parameter tuning has been carried out on the entire feature set, using a fixed validation set with 10 randomly selected patients for dep-man and 15 patients for dep-rem and man-rep. This resulted in a training validation split of 79 % (training)/21 % (validation) for dep-man, 85 %/15 % for dep-rem and 86 %/14 % for man-rem. Then, the model was fitted on the training set and evaluated (for the sole purpose of hyper-parameter tuning and feature selection) on the validation set. The following model configurations have been selected for each binary classification task:

- dep-man settings: $n_{estimators} = 40; n_{samples} = N_{man-dep};$
  $r_{features} = 0.5; n_{leaf} = 50$
- dep-rem settings: $n_{estimators} = 40; n_{samples} = {}^{N_{dep-rem}}/_{1.5};$
  $r_{features} = 0.7; n_{leaf} = 70$
- man-rem settings: $n_{estimators} = 40; n_{samples} = {}^{N_{man-rem}}/_{2};$
  $r_{features} = 0.1; n_{leaf} = 50$

where $n_{estimators}$ is the number of trees, which are being ensembled to the random forest, $n_{samples}$ is the maximum number of samples used for building individual trees, $r_{features}$ is the share of features available at each split and n_{leaf} is the minimum number of samples in each final leaf of a tree. $N_{man-dep}, N_{dep-rem}, N_{man-rem}$ are the whole training set sizes for each binary classification task. These model settings have then been used during (LOSO) cross-validation (see section 8.2.4).

In order to make the LOSO cross-validation loop more efficient, and the models simpler, the number of features has been trimmed down for each binary classification task. The feature selection process has been carried out based on feature importance for RF using scikit-learn toolbox for Python (Pedregosa *et al.*, 2012). Feature importance is obtained by going recursively through each tree and branch. Then, for each split, it is recorded how much the feature used for the split contributed to the model's improvement. This improvement is then added to the importance score of each feature.

Finally, the importance scores were normalized so that the scores of all features summed up to one. For each binary classification task, feature importance has been calculated to include only important features in the model. The importance score thresholds were kept low to give the RF model a possibility to choose from approximately 15-30 features in order to eliminate the possibility that the selected features might change depending on patients used for validation. Therefore, more features have been included, and the RF model decided which of them to select. Moreover, highly correlated features could also be removed.

8.2.4. Machine Learning Validation Process

Due to limited dataset size (number of episodes), the results have been validated using the LOSO cross-validation. In this scenario, the model is trained *N* times, where *N* is the number of patients. For each training step, the data from *N*-1 patients are used as a training set, and the results are validated on the patient who was left out – Leave One Subject Out. This way, all episodes of a single patient are used for validation and none for training at each step. This approach is more appropriate to evaluate the general capability of distinction between episodes than leaving out individual samples (days), which we did in the Cuesta-Frau *et al.* (2020) because it suppresses the similarity in consecutive days as well as similarity in episodes from one patient (specific regime).

Additionally, this represents the more demanding of two application scenarios of an actigraphy-based BD episode prediction model: predicting episodes for a new 'unseen' patient. The less demanding application would be to predict new episodes in a patient whose data were used for model training. While such prediction is expectedly easier, our dataset, where only a few patients have reported episodes of the same type, is not suitable for this task.

The validation set was further limited to include only patients with at least one episode of each class. And as there are significant differences between the number of episodes of each type (depression, mania, remission) in the dataset, the sizes of training data sets were equalised. Equalisation was done by random replication of samples (without repetition) from the under-represented class. When the difference was such that even doubling the under-represented class would still not reach equal class sizes, random samples were removed from the over-represented class. The equalisation of class sizes was done only for the training dataset, as in the validation set, it would increase the variation of results.

Additionally, the surrogates' analysis was applied to evaluate the models' results for individual patients. In this analysis, the episode labels were randomly shuffled 5000 times in the test-sets

(LOSO) patients. AUCs (AUC_{surrog}) were obtained for these randomly shuffled labels using the currently learned model. The AUC, with original labels, was compared to quantiles of the AUC_{surrog} empirical distribution to find whether it is significantly different (two-sided $p < 0.05$). The significant results were interpreted as the model's ability to distinguish episodes from a given patient was better than random (BTR). If the original AUC quantile was significantly low ($p < 0.025$) compared to the AUC_{surrog} empirical distribution, we marked that this patient has the opposite episode manifestation than expected.

8.2.5. Individual Features vs Subjective Relapses

The association of actigraphic features changes with the self-perceived worsened state was explored to find which features are most commonly associated with elevated or depressed mood (Goal 3). The strength of each of the features to recognise episodes of worsened state was tested for individual patients. The motivation was to find a possible existence of additional BD subtypes, which have different changes in circadian rhythm parameters, or an existence of more types of activity changes for episodes of the same kind. In order to work with a larger dataset for this task, the patients' state was assessed by the ASERT (see section 4.2.2) rather than by expert labels. The procedure then was as follows:

1. The actigraphy features were selected only for days where ASERTs were collected.
2. The ASERTs divided by a threshold separately for manic and depressive (and non-specific), using the relapse global model threshold values from models by (Anýž *et al.*, 2021, preprint) 5 of the manic subset of questions and 15 for depressive (& nonspecific) subsets of questions.
3. The classification strength was tested for each of the features separately, by its AUC, for patients where there were at least 5 samples (days) in each category.
4. The potential of individual features was then assessed as the percentage of people for whom the $|AUC - 0.5| > 0.2$. The direction of changes was assessed by the percentage of patients with $AUC > 0.5$ and $AUC < 0.5$

The features were marked as high potential features, when $|AUC - 0.5| > 0.3$ occurred in 30 % of patients where the classification could be tested.

8.3. Results

8.3.1. Dataset Information - Episodes

Out of the 98 CORE patients, 91 had at least one episode (in this chapter, the term episode also includes remissions) lasting 14 days. Out of the 91 patients, 50 patients experienced episodes of one type only (45 remissions, 5 depressions, 0 manias), 35 patients experienced episodes of two types, and 6 patients experienced episodes of all three types. There have been in total 56 depression episodes, 25 manic episodes, and 173 remission episodes, which were statistically compared (Goals 1). Multiple episodes of one type may be found for some patients. As the models (binary classifiers) were trained solely on patients with episodes in both tested classes, there were 10 patients with depression and mania, 15 patients with remission and mania, and 28 patients with remission and depression used for models training and testing (Goal 2).

8.3.2. Statistical Comparison

There were significant differences between remission and depression, remission and mania and depression and mania in most used actigraphic features. For a summary, see Table 8-1. The (Diff dataset) features are presented in the table because these features reduce the effect of inter-patient activity profile differences. Although statistical differences were significant, the effect sizes, measured as a standardized mean difference (SMD), were ***very small***, SMD < 0.2 for most cases. All except two remission-to-relapse differences reached at best only ***small effect sizes*** (SMD in range 0.2 - 0.5). The two features that reached ***medium effect size*** (SMD in range 0.5 – 0.8) were the $Acrophase_7$ between rem-man (z = -3.7, $p < 0.001$, SMD = 0.68) and $L5_{14}$ between rem-man (z = -11.3, $p < 0.001$, SMD = 0.51). For features estimated from both 7-day and 14-day windows, only the one with a larger effect size is presented here.

Table 8-1: Features (Diff) values during episodes and their strength in the classification of ASERT relapses in individual patients

Feature[+]	Feature values for relapses based on expert annotations (dataset 8.3.1) Median (IQR)			ASERT relapses (dataset 8.3.4) % of patients divided based on relapse predictability			
	Remission	Depression[‡‡]	Mania[‡‡]	Depression (AUC)		Mania (AUC)	
				<0.3 / >0.7[‡]	<0.5 / > 0.5	<0.3 / >0.7[‡]	<0.5 / > 0.5
Cosinor Analysis - 7 days estimation window							
Amplitude$_7$ (counts)	0.6 (38.7)	-6.4*** (47.5)	-4.1** (42.6)	**27% /3%**	68% /32%	2% /14%	33% /67%
Acrophase$_7$ (hours)	0.23 (3.15)	0.72* (3.95)	**-0.41$^{***/\#\#\#}$ (4.96)**	9% /11%	48% /52%	11% /9%	59% /41%
MESOR$_7$ (counts)	-0.7 (41.1)	***-14.5*** (51.3)***	***11.8$^{***/\#\#\#}$ (81.8)***	**43% /3%**	80% /20%	**2% /43%**	15% /85%
CQ$_7$ (Circadian Quotient)	0.00 (0.12)	0.02*** (0.12)	***-0.05$^{***/\#\#\#}$ (0.13)***	4% /13%	35% /65 %	**24% /6%**	71% /29%
MSE$_7$ (Mean Square Error of the fitted cosine)	-1987 (12608)	-5109*** (14821)	***5095$^{***/\#\#\#}$ (17477)***	**36% /6%**	69% /31%	**4% /32%**	25% /75%
GOF$_7$ (Goodness of Fit)	0.07 (5.64)	-0.40* (6.48)	-1.42$^{***/\#\#\#}$ (5.13)	19% /6%	62% /38%	9% /11%	48% /52%
Cosinor Analysis - 14 days estimation window							
Amplitude$_{14}$	-1.0 (32.1)	-3.7*** (39.8)	-6.7** (35.6)	***27% /2%***	70% /30%	8% /17%	37% /63%
Acrophase$_{14}$	0.33 (2.55)	***0.70 (4.05)***	***0.24* (3.97)***	10% /10%	43% /57%	12% /11%	56% /44%
MESOR$_{14}$	0.1 (39.4)	***-13.3*** (42.7)***	***18.7$^{***/\#\#\#}$ (76.7)***	**45% /4%**	72% /28%	**1% /35%**	16% /84%
CQ$_{14}$	0.00 (0.10)	0.01*** (0.10)	***-0.07$^{***/\#\#\#}$ (0.13)***	8% /15%	46% /54%	**31% /3%**	70% /30%
MSE$_{14}$	-1947 (10767)	-3985*** (14217)	***7217$^{***/\#\#\#}$ (17469)***	**32% /5**	68% /32%	**3% /34%**	26% /74%
GOF$_{14}$ (%)	-0.11 (4.50)	-0.23 (5.34)	***-1.18$^{***/\#\#\#}$ (4.55)***	19% /4%	67% /33%	11% /11%	50% /50%
Nonparametric circadian rhythm analysis (NPCRA) - 7 days estimation window							
IV$_7$ (interdaily variability)	-9.9E-3 (0.10)	-4.7E-3 (0.12)	8.0E-3** (0.12)	6% /21%	35% /65%	12% /2%	62% /38%
IS$_7$ (intradaily stability)	4.5E-3 (0.10)	2.4E-3 (0.10)	-16.0E-2$^{***/\#\#}$ (0.10)	12% /8%	61% /39%	9% /10%	46% /54%
M10$_7$ (most active 10 hours)	0.3 (60.7)	***-18.0*** (71.4)***	7.8$^{***/\#\#\#}$ (83.3)	**37% /3%**	74% /26%	**2% /32%**	17% /83%
M10-time$_7$ (M10$_7$ mid-time)	-0.01 (1.77)	0.07* (1.82)	0.13 (2.32)	6% /6%	46% /54%	9% /3%	54% /46%
L5$_7$ (Least active 5 hours)	-3.3 (13.6)	-7.0*** (19.6)	***3.5*** (33.1)***	***26% /3%***	68% /32%	**2% /33%**	27% /73%
L5-time$_7$ (L5$_7$ mid-time)	0.02 (1.00)	-0.22*** (1.18)	-0.02$^{\#}$ (1.47)	7% /10%	42% /58%	10% /6%	60% /40%
RA$_7$ (Relative Amplitude)	0.01 (0.06)	0.02* (0.07)	***-0.02$^{***/\#\#\#}$ (0.09)***	11% /11%	46% /54%	16% /7%	56 %/44%
NPCRA - 14 days estimation window							
IV$_{14}$	-1.3E-3 (0.08)	-5.6E-4 (0.11)	4.3E-3$^{**/\#}$ (0.08)	8% /21%	42% /58%	11% /1%	59% /41%
IS$_{14}$	2.4E-3 (0.09)	-3.0E-3 (0.08)	-18.0E-3$^{***/\#\#\#}$ (0.09)	19% /5%	67% /33%	10% /11%	50% /50%
M10$_{14}$	-1.0 (57.3)	***-12.7*** (63.6)***	***9.5$^{***/\#\#\#}$ (81.1)***	**39% /4%**	74% /26%	**2% /31%**	22% /78%
M10-time$_{14}$	-0.10 (1.54)	0.19*** (1.52)	0.05* (1.47)	12% /12%	46% /54%	6% /10%	56% /44%
L5$_{14}$	-2.8 (14.2)	-5.4*** (21.6)	**8.3$^{***/\#\#\#}$ (32.3)**	**27% /4%**	62% /38%	**2% /33%**	28% /72%
L5-time$_{14}$	0.02 (0.82)	-0.23*** (1.14)	-0.16*** (1.07)	12% /12%	44% /56%	8% /6%	56% /44%
RA$_{14}$	0.01 (0.06)	0.00 (0.07)	***-0.02$^{***/\#\#\#}$ (0.09)***	13% /15%	50% /50%	16% /4%	59% /41%
NPCRA - daily values							
M10	0.3 (100.8)	***-23.2*** (120.0)***	11.5$^{**/\#\#\#}$ (130.0)	***26% /2%***	77% /23%	2% /24%	22% /78%
M10-time	-0.16 (3.27)	-0.08 (3.00)	-0.08 (4.12)	4% /11%	39% /61%	6% /4%	56% /44%
L5	-3.0 (12.9)	-5.8*** (12.2)	***-2.4$^{\#\#\#}$ (15.9)***	13% /3%	66% /34%	0% /19%	27% /73%
L5-time	-0.10 (1.95)	-0.28** (2.18)	-0.03$^{\#\#}$ (2.05)	7% /3%	51% /49%	5% /1%	56% /44%
RA	0.01 (0.06)	0.02 (0.06)	0.01 (0.08)	8% /5%	56% /44%	6% /4%	49% /51%
RMSSD$_{M10}$ (in the M10 window)	-14.2 (332.1)	-28.9 (366.7)	-2.2 (345.7)	12% /2%	66% /34%	4% /4%	34% /66%
SD$_{M10}$	-70.8 (465.5)	-69.6 (457.7)	-60.5 (451.9)	9% /4%	57% /43%	6% /2%	52% /48%
Other nonparametric features – daily values							
ADA (Average daily activity)	-1.6 (67.0)	***-16.7*** (74.4)***	17.0$^{***/\#\#\#}$ (93.0)	**29% /3%**	78% /22%	**2% /31%**	18% /82%

Feature[+]	Remission	Depression[‡‡]	Mania[‡‡]	Depression (AUC)		Mania (AUC)	
				<0.3 / >0.7[‡]	<0.5 / > 0.5	<0.3 / >0.7[‡]	<0.5 / > 0.5
AQA_1 (Average activity between 0:00-6:00)	-11.6 (35.0)	-19.6*** (39.0)	***-7.1$^{*/\#\#\#}$ (75.0)***	14% /3%	69% /31%	1% /21%	28% /72%
AQA_2 (6:00-12:00)	9.9 (135.1)	***-22.0*** (142.9)***	6.4$^{\#\#\#}$ (192.2)	22% /3%	73% /27%	2% /24%	25% /75%
AQA_3 (12:00-18:00)	0.6 (127.0)	***-23.4*** (142.0)***	15.9$^{*/\#\#\#}$ (148.1)	25% /2%	70% /30%	2% /18%	23% /77%
AQA_4 (18:00-24:00)	-1.7 (113.7)	***-33.8*** (120.7)***	18.3$^{***/\#\#\#}$ (136.3)	18% /3%	69% /31%	2% /20%	27% /73%
DA_{high} (% of high activty)	-0.13 (7.51)	***-2.21*** (8.29)***	***1.28$^{***/\#\#\#}$ (10.65)***	***28% /1%***	79% /21%	***1% /27%***	22% /78%
$DA_{moderate}$ (% of moderate activty)	0.00 (6.05)	***-1.27*** (6.27)***	***1.65$^{***/\#\#\#}$ (6.71)***	14% /4%	64% /36%	3% /16%	31% /69%
$DA_{sedentary}$ (% of sedentary activty)	-0.23 (4.20)	-0.05 (4.18)	0.28* (6.17)	3% /11%	44% /56%	8% /4%	57% /43%
DA_{low} (% of low activty)	-0.78 (9.63)	***2.06*** (10.74)***	***-3.67$^{***/\#\#\#}$ (12.51)***	1% /23%	29% /71%	**30% /0%**	83% /17%
ACL^{++} (active day autocorelation 5 min lag)	0.004 (0.032)	***-0.001*** (0.041)***	0.008$^{*/\#\#\#}$ (0.037)	20% /4%	78% /22%	1% /18%	24% /76%
$RMSSD_{actday}$ (for active part of the day)	0.0 (31.0)	***-7.3*** (34.1)***	5.6$^{***/\#\#\#}$ (35.9)	23% /3%	71% /29%	0% /21%	26% /74%
Sleep based features - daily values							
SleOn (sleep onset)	-3.81 (4.63)	-4.66** (5.21)	-3.57$^{\#\#}$ (4.87)	11% /6%	62% /38%	1% /13%	30% /70%
Mid-sleep	-0.38 (1.54)	-0.49* (1.83)	-0.54 (2.16)	4% /8%	54% /46%	5% /5%	54% /46%
SleOFF (sleep offset)	-2.81 (3.63)	-2.73* (3.97)	-3.31$^{**/\#\#\#}$ (4.41)	4% /11%	41% /59%	17% /5%	69% /31%
SleDur (main daily – night - sleep duration)	-0.12 (2.32)	***0.39*** (2.66)***	***-0.92$^{***/\#\#\#}$ (2.70)***	2% /14%	32% /68%	***27% /2%***	77% /23%
$SleDur_{18}$ (sum of sleeps 18:00-18:00)	0.07 (2.42)	***0.75*** (2.87)***	***-0.84$^{***/\#\#\#}$ (2.98)***	3% /21%	22% /78%	***29% /0%***	80% /20%
$SleDur_{daily}$ (mid-night to midnight sum of sleeps)	0.10 (2.54)	***0.77*** (2.66)***	**-0.72$^{***/\#\#\#}$ (2.83)**	2% /20%	25% /75%	***25% /0%***	77% /23%
ISL (Immobile sleep)	0.02 (7.27)	0.26 (7.26)	0.65* (7.01)	4% /9%	49% /51%	7% /7%	53% /47%
RSL (Restless sleep)	-0.35 (1.57)	-0.32 (1.71)	-0.38 (1.53)	9% /11%	46% /54%	2% /6%	46% /54%
$RMSSD_{sleep}$	1.49 (36.02)	1.55 (34.94)	***-2.33$^{***/\#\#\#}$ (36.97)***	5% /4%	46% /54%	8% /6%	52% /48%
WASO (Wake After Sleep Onset)	-4.69 (12.05)	-4.45 (17.21)	-4.30$^{*/\#}$ (9.88)	4% /10%	47% /53%	6% /2%	52% /48%
APSO (Activity Prior Sleep Onset)	0.2 (132.8)	-18.7*** (138.8)	18.9$^{**/\#\#\#}$ (167.7)	8% /2%	58% /42%	4% /9%	29% /71%
AASO (Activity After Sleep Onset)	-3.4 (22.0)	-4.3 (21.8)	-5.3* (18.4)	9% /7%	50% /50%	4% /4%	52% /48%
APWU (Activity Prior Wake-Up)	-3.2 (21.4)	-2.4 (23.0)	-4.7$^{*/\#}$ (21.3)	5% /7%	44% /56%	5% /5%	51% /49%
AAWU (Activity After Wake-Up)	16.4 (122.6)	-12.1*** (153.8)	11.0$^{\#\#}$ (147.7)	17% /2%	73% /27%	4% /7%	30% /70%
APSO/AASO (sleep onset ratio %)	-0.05 (0.16)	-0.05 (0.19)	-0.07$^{***/\#\#\#}$ (0.14)	3% /3%	43% /57%	9% /2%	64% /36%
AAWU/APWU (sleep ofset ratio %)	-0.03 (0.11)	-0.03 (0.15)	-0.04 (0.11)	3% /7%	32% /68%	4% /3%	52% /48%
Explainable activity features - daily values							
ExAct	202.5 (422.2)	***107.3*** (446.9)***	301.7$^{***/\#\#\#}$ (510.4)	***26% /1%***	78% /22%	**2% /30%**	18% /82%
$ExAct_{active}$ (normlised by active day duration)	12.0 (22.8)	9.5*** (24.5)	15. 5$^{*/\#\#\#}$ (23.7)	18% /1%	71% /29%	2% /16%	26% /74%
Complexity analysis – entropy - daily values							
$SlopeEntropy_{M10}$ (in the M10 window)	0.31 (2.64)	0.55** (2.60)	0.07$^{\#\#}$ (2.82)	2% /13%	36% /64%	13% /0%	75% /25%

Statistical significance * < 0.05 ** <0.01 *** <0.001 for rem-dep and rem-man differences using Wilcoxon rank-sum test

Statistical significance $^{\#}$ < 0.05 $^{\#\#}$ <0.01 $^{\#\#\#}$ <0.001 for dep-man difference using Wilcoxon rank-sum test

[+]Feature calculations are described in Chapter 3 - section 3.5;

[‡‡] **Bold** values present remission-relapse differences with medium effect size (SMD 0.5 – 0.8) the ***bold italic*** presents small effect size (SMD 0.2 – 0.5), normal text presents values with SMD < 0.2.

[‡]In the ASERT relapses prediction part bold results represent features where ≥ 30 % patients have high predictive capability (AUC > 0.7 or < 0.3), and bold-italic where ≥ 25 % of patients have high predictive possibilities in one direction.

Features that differ during **remission** and **depression** with a small effect size (ordered by SMD) were:
DA_{high}[15] ($z = 8.5$, $p < 0.001$, SMD = 0.35), $MESOR_{14}$ ($z = 10.9$, $p < 0.001$, SMD = 0.33), DA_{low} ($z = -8.8$, $p < 0.001$, SMD = 0.32), ExAct ($z = 8.5$, $p < 0.001$, SMD = 0.31), $SleDur_{18}$ ($z = -8.4$, $p < 0.001$, SMD = 0.31), $SleDur_{daily}$ ($z = -8.3$, $p < 0.001$, SMD = 0.30), ADA ($z = 8.9$, $p < 0.001$, SMD = 0.30), $M10_7$ ($z = 8.7$, $p < 0.001$, SMD = 0.27), AQA_4 ($z = 7.8$, $p < 0.001$, SMD = 0.24), M10 ($z = 6.5$, $p < 0.001$, SMD = 0.22), $RMSSD_{daily}$ ($z = 7.3$, $p < 0.001$, SMD = 0.22), $SleDur_{night}$ ($z = -7.4$, $p < 0.001$, SMD = 0.22), AQA_2 ($z = 6.4$, $p < 0.001$, SMD = 0.21), $DA_{moderate}$ ($z = 5.7$, $p < 0.001$, SMD = 0.21), ACL ($z = 4.7$, $p < 0.001$, SMD = 0.21), and AQA_3 ($z = 5.6$, $p < 0.001$, SMD = 0.20).

The features with the largest differences between **remission** and **mania** were the already mentioned $Acrophase_7$ and $L5_{14}$. Those that reached at least a small effect size follows:
CQ14 ($z = 10.7$, $p < 0.001$, SMD = 0.45), RA_{14} ($z = 9.6$, $p < 0.001$, SMD = 0.41), $MESOR_{14}$ ($z = -8.2$, $p < 0.001$, SMD = 0.40), $SleDur_{daily}$ ($z = 7.1$, $p < 0.001$, SMD = 0.35), SleDur18 ($z = 7.1$, $p < 0.001$, SMD = 0.34), DA_{low} ($z = -8.8$, $p < 0.001$, SMD = 0.30), cos. MSE_{14} ($z = -9.9$, $p < 0.001$, SMD = 0.29), $DA_{moderate}$ ($z = 4.7$, $p < 0.001$, SMD = 0.24), $SleDur_{night}$ ($z = 7$, $p < 0.001$, SMD = 0.24), GOF ($z = 6.4$, $p < 0.001$, SMD = 0.23), AQA_1 ($z = -2.3$, $p = 0.02$, SMD = 0.22), $RMSSD_{sleep}$ ($z = 3.7$, $p < 0.001$, SMD = 0.22), $M10_{14}$ ($z = -5.2$, $p < 0.001$, SMD = 0.21), and DA_{high} ($z = -3.4$, $p < 0.001$, SMD = 0.20).

The differences between **depression** and **mania** were larger than differences between remission and relapses, as many of these differences had opposite directions. Therefore, few features achieved medium effect size. These features follow, ordered by SMD:
$MESOR_{14}$ ($z = 11.1$, $p < 0.001$, SMD = 0.66), $SleDur_{18}$ ($z = -10.6$, $p < 0.001$, SMD = 0.65), $SleDur_{daily}$ ($z = -10.7$, $p < 0.001$, SMD = 0.65), DA_{low} ($z = -10.3$, $p < 0.001$, SMD = 0.62), $L5_{14}$ ($z = 11.5$, $p < 0.001$, SMD = 0.57), DA_{high} ($z = 7.7$, $p < 0.001$, SMD = 0.55), $Acrophase_7$ ($z = 3.84$, $p < 0.001$, SMD = 0.54), CQ_{14} ($z = -10.9$, $p < 0.001$, SMD = 0.54), and ExAct ($z = 8.5$, $p < 0.001$, SMD = 0.50).

[15] For features name explanations see Table 8-1

8.3.3. Classification and Feature Selection

The results for both machine learning approaches (LRM and RF) are presented individually for each binary classification task (dep-man, dep-rem, and man-rem in Table 8-2). The overall performance of models is presented as average characteristics (accuracy, sensitivity, specificity, and AUC) obtained for all LOSO cross-validation test set (average performance over relevant patients).

Moreover, the performance for individual patients is presented based on the surrogates' analysis, mentioning the number of patients where the model reached significantly different performance ($p < 0.05$) than the surrogates. For LRM we have also included the count of the patients for whom the model performance would be improved by switching classes labels ($AUC < AUC_{surrog}$), therefore the episodes affect the activity profile in the opposite way.

Finally, for each model, we present the set of selected features. While for LRM, the removal of highly correlated features considerably improved the model performance, in the case of RF, the changes were negligible.

Table 8-2: Summary of final models evaluation global and individualised results

Model	dep-man‡	dep-rem‡	man-rem‡
Results of classification models			
Logistic Regression+	Accuracy: 0.61 Sensitivity: 0.65 Specificity: 0.55 AUC: 0.71	Accuracy: 0.61 Sensitivity: 0.60 Specificity: 0.52 AUC: 0.58	Accuracy: 0.61 Sensitivity: 0.49 Specificity: 0.64 AUC: 0.62
Random Forest	Accuracy: 0.63 Sensitivity: 0.86 Specificity: 0.38 AUC: 0.71	Accuracy: 0.57 Sensitivity: 0.43 Specificity: 0.65 AUC: 0.55	Accuracy: 0.70 Sensitivity: 0.53 Specificity: 0.79 AUC: 0.70
Results of surrogate analysis			
Logistic Regression	BTR: 7/10	BTR: 13/28	BTR: 9/15
Random Forest	BTR: 6/10	BTR: 12/28	BTR: 8/15

‡ the first mentioned episode type is taken as a positive class (sensitivity)
+as randomisation is used in the model training, the results varied by 3 % for all accuracy, sensitivity, and specificity
BTR means Better Than Random tested using surrogates on a 5 % significance level

Classification of Mania and Depression Episodes

For the dep-man classification, the LRM achieved an accuracy of 61 %, sensitivity of 65 %, and specificity of 55 %. Therefore, results were well balanced but only slightly better than guessing. For comparison, the RF achieved an accuracy of 63%, sensitivity of 86%, and 38% specificity. While the RF was better in detecting depression episodes, it was worse in detecting manic episodes. With a specificity lower than 50 %, the model predicts less than half of all manic episodes correctly. Both models reach the same AUC of 0.71. The LRM was better than

surrogates in 7 out of 10 patients (where 3 had AUC value lower than surrogates), and a similar result was obtained for the RF model, with 6 out of 10 patients being better than surrogates.

The LRM used 17 features (sorted by importance): $L5_{14}$, SUN, $M10_{7}$, RA_{14}, $Acrophase_{14}$, $L5\text{-}time_{7}$, $DA_{sedentary}$, $L5\text{-}time_{14}$, L5-time, $DA_{moderate}$, CQ_{14}, ISL, AQA_{4}, AQA_{1}, and AASO, while the RF used 15 features (with descending importance): $MESOR_{14}$ $SleDur_{18}$, SUN, CQ_{14}, IV_{14}, $M10_{7}$, $Amplitude_{14}$, $Acrophase_{14}$ $DA_{sedentary}$, $L5_{14}$, $L5\text{-}time_{14}$, ISL, $RMSSD_{daily}$, IS_{14}, and APSO.

Classification of Remission and Depression Episodes

In the classification of depression and remission in 28 patients, the LRM achieved an accuracy of 61 %, sensitivity of 60 %, and specificity of 52 % and AUC of 0.58. The RF achieved an accuracy of 57 %, sensitivity of 43 % and specificity of 65% and AUC 0.55. The resulting accuracies are very similar between the models, especially when we consider variability in LRM results by about 3%, caused by randomisation. The LRM is slightly better in detecting depression, while RF is better in the detection of remission.

The surrogate analysis of LRM found better than random distinguishing possibilities in 13 out of 28 patients (in 5 patients, the AUC was significantly lower than in surrogates). Similarly, for RF, in 12 out of 28 patients, it was easier to distinguish depressive days and remission days compared with the surrogates.

Selected features for LRM (in descending order descending) were: $Acrophase_{14}$, APSO, DA_{low}, $L5\text{-}time_{14}$, ADA, IS_{14}, MOON, DA_{high}, $M10_{14}$, IV_{14}, SleDur, CQ_{14}, $L5_{14}$, RA_{14}, and AAWU. Selected features for RF were: $MESOR_{7}$, $MESOR_{14}$, SUN, $Acrophase_{14}$, IS_{14}, RA_{14}, $L5\text{-}time_{14}$, $DA_{sedentary}$, $M10_{14}$, AQA_{1}, $L5\text{-}time_{7}$, CQ_{14}, IV_{14}, $Amplitude_{14}$, RA_{7}, $Acrophase_{7}$, APSO, $M10\text{-}time_{14}$, IS_{14}, $L5_{7}$, $Amplitude_{7}$, and SleDur.

Classification of remission and episodes of mania

Finally, for classifying mania and remission, the LRM achieved an accuracy of 61 %, sensitivity of 49 %, and specificity of 64 %., and AUC of 0.62. The RF model achieved an accuracy of 70 %, sensitivity of 53 %, specificity of 79 % and AUC of 0.70. In this task, therefore, RF clearly outperformed the LRM.

The LRM was successful in 9 out of 15 patients (with 2 having the AUC significantly lower than AUC_{surrog}). The RF was successful in 8 out of 15 patients. Both models were, therefore, better than random in more than half of the patients.

Selected features in this task for LRM (in descending order) were: L5-time$_{14}$, SUN, MOON, M10-time, SleDur$_{18}$, L5-time$_{7}$, AQA$_{2}$, ISL, DA$_{low}$, RMSSD$_{sleep}$, RA$_{7}$, L5$_{7}$, IV$_{14}$, IS$_{14}$, MESOR$_{7}$, DA$_{high}$, and SleDur The selected features for RF were: CQ$_{14}$, SleDur$_{daily}$, IV$_{7}$, IV$_{14}$, RA$_{14}$, DA$_{low}$, SUN, L5$_{14}$, SleDur$_{18}$, SleDur, RA$_{7}$, MESOR$_{7}$, RSL, L5$_{7}$, IS$_{14}$, MESOR$_{14}$, APWU, M10$_{7}$, M10-time$_{14}$, RMSSD$_{sleep}$, DA$_{sedentary}$, MOON, ISL, DA$_{moderate}$, SleOFF, M10-time$_{14}$.

8.3.4. Dataset Information - ASERTs

Individual feature's strength to recognise ASERT (section 4.2.2) relapses were studied using the whole ACTIBIPO 2 dataset (section 4.2). Out of 275 patients in the dataset, the inclusion criteria of at least five days with and without ASERT depression (sum of depressive and non-specific groups of ASERT score > 15) were fulfilled by 115 patients. The same criteria for ASERT mania (ASERT mania score > 5) were fulfilled by 129 patients.

8.3.5. Individual Features vs Subjective Relapses

The best features, considering at least moderate power (AUC > 0.7 or AUC < 0.3 – further referred as recognising ASERT states) for mood detection, were a successful predictor for more than 40 % of patients. These high success percentages were only observed in features estimated from windows (7- and 14-day). An example of a distribution of features associated with ASERT relapses in a randomly selected patient is shown in Figure 8.1.

The results for all features and all patients are presented in Table 8-1. Features that recognise ASERT relapses in the largest portion of patients are listed below. The percentage of patients with recognisable mood changes, the dominant feature change direction associated with the worsened state, and the percentage of patients with the change in this dominant direction are given for each feature, as well as the average AUC results for the dominant direction sub-group. When both 7-day and 14-day features have been successful, only the one with a higher percentage of patients with the recognisable state is presented.

For **depression,** MESOR$_{14}$ (48.7 %; decrease - 45.1 % [AUC 0.18), M10$_{14}$ (42.5 %; decrease - 39.3 % [AUC 0.18]), and cosinor MSE$_{7}$ (41.2 %; decrease - 36.0 % [AUC 0.20]), were the only three features that successful in more than 40 % of patients. Additionally, two features

were able to recognise ASERT depression relapse in 30+ % of patients: ADA (32.2 %; decrease - 28.7 % [AUC 0.21]), and $L5_{14}$ (31.0 %; decrease - 27.4 % [AUC 0.19]).

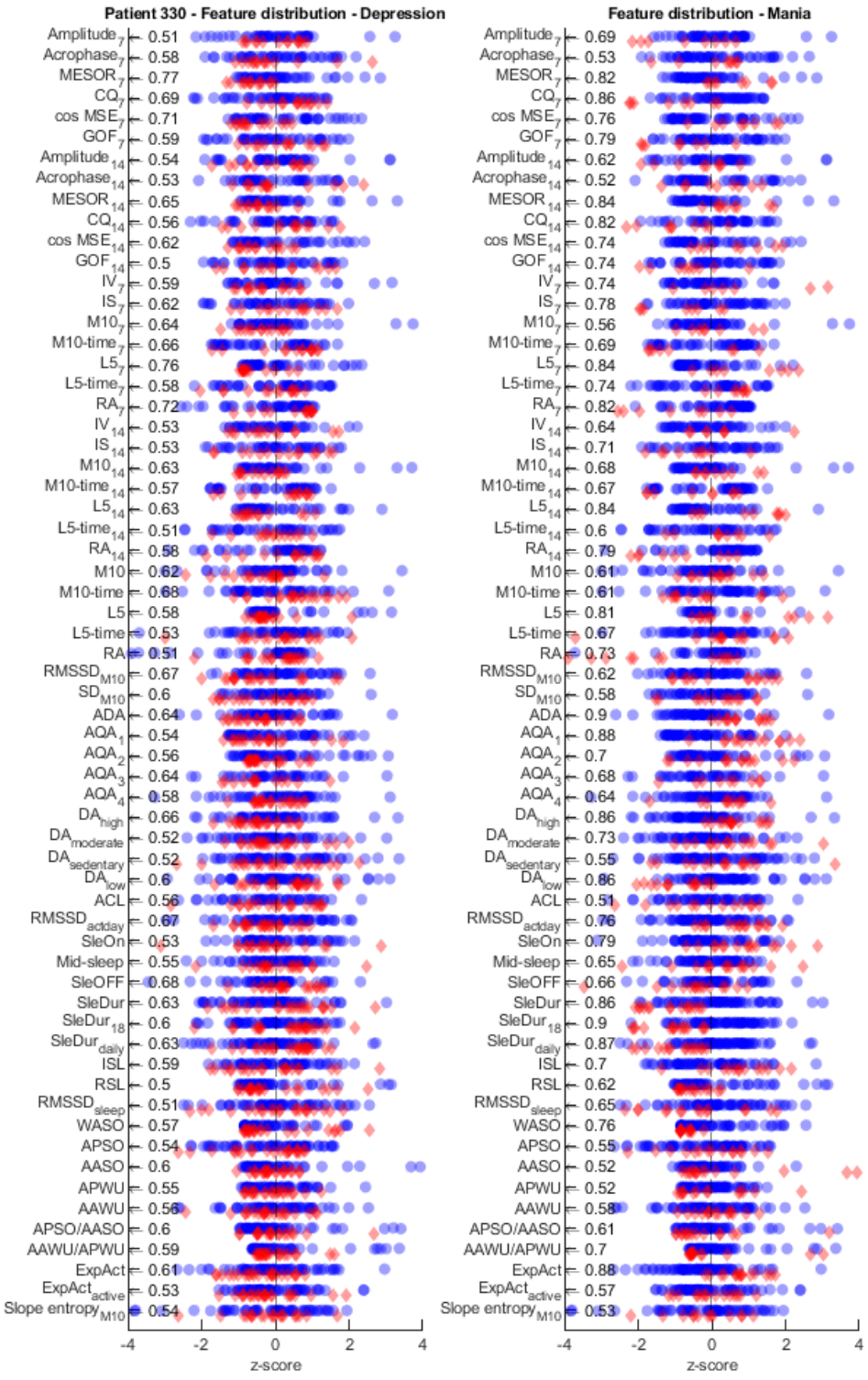


Figure 8.1 - Individual features distribution for ASERT relapses for patient ID330. The z-scores for features at the day of ASERT, the blue circles present non-relapse (depression in right and mania in left) ASERT score (for depression < 15 and for mania < 5), the red diamonds present ASERT relapses. Next to each feature name (for the explanation of feature names, see Table 8-1) is AUC for feature-based ASERT relapse classification.

For **mania**, only the $MESOR_7$ (44.9 %; increase - 43.3 % [AUC 0.82]) was successful in more than 40 % patients. Additionally, ASERT mania relapse was recognisable for 30+ % of patients in the following features:

MSE_{14} (36.0 %; increase - 33.9 % [AUC 0.78]), $L5_7$ (35.4 %; increase - 33.3 % [AUC 0.80]), CQ_{14} (34.4 %; lower 31.2 % [AUC 0.19]), $M10_7$ (33.9 %; increase - 32.3 % [AUC 0.80]), ADA (32.8 %; higher 31.5 % [AUC 0.81]), ExAct (31.0 %; increase - 30.5 % [AUC 0.80]), and DA_{low} (30.4 %; decrease - 30.4 % [AUC 0.19]).

Moreover, the following features had a highly specific association with increased manic symptoms, meaning the associated change was of one direction only:

$SleDur_{18}$ (decrease - 29.1 % [AUC 0.20]), $SleDur_{daily}$ (decrease - 24.8 % [AUC 0.20]), L5 (increase - 18.9 % [AUC 0.76]) and $SlopeEntropy_{M10}$ (decrease - 12.7 % [AUC 0.23]).

8.4. Discussion

We have evaluated state-induced physical activity changes, assessed by actigraphy, in BD outpatients. Additionally, we provided these actigraphy differences to machine learning models and trained them to distinguish between the remission and relapses of mania or depression. While the classification of relapses, based solely on actigraphic features, has already been assessed by Scott *et al.* (2017), our work is, to our knowledge, the first that explored the difference in outpatients rather than inpatients. The long-term actigraphy allowed us to measure activity during untreated (not hospitalised) relapses. Using these data, we were able to assess the differences between outpatient relapsed episodes and remission, which are more important for treatment supporting purposes. The relapses and remission differences in outpatients were also assessed by Cho *et al.* (2019), but their dataset was enriched by continuous heart-rate and illumination measurements.

Both of the models, which we used - the linear logistic regression (LRM) and nonlinear random forest (RF) - reached similar results for all three (dep-man, rem-dep, rem-man) classification tasks. The accuracy of distinction between relapses of **mania** and **depression** using a balanced dataset was approximately 60 %, with random forest having higher depression sensitivity (86 %) and lower specificity (38 %), and logistic regression with more balanced sensitivity (65 %) and specificity (55 %). For comparison, Scott *et al.* (2017) models reached an accuracy of 79 % on training data (and of 55 % using cross-validation) in

distinguishing depression, mania, and mixed states. The most commonly misclassified episode type in the Scotts' research was depression, where 42 % of episodes were classified as mania (the predominant episode type 15 to 12). The similarity between depression and manic episodes was also observed in our RF model, where many manic episodes were classified as depressions, as the predominant episode was depression (56 to 25). The differences in the ratio of mania and depression are probably caused by the fact that Scott was evaluating hospitalised patients. While depression is much more common in BD (Akiskal *et al.*, 2000; Látalová, 2010), it could be managed using a clinical approach, while manic episodes in many cases require hospitalisations.

The accuracy of differentiation between remission and relapses was around 60 % for both types of clinical episodes (slightly better – about 70 % - for RF and mania). These results are worse than approximately 80 % obtained by Cho *et al.* (2019). But as we had balanced our training sets, a comparison of relapse detection accuracies may be misleading. When we compare sensitivity instead, the LRM depression sensitivity of 60 % resembles the sensitivity achieved by Chos' model for BD-II patients (64 %), and it is higher than depression sensitivity for BD-I patients (25 %). Concerning days with mania (or hypomania), our RF model's sensitivity of 53 % is comparable to 70 % sensitivity for hypomania (BD-II) and 21 % sensitivity for mania (BD-I) achieved by Chos' models. There are two notable differences between our and Chos' approaches. Firstly, our binary classification included only remissions and relapses of one type, while Cho evaluated one type of episode days against all other data days. Second, our models were tested using unseen patients, while Chos' models were tested using unseen days. Unfortunately, the information about patients' BD type (BD-I and BD-II) was not yet known for our sample.

The results from both models are slightly worse than those from our previous study (Cuesta-Frau *et al.*, 2020), where an accuracy of about 70 % was achieved using the slope entropy feature only. As slope entropy wasn't selected for any of our models, we have to conclude that the drop in the accuracy is caused mainly by different length of segments for slope entropy estimation (the daily M10 segment vs the longest segment), and by the more demanding LOSO validation process used for the models in this study.

Concerning statistical differences in individual features as presented in Table 8-1, it has to be noted that while many differences were highly significant, the classification was not simple. Therefore, comparative studies (Krane-Gartiser *et al.*, 2014; Gershon *et al.*, 2016) would

benefit from incorporating at least a simple classifier to evaluate whether the differences are feasible for classification.

The statistical difference and feature selection by classifiers provided us with a large amount of information about typical behaviour changes during relapses. The most apparent change was in the overall activity (MESOR, also supported by ADA and ExAct), which was higher during mania and lowered during the depression. With Amplitude being lower for both depression and mania, combined with much shorter sleep in mania and longer in depression, the daily profile changes seem clear.

During the **depression,** the physical activity profile (see Figure 8.2) is low, starting with longer sleep. Then it increases slowly into the active part of the day, where it is significantly lower than during remission, as can be deduced from lowering of M10 features as well as in all parts of the day ($AQA_{1\text{-}4}$ and $ExAct_{active}$ – an activity score per active day hour). Still, the activity profile follows a similar daily profile (GOF is just slightly smaller, and CQ is higher) as in remission. These results correspond to lower overall activity, with later onset and low evening activity observed by Gershon *et al.* (2016). The sleep is longer but similar in quality to remission (ISL, RSL, $RMSSD_{sleep}$).

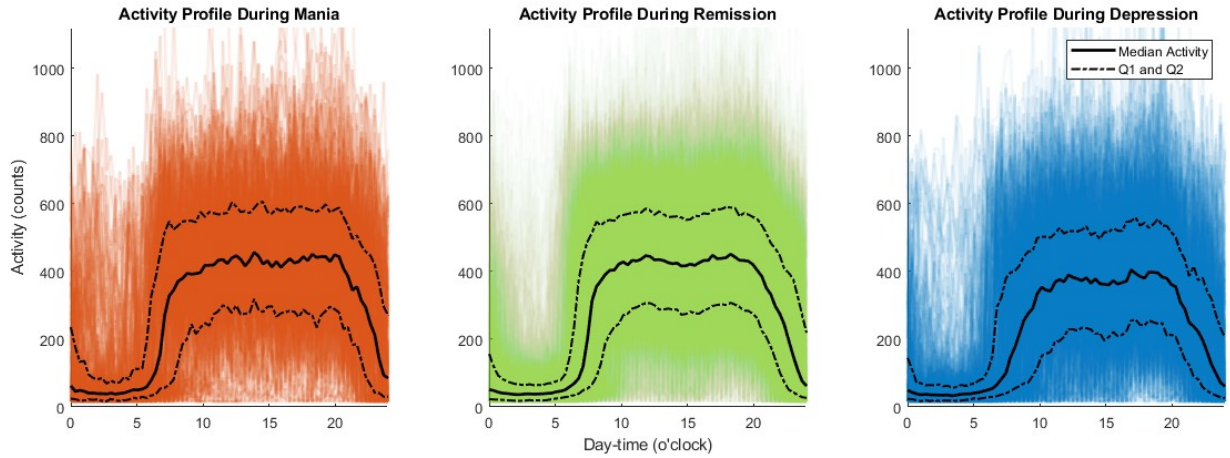


Figure 8.2 - Activity profiles during mania, remission, and depression

During **mania,** the activity profile (see Figure 8.2) is flat. In this case, it is caused by a much shorter sleep ($SleDur_{main}$, DA_{low}). Sleep shortening is caused mainly by earlier wake-up (SleOFF), as is supported by earlier Acrophase. During the day, the activity varies around high

values with probably multiple peaks, as is given by lowering of cosinor GOF and IS while increasing IV and $RMSSD_{actday}$. The daily profile (AQA_{1-4}, M10, L5) is increased, especially during the evening (AQA4, APSO). The physical activity increase in manias is not as large as its decrease observed during depressions. While the sleep is short, it is also less disturbed ($RMSSD_{sleep}$, AASO, APWU).

Contrary to our findings, Krane-Gartiser *et al.* (2014) didn't observe increased activity in hospitalised patients with mania in the evening. We also cannot confirm a decreased activity during the morning of BD patients in mania (AAWU). Our results only confirm that there was no significant increase of activity during two hours after wake-up (AAWU) in manic patients. Our results confirm a decreased activity in patients with depression during the morning, but our sample also showed their significantly decreased activity in the evening. The differences between our and Krane-Gartisers' samples and analyses are that, her volunteers were hospitalised patients, and the evaluated periods were 64 minutes during morning and evening (while AAWU and APSO were 120 min long), and that Krane-Gartiser compared the activities to HC while we compared it to patients' remission.

The obtained subset of features with significant changes between remission and relapses, as well as that selected by the models, enhance our knowledge about relapse manifestation in BD outpatients in detail that was not yet available. Interestingly, the frequency with which the models selected the day length (the SUN feature) supports the expected seasonality of BD relapses (Geoffroy, Bellivier, *et al.*, 2014; Bakstein *et al.*, 2019; Fellinger *et al.*, 2019).

The remission-depression differences include most of the differences observed between remitted BD patients and healthy controls (HC) (see Chapter 7). These differences were namely significant decreases in M10, ADA, and APSO, and increases in SleDur and CQ were observed. The only difference observed between BD patients and HC, which was not significant between remission and depression, was the AASO. These similarities would indicate, that the BD vs HC differences are at least partly nourished by persisting symptoms of depression.

While the described changes in physical activity associated with depression or mania seem clear, our analysis on the association between actigraphic features and mood shows that the changes occur only in a subset of patients. The results based on ASERT (section 4.2.2) self-evaluated mood showed that even the most significant differences are individually confirmed

only in approximately one-third[16] of the patients. For example, in the most consistently changing feature, the MESOR, the change was significantly (AUC > 0.7 or AUC < 0.3) associated with ASERT depression relapse in 49 % of patients and just 45 % in the expected direction. While AUC > 0.7 (or AUC < 0.3) could be reached in few patients by chance, globally, the results obtained from this analysis confirmed the significant differences from the statistical comparison. The LRM surrogate analysis also supports the hypothesis of untypical patients whose physical activity and circadian rhythm changes with mood are opposite to most BD patients.

Additionally, while there are features with only one significant direction of change for ASERT mania, for ASERT depression, there are always few patients where changes are opposite. These untypical patients or relapses could be caused by different subtypes of bipolar disorder or episodes of depression. This is consistent with a theory that divides bipolar depressions into two kinds: i) a more common depression characterised by hyporeactivity and motor retardation, and ii) agitated depression, which is rarer but riskier, as it is associated with higher suicidality prevalence (Akiskal, 2005; Henry *et al.*, 2007). The different types of bipolar depression (and mania) are also supported by Cho *et al.* (2019), where the predictability in BD-II was much better than in BD-I.

The high number of patients without any apparent activity change leads to the suggestion that both subtypes of BD depressions may evolve in one patient. Additionally, patients with 'stable' activity profiles may project their elevated moods into other activities. Matthews *et al.* (2017) conducted a survey of BD patients concentrating on their use of technology during self-perceived mood episodes, and found many links between the technology usage and mood state most often associated with overuse during manic-like states and reduced usage during the depressed mood (Matthews *et al.*, 2017). This also corresponds with typical changes of behaviour like goal-oriented attitude and lack of self-control in mania. When combined with technology, these may manifest in other ways, like playing computer games, excessive shopping, or increased socialisation evident from changes in instant messaging with friends (Urošević *et al.*, 2008). While such activities may be extreme, they would not affect the physical activity measured by actigraph.

Therefore, we expect that by exploring patients' and episodes' metadata, where the mood is associated with activity, we may recognise a sub-group of patients, where the activity clearly

[16] Similarly the mood stabilisation by lithium is effective in approximately one-third of patients (Hui *et al.*, 2019)

changes with mood. These patients would then benefit from automatic relapse detection. Moreover, such exploration could lead to finding new BD subtypes or episodes' subtypes. We believe that the addition of other passively collected features as keystrokes or voice analysis during calls (see section 2.4.2) could enlarge the subgroup where the episodes could be detected (Matthews *et al.*, 2017). While heart rate is also a promising feature (Cho *et al.*, 2019; Zebin, Peek and Casson, 2019), its collection using the same wearable would most probably limit the battery life. This would require a short charging period, which may lead to a significantly reduced amount of collected data, as is the case of Chos' study.

8.5. Limitations

Results of this study need to be interpreted considering the following limitations:

First, the presented results are based on the study that has just ended, and some additional information is still being collected. Especially datasets including used medications and additional meta-information like notable life-events data are not completed yet.

Second, although our sample size is more extensive than in most comparable studies (Krane-Gartiser *et al.*, 2014; Gershon *et al.*, 2016; Scott, Vaaler, *et al.*, 2017; Cho *et al.*, 2019), the final models are trained based on episodes from a few patients (10 for depression-mania comparison, 28 for depression-remission comparison, and 15 for mania-remission comparison). The number of patients with relapses is low, because relapses are relatively rare (especially mania). Moreover, we removed all hospitalisations, as well as episodes, where there was an excessive amount of missing data. While we enlarged the sample size by also using the ASERT relapse-induced changes in actigraphic features (this dataset includes more than 100 patients with at least 5 or more ASERT relapses and remission), results from this secondary analysis are primarily exploratory.

Third, our analysis didn't include mixed states as none of the patients fulfilled our relapse criteria for both mania and depression. Nonetheless, there were episodes with both elevated scores in MADRS and YMRS. Although the monthly collected scales were not used for treatment adjustment, their collection could already affect patients' self-perceived state.

Fourth, all patients were undergoing classical treatment by their caring physician. These treatments include different types of medication for which effects we used no correction. For

a more detailed discussion of the effects of medication on physical activity and sleep, see limitations section 7.5.

Fifth, while BD is typical by having many comorbidities (section 2.1), the possible effects of these comorbidities were not corrected. Comorbidities could cause some of the atypical changes observed among patients.

As above mentioned points three and four represent some of the highly probable causes of atypical changes in actigraphic features connected with BD relapses, they shall be the first to be explored once the dataset is completed.

8.6. Conclusions

To our knowledge, this is the first study that compared actigraphy data from BD outpatients during relapses and evaluated relapse classification possibilities validated on unseen patients. The relapses manifestation into BD patients' physical activity in their natural environment was described in finer detail than ever before. While the statistical differences between many relapse states are highly significant, the classification accuracies are only slightly above chance. A probable cause for this is the existence of different subtypes of BD patients or relapses, as the typical statistically significant changes in actigraphic features associated with relapses were found only in a subset of patients. Still, these observations are based on the evaluation of a limited number of relapses and are not compensated for used medication and other comorbidities. Therefore, future analyses have to be conducted to validate these results. Apriori recognition of these subtypes, if validated, could lead to a better understanding of the different manifestations of BD symptoms. Moreover, these patients and their physicians may largely benefit from treatment supporting tool, providing alerts for risky behaviour.

9. Summary and Future Research

In the presented thesis, I have participated in the collection of the largest actigraphy dataset from patients suffering from BD (Chapter 4). Based on the specifics of the long-term actigraphy data, I have proposed several new actigraphic features and suggested several updates and clarifications on estimating the non-parametric features (Chapter 3). Additionally, I have evaluated the validity of the features, describing circadian rhythmicity, when they are based on recordings affected by missing values (Chapter 5). Furthermore, I have suggested and validated several machine learning models applied to actigraphy based on sleep and circadian features in order to detect behavioural changes connected with BD (Chapter 7) and its symptomatic episodes (Chapter 8). Moreover, I have tested the feasibility and stability of chronotype estimation using actigraphy (Chapter 6). The changes in circadian rhythm and sleep parameters that were analysed, may deepen our knowledge of behavioural changes in BD and contribute to the distinction of clinical BD manifestation and its relapse episodes subtypes.

9.1. Thesis Contributions

- I have reimplemented the cosinor and non-parametric actigraphy circadian features (Chapter 3) while clarifying the estimation process of several non-parametric features and evaluating their robustness to missing data (Chapter 5). I have also added additional features that describe sleep onset and offset, and variability of activity during the day, which may contribute to the long-term actigraphy research. The features were used in a study evaluating the association of circadian rhythm with success in weight reduction program published in the Journal of *BioPsychoSocial Medicine* (IF(2019) = 0.9), with my considerable contribution. One of the added features - the explainable activity – was additionally trained to distinguish among different levels of activity. The explainable activity is incorporated as a part of the patients' micro-education system, which is going to be focused on in a clinical study performed by NIMH in 2021.
- A method of objective estimation of chronotype, and social jetlag was suggested, using actigraphic features. I have analysed the data in order to determine the required length

of actigraphy recording and stability of different chronotype-related actigraphic features. The results were then compared to widely used clinical chronotype questionnaires (MEQ and MCTQ). The results show that specific features, such as Acrophase are more suited for chronotyping. Duration of actigraphy observation was suggested for future chronotype research. We also pointed out the advantage of objective chronotyping, where we are showing that extreme questionnaire-based chronotype values are often inaccurately overestimated. The actigraphy-based chronotype stability was found similar to that of MCTQ-based chronotype, while it was substantially smaller than for the MEQ-based chronotype. The results presented in Chapter 6 are about to be submitted to an impacted journal *Sleep* (IF(2019) = 4.8).

- A method for classification of non-relapsed BD patients, and sex- and age-matched healthy controls (Chapter 7) was developed and evaluated on a dataset collected for this purpose, containing three months of continuous actigraphy. This is one of the longest continuous actigraphy recordings, which we have used mainly to evaluate the long-term variability in the circadian rhythm and sleep features. The model, which has been published in *CNS Spectrums* (IF(2019) = 3.4) (Schneider *et al.*, 2020), was able to classify the patients with 88 % accuracy (79 % using only features with low dependency on working status) using actigraphy alone. These results are better (resp. similar for low work status dependant features) than those achieved by similar studies.
- Several machine learning techniques were applied to actigraphy data in order to distinguish among BD episodes in the patients' natural environment (Chapter 8). In a study published in *Entropy* (IF(2019) = 2.5) (Cuesta-Frau *et al.*, 2020), with my considerable contribution, different entropy measures were tested. It was found that the slope entropy is best for distinguishing among patients' episodes. The approaches using actigraphic features alone, together with our suggested validation process, provided classification accuracy about 60 %, which is just slightly better than chance. In contrast, the statistical comparison of individual features was highly significant ($p < 0.001$) for many of them. The combination of these highly significant differences and low classification accuracy supports a theory suggesting the existence of multiple subtypes of relapses and possibly bipolar disorders, which should be assessed and tested in future work.

9.2. Future Work

In my future work, I would like to improve the classification of BD symptomatic episodes. The most promising source of improvement could be based on the clustering of clinical episodes (Cassidy, 2001; Akiskal, 2005; Henry *et al.*, 2007) and patients (Akiskal and Pinto, 1999; Akiskal *et al.*, 2000; Akiskal, 2002; Ghaemi, 2013) into subtypes either by using a combination of meta-information, such as clinical and sociodemographic data, actigraphic features, and both supervised (BD-I-VI) and unsupervised clustering techniques (beyond the clinical BD types). I would also like to apply the knowledge and methods achieved from this work to other mental and affective disorders. In addition to that, I would like to include other behavioural and physiological measurements.

I consider the main steps of my future work:

- To include corrections for pharmaco-therapy and comorbidities into models presented in Chapter 8, and publish it as a journal article.
- To explore the possibility of recognising different subtypes of depression and mania episodes (Cassidy, 2001; Akiskal, 2005; Henry *et al.*, 2007) from actigraphic features, and their typical changes during clinical episodes, and to include these episodes subtypes' into the models to improve episodes recognition possibilities.
- To explore the possibilities to discover/recognise different BD subtypes. In the (Cho *et al.*, 2019) study, there are clear differences between possibilities of episode prediction between BD-I and BD-II. Thus, it would be interesting to see whether the defined BD subtypes could be distinguished from actigraphy, and mainly from circadian rhythm alternation during clinical episodes, and to explore further possibilities of differentiation of the broad spectrum of bipolar disorders (Akiskal and Pinto, 1999; Ghaemi, 2013) using additional analysis and clustering of patients based on actigraphic features, medication, and other meta-information in order to discover new BD subtypes.
- To use the methods developed for BD patients to monitor and evaluate other affective disorders, primarily schizophrenia and borderline personality disorder. Some of the feedback provided during our analyses could also be used for the benefit of the healthy population, as it could help them to establish and maintain an active and regular lifestyle (Schneider, 2021).

- To include other digital phenotyping measures (see section 2.4.2) to improve the recognition of clinical episodes. Presently, we are working with students on implementation of Beiwe and MindLamp applications which are designed to provide holistic clinical platform for behavioural data monitoring with strong emphasis on data security and privacy. Additionally, a separate heart rate or similar vital sensors could be incorporated to access yet another promising source of information.

9.3. List of Candidate Publications

Autor contribution are given based on the V3S database. Citation counts according to ISI Web of Science (WoS) are valid as of 27th of April 2021.

9.3.1. Impacted Journals Publications Related to the Thesis

Schneider, J., Bakštein, E., Kolenič, M., Vostatek, P., Correl, Christoph U., Novák, D. and Španiel, F. (2020) 'Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls', ***CNS Spectrums***, pp. 1–11. doi: 10.1017/S1092852920001777.

IF(2019) = 3.356 , Q2 (51/155), Contribution: 50 %, 0 WoS citations

Cuesta-Frau, D., **Schneider, J.**, Bakštein, E., Vostatek, P., Španiel, F. and Novák, D. (2020) 'Classification of Actigraphy Records from Bipolar Disorder Patients Using Slope Entropy: A Feasibility Study', ***Entropy***, 22(11), p. 1243. doi: 10.3390/e22111243.

IF(2019) = 2,494, Q2 (33/85), Contribution: 17 % , 0 WoS citations

Fárková, E., **Schneider, J.**, Šmotek, M., Bakštein, E., Herlesová, J., Kopřivová, J., Šrámková, P., Pichlerová D., and Fried, M. (2019) 'Weight loss in conservative treatment of obesity in women is associated with physical activity and circadian phenotype: A longitudinal observational study', ***BioPsychoSocial Medicine***, 13(1), pp. 1–10. doi: 10.1186/s13030-019-0163-2.

IF(2019) = 0.904, Q3 (101/138), Contribution: 11 %, 0 WoS citations

9.3.2. Conference Reports Related to the Thesis

Fárková, E., Šmotek, M., Herlesová, J., **Schneider, J.,** Bakštein, E., and Kopřivová, J. (2018) 'The role of chronotype and sleep hygiene in the treatment of obesity.' *Journal of Sleep Research*. 316(2018), ISSN 1365-2869.

Contribution = 25 %, 0 WOS citations

Fárková, E., **Schneider, J.,** Bakštein, E., and Kopřivová, J. (2019) 'Objectivization of chrono-biological parameters using actigraphy', *Sleep Medicine*, 64(2019), p. S110. doi: 10.1016/j.sleep.2019.11.301.

Contribution= 25 %

9.3.3. Impacted Journal Publication and Selected Conference Reports Unrelated to the Thesis

Bakštein, E., Sieger, T., Wild, J., Novák, D., **Schneider, J.**, Vostatek, P., Urgošík, D. and Jech, R. (2017) 'Methods for automatic detection of artifacts in microelectrode recordings', ***Journal of Neuroscience Methods***, 290, pp. 39–51. doi: 10.1016/j.jneumeth.2017.07.012.

IF(2017) = 2.668, Q3 (154/261), Contribution: 5 %, 8 WoS citations

Schneider, J., Novak, D. and Jech, R. (2015) 'Optimization of Parkinson Disease treatment combining anti-Parkinson drugs and deep brain stimulation using patient diaries', *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem, pp. 3444–3447. doi: 10.1109/EMBC.2015.7319133.

Contribution: 90 %, 2 WoS citations

Bakštein, E., **Schneider, J.**, Sieger, T., Novák, D., Wild, J., and Jech, R. (2015) 'Supervised segmentation of microelectrode recording artifacts using power spectral density', *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem, pp. 1524–1527. doi: 10.1109/EMBC.2015.7318661.

Contribution: 35 %, 6 WOS citations

Smejkal, V., Sieger, L., Kodl, J. Novák, D. and **Schneider, J.** (2015) 'The dynamic biometric signature — Is the biometric data in the created signature constant?', in *2015 International Carnahan Conference on Security Technology (ICCST)*. IEEE, pp. 385–390. doi: 10.1109/CCST.2015.7389715.

Contribution: 20 %, 4 WoS citations

Smejkal, V., Kodl, J., Kodl, J. Jr., Novák, D., and **Schneider, J.** (2015) 'Strong Identification and Authentication Using Dynamic Biometric Signature', in *Lecture Notes in Electrical Engineering*, pp. 1245–1252. doi: 10.1007/978-3-662-45402-2_175.

Contribution: 20 %

Schneider, J. and Janča, R. (2013) 'Mutual Phase Spectrum Based Method for Epileptic Spike Tracking', *POSTER 2013 - 17th International Student Conference on Electrical Engineering*, pp. 1–5. Available at: http://radio.feld.cvut.cz/conf/poster2013/.

Contribution: 70 %

References

Abbott, S. M., Malkani, R. G. and Zee, P. C. (2020) 'Circadian disruption and human health: A bidirectional relationship', *European Journal of Neuroscience*, 51(1), pp. 567–583. doi: 10.1111/ejn.14298.

Acebo, C. *et al.* (1999) 'Estimating sleep patterns with activity monitoring in children and adolescents: how many nights are necessary for reliable measures?', *Sleep*, 22(1), pp. 95–103. doi: 10.1093/sleep/22.1.95.

Achtyes, E. D. *et al.* (2015) 'Validation of computerized adaptive testing in an outpatient nonacademic setting: The VOCATIONS trial', *Psychiatric Services*, 66(10), pp. 1091–1096. doi: 10.1176/appi.ps.201400390.

Adler, M. *et al.* (2008) 'Development and validation of the affective self rating scale for manic, depressive, and mixed affective states', *Nordic Journal of Psychiatry*, 62(2), pp. 130–135. doi: 10.1080/08039480801960354.

Aili, K. *et al.* (2017) 'Reliability of Actigraphy and Subjective Sleep Measurements in Adults: The Design of Sleep Assessments.', *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, 13(1), pp. 39–47. doi: 10.5664/jcsm.6384.

Akiskal, H. S. *et al.* (2000) 'Re-evaluating the prevalence of and diagnostic composition within the broad clinical spectrum of bipolar disorders', *Journal of Affective Disorders*, 59, pp. S5–S30. doi: 10.1016/S0165-0327(00)00203-2.

Akiskal, H. S. (2002) 'The bipolar spectrum--the shaping of a new paradigm in psychiatry.', *Current psychiatry reports*, 4(1), pp. 1–3. doi: 10.1007/s11920-002-0001-1.

Akiskal, H. S. (2005) 'The dark side of bipolarity: detecting bipolar depression in its pleomorphic expressions', *Journal of Affective Disorders*, 84(2–3), pp. 107–115. doi: 10.1016/j.jad.2004.06.003.

Akiskal, H. S. and Pinto, O. (1999) 'The evolving bipolar spectrum. Prototypes I, II, III, and IV.', *The Psychiatric clinics of North America*, 22(3), pp. 517–34, vii. doi: 10.1016/s0193-953x(05)70093-9.

Alloy, L. B. *et al.* (2017) 'Circadian Rhythm Dysregulation in Bipolar Spectrum Disorders', *Current Psychiatry Reports*. Current Psychiatry Reports, 19(4), p. 21. doi: 10.1007/s11920-017-0772-z.

Altman, D. G. and Bland, J. M. (1983) 'Measurement in Medicine: The Analysis of Method Comparison Studies', *The Statistician*, 32(3), p. 307. doi: 10.2307/2987937.

Altman, E. G. *et al.* (1994) 'The clinician-administered rating scale for mania (CARS-M): Development, reliability, and validity', *Biological Psychiatry*, 36(2), pp. 124–134. doi: 10.1016/0006-3223(94)91193-2.

Altman, E. G. *et al.* (1997) 'The altman self-rating Mania scale', *Biological Psychiatry*, 42(10), pp. 948–955. doi: 10.1016/S0006-3223(96)00548-3.

Ancoli-Israel, S. *et al.* (2003) 'The role of actigraphy in the study of sleep and circadian rhythms.', *Sleep*, 26(3), pp. 342–392.

Andreazza, A. C., Duong, A. and Young, L. T. (2018) 'Bipolar Disorder as a Mitochondrial Disease', *Biological Psychiatry*. Society of Biological Psychiatry, 83(9), pp. 720–721. doi: 10.1016/j.biopsych.2017.09.018.

Angst, J. (1998) 'The emerging epidemiology of hypomania and bipolar II disorder', *Journal of Affective Disorders*, 50(2–3), pp. 143–151. doi: 10.1016/S0165-0327(98)00142-6.

Anýž, J. *et al.* (2020) 'Validation of the Aktibipo SElf-RaTing (ASERT) questionnaire for digital self-assessment of mood and relapse-detection in bipolar disorder: Observational Study', *Preprint*. doi: 10.2196/preprints.26348.

APA (2013) *Diagnostic and statistical manual of mental disorders : DSM-5*, *American Psychiatric Association*. Arlington, VA Washington, D.C.: American Psychiatric Association.

Bakstein, E. *et al.* (2019) 'Cross-sectional and within-subject seasonality and regularity of hospitalizations – a population study in mood disorders and schizophrenia', *Bipolar Disorders*, pp. 1–9. doi: 10.1111/bdi.12884.

Baldessarini, R. J. *et al.* (2007) 'Effects of treatment latency on response to maintenance treatment in manic-depressive disorders', *Bipolar Disorders*, 9(4), pp. 386–393. doi: 10.1111/j.1399-5618.2007.00385.x.

Bandt, C. and Pompe, B. (2002) 'Permutation Entropy: A Natural Complexity Measure for Time Series', *Physical Review Letters*, 88(17), p. 174102. doi: 10.1103/PhysRevLett.88.174102.

Barnett, I. *et al.* (2018) 'Relapse prediction in schizophrenia through digital phenotyping: a pilot study', *Neuropsychopharmacology*. Springer US, 43(8), pp. 1660–1666. doi: 10.1038/s41386-018-0030-z.

Barrigón, M. L. *et al.* (2017) 'User profiles of an electronic mental health tool for ecological momentary assessment: MEmind', *International Journal of Methods in Psychiatric Research*, 26(1), p. e1554. doi: 10.1002/mpr.1554.

Bass, J. (2012) 'Circadian topology of metabolism', *Nature*, 491(7424), pp. 348–356. doi: 10.1038/nature11704.

Bäuml, J. *et al.* (2006) 'Psychoeducation: A basic psychotherapeutic intervention for patients with schizophrenia and their families', *Schizophrenia Bulletin*, 32(SUPPL.1), pp. 1–9. doi: 10.1093/schbul/sbl017.

Bech, P., Rafaelsen, O. J., Kramp, P., & Bolwig, T. G. (1974) 'The Mania Rating Scale : Scale Construction and Inter-observer Agreement', *British Association for Psychopharmacology*, 11, pp. 430–431.

Bei, B. *et al.* (2016) 'Beyond the mean: A systematic review on the correlates of daily intraindividual variability of sleep/wake patterns', *Sleep Medicine Reviews*. Elsevier Ltd, 28, pp. 108–124. doi: 10.1016/j.smrv.2015.06.003.

Beiwinkel, T. *et al.* (2016) 'Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study', *JMIR Mental Health*, 3(1), p. e2. doi: 10.2196/mental.4560.

Bellivier, F. *et al.* (2015) 'Sleep- and circadian rhythm–associated pathways as therapeutic targets in bipolar disorder', *Expert Opinion on Therapeutic Targets*, 19(6), pp. 747–763. doi: 10.1517/14728222.2015.1018822.

Bellone, G. J. *et al.* (2016) 'Comparative analysis of actigraphy performance in healthy young subjects', *Sleep Science*. Elsevier, 9(4), pp. 272–279. doi: 10.1016/j.slsci.2016.05.004.

Berk, M. *et al.* (2007) 'The Bipolar Depression Rating Scale (BDRS): Its development, validation and utility', *Bipolar Disorders*, 9(6), pp. 571–579. doi: 10.1111/j.1399-5618.2007.00536.x.

Bernstein, I. H. *et al.* (2010) 'The quick inventory of depressive symptomatology (clinician and self-report versions) in patients with bipolar disorder', *CNS Spectrums*, 15(6), pp. 367–373. doi: 10.1017/S1092852900029230.

Born, C. *et al.* (2014) 'Saving time and money: A validation of the self ratings on the prospective NIMH life-chart method (NIMH-LCM)', *BMC Psychiatry*, 14(1), pp. 1–7. doi: 10.1186/1471-244X-14-130.

Breiman, L. (2001) ‘Random Forests’, in. Berkley: Statistics Department University of California, pp. 1–33. doi: doi.org/10.1023/A:1010933404324.

Bullock, B. and Murray, G. (2014) ‘Reduced Amplitude of the 24 Hour Activity Rhythm’, *Clinical Psychological Science*, 2(1), pp. 86–96. doi: 10.1177/2167702613490158.

Calogiuri, G., Weydahl, A. and Roveda, E. (2011) ‘Effects of Sleep Loss and Strenuous Physical Activity on the Rest-Activity Circadian Rhythm: A Study on 500 km and 1,000 km Dogsled Racers’, *Biological Research For Nursing*, 13(4), pp. 409–418. doi: 10.1177/1099800410392021.

Cao, B. *et al.* (2017) ‘DeepMood: Modeling mobile phone typing dynamics for mood detection’, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1296, pp. 747–755. doi: 10.1145/3097983.3098086.

Carr, O. *et al.* (2018) ‘Variability in phase and amplitude of diurnal rhythms is related to variation of mood in bipolar and borderline personality disorder’, *Scientific Reports*. Springer US, 8(1), pp. 1–11. doi: 10.1038/s41598-018-19888-9.

Cassidy, F. (2001) ‘Subtypes of Mania Determined by Grade of Membership Analysis’, *Neuropsychopharmacology*, 25(3), pp. 373–383. doi: 10.1016/S0893-133X(01)00223-8.

Cellini, N. *et al.* (2013) ‘Direct comparison of two actigraphy devices with polysomnographically recorded naps in healthy young adults’, *Chronobiology International*, 30(5), pp. 691–698. doi: 10.3109/07420528.2013.782312.

Cerimele, J. M. *et al.* (2019) ‘Systematic Review of Symptom Assessment Measures for Use in Measurement-Based Care of Bipolar Disorders’, *Psychiatric Services*, 70(5), pp. 396–408. doi: 10.1176/appi.ps.201800383.

Chakrabarti, S. (2016) ‘Treatment-adherence in bipolar disorder: A patient-centred approach’, *World Journal of Psychiatry*, 6(4), p. 399. doi: 10.5498/wjp.v6.i4.399.

Cho, C. H. *et al.* (2019) ‘Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study’, *Journal of Medical Internet Research*, 21(4). doi: 10.2196/11029.

Claes, J. *et al.* (2017) ‘Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities’, *Journal of Medical Engineering and Technology*. Informa UK Ltd., 41(6), pp. 480–485. doi: 10.1080/03091902.2017.1333166.

Colombo, C., Fossati, A. and Colom, F. (2012) ‘Bipolar disorder’, *Depression Research and Treatment*, 2012. doi: 10.1155/2012/525837.

Cornelissen, G. (2014) ‘Cosinor-based rhythmometry’, *Theoretical Biology and Medical Modelling*. Theoretical Biology and Medical Modelling, 11(1), p. 16. doi: 10.1186/1742-4682-11-16.

De Crescenzo, F. *et al.* (2017) ‘Actigraphic features of bipolar disorder: A systematic review and meta-analysis’, *Sleep medicine reviews*. Elsevier Ltd, 33, pp. 58–69. doi: 10.1016/j.smrv.2016.05.003.

Crowley, S. J. *et al.* (2006) ‘Estimating Dim Light Melatonin Onset (DLMO) Phase in Adolescents Using Summer or School-Year Sleep/Wake Schedules’, *Sleep*, 29(12), pp. 1632–1641. doi: 10.1093/sleep/29.12.1632.

Crowley, S. J. *et al.* (2016) ‘Estimating the dim light melatonin onset of adolescents within a 6-h sampling window: the impact of sampling rate and threshold method’, *Sleep Medicine*, 20(3), pp. 59–66. doi: 10.1016/j.sleep.2015.11.019.

Cuesta-Frau, D. *et al.* (2019) ‘Influence of Duodenal–Jejunal Implantation on Glucose Dynamics: A Pilot Study Using Different Nonlinear Methods’, *Complexity*, 2019, pp. 1–10. doi: 10.1155/2019/6070518.

Cuesta-Frau, D. (2019) 'Slope Entropy: A New Time Series Complexity Estimator Based on Both Symbolic Patterns and Amplitude Information', *Entropy*, 21(12), p. 1167. doi: 10.3390/e21121167.

Cuesta-Frau, D. *et al.* (2020) 'Classification of Actigraphy Records from Bipolar Disorder Patients Using Slope Entropy: A Feasibility Study', *Entropy*, 22(11), p. 1243. doi: 10.3390/e22111243.

Denicoff, K. D. *et al.* (2000) 'Validation of the prospective NIMH-life-chart method (NIMH-LCM(TM)-p) for longitudinal assessment of bipolar illness', *Psychological Medicine*, 30(6), pp. 1391–1397. doi: 10.1017/S0033291799002810.

Dennehy, E. B. *et al.* (2004) 'Development of the Brief Bipolar Disorder Symptom Scale for patients with bipolar disorder', *Psychiatry Research*, 127(1–2), pp. 137–145. doi: 10.1016/j.psychres.2004.02.009.

Dome, Rihmer and Gonda (2019) 'Suicide Risk in Bipolar Disorder: A Brief Review', *Medicina*, 55(8), p. 403. doi: 10.3390/medicina55080403.

Dong, M. *et al.* (2019) 'Prevalence of suicide attempts in bipolar disorder: A systematic review and meta-analysis of observational studies', *Epidemiology and Psychiatric Sciences*. doi: 10.1017/S2045796019000593.

Fadlallah, B. *et al.* (2013) 'Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information', *Physical Review E*, 87(2), p. 022911. doi: 10.1103/PhysRevE.87.022911.

Faedda, G. L. *et al.* (2016) 'Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls', *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 57(6), pp. 706–716. doi: 10.1111/jcpp.12520.

Fagiolini, A. *et al.* (2003) 'Obesity as a correlate of outcome in patients with bipolar I disorder', *American Journal of Psychiatry*, 160(1), pp. 112–117. doi: 10.1176/appi.ajp.160.1.112.

Fárková, E. *et al.* (2019) 'Weight loss in conservative treatment of obesity in women is associated with physical activity and circadian phenotype: A longitudinal observational study', *BioPsychoSocial Medicine*, 13(1), pp. 1–10. doi: 10.1186/s13030-019-0163-2.

Fárková, E. *et al.* (2020) 'Comparison of Munich Chronotype Questionnaire (MCTQ) and Morningness-Eveningness Questionnaire (MEQ) Czech version', *Chronobiology International*. Taylor & Francis, 37(11), pp. 1591–1598. doi: 10.1080/07420528.2020.1787426.

Faurholt-Jepsen, M., Busk, J., *et al.* (2019) 'Objective smartphone data as a potential diagnostic marker of bipolar disorder', *Australian and New Zealand Journal of Psychiatry*, 53(2), pp. 119–128. doi: 10.1177/0004867418808900.

Faurholt-Jepsen, M., Geddes, J. R., *et al.* (2019) 'Reporting guidelines on remotely collected electronic mood data in mood disorder (eMOOD)—recommendations', *Translational Psychiatry*. Springer US, 9(1), p. 162. doi: 10.1038/s41398-019-0484-8.

FDA (2006) 'Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance.', *Health and quality of life outcomes*, 4, p. 79. doi: 10.1186/1477-7525-4-79.

Fellinger, M. *et al.* (2019) 'Seasonality in bipolar disorder: Effect of sex and age', *Journal of Affective Disorders*, 243(June 2018), pp. 322–326. doi: 10.1016/j.jad.2018.09.073.

Frank, E. (2007) 'Interpersonal and social rhythm therapy: A means of improving depression and preventing relapse in bipolar disorder', *Journal of Clinical Psychology*, 63(5), pp. 463–473. doi: 10.1002/jclp.20371.

Geoffroy, P. A., Etain, B., *et al.* (2014) 'Circadian biomarkers in patients with bipolar disorder: promising putative predictors of lithium response', *International Journal of Bipolar Disorders*, 2(1),

pp. 5–8. doi: 10.1186/2194-7511-2-5.

Geoffroy, P. A., Bellivier, F., *et al.* (2014) 'Seasonality and bipolar disorder: A systematic review, from admission rates to seasonality of symptoms', *Journal of Affective Disorders*. Elsevier, 168, pp. 210–223. doi: 10.1016/j.jad.2014.07.002.

Geoffroy, P. A., Boudebesse, C., *et al.* (2014) 'Sleep in remitted bipolar disorder: A naturalistic case-control study using actigraphy', *Journal of Affective Disorders*. Elsevier, 158, pp. 1–7. doi: 10.1016/j.jad.2014.01.012.

Geoffroy, P. A. *et al.* (2015) 'Sleep in patients with remitted bipolar disorders: A meta-analysis of actigraphy studies', *Acta Psychiatrica Scandinavica*, 131(2), pp. 89–99. doi: 10.1111/acps.12367.

Gershon, A. *et al.* (2012) 'Restless Pillow, Ruffled Mind: Sleep and Affect Coupling in Interepisode Bipolar Disorder.', *Journal of Abnormal Psychology*, 121(4), pp. 863–873. doi: 10.1037/a0028233.

Gershon, A. *et al.* (2016) 'Daily Actigraphy Profiles Distinguish Depressive and Interepisode States in Bipolar Disorder', *Clinical Psychological Science*, 4(4), pp. 641–650. doi: 10.1177/2167702615604613.

Gershon, A. *et al.* (2018) 'Subjective versus objective evening chronotypes in bipolar disorder', *Journal of Affective Disorders*. Elsevier B.V., 225(June 2017), pp. 342–349. doi: 10.1016/j.jad.2017.08.055.

Ghaemi, S. N. (2013) 'Bipolar Spectrum: A Review of the Concept and a Vision for the Future', *Psychiatry Investigation*, 10(3), p. 218. doi: 10.4306/pi.2013.10.3.218.

Gideon, J., Provost, E. M. and McInnis, M. (2016) 'Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder', in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2359–2363. doi: 10.1109/ICASSP.2016.7472099.

Gilgen-Ammann, R., Schweizer, T. and Wyss, T. (2020) 'Accuracy of distance recordings in eight positioning-enabled sport watches: Instrument validation study', *JMIR mHealth and uHealth*, 8(6). doi: 10.2196/17118.

Gold, A. and Sylvia, L. (2016) 'The role of sleep in bipolar disorder', *Nature and Science of Sleep*, Volume 8, pp. 207–214. doi: 10.2147/NSS.S85754.

Gonçalves, B. S. B. *et al.* (2014) 'Nonparametric methods in actigraphy: An update', *Sleep Science*, 7(3), pp. 158–164. doi: 10.1016/j.slsci.2014.09.013.

Gonçalves, B. S. B. *et al.* (2015) 'A fresh look at the use of nonparametric analysis in actimetry', *Sleep Medicine Reviews*. Elsevier Ltd, 20, pp. 84–91. doi: 10.1016/j.smrv.2014.06.002.

González-Pinto, A. *et al.* (2009) 'Validity and reliability of the Hamilton Depression Rating Scale (5 items) for manic and mixed bipolar disorders', *Journal of Nervous and Mental Disease*, 197(9), pp. 682–686. doi: 10.1097/NMD.0b013e3181b3b3a0.

Gonzalez, J. M. *et al.* (2008) 'Development of the Bipolar Inventory of Symptoms Scale: concurrent validity, discriminant validity and retest reliability', *International Journal of Methods in Psychiatric Research*, 17(4), pp. 198–209. doi: 10.1002/mpr.262.

Gonzalez, R. *et al.* (2018) 'The association between mood state and chronobiological characteristics in bipolar I disorder: a naturalistic, variable cluster analysis-based study', *International Journal of Bipolar Disorders*. Springer Berlin Heidelberg, 6(1), p. 5. doi: 10.1186/s40345-017-0113-5.

Grierson, A. B. *et al.* (2016) 'Circadian rhythmicity in emerging mood disorders: state or trait marker?', *International journal of bipolar disorders*. Springer Berlin Heidelberg, 4(1), p. 3. doi: 10.1186/s40345-015-0043-z.

Gruber, J. *et al.* (2009) 'Sleep functioning in relation to mood, function, and quality of life at entry to the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD)', *Journal of Affective Disorders*. Elsevier B.V., 114(1–3), pp. 41–49. doi: 10.1016/j.jad.2008.06.028.

Grünerbl, A. *et al.* (2015) 'Smartphone-based recognition of states and state changes in bipolar disorder patients', *IEEE Journal of Biomedical and Health Informatics*, 19(1), pp. 140–148. doi: 10.1109/JBHI.2014.2343154.

Gupta, S. and Pati, A. K. (1994) 'Characteristics of circadian rhythm in six variables of morning active and evening active healthy human subjects.', *Indian journal of physiology and pharmacology*, 38(2), pp. 101–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1027738.

Gustavsson, A. *et al.* (2011) 'Cost of disorders of the brain in Europe 2010', *European Neuropsychopharmacology*. Elsevier B.V., 21(10), pp. 718–779. doi: 10.1016/j.euroneuro.2011.08.008.

Harvey, A. G. *et al.* (2005) 'Sleep-related functioning in euthymic patients with bipolar disorder, patients with insomnia, and subjects without sleep problems', *American Journal of Psychiatry*, 162(1), pp. 50–57. doi: 10.1176/appi.ajp.162.1.50.

Hastie, T., Tibshirani, R. and Friedman, J. (2017) *The Elements of Statistical Learning The Elements of Statistical Learning*. doi: 10.1198/jasa.2004.s339.

Henriksen, A. *et al.* (2020) 'Validity of the polar M430 activity monitor in free-living conditions: Validation study', *JMIR Formative Research*, 3(3). doi: 10.2196/14438.

Henry, C. *et al.* (2007) 'Evidence for Two Types of Bipolar Depression Using a Dimensional Approach', *Psychotherapy and Psychosomatics*, 76(6), pp. 325–331. doi: 10.1159/000107559.

Hernando, D. *et al.* (2018) 'Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects', *Sensors (Switzerland)*, 18(8). doi: 10.3390/s18082619.

Hlaváč, V. (2020) *Actigraphy Data Analysis in Bipolar Disorder Patients*. Czech Technical University in Prague.

Hofman, M. A. (1950) 'Human Circadian Timing System', in *Encyclopedia of Neuroscience*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1869–1873. doi: 10.1007/978-3-540-29678-2_2264.

Holm, S. (1979) 'Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure', *Source: Scandinavian Journal of Statistics Scand J Statist*, 6(6), pp. 65–70.

Honma, K. *et al.* (1992) 'Seasonal variation in the human circadian rhythm: dissociation between sleep and temperature rhythm', *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 262(5), pp. R885–R891. doi: 10.1152/ajpregu.1992.262.5.R885.

Horne, J. A. and Ostberg, O. (1976) 'A self assessment questionnaire to determine Morningness Eveningness in human circadian rhythms', *International Journal of Chronobiology*, pp. 97–110.

Hossain, S. *et al.* (2019) 'Medical and Psychiatric Comorbidities in Bipolar Disorder: Insights from National Inpatient Population-based Study', *Cureus*, 11(9). doi: 10.7759/cureus.5636.

Hsin, H. *et al.* (2018) 'Transforming Psychiatry into Data-Driven Medicine with Digital Measurement Tools', *npj Digital Medicine*. Springer US, 1(1), p. 37. doi: 10.1038/s41746-018-0046-0.

Huang, C. L. *et al.* (2003) 'Patient- and family-rated scale for bipolar disorder symptoms: Internal state scale', *Kaohsiung Journal of Medical Sciences*, 19(4), pp. 170–175. doi: 10.1016/s1607-551x(09)70467-x.

Hui, T. P. *et al.* (2019) 'A systematic review and meta-analysis of clinical predictors of lithium response in bipolar disorder', *Acta Psychiatrica Scandinavica*, 140(2), pp. 94–115. doi:

10.1111/acps.13062.

Hwang, J. Y. *et al.* (2017) 'Comparison of the effects of quetiapine XR and lithium monotherapy on actigraphy-measured circadian parameters in patients with bipolar II depression', *Journal of Clinical Psychopharmacology*, 37(3), pp. 351–354. doi: 10.1097/JCP.0000000000000699.

Jacobson, N. C., Summers, B. and Wilhelm, S. (2020) 'Digital biomarkers of social anxiety severity: Digital phenotyping using passive smartphone sensors', *Journal of Medical Internet Research*, 22(5), pp. 1–10. doi: 10.2196/16875.

Janney, C. A. *et al.* (2014) 'Are adults with bipolar disorder active? Objectively measured physical activity and sedentary behavior using accelerometry', *Journal of Affective Disorders*. Elsevier, 152–154(1), pp. 498–504. doi: 10.1016/j.jad.2013.09.009.

Jarque, C. M. and Bera, A. K. (1987) 'A Test for Normality of Observations and Regression Residuals', *International Statistical Review / Revue Internationale de Statistique*, 55(2), p. 163. doi: 10.2307/1403192.

Jean-Louis, G. *et al.* (2000) 'Sleep duration, illumination, and activity patterns in a population sample: effects of gender and ethnicity.', *Biological psychiatry*, 47(10), pp. 921–7. doi: 10.1016/s0006-3223(99)00169-9.

Jones, S. H., Hare, D. J. and Evershed, K. (2005) 'Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder', *Bipolar Disorders*, 7(2), pp. 176–186. doi: 10.1111/j.1399-5618.2005.00187.x.

Juda, M., Vetter, C. and Roenneberg, T. (2013a) 'Chronotype Modulates Sleep Duration, Sleep Quality, and Social Jet Lag in Shift-Workers', *Journal of Biological Rhythms*, 28(2), pp. 141–151. doi: 10.1177/0748730412475042.

Juda, M., Vetter, C. and Roenneberg, T. (2013b) 'The Munich ChronoType Questionnaire for Shift-Workers (MCTQ Shift)', *Journal of Biological Rhythms*, 28(2), pp. 130–140. doi: 10.1177/0748730412475041.

Judd, L. L. (2002) 'The Long-term Natural History of the Weekly Symptomatic Status of Bipolar I Disorder', *Archives of General Psychiatry*, 59(6), pp. 530–537. doi: 10.1001/archpsyc.59.6.530.

Judd, L. L. *et al.* (2003) 'Long-term symptomatic status of bipolar I vs. bipolar II disorders', *International Journal of Neuropsychopharmacology*, 6(2), pp. 127–137. doi: 10.1017/S1461145703003341.

Judd, L. L. and Akiskal, H. S. (2003) 'The prevalence and disability of bipolar spectrum disorders in the US population: re-analysis of the ECA database taking into account subthreshold cases', *Journal of Affective Disorders*, 73(1–2), pp. 123–131. doi: 10.1016/S0165-0327(02)00332-4.

Kaplan, K. A. *et al.* (2012) 'Evaluating sleep in bipolar disorder: Comparison between actigraphy, polysomnography, and sleep diary', *Bipolar Disorders*, 14(8), pp. 870–879. doi: 10.1111/bdi.12021.

Karam, Z. N. *et al.* (2014) 'Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, pp. 4858–4862. doi: 10.1109/ICASSP.2014.6854525.

Kaufmann, C. N. *et al.* (2018) 'Daytime midpoint as a digital biomarker for chronotype in bipolar disorder', *Journal of Affective Disorders*. Elsevier B.V., 241(August), pp. 586–591. doi: 10.1016/j.jad.2018.08.032.

Kessing, L. V. *et al.* (2015) 'Causes of decreased life expectancy over the life span in bipolar disorder', *Journal of Affective Disorders*. Elsevier, 180, pp. 142–147. doi: 10.1016/j.jad.2015.03.027.

Khan, A. and Anwar, Y. (2019) 'Framework to Predict Bipolar Episodes', in. Springer International Publishing, pp. 412–425. doi: 10.1007/978-3-030-01057-7_33.

Kosmadopoulos, A. *et al.* (2014) 'Alternatives to polysomnography (PSG): A validation of wrist actigraphy and a partial-PSG system', *Behavior Research Methods*, 46(4), pp. 1032–1041. doi: 10.3758/s13428-013-0438-7.

Krane-Gartiser, K. *et al.* (2014) 'Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder', *PLoS ONE*, 9(2). doi: 10.1371/journal.pone.0089574.

Krane-Gartiser, K. *et al.* (2019) 'Which actigraphic variables optimally characterize the sleep-wake cycle of individuals with bipolar disorders?', *Acta Psychiatrica Scandinavica*, 139(3), pp. 0–2. doi: 10.1111/acps.13003.

Krüger, S. *et al.* (2010) 'The Observer-Rated Scale for Mania (ORSM): development, psychometric properties and utility', *Journal of Affective Disorders*. Elsevier B.V., 122(1–2), pp. 179–183. doi: 10.1016/j.jad.2009.07.022.

Látalová, K. (2010) *Bipolární afektivní porucha*. 1st edn. Praha: Grada. Available at: http://marefateadyan.nashriyat.ir/node/150.

Lecrubier, Y. *et al.* (1997) 'The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI', *European Psychiatry*. Éditions scientifiques et médicales Elsevier, Paris, 12(5), pp. 224–231. doi: 10.1016/S0924-9338(97)83296-8.

Lee, J. H. *et al.* (2014) 'Reliability and validity of the Korean version of Morningness–Eveningness Questionnaire in adults aged 20–39 years', *Chronobiology International*, 31(4), pp. 479–486. doi: 10.3109/07420528.2013.867864.

Lehnkering, H. *et al.* (2006) 'Actigraphic Investigations On The Activity-Rest Behavior Of Right- and Left-Handed Students', *Chronobiology International*, 23(3), pp. 593–605. doi: 10.1080/07420520600724094.

Levandovski, R. *et al.* (2011) 'Depression Scores Associate With Chronotype and Social Jetlag in a Rural Population', *Chronobiology International*, 28(9), pp. 771–778. doi: 10.3109/07420528.2011.602445.

Lötjönen, J. *et al.* (2003) 'Automatic Sleep-Wake and Nap Analysis with a New Wrist Worn Online Activity Monitoring Device Vivago WristCare®', *Sleep*, 26(1), pp. 86–90. doi: 10.1093/sleep/26.1.86.

Macfadden, W. *et al.* (2009) 'A randomized, double-blind, placebo-controlled study of maintenance treatment with adjunctive risperidone long-acting therapy in patients with bipolar I disorder who relapse frequently', *Bipolar Disorders*, 11(8), pp. 827–839. doi: 10.1111/j.1399-5618.2009.00761.x.

Manis, G., Aktaruzzaman, M. and Sassi, R. (2017) 'Bubble Entropy: An Entropy Almost Free of Parameters', *IEEE Transactions on Biomedical Engineering*, 64(11), pp. 2711–2718. doi: 10.1109/TBME.2017.2664105.

Matikainen-Ankney, B. A. *et al.* (2019) 'Rodent activity detector (RAD), an open source device for measuring activity in rodent home cages', *eNeuro*, 6(4), pp. 1–9. doi: 10.1523/ENEURO.0160-19.2019.

Matthews, M. *et al.* (2017) 'The double-edged sword: A mixed methods study of the interplay between bipolar disorder and technology use', *Computers in Human Behavior*. Elsevier Ltd, 75, pp. 288–300. doi: 10.1016/j.chb.2017.05.009.

McGowan, N. M. *et al.* (2020) 'Actigraphic patterns, impulsivity and mood instability in bipolar disorder, borderline personality disorder and healthy controls', *Acta Psychiatrica Scandinavica*, 141(4), pp. 374–384. doi: 10.1111/acps.13148.

McInnis, M., Gideon, J. and Mower Provost, E. (2017) 'Digital Phenotyping In Bipolar Disorder', *European Neuropsychopharmacology*, 27, p. S440. doi: 10.1016/j.euroneuro.2016.09.502.

McIntyre, R. S. *et al.* (2020) 'Bipolar disorders', *The Lancet*. Elsevier Ltd, 396(10265), pp. 1841–1856. doi: 10.1016/S0140-6736(20)31544-0.

McKenna, B. S., Drummond, S. P. A. and Eyler, L. T. (2014) 'Associations between circadian activity rhythms and functional brain abnormalities among euthymic bipolar patients: A preliminary study', *Journal of Affective Disorders*. Elsevier, 164, pp. 101–106. doi: 10.1016/j.jad.2014.04.034.

McPartland, R. J., Kupfer, D. J. and Gordon Foster, F. (1976) 'The movement-activated recording monitor: A third-generation motor-activity monitoring system', *Behavior Research Methods & Instrumentation*, 8(4), pp. 357–360. doi: 10.3758/BF03201791.

Mecacci, L. *et al.* (1986) 'The relationships between morningness-eveningness, ageing and personality', *Personality and Individual Differences*, 7(6), pp. 911–913. doi: 10.1016/0191-8869(86)90094-2.

Meltzer, L. J. *et al.* (2012) 'Use of actigraphy for assessment in pediatric sleep research', *Sleep Medicine Reviews*, 16(5), pp. 463–475. doi: 10.1016/j.smrv.2011.10.002.

Merikangas, K. R. *et al.* (2011) 'Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative', *Archives of General Psychiatry*, 68(3), p. 241. doi: 10.1001/archgenpsychiatry.2011.12.

Merikangas, K. R. *et al.* (2019) 'Real-time Mobile Monitoring of the Dynamic Associations among Motor Activity, Energy, Mood, and Sleep in Adults with Bipolar Disorder', *JAMA Psychiatry*, 76(2), pp. 190–198. doi: 10.1001/jamapsychiatry.2018.3546.

Milhiet, V. *et al.* (2011) 'Circadian biomarkers, circadian genes and bipolar disorders', *Journal of Physiology-Paris*. Elsevier Ltd, 105(4–6), pp. 183–189. doi: 10.1016/j.jphysparis.2011.07.002.

Di Milia, L. *et al.* (2013) 'Reviewing the Psychometric Properties of Contemporary Circadian Typology Measures', *Chronobiology International*, 30(10), pp. 1261–1271. doi: 10.3109/07420528.2013.817415.

Millar, A., Espie, C. A. and Scott, J. (2004) 'The sleep of remitted bipolar outpatients: A controlled naturalistic study using actigraphy', *Journal of Affective Disorders*, 80(2–3), pp. 145–153. doi: 10.1016/S0165-0327(03)00055-7.

Minors, D. S. and Waterhouse, J. M. (1988) 'Mathematical and statistical analysis of circadian rhythms', *Psychoneuroendocrinology*, 13(6), pp. 443–464. doi: 10.1016/0306-4530(88)90030-3.

Mishra, T. *et al.* (2020) 'Pre-symptomatic detection of COVID-19 from smartwatch data', *Nature Biomedical Engineering*. Springer US, 4(12), pp. 1208–1220. doi: 10.1038/s41551-020-00640-6.

Miziou, S. *et al.* (2015) 'Psychosocial treatment and interventions for bipolar disorder: A systematic review', *Annals of General Psychiatry*. BioMed Central, 14(1), pp. 1–11. doi: 10.1186/s12991-015-0057-z.

Monk, T. H. *et al.* (2000) 'The sleep of healthy people--a diary study.', *Chronobiology international*, 17(1), pp. 49–60. doi: 10.1081/cbi-100101031.

Montgomery, S. A. and Åsberg, M. (1979) 'A New Depression Scale Designed to be Sensitive to Change', *British Journal of Psychiatry*, 134(4), pp. 382–389. doi: 10.1192/bjp.134.4.382.

Monti, J. M. (2016) 'The effect of second-generation antipsychotic drugs on sleep parameters in patients with unipolar or bipolar disorder', *Sleep Medicine*. Elsevier B.V., 23, pp. 89–96. doi: 10.1016/j.sleep.2016.04.020.

Muaremi, A. *et al.* (2014) 'Assessing Bipolar Episodes Using Speech Cues Derived from Phone Calls', in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, pp. 103–114. doi: 10.1007/978-3-319-11564-1_11.

Mullin, B. C., Harvey, A. G. and Hinshaw, S. P. (2011) 'A preliminary study of sleep in adolescents with bipolar disorder, ADHD, and non-patient controls', *Bipolar Disorders*, 13(4), pp. 425–432. doi: 10.1111/j.1399-5618.2011.00933.x.

Murray, G. and Harvey, A. (2010) 'Circadian rhythms and sleep in bipolar disorder', *Bipolar Disorders*, 12(5), pp. 459–472. doi: 10.1111/j.1399-5618.2010.00843.x.

Ng, T. H. *et al.* (2015) 'Sleep-wake disturbance in interepisode bipolar disorder and high-risk individuals: A systematic review and meta-analysis', *Sleep Medicine Reviews*. Elsevier Ltd, 20, pp. 46–58. doi: 10.1016/j.smrv.2014.06.006.

Nowrouzi, B. *et al.* (2016) 'Admixture analysis of age at onset in first episode bipolar disorder', *Journal of Affective Disorders*. Elsevier, 201, pp. 88–94. doi: 10.1016/j.jad.2016.04.006.

Onnela, J. P. and Rauch, S. L. (2016) 'Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health', *Neuropsychopharmacology*. Nature Publishing Group, 41(7), pp. 1691–1696. doi: 10.1038/npp.2016.7.

Orsolini, L., Fiorani, M. and Volpe, U. (2020) 'Digital Phenotyping in Bipolar Disorder: Which Integration with Clinical Endophenotypes and Biomarkers?', *International Journal of Molecular Sciences*, 21(20), p. 7684. doi: 10.3390/ijms21207684.

Ortiz, A., Bradler, K. and Hintze, A. (2018) 'Episode forecasting in bipolar disorder: Is energy better than mood?', *Bipolar Disorders*, (December 2017), pp. 1–6. doi: 10.1111/bdi.12603.

Palmius, N. *et al.* (2017) 'Detecting Bipolar Depression From Geographic Location Data', *IEEE Transactions on Biomedical Engineering*, 64(8), pp. 1761–1771. doi: 10.1109/TBME.2016.2611862.

Pedregosa, F. *et al.* (2012) 'Scikit-learn: Machine Learning in Python', *Journal ofMachine Learning Research*, 12, pp. 2825–2830. doi: 10.1289/EHP4713.

Perna, G. *et al.* (2018) 'The revolution of personalized psychiatry: Will technology make it happen sooner?', *Psychological Medicine*, 48(5), pp. 705–713. doi: 10.1017/S0033291717002859.

Pies, R. (2007) 'How "objective" are psychiatric diagnoses?: (guess again).', *Psychiatry (Edgmont (Pa. : Township))*, 4(10), pp. 18–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20428307.

Plante, D. T. and Winkelman, J. W. (2008) 'Sleep Disturbance in Bipolar Disorder: Therapeutic Implications', *American Journal of Psychiatry*, 165(7), pp. 830–843. doi: 10.1176/appi.ajp.2008.08010077.

Portaluppi, F., Smolensky, M. H. and Touitou, Y. (2010) 'Ethics and methods for biological rhythm research on animals and human beings.', *Chronobiology international*, 27(9–10), pp. 1911–29. doi: 10.3109/07420528.2010.516381.

Redmond, D. P. and Hegge, F. W. (1985) 'Observations on the design and specification of a wrist-worn human activity monitoring system', *Behavior Research Methods, Instruments, & Computers*, 17(6), pp. 659–669. doi: 10.3758/BF03200979.

Reid, K. J. (2019) 'Assessment of Circadian Rhythms', *Neurologic Clinics*, 37(3), pp. 505–526. doi: 10.1016/j.ncl.2019.05.001.

Richman, S. J. and Moorman, J. R. (2000) 'Physiological time-series analysis using approximate entropy and sample entropy', *Am. J. Physiol.*, 278, pp. H2039–H2049.

Ritter, P. S. *et al.* (2012) 'The characteristics of sleep in patients with manifest bipolar disorder, subjects at high risk of developing the disease and healthy controls', *Journal of Neural Transmission*, 119(10), pp. 1173–1184. doi: 10.1007/s00702-012-0883-y.

Roenneberg, T. *et al.* (2004) 'A marker for the end of adolescence', *Current Biology*, 14(24), pp. R1038–R1039. doi: 10.1016/j.cub.2004.11.039.

Roenneberg, T. *et al.* (2007) 'Epidemiology of the human circadian clock', *Sleep Medicine Reviews*, 11(6), pp. 429–438. doi: 10.1016/j.smrv.2007.07.005.

Roenneberg, T. *et al.* (2012) 'Social Jetlag and Obesity', *Current Biology*, 22(10), pp. 939–943. doi: 10.1016/j.cub.2012.03.038.

Roenneberg, T. (2015) 'Having Trouble Typing? What on Earth Is Chronotype?', *Journal of Biological Rhythms*, 30(6), pp. 487–491. doi: 10.1177/0748730415603835.

Roenneberg, T. *et al.* (2019) 'Why Should We Abolish Daylight Saving Time?', *Journal of Biological Rhythms*, 34(3), pp. 227–230. doi: 10.1177/0748730419854197.

Roenneberg, T., Wirz-Justice, A. and Merrow, M. (2003) 'Life between Clocks: Daily Temporal Patterns of Human Chronotypes', *Journal of Biological Rhythms*, 18(1), pp. 80–90. doi: 10.1177/0748730402239679.

Ryu, H. *et al.* (2018) 'Validation of the Munich ChronoType Questionnaire in Korean Older Adults*', *Psychiatry Investigation*, 15(8), pp. 775–782. doi: 10.30773/pi.2018.04.09.

Sachs, G. S., Guille, C. and McMurrich, S. L. (2002) 'A clinical monitoring form for mood disorders', *Bipolar Disorders*, 4(5), pp. 323–327. doi: 10.1034/j.1399-5618.2002.01195.x.

Sadeh, A. *et al.* (1995) 'The Role of Actigraphy in the Evaluation of Sleep Disorders', *Sleep*, 18(4), pp. 288–302. doi: 10.1093/sleep/18.4.288.

Sadeh, A. (2011) 'The role and validity of actigraphy in sleep medicine: An update', *Sleep Medicine Reviews*. Elsevier Ltd, 15(4), pp. 259–267. doi: 10.1016/j.smrv.2010.10.001.

Saghir, N. *et al.* (2020) 'A comparison of manual electrocardiographic interval and waveform analysis in lead 1 of 12-lead ECG and Apple Watch ECG: A validation study', *Cardiovascular Digital Health Journal*. Elsevier Inc., 1(1), pp. 30–36. doi: 10.1016/j.cvdhj.2020.07.002.

Salvatore, P. *et al.* (2008) 'Circadian activity rhythm abnormalities in ill and recovered bipolar I disorder patients', *Bipolar Disorders*, 10(2), pp. 256–265. doi: 10.1111/j.1399-5618.2007.00505.x.

Santisteban, J. A., Brown, T. G. and Gruber, R. (2018) 'Association between the Munich Chronotype Questionnaire and Wrist Actigraphy', *Sleep Disorders*, 2018, pp. 1–7. doi: 10.1155/2018/5646848.

Schneider, J. *et al.* (2020) 'Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls', *CNS Spectrums*, pp. 1–11. doi: 10.1017/S1092852920001777.

Schneider, J. (2021) 'Continuous Actigraphy for Active Balanced Lifestyle Maintenance', *LIFMAT2021*. Prague, 1(2021), p. A1.

Scott, J. (2011) 'Clinical parameters of circadian rhythms in affective disorders', *European Neuropsychopharmacology*. Elsevier B.V. and ECNP, 21(SUPPL.4), pp. S671–S675. doi: 10.1016/j.euroneuro.2011.07.006.

Scott, J., Vaaler, A. E., *et al.* (2017) 'A pilot study to determine whether combinations of objectively measured activity parameters can be used to differentiate between mixed states, mania, and bipolar depression', *International Journal of Bipolar Disorders*. Springer Berlin Heidelberg, 5(1). doi: 10.1186/s40345-017-0076-6.

Scott, J., Murray, G., *et al.* (2017) 'Activation in Bipolar Disorders', *JAMA Psychiatry*, 74(2), p. 189. doi: 10.1001/jamapsychiatry.2016.3459.

Sebela, A. *et al.* (2019) 'Decreased need for sleep as an endophenotype of bipolar disorder: an actigraphy study', *Chronobiology International*. Taylor & Francis, 36(9), pp. 1227–1239. doi: 10.1080/07420528.2019.1630631.

Shou, H. *et al.* (2017) 'Dysregulation of objectively assessed 24-hour motor activity patterns as a

potential marker for bipolar I disorder: results of a community-based family study', *Translational Psychiatry*. Nature Publishing Group, 7(8), p. e1211. doi: 10.1038/tp.2017.136.

Shugar, G. *et al.* (1992) 'Development, use, and factor analysis of a self-report inventory for mania', *Comprehensive Psychiatry*, 33(5), pp. 325–331. doi: 10.1016/0010-440X(92)90040-W.

Smith, C. *et al.* (2020) 'ActiGraph GT3X+ and Actical Wrist and Hip Worn Accelerometers for Sleep and Wake Indices in Young Children Using an Automated Algorithm: Validation With Polysomnography', *Frontiers in Psychiatry*, 10(January), pp. 1–12. doi: 10.3389/fpsyt.2019.00958.

Smith, M. T. *et al.* (2018) 'Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment', *Journal of Clinical Sleep Medicine*, 14(07), pp. 1209–1230. doi: 10.5664/jcsm.7228.

Sokolove, P. G. *et al.* (1977) 'A circadian rhythm in the locomotive behaviour of the giant garden slug Limax maximus.', *Journal of Experimental Biology*, 66(1), pp. 47–64.

Sokolove, P. G. and Bushell, W. N. (1978) 'The chi square periodogram: Its utility for analysis of circadian rhythms', *Journal of Theoretical Biology*, 72(1), pp. 131–160. doi: 10.1016/0022-5193(78)90022-X.

Van Someren, E. J. *et al.* (1999) 'Bright light therapy: improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods.', *Chronobiology international*, 16(4), pp. 505–18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10442243.

van Someren, E. J. W. *et al.* (1996) 'Circadian rest—activity rhythm disturbances in alzheimer's disease', *Biological Psychiatry*, 40(4), pp. 259–270. doi: 10.1016/0006-3223(95)00370-3.

Van Someren, E. J. W. *et al.* (1997) 'Indirect bright light improves circadian rest-activity rhythm disturbances in demented patients', *Biological Psychiatry*, 41(9), pp. 955–963. doi: 10.1016/S0006-3223(97)89928-3.

Sridhar, G. R. and Sanjana, N. S. N. (2016) 'Sleep, circadian dysrhythmia, obesity and diabetes', *World Journal of Diabetes*, 7(19), p. 515. doi: 10.4239/wjd.v7.i19.515.

St-Amand, J. *et al.* (2013) 'Sleep disturbances in bipolar disorder during remission', *Journal of Affective Disorders*, 146(1), pp. 112–119. doi: 10.1016/j.jad.2012.05.057.

Stahl, E. A. *et al.* (2019) 'Genome-wide association study identifies 30 loci associated with bipolar disorder', *Nature Genetics*, 51(5), pp. 793–803. doi: 10.1038/s41588-019-0397-8.

Steel, Z. *et al.* (2014) 'The global prevalence of common mental disorders: A systematic review and meta-analysis 1980-2013', *International Journal of Epidemiology*, 43(2), pp. 476–493. doi: 10.1093/ije/dyu038.

Tarassenko, L. and Greenhalgh, T. (2020) 'Question: Should smartphone apps be used clinically as oximeters? Answer: No. - CEBM', *Cebm*. Available at: https://www.cebm.net/covid-19/question-should-smartphone-apps-be-used-as-oximeters-answer-no/.

Tazawa, Y. *et al.* (2019) 'Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis', *Journal of Affective Disorders*. Elsevier B.V., 253(April), pp. 257–269. doi: 10.1016/j.jad.2019.04.087.

Thun, E. *et al.* (2012) 'An Actigraphic Validation Study of Seven Morningness-Eveningness Inventories', *European Psychologist*, 17(3), pp. 222–230. doi: 10.1027/1016-9040/a000097.

Tohen, M. *et al.* (2009) 'The International Society for Bipolar Disorders (ISBD) Task Force report on the nomenclature of course and outcome in bipolar disorders.', *Bipolar disorders*, 11(5), pp. 453–73. doi: 10.1111/j.1399-5618.2009.00726.x.

Tomlinson, S. *et al.* (2018) 'Accuracy of Smartphone-Based Pulse Oximetry Compared with Hospital-Grade Pulse Oximetry in Healthy Children', *Telemedicine and e-Health*, 24(7), pp. 527–535. doi: 10.1089/tmj.2017.0166.

Towbin, K. *et al.* (2013) 'Differentiating Bipolar Disorder–Not Otherwise Specified and Severe Mood Dysregulation', *Journal of the American Academy of Child & Adolescent Psychiatry*, 52(5), pp. 466–481. doi: 10.1016/j.jaac.2013.02.006.

Trivedi, M. H. *et al.* (2004) 'The Inventory of Depressive Symptomatology, clinician rating (IDS-C) and self-report (IDS-SR), and the Quick Inventory Depressive Symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: A psychome', *Psychological Medicine*, 34(1), pp. 73–82. doi: 10.1017/S0033291703001107.

Urošević, S. *et al.* (2008) 'Dysregulation of the behavioral approach system (BAS) in bipolar spectrum disorders: Review of theory and evidence', *Clinical Psychology Review*, 28(7), pp. 1188–1205. doi: 10.1016/j.cpr.2008.04.004.

Vancampfort, D. *et al.* (2017) 'Sedentary behavior and physical activity levels in people with schizophrenia, bipolar disorder and major depressive disorder: a global systematic review and meta-analysis', *World Psychiatry*, 16(3), pp. 308–315. doi: 10.1002/wps.20458.

Vitale, J. A. *et al.* (2015) 'Chronotype influences activity circadian rhythm and sleep: Differences in sleep quality between weekdays and weekend', *Chronobiology International*, 32(3), pp. 405–415. doi: 10.3109/07420528.2014.986273.

Vostatek, P. (2018) 'Mindpax Sleep Classification', *Mindpax White-papers*, pp. 1–6. Available at: https://www.mindpax.me/assets/docs/Validation_study_appendix.pdf.

Wahl, Y. *et al.* (2017) 'Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions', *Frontiers in Physiology*, 8(SEP). doi: 10.3389/fphys.2017.00725.

Walker, W. H. *et al.* (2020) 'Circadian rhythm disruption and mental health', *Translational Psychiatry*. Springer US, 10(1). doi: 10.1038/s41398-020-0694-0.

Wang, P. S. *et al.* (2005) 'Twelve-Month Use of Mental Health Services in the United States', *Archives of General Psychiatry*, 62(6), p. 629. doi: 10.1001/archpsyc.62.6.629.

Wang, R. *et al.* (2017) 'Accuracy of wrist-worn heart rate monitors', *JAMA Cardiology*, 2(1), pp. 104–106. doi: 10.1001/jamacardio.2016.3340.

Witting, W. *et al.* (1990) 'Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease', *Biological Psychiatry*, 27(6), pp. 563–572. doi: 10.1016/0006-3223(90)90523-5.

Wittmann, M. *et al.* (2006) 'Social jetlag: Misalignment of biological and social time', *Chronobiology International*, 23(1–2), pp. 497–509. doi: 10.1080/07420520500545979.

Young, R. C. *et al.* (1978) 'A Rating Scale for Mania: Reliability, Validity and Sensitivity', *British Journal of Psychiatry*, 133(5), pp. 429–435. doi: 10.1192/bjp.133.5.429.

Zavada, A. *et al.* (2005) 'Comparison of the Munich Chronotype Questionnaire with the Horne-Ostberg's Morningness-Eveningness Score.', *Chronobiology international*, 22(2), pp. 267–78. doi: 10.1081/cbi-200053536.

Zebin, T., Peek, N. and Casson, A. J. (2019) 'Physical activity based classification of serious mental illness group participants in the UK Biobank using ensemble dense neural networks', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 1251–1254. doi: 10.1109/EMBC.2019.8857532.

Supplementary Materials:

Chapter 5 - Supplementary materials

Table S-1: Reliability of feature estimation - The modelled variation of estimation error for data missing in blocks

Feature	Window size	Miss and block model (Eq. 5.7)			Predicted estimation error variation ($\widehat{\tilde{E}_{F_i}}$)											
		Coeff miss	Coeff blocks	R^2 adj.	5% 4 blocks	5% 6 blocks	5% 10blocks	10% 4 blocks	10% 6 blocks	10% 10blocks	15% 4 blocks	15% 6 blocks	15% 10blocks	20% 4 blocks	20% 6 blocks	20% 10blocks
M10	One day	1.0154	-0.2437	0.690	4.1022	3.6149	2.6403	9.1790	8.6917	7.7171	14.2558	13.7685	12.7939	**19.3326**	**18.8453**	**17.8707**
L5		0.4258	-0.1231	0.565	1.6366	1.3904	0.8980	**3.7657**	**3.5195**	**3.0270**	**5.8948**	**5.6486**	**5.1561**	**8.0239**	**7.7777**	**7.2852**
M10-time		0.0549	-4.74E-3	0.739	0.2554	0.2460	0.2270	**0.5298**	**0.5204**	**0.5014**	**0.8042**	**0.7948**	**0.7758**	**1.0786**	**1.0692**	**1.0502**
L5-time		0.0504	8.10E-3	0.694	0.2843	0.3005	0.3329	**0.5361**	**0.5523**	**0.5847**	**0.7880**	**0.8042**	**0.8366**	**1.0399**	**1.0561**	**1.0885**
RA		1.38E-3	-2.99E-4	0.681	5.71E-3	5.11E-3	3.92E-3	**0.0126**	**0.0120**	0.0108	**0.0195**	**0.0189**	**0.0177**	**0.0264**	**0.0258**	**0.0246**
RMSSD$_{M10}$		5.6416	**3.9474**	**0.725**	**43.9976**	**51.8924**	**67.6818**	**72.2058**	**80.1006**	**95.8900**	**100.414**	**108.309**	**124.098**	**128.622**	**136.517**	**152.306**
MESOR$_7$	Seven days	0.3610	-0.0425	0.908	1.6350	1.5500	1.3800	3.4400	3.3550	3.1850	5.2450	5.1600	4.9900	**7.0500**	**6.9650**	**6.7950**
Amplitude$_7$		0.4478	-0.0462	0.911	2.0544	1.9621	1.7775	4.2935	4.2012	4.0165	**6.5325**	**6.4402**	**6.2556**	**8.7716**	**8.6793**	**8.4946**
Acrophase$_7$		3.41E-3	-6.24E-4	0.881	0.0146	0.0133	0.0108	0.0316	0.0304	0.0279	0.0487	0.0474	0.0449	0.0657	0.0645	0.0620
CQ$_7$		1.50E-3	-1.32E-4	0.916	6.96E-3	6.70E-3	6.17E-3	0.0145	0.0142	0.0137	**0.0219**	**0.0217**	**0.0211**	**0.0294**	**0.0292**	**0.0286**
GOF$_7$		0.0783	-6.62E-3	0.908	0.3652	0.3520	0.3255	0.7570	0.7437	0.7172	**1.1487**	**1.1354**	**1.1090**	**1.5404**	**1.5272**	**1.5007**
M10$_7$		0.5756	-0.0665	0.922	2.6119	2.4788	2.2127	5.4900	5.3569	5.0907	8.3681	8.2350	7.9688	**11.2462**	**11.1131**	**10.8469**
L5$_7$		0.3356	-0.0387	0.890	1.5231	1.4457	1.2910	**3.2009**	**3.1235**	**2.9688**	**4.8787**	**4.8013**	**4.6466**	**6.5565**	**6.4791**	**6.3244**
M10-time$_7$		0.0349	5.79E-3	0.753	0.1974	0.2090	0.2322	**0.3717**	**0.3833**	**0.4065**	**0.5460**	**0.5576**	**0.5807**	**0.7203**	**0.7319**	**0.7550**
L5-time$_7$		0.0221	2.26E03	0.851	0.1194	0.1239	0.1329	**0.2297**	**0.2342**	**0.2432**	**0.3400**	**0.3445**	**0.3535**	**0.4503**	**0.4548**	**0.4638**
RA$_7$		1.01E-3	-9.18E-5	0.903	4.68E-3	4.50E-3	4.13E-3	**9.73E-3**	**9.55E-3**	**9.18E-3**	**0.0148**	**0.0146**	**0.0142**	**0.0198**	**0.0196**	**0.0193**
IV$_7$		1.22E-3	2.66E-4	0.880	7.17E-3	7.70E-3	8.76E-3	0.0133	**0.0138**	**0.0149**	**0.0194**	**0.0199**	**0.0210**	**0.0255**	**0.0260**	**0.0271**
IS$_7$		1.56E-3	-1.70E-4	0.892	7.11E-3	6.77E-3	6.09E-3	**0.0149**	**0.0146**	**0.0139**	**0.0227**	**0.0224**	**0.0217**	**0.0305**	**0.0301**	**0.0295**
MESOR$_{14}$	Fourteen days	0.2790	-0.0176	0.926	1.3246	1.2894	1.2190	2.7196	2.6844	2.6140	4.1146	4.0794	4.0090	5.5096	5.4744	5.4040
Amplitude$_{14}$		0.3408	-0.0136	0.931	1.6494	1.6222	1.5677	3.3534	3.3261	3.2717	**5.0573**	**5.0300**	**4.9756**	**6.7612**	**6.7340**	**6.6795**
Acrophase$_{14}$		4.31E-3	-5.52E-4	0.820	0.0193	0.0182	0.0160	0.0409	0.0398	0.0376	0.0624	0.0613	0.0591	0.0840	0.0829	0.0807
CQ$_{14}$		1.12E-3	-9.82E-6	0.931	5.55E-3	5.53E-3	5.49E-3	0.0111	0.0111	0.0111	**0.0167**	**0.0167**	**0.0167**	**0.0223**	**0.0223**	**0.0222**
GOF$_{14}$		0.0571	2.64E-4	0.924	0.2865	0.2870	0.2880	0.5719	0.5724	0.5734	**0.8573**	**0.8578**	**0.8589**	**1.1427**	**1.1432**	**1.1443**
M10$_{14}$		0.4425	-0.0346	0.922	2.0740	2.0048	1.8662	4.2866	4.2174	4.0788	6.4992	6.4300	6.2914	**8.7118**	**8.6426**	8.5040
L5$_{14}$		0.2453	-0.0102	0.915	1.1859	1.1654	1.1246	**2.4126**	**2.3921**	**2.3513**	**3.6393**	**3.6188**	**3.5780**	**4.8660**	**4.8456**	**4.8047**
M10-time$_{14}$		0.0353	0.0136	0.595	**0.2311**	**0.2582**	**0.3126**	**0.4078**	**0.4349**	**0.4893**	**0.5845**	**0.6117**	**0.6660**	**0.7612**	**0.7884**	**0.8427**
L5-time$_{14}$		0.0195	2.61E-3	0.844	0.1079	0.1132	0.1236	**0.2054**	**0.2107**	**0.2211**	**0.3029**	**0.3082**	**0.3186**	**0.4004**	**0.4057**	**0.4161**
RA$_{14}$		8.17E-4	-3.29E-5	0.924	3.95E-3	3.89E-3	3.76E-3	**8.04E-3**	**7.97E-3**	**7.84E-3**	**0.0121**	**0.0121**	**0.0119**	**0.0162**	**0.0161**	**0.0160**
IV$_{14}$		9.03E-4	1.47E-4	0.914	5.10E-3	5.40E-3	5.99E-3	9.62E-3	9.91E-3	0.0105	**0.0141**	**0.0144**	**0.0150**	**0.0187**	**0.0189**	**0.0195**
IS$_{14}$		9.83E-4	-9.00E-6	0.927	4.88E-3	4.86E-3	4.83E-3	9.79E-3	9.78E-3	9.74E-3	**0.0147**	**0.0147**	**0.0147**	**0.0196**	**0.0196**	**0.0196**

Bold red shaded text indicates the missing data-points setting, where the standard deviation (variation) of estimation error (EE) reaches 40+ % of the natural $LTTV_{SD}$ of a feature for the patient with median stability (Table 5-1). **Bold** text indicates the setting where the EE variation reaches 20+ % of the natural LTTV

Chapter 6 - Supplementary materials

Table S-2: Impact of window length on chronotype estimation

1) MEQ

Feature	Window (weeks)	TRAIN: beta1		TRAIN: R-squared		TEST: MAE	
		mean	**std**	**mean**	**std**	**mean**	**std**
$MSFsc_{act}$	1	-2.666	0.353	0.164	0.025	6.581	0.931
	2	-3.640	0.558	0.192	0.039	6.606	0.755
	3	-4.416	0.560	0.265	0.050	6.376	0.419
	4	**-4.249**	**0.450**	**0.260**	**0.043**	**6.287**	**0.514**
	5	-4.137	0.343	0.244	0.037	6.303	0.624
	6	-4.437	0.409	0.269	0.043	6.368	0.636
M10-time	1	-2.497	0.287	0.152	0.028	6.470	0.953
	2	-3.266	0.416	0.204	0.042	6.366	0.710
	3	-3.693	0.402	0.254	0.047	6.113	0.559
	4	**-3.871**	**0.267**	**0.289**	**0.038**	**5.869**	**0.498**
	5	-3.826	0.322	0.281	0.043	6.037	0.724
	6	-3.780	0.360	0.281	0.050	6.125	0.850
L5-time	1	-2.536	0.372	0.085	0.016	6.831	0.892
	2	-3.888	0.445	0.156	0.022	6.554	0.965
	3	-5.163	0.460	0.236	0.026	6.340	0.879
	4	-5.316	0.402	0.264	0.028	6.183	0.951
	5	-5.656	0.335	0.293	0.022	6.124	0.878
	6	**-5.994**	**0.337**	**0.313**	**0.021**	**6.125**	**0.781**
Acrophase	1	-5.727	0.521	0.335	0.037	5.871	0.886
	2	-5.591	0.282	0.335	0.032	5.825	0.545
	3	**-5.884**	**0.377**	**0.372**	**0.042**	**5.608**	**0.453**
	4	-5.716	0.301	0.361	0.032	5.708	0.579
	5	-5.527	0.249	0.356	0.030	5.737	0.607
	6	-5.681	0.290	0.365	0.034	5.691	0.682
Mid-sleep	1	-2.704	1.614	0.144	0.080	7.067	0.631
	2	-5.277	0.490	0.282	0.024	6.087	0.897
	3	-5.654	0.600	0.313	0.033	6.025	0.922
	4	-5.418	0.635	0.278	0.031	6.215	0.987
	5	-5.907	0.782	0.333	0.041	5.999	0.983
	6	**-6.400**	**0.705**	**0.359**	**0.042**	**5.995**	**1.148**

Bold row text marks window length with lowest MAE represents for each feature

The table continues on the next page

2) MCTQ-MSFsc

Feature	Window (weeks)	TRAIN: beta1		TRAIN: R-squared		TEST: MAE	
		mean	**std**	**mean**	**std**	**mean**	**std**
$MSFsc_{act}$	1	0.403	0.037	0.289	0.035	0.703	0.133
	2	0.572	0.045	0.368	0.060	0.640	0.096
	3	0.653	0.045	0.446	0.045	0.594	0.086
	4	0.628	0.032	0.438	0.037	0.590	0.125
	5	0.617	0.034	0.420	0.040	0.601	0.124
	6	**0.658**	**0.035**	**0.471**	**0.034**	**0.569**	**0.089**
M10-time	1	0.316	0.025	0.186	0.016	0.720	0.143
	2	0.404	0.062	0.240	0.049	0.712	0.098
	3	0.415	0.065	0.246	0.052	0.715	0.091
	4	0.426	0.068	0.268	0.055	0.712	0.075
	5	0.436	0.063	0.279	0.051	0.701	0.068
	6	**0.437**	**0.061**	**0.287**	**0.046**	**0.691**	**0.086**
L5-time	1	0.368	0.058	0.138	0.030	0.736	0.139
	2	0.496	0.082	0.196	0.045	0.722	0.111
	3	0.625	0.068	0.266	0.033	0.691	0.131
	4	0.671	0.057	0.324	0.019	0.670	0.141
	5	0.694	0.061	0.339	0.025	0.668	0.150
	6	**0.729**	**0.062**	**0.356**	**0.022**	**0.665**	**0.155**
Acrophase	1	0.707	0.047	0.396	0.045	0.613	0.134
	2	0.681	0.060	0.384	0.055	0.625	0.082
	3	0.701	0.051	0.407	0.048	0.613	0.073
	4	0.684	0.047	0.399	0.040	0.612	0.081
	5	0.662	0.042	0.394	0.039	0.611	0.105
	6	**0.679**	**0.041**	**0.402**	**0.037**	**0.605**	**0.113**
Mid-sleep	1	0.397	0.240	0.238	0.132	0.780	0.202
	2	0.677	0.086	0.364	0.073	0.652	0.147
	3	0.722	0.068	0.398	0.053	0.635	0.122
	4	0.714	0.056	0.377	0.043	0.639	0.125
	5	0.743	0.051	0.408	0.035	0.609	0.123
	6	**0.830**	**0.083**	**0.455**	**0.061**	**0.615**	**0.107**

Bold row text marks window length with lowest MAE represents for each feature
The table continues on the next page

3) MCTQ-SJL

Feature	Window (weeks)	TRAIN: beta1		TRAIN: R-squared		TEST: MAE	
		mean	**std**			**mean**	**std**
$SJLrel_{act}$	1	0.161	0.051	0.675	0.153	0.055	0.016
	2	0.275	0.053	0.663	0.145	0.069	0.020
	3	0.466	0.081	0.632	0.116	0.166	0.032
	4	**0.497**	**0.059**	**0.622**	**0.116**	**0.188**	**0.024**
	5	0.517	0.050	0.626	0.144	0.177	0.020
	6	0.527	0.063	0.637	0.139	0.163	0.025
$M10\text{-}time_{diff}$	1	0.078	0.014	0.671	0.180	0.055	0.021
	2	0.092	0.025	0.692	0.145	0.057	0.028
	3	0.088	0.017	0.693	0.149	0.034	0.011
	4	0.100	0.018	0.694	0.143	0.040	0.013
	5	0.111	0.021	0.686	0.126	0.048	0.015
	6	**0.142**	**0.019**	**0.687**	**0.141**	**0.069**	**0.016**
$L5\text{-}time_{diff}$	1	0.009	0.024	0.697	0.165	0.002	0.003
	2	0.112	0.072	0.707	0.176	0.033	0.026
	3	0.185	0.063	0.683	0.180	0.044	0.027
	4	0.270	0.075	0.686	0.183	0.064	0.032
	5	0.188	0.068	0.675	0.184	0.036	0.027
	6	**0.214**	**0.069**	**0.674**	**0.196**	**0.034**	**0.022**
$Mid\text{-}sleep_{diff}$	1	0.047	0.047	0.705	0.186	0.019	0.031
	2	0.162	0.113	0.690	0.186	0.069	0.074
	3	**0.223**	**0.064**	**0.647**	**0.160**	**0.086**	**0.047**
	4	0.232	0.070	0.662	0.169	0.065	0.037
	5	0.129	0.085	0.672	0.162	0.024	0.017
	6	0.110	0.089	0.671	0.190	0.018	0.012

Bold row text marks window length with lowest MAE for each feature

Table S-3: Chronotyping results with confounders AGE and BMI

1) MEQ		TRAIN: β coeff.			TRAIN: R-squared		TEST: MAE	
Feature	***Window (weeks)***	***Feat***	***AGE***	***BMI***	***mean***	***SD***	***mean***	***SD***
Acrophase	3	-5.840	0.083	-0.126	0.389	0.030	5.621	0.465
$MSFsc_{acti}$	4	-3.546	0.180	-0.008	0.285	0.035	6.133	0.625
M10-time	4	-3.603	0.185	-0.129	0.327	0.028	5.873	0.521
L5-time	6	-5.332	0.242	-0.053	0.365	0.016	5.968	0.915
Mid-sleep	6	-5.707	0.176	-0.047	0.385	0.031	5.936	0.988

2) MCTQ-MSFsc		TRAIN: β coeff.			TRAIN: R-squared		TEST: MAE	
Feature	***Window***	***Feat***	***AGE***	***BMI***	***mean***	***SD***	***mean***	***SD***
Acrophase	6	*0.615*	-0.020	0.012	*0.429*	*0.044*	*0.595*	*0.093*
$MSFsc_{acti}$	6	*0.558*	-0.027	0.003	*0.515*	*0.043*	*0.546*	*0.076*
M10-time	6	*0.365*	-0.033	0.012	*0.358*	*0.058*	*0.656*	*0.055*
L5-time	6	0.614	-0.037	-0.037	0.443	0.028	0.611	0.132
Mid-sleep	6	0.718	-0.027	0.004	0.495	0.071	0.591	0.092

3) MCTQ-SJLrel		TRAIN: β coeff.			TRAIN: R-squared		TEST: MAE	
Feature	***Window***	***Feat***	***AGE***	***BMI***	***mean***	***SD***	***mean***	***SD***
$SJLrel_{acti}$	4	0.469	-0.027	0.012	0.218	0.033	0.658	0.083
$Mid\text{-}sleep_{diff}$	*3*	0.214	-0.029	0.016	0.138	0.037	0.711	0.099
$M10\text{-}time_{diff}$	*6*	0.105	-0.029	0.012	0.126	0.010	0.727	0.129
$L5\text{-}time_{diff}$	*6*	0.283	-0.032	0.017	0.142	0.045	0.714	0.132

Chapter 8 - Supplementary materials

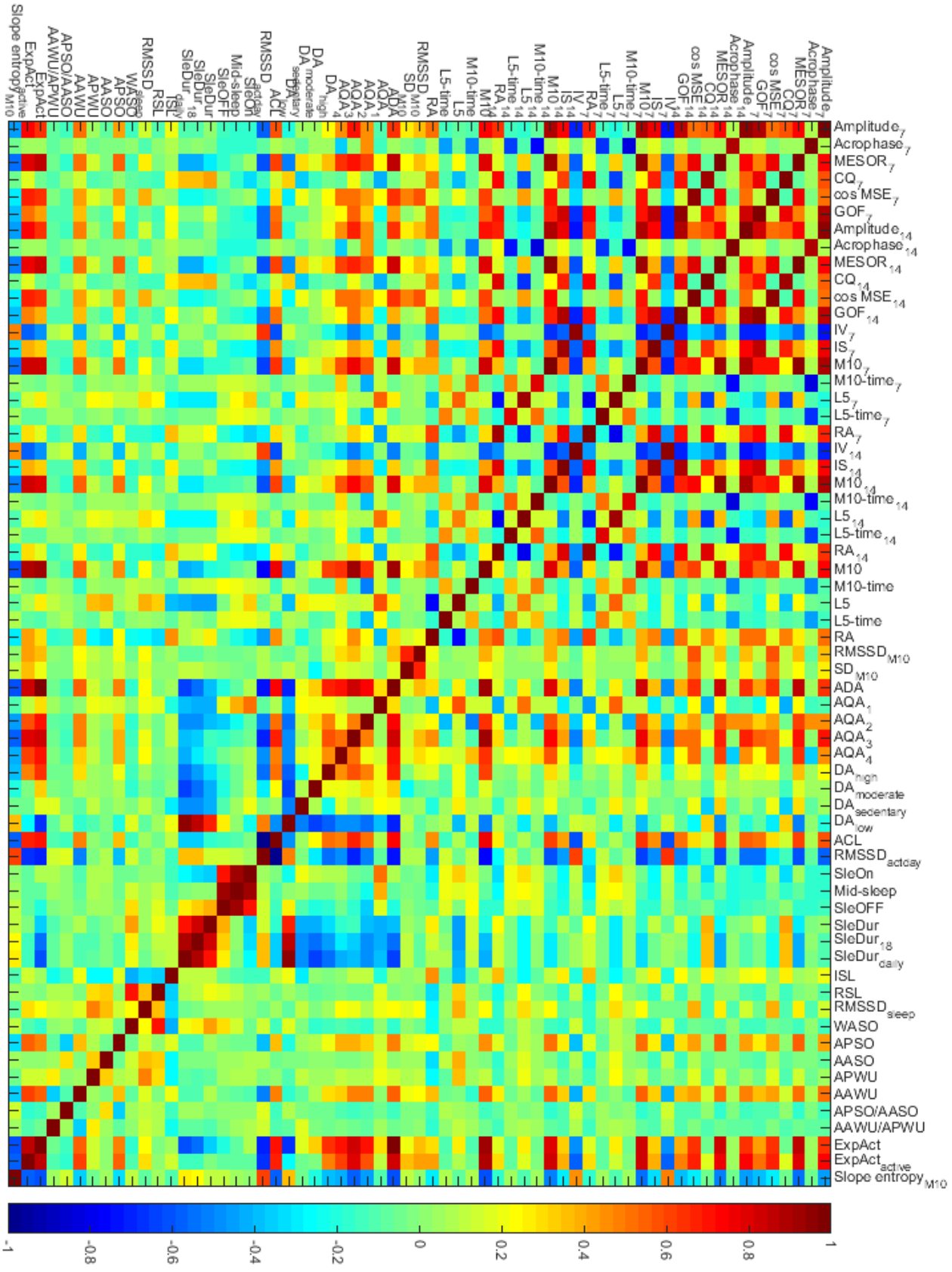


Figure S.1 - Correlation between actigraphic features represented as colours see colorbar at the bottom of the image

Table S-4: Actigraphic features during relapses

Feature	Remission Mean (SD)	Depression Mean (SD)	Mania Mean (SD)	Anova
Cosinor Analysis - 7 days estimation window				
Amplitude$_7$	209.50 (55.27)	188.01*** (55.25)	196.05$^{***/\#}$ (54.96)	F = 48.31 p < 0.0001
Acrophase$_7$	14.87 (1.55)	15.22*** (1.99)	15.04 (1.63)	F = 13.98 p < 0.0001
MESOR$_7$	298.16 (64.11)	263.16*** (60.28)	311.82$^{***/\#\#\#}$ (64.34)	F = 109.96 p < 0.0001
CQ$_7$	0.71 (0.14)	0.72 (0.16)	0.63$^{***/\#\#\#}$ (0.14)	F = 54.33 p < 0.0001
MSE$_7$	89756 (24367)	86327*** (24958)	94156$^{**/\#\#\#}$ (23911)	F = 13.06 p < 0.0001
GOF$_7$	20.18 (7.19)	17.65*** (7.18)	17.48*** (6.89)	F = 50.23 p < 0.0001
Cosinor Analysis - 14 days estimation window				
Amplitude$_{14}$	207.31 (52.68)	186.72*** (52.49)	193.40$^{***/\#}$ (53.03)	F = 49.73 p < 0.0001
Acrophase$_{14}$	14.89 (1.36)	15.22*** (1.99)	14.96$^{\#}$ (1.50)	F = 13.60 p < 0.0001
MESOR$_{14}$	298.30 (62.57)	265.28*** (57.81)	312.32$^{***/\#\#\#}$ (63.06)	F = 105.39 p < 0.0001
CQ$_{14}$	0.70 (0.13)	0.71 (0.15)	0.62$^{***/\#\#\#}$ (0.15)	F = 57.47 p < 0.0001
MSE$_{14}$	90693 (24260)	87802** (24904)	95306$^{***/\#\#\#}$ (23734)	F = 11.71 p < 0.0001
GOF$_{14}$ (%)	19.67 (6.74)	17.26*** (6.85)	17.04*** (6.40)	F = 52.24 p < 0.0001
Nonparametric circadian rhythm analysis (NPCRA) - 7 days estimation window				
IV$_7$	0.47 (0.12)	0.51*** (0.13)	0.47$^{\#\#\#}$ (0.29)	F = 23.05 p < 0.0001
IS$_7$	0.52 (0.13)	0.47*** (0.13)	0.47*** (0.15)	F = 45.39 p < 0.0001
M10$_7$	446.86 (93.42)	400.51*** (88.91)	449.19$^{\#\#\#}$ (96.07)	F = 77.43 p < 0.0001
M10-time$_7$	14.80 (1.85)	15.29*** (2.05)	14.91$^{\#\#}$ (2.28)	F = 18.54 p < 0.0001
L5$_7$	59.05 (32.06)	58.46 (39.29)	77.11$^{***/\#\#\#}$ (46.00)	F = 41.98 p < 0.0001
L5-time$_7$	3.28 (1.40)	3.51*** (2.06)	3.16$^{\#\#}$ (1.56)	F = 7.94 p = 0.0004
RA$_7$	0.77 (0.10)	0.75*** (0.12)	0.71$^{***/\#\#\#}$ (0.13)	F = 39.93 p < 0.0001
NPCRA - 14 days estimation window				
IV$_{14}$	0.47 (0.10)	0.51*** (0.12)	0.50*** (0.09)	F = 41.74 p < 0.0001
IS$_{14}$	0.49 (0.12)	0.45*** (0.12)	0.44*** (0.13)	F = 53.87 p < 0.0001
M10$_{14}$	443.49 (90.28)	399.82*** (83.31)	445.36$^{\#\#\#}$ (93.29)	F = 74.28 p < 0.0001
M10-time$_{14}$	14.78 (1.78)	15.34*** (1.91)	15.02$^{*/\#}$ (2.08)	F = 26.95 p < 0.0001
L5$_{14}$	61.93 (32.33)	61.81 (43.21)	82.64$^{***/\#\#\#}$ (45.64)	F = 51.09 p < 0.0001
L5-time$_{14}$	3.32 (1.24)	3.54*** (1.90)	3.04$^{***/\#\#\#}$ (1.44)	F = 15.65 p < 0.0001
RA$_{14}$	0.76 (0.10)	0.74*** (0.12)	0.69$^{***/\#\#\#}$ (0.15)	F = 50.64 p < 0.0001
NPCRA - daily values				
M10	469.93 (118.95)	426.39*** (118.56)	474.79$^{\#\#\#}$ (124.48)	F = 42.03 p < 0.0001
M10-time	14.90 (2.72)	15.28*** (2.92)	15.10 (3.44)	F = 5.46 p = 0.0043
L5	45.18 (27.29)	42.41** (21.87)	53.32$^{***/\#\#\#}$ (43.77)	F = 18.15 p < 0.0001
L5-time	3.04 (2.18)	3.18 (2.62)	3.04 (2.55)	F = 1.05 p = 0.3494
RA	0.82 (0.08)	0.81* (0.09)	0.80*** (0.11)	F = 7.98 p = 0.0003
RMSSD$_{M10}$	1979 (375)	2031*** (405)	1974$^{\#}$ (337)	F = 6.00 p = 0.0025
SD$_{M10}$	2124 (534)	2192** (533)	2137 (475)	F = 4.82 p = 0.0081
Other nonparametric features – daily values				
ADA (0:00-24:00)	298.05 (78.69)	262.09*** (73.47)	310.50$^{**/\#\#\#}$ (84.61)	F = 74.29 p < 0.0001
AQA$_1$ (0:00:6:00)	79.60 (68.60)	80.68 (86.75)	102.87$^{***/\#\#\#}$ (84.41)	F = 15.07 p < 0.0001
AQA$_2$ (6:00-12:00)	345.96 (138.01)	285.14*** (126.95)	358.95$^{\#\#\#}$ (145.33)	F = 65.48 p < 0.0001
AQA$_3$ (12:00-18:00)	431.47 (131.73)	387.46*** (132.24)	433.99$^{\#\#\#}$ (136.16)	F = 34.37 p < 0.0001
AQA$_4$ (18:00-24:00)	335.85 (125.68)	297.68*** (128.58)	349.20$^{\#\#\#}$ (139.30)	F = 31.44 p < 0.0001
DA$_{high}$ (%)	25 (6)	22*** (6)	26$^{***/\#\#\#}$ (8)	F = 55.10 p < 0.0001
DA$_{moderate}$ (%)	25 (5)	24*** (6)	26$^{***/\#\#\#}$ (6)	F = 30.26 p < 0.0001
DA$_{sedentary}$ (%)	14 (6)	14 (6)	15$^{***/\#\#\#}$ (6)	F = 9.94 p < 0.0001
DA$_{low}$ (%)	36 (9)	40*** (0.11)	32$^{***/\#\#\#}$ (0.11)	F = 75.69 p < 0.0001
ACL (auto-corr. lag)	0.90 (0.04)	0.88*** (0.04)	0.90$^{\#\#\#}$ (0.03)	F = 89.39 p < 0.0001
RMSSD$_{actday}$	269.21 (37.34)	262.99*** (39.11)	272.94$^{\#\#\#}$ (36.63)	F = 11.14 p < 0.0001
Sleep based features - daily values				
SleOn	-0.23 (3.87)	-0.03 (4.32)	1.07$^{***/\#\#\#}$ (5.34)	F = 15.10 p < 0.0001
Mid-sleep	3.85 (3.04)	4.33*** (3.48)	4.34** (4.21)	F = 8.55 p = 0.0002
SleOFF	8.20 (3.67)	8.98*** (4.27)	8.28$^{\#}$ (5.09)	F = 11.49 p < 0.0001
SleDur (main daily sleep)	8.56 (3.10)	9.24*** (3.60)	7.39$^{***/\#\#\#}$ (3.48)	F = 39.62 p < 0.0001
SleDur$_{18}$ (sum of sleeps 18:00-18:00)	9.03 (2.47)	9.83*** (2.76)	7.90$^{***/\#\#\#}$ (2.64)	F = 71.86 p < 0.0001
SleDur$_{daily}$ (mid-night to midnight sum of sleeps)	9.01 (2.50)	9.82*** (2.78)	7.80$^{***/\#\#\#}$ (2.75)	F = 75.21 p < 0.0001

Feature	Remission Mean (SD)	Depression Mean (SD)	Mania Mean (SD)	Anova
ISL (Immobile sleep)	0.77 (0.09)	0.76** (0.10)	0.77 (0.11)	F = 4.52 p = 0.0109
RSL (Restless sleep)	0.03 (0.02)	0.03 (0.02)	0.02$^{**/\#}$ (0.02)	F = 4.06 p = 0.0173
RMSSD$_{sleep}$	162.05 (40.70)	165.66* (41.75)	154.26$^{***/\#\#\#}$ (35.20)	F = 9.61 p < 0.0001
WASO	12.60 (26.90)	18.98*** (40.10)	10.22$^{\#\#\#}$ (23.03)	F = 16.16 p < 0.0001
APSO (Activity Prior Sleep Onset)	361.58 (125.69)	315.59*** (122.84)	368.33$^{\#\#\#}$ (132.02)	F = 42.86 p < 0.0001
AASO (Activity After Sleep Onset)	51.80 (25.29)	51.61 (22.61)	47.94$^{**/\#}$ (30.73)	F = 3.60 p = 0.0274
APWU (Activity Prior Wake-Up)	58.65 (24.85)	58.32 (22.13)	55.74* (26.02)	F = 2.17 p = 0.1138
AAWU (Activity After Wake-Up)	459.42 (133.95)	406.57*** (138.81)	430.93$^{***/\#\#}$ (134.10)	F = 47.45 p < 0.0001
APSO/AASO (sleep onset ratio)	0.22 (0.22)	0.26*** (0.22)	0.19$^{*/\#\#\#}$ (0.20)	F = 13.14 p < 0.0001
AAWU/APWU (sleep offset ratio)	0.18 (0.18)	0.22*** (0.22)	0.19$^{\#}$ (0.17)	F = 9.61 p < 0.0001
Explainable activity features - daily values				
ExAct	1319 (426)	1140*** (390)	1371$^{*/\#\#\#}$ (475)	F = 61.43 p < 0.0001
ExAct$_{active}$	88.26 (23.31)	80.98*** (22.57)	86.00$^{\#\#\#}$ (25.91)	F = 28.71 p < 0.0001
Complexity analysis – entropy - daily values				
SlopeEntropy$_{M10}$ $^{++}$	21.83 (2.61)	22.53*** (2.55)	21.53$^{*/\#\#\#}$ (3.12)	F = 25.88 p < 0.0001

Statistical significance * < 0.05 ** <0.01 *** <0.001 for rem-dep and rem-man differences using *t*-test

Statistical significance $^{\#}$ < 0.05 $^{\#\#}$ <0.01 $^{\#\#\#}$ <0.001 for dep-man difference using Wilcoxon rank-sum test

$^{+}$Feature calculations are described in Chapter 3 - section 3.5;

$^{++}$Slope entropy was estimated based on equations presented in (Cuesta-Frau *et al.*, 2020) for each day in the M10 window.