THE UNIVERSITY *of* EDINBURGH

Clinically-Interpretable and Large-Scale Machine Learning to Monitor Mood Disorders with Wearables

Filippo Corponi

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2024

Abstract

Mood Disorders (MDs) are common and severe psychiatric conditions with a relapsing-remitting course. Timely intervention during impending episodes improves outcomes, but pre-emptive measures are limited due to infrequent patient reviews and limited symptom reporting from patients. MDs involve changes in energy levels, circadian rhythms, and neurovegetative functions, correlating with changes in physiological data, like acceleration and galvanic skin conductance. Personal sensing, leveraging data from wearables, offers a way to monitor MDs remotely with objective biomarkers, which psychiatry currently lacks, relying mainly on clinical observation and patient self-reports. AI can harness wearable data to realise remote monitoring. I outline a unifying perspective on personal sensing for MDs and make original contributions to the field, using a prospective, observational cohort (TIMEBASE/INTREPIBD), recorded with an Emaptica E4 device

Our first contribution focuses on Heart Rate Variability (HRV), an indicator of the autonomic nervous system functionality. I find HRV increases as symptoms subside after acute episodes, suggesting it as a potential symptom improvement biomarker. Due to limited HRV study samples, I use Bayesian statistics to propose an interpretable probabilistic model, explaining the HRV data generating process. Longitudinal HRV data collection is indeed labour-intensive, often resulting in small samples that undermine frequentist statistics reliability.

Personal sensing research typically attempted to detect the mere presence of acute episodes or the total score on a psychometric scale, missing actionable clinical information. I propose inferring all symptoms from two popular scales assessing the full MD symptom spectrum, akin to a concept bottleneck. This approach ensures AI output is interpretable, recognizing that different symptom combinations require varied therapeutic strategies. I develope a model for this task and investigate key AI challenges.

Lastly, to address labelled data scarcity in AI systems for personal sensing, I gather open-access datasets using the E4 wearable, regardless of the task they are concerned with, and make such collection publicly available. I propose a Transformer model tailored to the E4 and show that self-supervised learning, repurposing unlabelled data to learn useful representations through surrogate tasks, is viable in personal sensing. This method outperforms fully supervised models, whether using deep learning or classical machine learning with hand-crafted features.

Lay Summary

Mood disorders are common and severe psychiatric conditions characterized by recurring disease episodes. Unfortunately, catching new episodes early is challenging because mental healthcare currently depends on infrequent check-ups, and patients may not notice or report early warning signs due to limited illness insight. Presently, clinical decision-making in mood disorders, like many other mental health conditions, relies nearly entirely on patient self-reports and doctor clinical judgment. Unlike other medical disciplines, there are no objective tests, like a blood test for diabetes in endocrinology.

Thankfully, modern technology provides new methods to study and monitor mood disorders. Wearables, especially wrist-worn devices, record data on energy expenditure, sleep, and heart rate, which are altered in patients with mood disorders, especially during acute episodes. This data is recorded near-continuously as patients go about their daily lives.

Artificial Intelligence (AI), a discipline focused on giving machines abilities typically seen only in intelligent animals, can analyse wearable data to advance our understanding of mood disorders. For example, by studying variations in heart rate, we can obtain objective mental state indicators, or "biomarkers". Moreover, the real-time nature of data collection with wearables provides unprecedented opportunities to monitor mood disorders "in the wild", as patients conduct their daily activities. This could be revolutionary, enabling early detection and early intervention.

In this thesis, I explore how AI can utilize wearable data to improve outcomes in mood disorders. There is a tension between simple models, easily interpretable by clinicians, and large-scale models, which may be better at detecting abnormal mood states but can often be too complex to be transparent. This dissertation investigates both approaches. I show how aligning large models with clinical practices and expectations is important for model outputs to be actionable. Finally, I address a fundamental challenge in AI applications to mental health, i.e. the limited size of datasets.

Acknowledgements

I would first like to thank my parents for their unwavering support. Without them, I could not have come this far. I would also like to thank my partner for her loving kindness and light-hearted nature, which complements my own. She patiently endured my long hours in front of the computer and always supported me.

Special thanks to Antonio, my AI supervisor. He graciously welcomed me as a student after an unconventional start to my PhD and despite my having an undergraduate background unusual for someone at the School of Informatics. Antonio provided relentless guidance, encouraged my academic growth, and was always engaged with the research questions I was eager to explore. I am also grateful to my biomedical supervisors, Heather and Stephen, for their strategic advice and prompt support whenever it was required.

My appreciation extends to Sharon for her support and mentorship in clinical practice and to the entire Community Mental Health Team at Cambridge Street House for their collaborative spirit.

During my PhD, I worked closely with my fellow PhD student Bryan, from whom I learned a lot about programming. I am thankful for his contributions to my research. Likewise, my gratitude goes to my collaborators in Barcelona, especially Diego and Gerard, for including me in the TIMEBASE/INTREPIBD project.

I would also like to thank Diego and Ian, the current and former directors of the CDT BAI, for their support to the CDT programme, and to Ekaterina and Isabelle, CDT BAI coordinators, for their constant readiness to assist with any request.

Finally, I extend my thanks to the contributors and maintainers of the various open-source frameworks that I used throughout my research. The machine learning community's commitment to open-source publication and software is exemplary, a practice I hope will spread throughout the clinical research community.

List of Publications

Listed below are the original works used in this dissertation. * denotes equal contributions. A complete list of my publications is available from Google Scholar.

- **Corponi, F.**, Li, B. M., Anmella, G., Valenzuela-Pascual, C., Pacchiarotti, I., Valentí, M., Grande, I., Benabarre A., Garriga M., Vieta E., Lawrie, S., Whalley, H., Hidalgo-Mazzei, D., Vergari, A. (2024). Does heart rate variability change over acute episodes of bipolar disorder? A Bayesian analysis. *Preprint* under review in *npj Mental Health Research.*

- **Corponi, F.** *, Li, B. M. *, Anmella, G., Mas, A., Pacchiarotti, I., Valentí, M., Grande, I., Benabarre A., Garriga M., Vieta E., Lawrie, S., Whalley, H., Hidalgo-Mazzei, D., Vergari, A. (2024). Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. *Translational Psychiatry*, 14(1), 161.

- **Corponi, F.**, Li, B. M., Anmella, G., Valenzuela-Pascual, C., Mas, A., Pacchiarotti, I., Valentí, M., Grande, I., Benabarre A., Garriga M., Vieta E., Young, H., Lawrie, S., Whalley, H., Hidalgo-Mazzei, D., Vergari, A. (2024). Wearable Data From Subjects Playing Super Mario, Taking University Exams, or Performing Physical Exercise Help Detect Acute Mood Disorder Episodes via Self-Supervised Learning: Prospective, Exploratory, Observational Study. *JMIR mHealth and uHealth*, 12, e55094.

Other works relevant to this dissertation but not included in it are:

- Li, B. M. *, **Corponi, F.** *, Anmella, G., Mas, A., Sanabra, M., Hidalgo-Mazzei, D., and Vergari, A. (2022). Inferring mood disorder symptoms from multivariate time-series sensory data. In *NeurIPS 2022 Workshop on Learning from Time Series for Health.*

- Anmella, G. *, **Corponi, F.** *, Li, B. M. *, Mas, A., Sanabra, M., Pacchiarotti, I., Valentí, M., Grande, I., Benabarre, A., Giménez-Palomo, A., Garriga, M., Agasi, I., Bastidas, A., Cavero, M., Fernández-Plaza, T., Arbelo, N., Bioque, M., García-Rizo, C., Verdolini, N., Madero, S., Murru, A., Amoretti, S., Martínez-Aran, A., Ruiz, V., Fico, G., De Prisco, M., Oliva, V., Solanes, A., Radua, J., Samalin, L., Young, H., Vieta, E., Vergari, A., Hidalgo-Mazzei, D. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating

and model development study. *JMIR mHealth and uHealth*, 11(1), e45405.

- Anmella, G., **Corponi, F.** ⋆, Li, B. M. ⋆, Mas, A., Garriga, M., Sanabra, M., Pacchiarotti, I., Valentí, M., Grande, I., Benabarre, A., Giménez-Palomo, A., Agasi, I., Bastidas, A., Cavero, M., Bioque, M., García-Rizo, C., Madero, S., Arbelo, N., Murru, A., Amoretti, S., Martínez-Aran, A., Ruiz, V., Rivas, Y., Fico, G., De Prisco, M., Oliva, V., Solanes, A., Radua, J., Samalin, L., Young, H., Veragari, A., Vieta, E., Hidalgo-Mazzei, D. Identifying digital biomarkers of illness activity and treatment response in bipolar disorder with a novel wearable device (TIMEBASE): protocol for a pragmatic observational clinical study. *BJPsych Open*. 2024;10(5):e137. doi:10.1192/bjo.2024.716

- Anmella, G., Mas, A., Sanabra, M., Valenzuela-Pascual, C., Valentí, Marc, Pacchiarotti, I., Benabarre, A., Grande, I., De Prisco, M., Oliva, V., Fico, G., Giménez-Palomo, A., Bastidas, A., Agasi, I., Young, A., Garriga, **Corponi, F.**, Li, B. M., de Looff, P., Vieta, E., Hidalgo-Mazzei, D. Electrodermal activity in bipolar disorder: Differences between mood episodes and clinical remission using a wearable device in a real-world clinical setting. *Journal of Affective Disorders*, 345, 43-50.

- Valenzuela-Pascual, C., Mas, A., Borràs, R., Anmella, G., Sanabra, M., González-Campos, M., Valentí, M., Pacchiarotti, I., Benabarre, A., Grande, I., De Prisco, M. Oliva, V., Bastidas, A., Agasi, I., Young, H., Garriga, M., Murru, A., **Corponi, F.**, Li, B. M., de Loof, P., Vieta, E., Hidalgo-Mazzei, D. Sleep–wake variations of electrodermal activity in bipolar disorder. doi: https://doi.org/10.1111/acps.13718 *Acta Psychiatrica Scandinavica*.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Filippo Corponi*)

Table of Contents

Acronyms

AI Artificial Intelligence. 1, 5, 6, 9, 11, 12, 16, 17, 23, 24, 27, 62, 63, 95, 99, 122, 124, 126, 127

ANN Neural Network. 1, 21, 26, 28, 65, 67, 95, 96, 123

ANS Autonomic Nervous System. 1, 14–17, 39, 123

BD Bipolar Disorder. 1, 6, 10–13, 19, 30, 31, 34, 36, 38, 41, 43–45, 123, 124

BVP Blood Volume Pressure. 1, 34

CGI-S Clinical Global Impressions-Severity. 1, 31, 32

CV Computer Vision. 1, 95, 96

DL Deep Learning. 1, 21, 64, 65, 67

DSM-5 Diagnostic and Statistical Manual 5^{th}. 1, 6, 11, 14, 24, 30–32

ECG Electrocardiogram. 1, 16, 39

EDA Electrodermal Activity. 1, 16, 34

GPS Global Positioning System. 1, 8, 15, 16

HC Healthy Control. 1, 9, 10, 19, 29, 34, 36, 38

HDI-95 Highest Density Interval 95. 1, 44

HDRS Hamilton Depression Rating Scale. 1, 13, 24, 30–32, 62, 68

HR Heart Rate. 1, 34

HRV Heart Rate Variability. 1, 10, 11, 16, 34, 39–41, 43–45, 123

IBI Interbeat Intervals. 1, 34, 40

ICD-11 International Classification of Diseases 11th Revision. 1, 6

iid Independent and Identically Distributed. 1, 18, 19

IPAQ Short Scale International Physical Activity Questionnaire. 1, 31–33

LSTM Long Short-Term Memory. 1, 65

MAP Maximum A Posterior. 1

MD Mood Disorder. 1, 5–17, 19, 20, 22–25, 27, 36, 40, 62, 63, 95, 99, 100, 122–126

MDD Major Depressive Disorder. 1, 6, 12, 14, 19, 30, 31, 34, 36, 38, 40, 41

MDE Major Depressive Episode. 1, 12, 13, 24, 30, 31, 38, 63

ME Manic Episode. 1, 13, 24, 30, 31, 36, 38, 41, 125

MET-min Metabolic Equivalent of Task-minutes. 1, 33

ML Machine Learning. 1, 17, 18, 20–23, 26–28, 64, 68, 95, 126

MLP Multilayer Perceptron. 1, 65, 67

MR Magnetic Resonance. 1, 14

MSE Mean Squared Error. 1, 96, 97

NHST Null Hypothesis Significance Testing. 1, 41, 42, 44

NICE National Institute for Health and Care Excellence. 1, 10, 14

NLP Natural Language Processing. 1, 66, 96, 99

NN Normal-to-Normal intervals. 1, 40

PD Probability of Direction. 1, 44

PDD Persistent Depressive Disorder. 1, 6, 12

PPG Photoplethysmography. 1, 16, 34, 39

RMSSD Root-Mean-Square of Successive Differences. 1, 40, 44

RNN Recurrent Neural Networks. 1, 65, 66

ROPE Region of Practical Equivalence. 1, 44

SCID-5-RV Structured Clinical Interview for DSM-5. 1, 30, 31

SDNN Standard Deviation of NN Intervals. 1, 40

SOFAS Social and Occupational Functioning Assessment Scale. 1, 31, 32

SSL Self-Supervised Learning. 1, 96, 100, 122, 123

TEMP Temperature. 1, 17, 34

YMRS Young Mania Rating Scale. 1, 13, 24, 30–32, 68

Chapter 1

Introduction

Mental ill-health has increasingly been identified as a global crisis, with roughly one in four people grappling with a mental condition [1] and total costs estimated at about USD 5 trillion worldwide in 2019 [2]. Furthermore, the aftermath of the COVID-19 pandemic [3] and geopolitical instability [4, 5] are exacerbating determinants of poor mental health. In the UK specifically, mental ill-health is the single largest cause of disability [6] and its annual costs amounted to 117.9 GBP billion in 2019 [7], i.e. approximately 5% of the national GDP. Mood disorders (MDs) account for a substantial portion of the global and domestic mental health burden [8, 9]. They are relapsing-remitting conditions, featuring disturbances in mood, energy levels, sleep, and cognition [10]. Such symptoms lie on a spectrum, going from depression to mania. The former is characterized by sadness, loss of interest in previously enjoyed activities, and feelings of worthlessness; on the other hand, elevated or irritable mood, increased energy, and hyperactivity are core features of mania [11].

In the face of rising demand for mental healthcare, the number of mental health specialists is shrinking [12] and innovation in psychiatry is stagnant, as decades of research in neuroscience and genetics has translated into hardly any novel clinical intervention [13, 14]. Artificial intelligence (AI) is a discipline focused on developing machines capable of performing tasks that traditionally only intelligent beings could perform. It has witnessed remarkable advancements over the past decade, and it promises to revolutionize healthcare. For example, it could help analyse vast amounts of data from a patient to predict outcomes or optimize therapeutic strategies [15]. The increasing adoption of wearable devices (or wearables), in particular wrist-worn devices, including

smartwatches and various types of wristbands, provides an abundant stream of clinically relevant data. At the cross-section of AI, digital technologies, and clinical science, a new, promising paradigm in mental healthcare, named digital phenotyping [16], could help mitigate the current mental health crisis.

1.1 Mood Disorders

MDs encompass a broad spectrum of mental health conditions, where mood disturbances stand out as one of the predominant clinical features. Such disturbances, running the whole gamut from depression to mania, are not merely transient emotional conditions, but pervasive and persistent states that substantially impair an individual's ability to function. Diagnostic criteria for MDs are found in the Diagnostic and Statistical Manual of Mental Disorders 5th Edition Text Revision (DSM-5)[10] and the International Classification of Diseases 11th Revision (ICD-11)[17]. MDs are distinguished into depressive disorders and bipolar and related disorders. The depressive disorders include Major Depressive Disorder (MDD) and Persistent Depressive Disorder (PDD) where the latter, previously known as dysthymia, involves a low-grade but chronic form of depression. Bipolar disorders (BDs), on the other hand, encompass type I BD, type II BD, and Cyclothymic Disorder, also known as Cyclothymia. The duration and the severity of the mood swings in Cyclothymia are milder than in full-blown BD and similarly, the impact on functionality is comparatively modest.

MDD and BDs are the most severe forms of MDs (Section 2.1). With a lifetime prevalence of around 10% [18], a typical onset in early adulthood [19], and a severe, relapsing-remitting course [20, 21], they constitute one of the world's greatest public health problems, with significant direct and indirect costs [9]. Their course is characterized by acute episodes followed by periods of symptom remission, referred to as euthymia in medical parlance. However, while some patients manage to return to their pre-morbid (symptoms-free) functioning, a significant proportion continues to struggle with one or more so-called residual symptoms after the resolution of the acute episode [22, 23]. Crucially, a longer duration of untreated acute illness has been consistently associated with worse outcomes, including treatment resistance, residual symptoms, and functional decline. Thus, timely recognition and early intervention in impending mood episodes is of paramount importance [24].

The contraction of the mental healthcare workforce coupled with an increasing preva-

lence of MDs has been putting a strain on access to services specialized in MDs. Furthermore, active monitoring of MDs by healthcare services currently relies on medical appointments, typically scheduled only every few months, where a clinician looks for indications of mood instability during the appointment as well as in the patient's daily life in the time between appointments, retrospectively. Crucially, doctor assessments heavily rely on clinical observation and patient self-reports, with no objective and measurable disease signatures to aid clinical decision-making. This is problematic as patient retrospective self-reports are subject to a number of biases, such as recall or social desirability, whose degree might vary based on current mood state. Furthermore, as limited patient insight is commonly observed in acute mood episodes, the delay between exacerbation onset and medical services acknowledging the unfolding episode is frequent, thereby reducing the scope for timely interventions [25, 26].

Besides abnormal emotional states, MDs' manifestations include disruptions in energy levels, motor activity, sleep, and vegetative functions. These can take different forms, often in opposite directions, e.g. insomnia or increased sleep (hypersomnia), psychomotor agitation or retardation [10]. Such symptoms have been shown to translate into changes in physiological parameters [27, 28, 29]. This observation, along with the increasing adoption of wrist-worn devices sensing physiological data, creates new opportunities for monitoring in MDs, taking it outside the doctor's office and grounding it on measurable and objective patterns of physiological data [30].

1.2 Digital Phenotyping

Digital phenotyping relies on digital technologies, such as smartphones, wearables, and sensors, for a near-continuous collection of a subject's digital footprint, in the context of their daily life [31]. By mining these digital traces, digital phenotyping promises to uncover signatures associated with altered mental states. Symptom monitoring could therefore be delivered during a patient's daily life, in their ecological environment, bridging the gap between medical appointments, typically offered only every few months. Furthermore, it could leverage objective measurements, aiding clinical judgement in psychiatry. The growing rate of technology access and wearables ownership [32], along with the good level of acceptability of wearable-based monitoring expressed by psychiatric populations [33], suggest an ideal environment for the implementation of digital phenotyping. This potential has been acknowledged by the World Health Organization

in a recent report [34] stating that digital interventions are the most promising way to reduce the global mental health burden.

There are two approaches to data collection in digital phenotyping. **Active data collection**, also known as ecological momentary assessment [35], refers to smartphone-based surveys or tests, which a participant can complete in response to a prompt or spontaneously. On the other hand, **passive data collection**, or personal sensing [36], exploits digital traces generated without active and deliberate participation from an individual, reducing patient burden related to active data collection. Furthermore, passive data is robust to biases affecting smartphone-based questionnaires, but such "objectivity" does not *ipso facto* imply it can be exploited to accurately monitor abnormal mental states, e.g. because of high noise or low signal in the data [30]. Examples of passively collected data include tracking Global Positioning System (GPS) location, monitoring physical activity levels through accelerometers, and analysing speech and smartphone usage patterns.

The two aforementioned approaches are not mutually exclusive, and some studies collected active and passive data in tandem. However, a major barrier towards long-term implementations of smartphone-based surveys is compliance and response quality degradation over time, as it may be impractical for patients to engage for prolonged periods of time [37]. On the other hand, personal sensing is less affected by this limitation, but, especially in the case of certain modalities, e.g. speech or social media, it may still be associated with suboptimal adherence and acceptability, due to privacy and security concerns as well as perceived surveillance [38, 39]. In this regard, some studies [40, 41] point to better user perception around physiological data, e.g. motor activity and sleep patterns from accelerometers.

1.3 Sensing Physiological Data

The interest in physiological data in the context of MDs has a long-standing tradition dating back to the XIX century [42] when measurements could only be taken with cumbersome equipment in the laboratory. The earliest investigations into activity and sleep patterns using actigraphy – a compact, lightweight device incorporating an accelerometer, typically worn on the wrist or ankle – date back to the 1980s [43]. The resurgence of interest in physiological data and the emergence of personal sensing as a field of research stems from the convergence of several innovations [44]. Sensor technology

has undergone significant advancements over the past decades, resulting in sensors that are smaller, lighter, and more accurate. Moreover, they have become increasingly pervasive due to the widespread adoption of smartphones and smartwatches, which are equipped with an array of sensors, enabling the collection of multiple data modalities during daily life. Concurrently, data collection and storage capabilities have undergone dramatic improvements, facilitating the accumulation of larger datasets [45], such as the UK Biobank [46] and the ABCD study [47]. While both include actigraphy, neither was collected for MDs specifically: recordings do not monitor any clinically significant event (e.g. disease exacerbation), span only a few consecutive days picked at random, and are not accompanied by specialist assessments of the wearers. A novel, longitudinal dataset addressing these limitations is iden**T**ifying d**I**gital bio**M**arkers of illn**E**ss activity in **B**ipolAr di**S**ord**E**r/**I**dentifying digital biomarkers of ill**N**ess activity and **T**reatment **RE**s**P**onse **I**n **B**ipolar **D**isorder (TIMEBASE/INTREPIBD) [48] (Chapter 3). Finally, recent advancements in AI have equipped researchers with novel data analytics techniques for mining this wealth of data, e.g. to predict outcomes at the level of single patients or cluster heterogeneous conditions such MDs into subgroups [49].

Over the past decade, passive sensing data has been explored for the identification of digital biomarkers, and objective and measurable indicators of a biological process, i.e. a mental health status in a psychiatric context. Unlike other branches of medicine, psychiatry currently lacks biomarkers to inform clinical decision-making [50]. For instance, while blood sugar levels are utilized in diagnosing and managing diabetes, no such biomarkers exist for MDs. Consequently, researchers extracted handcrafted features from raw physiological data, guided by clinical knowledge, and investigated group-level associations, typically using simple models and frequentist statistics. These associations are examined across patients and healthy controls (HCs), or across different illness stages (e.g. [51, 52]). As an alternative approach, embracing the individual-level orientation of precision psychiatry, where interventions are tailored to the specific characteristics of each patient, researchers have more recently turned to AI techniques to infer clinical outcomes from passive sensing data at the level of individual patients (e.g. [53, 54]). These endeavours, considering the black-box nature of large(r) AI models, do not typically prioritize a deep understanding of the pathophysiology underpinning MDs.

1.4 Main Contributions and Structure of the Thesis

The field of wearables and, more broadly, digital technologies are credited with the potential to revolutionize mental healthcare. However, this potential remains largely untapped to this day. Commercially available applications focus on "wellness" rather than "health" solutions, often bypassing stringent regulations. These products typically lack robust evidence for their effectiveness and, in some cases, have even caused harm to users [55]. Despite these concerns, there's reason for cautious optimism. Wearables have recently gained recognition for their role in managing blood glucose for diabetes and remotely monitoring Parkinson's disease, as evidenced by recommendations from the National Institute for Health and Care Excellence (NICE) [56, 57]. However, applying wearables to mental health, particularly MDs, presents unique challenges.

This thesis delves into passive sensing for MDs. It is structured around three original works, constituting its core original contribution, each exploring a distinct aspect of the field. Beyond the individual studies, this thesis offers a **unifying perspective** on passive sensing in MDs, and it identifies key **opportunities and challenges** in the development of clinical decision support tools leveraging passive sensing for MDs. Each publication-based chapter contextualizes the respective original study in the surrounding literature before including the associated paper as it appears in the respective venue. The rest of this dissertation is structured as follows:

Chapter 2 - The necessary background for this dissertation is provided. I argue the case for passive sensing of physiological data with wrist-worn devices as the most suitable approach to remote monitoring in MDs. I then outline the main challenges and research frontiers in the field, to which the original studies presented in the subsequent chapters provide contributions.

Chapter 3 - I present TIMEBASE/INTREPIBD, which my original research relies on. This is a longitudinal, observational, exploratory single-centre study, recruiting three groups of participants recorded with an Empatica E4 [58] device: A) patients on an acute mood episode, B) patients with a historical MD diagnosis but clinically stable at the moment of study admittance and C) HCc.

Chapter 4 - I study patterns of change in Heart Rate Variability (HRV) over acute episodes in BD, proposing an interpretable Bayesian probabilistic model, attempting to explain the HRV generating process. I introduce Bayesian statistics to the field of biomarkers study in personal sensing for MDs and illustrate its benefits, especially with

small samples. Findings indicate that HRV increases as symptom severity improves, but there are no polarity-specific patterns (i.e., depressive vs. manic).

Chapter 5 - I explore the inference of mood states from personal sensing data and introduce a new task: predicting all items from two popular psychometric scales that assess symptoms of depression and mania (Appendix A). This task better aligns with daily clinical practice, as different symptom combinations, requiring different therapeutical and management approaches, may underlie the same diagnosis. Thus, reducing MDs to a single label (e.g., a diagnosis or the total score on a scale) misses actionable clinical information. Furthermore, the single label can be recovered from scores on psychometric questionnaire items, which adds to the model's interpretability, as the model can indeed be interrogated as to which symptoms (concepts) motivated a given output.

Chapter 6 - I take on one of the key challenges in deploying modern AI systems in personal sensing, that is scarcity of annotated data. I curate the largest open-access collection of datasets recorded with the E4 device. I show that pre-training an E4-tailored model on this (relatively) large corpus of data in a self-supervised fashion leads to a significant boost in performance in a downstream task of acute episode detection. Results show that self-supervised pre-training followed by fine-tuning outperforms end-to-end supervised learning.

Chapter 7 - I synthesize the contributions of this thesis, discuss its limitations, and outline future research directions for advancing personal sensing in MDs.

The original works in Chapter 4, Chapter 5, and Chapter 6 represent my main recent research output. I herewith acknowledge other works on personal sensing for MDs I co-authored: a) Anmella et al. 59, presenting the TIMEBASE/INTREPIBD protocol; b) Anmella et al. 48, preliminary analyses on inferring DSM-5 diagnoses from wearable data; c) Li et al. 60, a workshop paper presented at Learning from Time Series for Health workshop during NeurIPS2022, laying the groundwork for data preprocessing in TIMEBASE/INTREPIBD and regression of manic and depressive symptoms from wearable data; d) Valenzuela-Pascual et al. 61, showing that electrodermal activity, a proxy for sympathetic nervous system activity (Section 2.2.1) can help distinguish clinical phases in BD, especially if measured during wake rather than during sleep.

Chapter 2

Background

This chapter gives an overview of the primary research frontiers in personal sensing for MDs and lays the groundwork for the original research presented in **Chapter 4**, **Chapter 5**, and **Chapter 6**. Through the present chapter, it becomes evident that the barriers to implementing passive sensing for MDs lie at the intersection of clinical science, computer science, and engineering. While a multidisciplinary approach holds promise, these fields have yet to be fully integrated. Furthermore, while some open problems presented here are specific to personal sensing in MDs, others are common across various applications of digital health and AI in healthcare.

2.1 Major Depressive Disorder & Bipolar Disorders

Within the MDs spectrum, MDD and BDs have the largest impact on public health [3]. We henceforth use the term MDs for referring to these two conditions, as our investigation will not be concerned with PDD or cyclothymia (Section 1.1).

MDD is characterized by one or more major depressive episodes (MDEs), defined by at least two weeks of pervasive low mood or anhedonia (decreased interest in previously enjoyed activities), accompanied by other symptoms such as feelings of guilt or worthlessness, lack of energy, poor concentration, appetite changes, psychomotor retardation or agitation, sleep disturbances, and suicidal thoughts [21]. With a life prevalence of around 12% and an average onset age between 25 and 32 years, MDD is expected to become the first cause of disease burden by 2023. Only little difference in prevalence has been found based on racial background and socioeconomic status,

whereas rates in females are almost double that in males [21].

BDs as well can involve MDEs, but the occurrence of (hypo)mania is required for their diagnosis. In particular, type I BD is defined by the occurrence of at least one manic episode (ME), which is a period of abnormally elevated, expansive, or irritable mood lasting at least one week, accompanied by increased energy and activity, and significant functional impairment. Other symptoms may include inflated self-esteem, decreased need for sleep, talkativeness, flight of ideas, distractibility, increased goal-directed activity, and risky behaviours. Type II BD, on the other hand, involves at least one hypomanic episode (a milder form of mania causing relatively less functional impairment) and one MDE [10]. BDs affect >1% of the global population, showing a relatively equal distribution across sex, ethnicity, and urban compared to rural areas, and their typical onset age is between 20 and 25 years of age [20].

Symptoms of depression and (hypo)mania are assessed using standardized questionnaires, whose score is used to measure clinical outcomes in clinical trials and research studies [62]. For depression, the Hamilton Depression Rating Scale (HDRS) [63] is frequently used, while (hypo)mania is often evaluated with the Young Mania Rating Scale (YMRS) [64]. These questionnaires consist of Likert-type ordinal items, where clinicians rate the severity or frequency of symptoms on a scale (e.g., 0 to 4, where 0 might indicate the absence of a symptom and 4 indicates severe presence). For example, an HDRS item assesses the severity of "Agitation", with responses ranging from 0 ("None") to 4 ("Hand wringing, nail-biting, hair-pulling, biting of lips."). The full HDRS and YMRS questionnaires are reported in Appendix A. As they are not used in routine clinical practice and require rater training, these questionnaires are particularly expensive to acquire (Chapter 6), especially in longitudinal studies where a patient is assessed at multiple time points. Of note, the total score on these questionnaires is typically binned to define severity brackets [65]. However, the same total score, even within the same nosographic category (e.g. MDE), can be realized from different symptom combinations, which are acted on differently in the clinical practice, for example recommending one particular antidepressant over other options (Chapter 5).

2.1.1 Elusive Biological Underpinnings

MDs are complex conditions that cannot be fully explained by any one single established biological or environmental pathway. Instead, they seem to be caused by a combination of genetic, environmental, psychological and biological factors [20, 21]. Furthermore,

evidence suggests that even within a single DSM-5 nosographic construct, e.g. MDD, heterogeneous conditions with distinct biological underpinnings may be conflated [66]. An analogy from internal medicine can illustrate this point. Fever can arise from various underlying illnesses, e.g. an infectious disease or a tumour, which can be revealed and followed up with different examinations. Similarly, diverse biological processes might underlie conditions that currently fall under the same construct, such as MDD.

In medicine and science at large, the introduction of novel measurement techniques has historically catalysed advancements in understanding biological processes, leading to groundbreaking discoveries. This phenomenon is exemplified by the impact of microscopy and magnetic resonance (MR) imaging on medical research [67, 68]. Similarly, passive sensing has ushered in unprecedented opportunities to gather measurements from patients as they navigate their daily lives. Researchers are now leveraging this tool to illuminate various aspects of biological functions in MDs. For instance, passive sensing enables the monitoring of the state of the autonomic nervous system (ANS) over acute mood episodes, which may correlate with the progression to remission [51, 69]. Consequently, there is optimism that personal sensing could advance our understanding of MDs, for example revealing different clinical subtypes of MDs, associated with specific patterns of physiological data.

To this day, however, as the precise biological mechanisms underlying the development and progression of MDs remain poorly understood [20, 70], biological justification for what modalities or features should be used for monitoring MDs is limited. The picture is quite different for other medical conditions, for example, diabetes mellitus, where NICE recommendations have recently included the use of wearables, specifically hybrid closed loop systems for managing blood glucose levels in type 1 diabetes [57]. Diabetes affects multiple body systems, yet its core pathology involves either an autoimmune reaction targeting insulin-producing pancreatic cells (type I) or reduced sensitivity to insulin signals in peripheral cells (type II). These mechanisms both result in dysregulated glucose levels, driving the clinical manifestations of diabetes [71]. Unfortunately, MDs, initially linked to imbalanced neurotransmitter levels and more recently to brain circuits organization and function [72], lack such a clear (causal) model informing personal sensing.

2.2 The Case for Wrist-Worn Devices

In the realm of passive monitoring for mental health, including MDs, various devices and sensors have been used, with the wearable market experiencing rapid expansion and introducing new technologies at a swift pace. Smartphones and smartwatches emerge as the most prevalent wearables in research studies [73, 74, 75, 76, 77]. This popularity aligns with high adoption rates in the general population; for instance, in the US, approximately 90% of adults owned a smartphone in 2023 [78], with over 20% owning a smartwatch or fitness tracker according to a 2019 report [79]. Smartphones offer a wealth of passive data, such as call logs, app usage (including social media and text data), and speech data. However, due to the sensitivity of such data, users' willingness to share it for ongoing remote monitoring in clinical settings may be limited, potentially impacting compliance. Conversely, less sensitive smartphone data, such as touchscreen typing/tapping patterns and GPS data, have shown promising results in personal sensing applications [80, 81, 82]. Although smartphones possess accelerometers capable of recording activity levels in principle, studies suggest such data may be noisy and unreliable as users do not consistently carry smartphones on their bodies [83].

In contrast, wrist-worn devices capture various physiological data modalities [77]. Given the correlation between changes in physiological parameters and MDs [27, 28, 29], the relatively fewer privacy concerns compared to other data modalities, ease of use, and high adoption rates among both the general and psychiatric populations, wrist-worn devices have increasingly become the focus of personal sensing in MDs [73, 74, 75, 76, 77]. Consequently, this thesis will focus on wrist-worn devices. It is still worth mentioning other devices with arguably limited potential for ongoing remote monitoring. Indeed, they are rather cumbersome for daily use and do not offset this limitation against a provenly stronger signal for MDs monitoring. Smart fabrics, i.e. garments embedded with various sensors, and smart patches have been developed for monitoring cardiac, respiratory, and perspiratory activity, which serve as proxies for ANS activity [75]. Additionally, smart headsets and goggles equipped with sensors for electro-oculography and electroencephalography have been utilized in research across various mental health conditions, including schizophrenia, depression, and post-traumatic stress disorder [84].

2.2.1 Physiological Data Modalities

Acceleration & GPS - The array of modalities available through wearables for research into biomarkers and personalized AI systems in the context of MDs is broad, as the technology is moving at a fast pace. Acceleration, which measures the rate of change of velocity in terms of both speed and direction, is tracked with most wearables, outputted as raw data or some derived features. Its utility in studying activity and sleep patterns is well-established, with decades of actigraphy research supporting its relevance in MDs [29]. Some researchers have even proposed interpreting MDs as disturbances of energy rather than primarily mood-related phenomena [85, 86].

GPS data, consisting of time-stamped longitudes and latitudes, although not directly capturing physiological information, serves as a proxy for physical activity and social interactions. For instance, it can track the time spent at home relative to other locations. The study of GPS data in MDs is fairly recent, and its unique role relative to acceleration, which is more commonly collected with wrist-worn devices, is yet to be ascertained [82].

Photoplethysmography & Electrocardiogram - Both photoplethysmography (PPG) and electrocardiogram (ECG) sensors assess various aspects of heart health, which in turn reflects the state of the autonomic nervous system (ANS). However, their methodologies differ significantly. PPG operates by indirectly estimating changes in blood flow, illuminating the skin and measuring the absorption or reflection of light by blood vessels. In contrast, ECG employs electrodes to directly detect the electrical activity of the heart muscle, making it the gold standard for heart monitoring. ECG's size and power requirements, however, typically limit its presence in commercial wearables [87].

Monitoring heart health is particularly relevant for individuals with MDs, who experience a higher burden of cardiovascular co-morbidities compared to the general population [88]. More specifically to mental health, research suggests that an imbalance in the ANS is common across various psychiatric conditions. This has fuelled interest in HRV as a potential biomarker for MDs (Chapter 4). HRV measures the variation in time intervals between heartbeats, providing an indirect assessment of ANS function [89].

Electrodermal Activity - Electrodermal activity (EDA) sensors measure changes in electrical skin conductance, reflecting sweat gland activity. These changes are influenced

by the sympathetic branch of the ANS, which plays a key role in the body's fight-or-flight response. Electrodermal hypo-activity has been linked to depression and suicidal risk [28].

Temperature - While less commonly explored compared to other physiological data, skin temperature (TEMP) may also hold promise as a potential biomarker in MDs, where higher values during wake time were associated with greater depression symptom severity. The reasons behind this connection are still being investigated, and it is uncertain whether it results from increased metabolic heat production, decreased ability to induce thermoregulatory cooling, or a combination of both [90].

2.3 Machine Learning Approaches

In addition to its role in advancing our understanding of MDs, personal sensing is being explored for its potential to enhance symptom monitoring, thereby enabling timely interventions. Machine learning (ML) approaches have increasingly been utilized towards this end. ML is a subset within the broader discipline of AI, aimed at creating intelligent systems. It focuses specifically on enabling computers to learn from data without explicit programming, while AI encompasses other approaches too, e.g. logic and rule-based systems. Due to ML recent popularity and success in various domains, the terms AI and ML are sometimes used interchangeably, as we will do in this dissertation [91, 92]. The growing interest in deploying ML in healthcare is evident from the estimated global market size of USD 19.27 billion in 2023 for ML-enabled healthcare applications, projected to grow at a compound annual growth rate of 38.5% from 2024 to 2030 [93].

2.3.1 Supervised Learning

In the realm of personal sensing, supervised learning, a key paradigm within ML, has been extensively utilized [94, 95, 96, 97, 98, 99, 100]. In a supervised learning setting, there are two spaces of objects X and Y and we would like to learn a function, referred to as hypothesis, $h : X \mapsto Y$, which outputs a prediction $\hat{y} \in Y$, given $x \in X$. To achieve this, we have a training set $D = \{(x_i, y_i)\}_{i=1}^{N}$ at our disposal, where $x_i \in X$ is the input and $y_i \in Y$ is the correct output we would like $h(x_i)$ to produce [101]. In personal sensing, x is a wearable recording, concretely a multi-variate time-series (Figure 2.1), i.e. a collection of univariate time-series vectors, each coming from one of the sensors

in the device (Section 5.3). On the other hand, y represents the corresponding mood state of the wearer: this could be a scalar, e.g. acute episode vs euthymia, or a vector, e.g. severity of manic-depressive symptoms (Chapter 5).

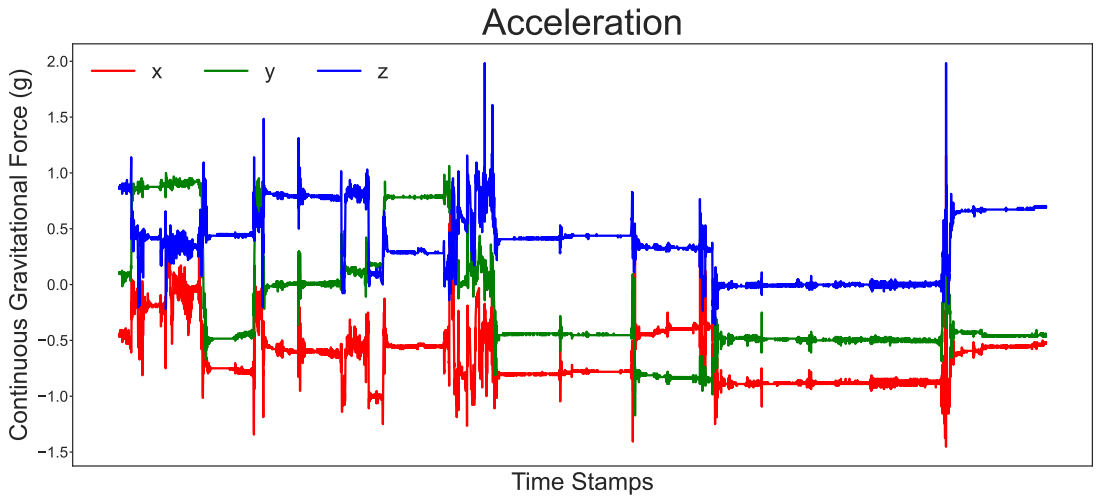


Figure 2.1: **Triaxial acceleration.** An example of a multi-variate time-series input, x, recorded with the accelerometer from Empatica E4 [102], the device used in TIMEBASE/INTREPIBD (Section 3.3). On the x-axis are time stamps, and on the y-axis are gravitational force (g) values across the three orthogonal space directions. Research grade devices, such as Empatica E4, typically provide other sensory modalities along with acceleration (Section 2.2.1 and Section 3.3). These may be sampled at different frequencies (Hz). So, time series from different sensors may have unequal lengths.

Many ML algorithms, including those presented in the chapters that follow, rely on the principle of **Empirical Risk Minimization** to learn h [103]. The (x_i, y_i) tuples are assumed to be independent and identically distributed (iid) samples from a joint probability distribution $P(x,y)$. While this assumption may not hold for time-series data, where temporal dependencies exist, it is still applied across various state-of-the-art algorithms for time-series [104]. A non-negative real-valued *loss function*, $\mathcal{L} : Y \times Y \mapsto [0, +\inf)$, measures the quality of the prediction $\hat{y}$, i.e. how far off it is from the correct output y. The risk associated with the hypothesis $h(x)$ is defined as the expectation of the loss function, taken over the joint probability $P(x,y)$:

$$\mathcal{R}_{P(x,y)}(h) = \mathbf{E}_{P(x,y)}[\mathcal{L}(h(x),y)] = \int \mathcal{L}(h(x),y)dP(x,y)$$

The optimal hypothesis $h(x)$ minimizes $\mathcal{R}_{P(x,y)}(h)$. Unfortunately, we do not have access to $P(x,y)$, only a collection of iid samples drawn from it, which constitute our training set D. Therefore, we select a hypothesis $h(x)$ that minimizes an estimate of $\mathcal{R}_{P(x,y)}(h)$, called the empirical risk $\mathcal{R}_{emp}(h)$, calculated by replacing the integral with an average taken over samples in D:

$$\mathcal{R}_{emp}(h) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(h\left(x_i\right), y_i\right)$$

The law of large numbers indicates that $\mathcal{R}_{emp}(h) \simeq \mathcal{R}_{P(x,y)}(h)$ when N is large. However, there is a caveat. There exist multiple functions $h(x)$ such that $\mathcal{R}_{emp}(h) = 0$, some of which are irregular and do not generalize well to future data points. This is problematic because we seek good predictions on future, unseen data, not included in the training set D. To mitigate this, $\mathcal{R}_{emp}(h)$ is minimized over a restricted class of functions $\mathcal{F}$, encoding some regularity we expect h to possess. Regularization is one method that modifies $\mathcal{L}$ to constrain $\mathcal{F}$. While $\mathcal{L}$ minimization is the objective guiding how h is learned, it is sometimes more convenient to use a different function of $h(x)$ and y, known as *metric*, to summarize the performance on a given task and compare different hypotheses. Metrics are designed to be human-interpretable, encoding insights into the task at hand, and may allow for greater design flexibility than the loss function, which, for example, may need to be differentiable with respect to $h(x)$'s parameters.

There are different tasks within supervised learning, based on the form of the output y [101]. Regression is concerned with predicting a continuous output variable, such as the total score on a psychometric scale. In contrast, classification deals with categorical variables and can be further distinguished into binary classification, where only two classes are possible (e.g., acute state vs. euthymia), and multi-class classification, where there are more than two classes (e.g. MDD, BD, and HC). Some output variables y exist on a discrete scale with a meaningful order between different values. The corresponding supervised task, sitting between classification and regression, is called ordinal classification (or ordinal regression), such as when the total score on a psychometric scale is binned into severity bands (absent, mild, and severe symptomatology). In all the tasks mentioned above, y can be represented with a single scalar. However, there are instances where y is a vector. For example, in multi-label classification, an input can belong to multiple classes simultaneously, such as a patient with an MD who may

also have various physical comorbidities. The different forms y can take are typically reflected in the functional form of the loss and the metric.

2.3.2 Unsupervised Learning

Unsupervised learning is another critical paradigm within ML, where the goal is to discover patterns or structures in data [101]. Unlike supervised learning, unsupervised learning deals with a dataset $D = \{(x_i)\}_{i=1}^{N}$, where $x_i \in X$ and no corresponding y_i target values are used to supervise the training process. While labels do not inform the training objective, they may be still required to validate or examine the insights from unsupervised methods. Examples of tasks under this learning paradigm, used in the personal sensing literature or in the chapters that follow, include anomaly detection, clustering, and dimensionality reduction.

In **anomaly detection**, the objective is to identify items or events that differ significantly from the majority of the data. In the context of personal sensing for MDs, an acute episode can be viewed as a departure from a patient's baseline patterns of physiological data [105]. Formally, given a set of observations $\{(x_i)\}_{i=1}^{N}$, we aim to learn a function $f : X \mapsto \mathbb{R}$ that assigns an anomaly score to each x_i. Anomalies are then identified as those observations with scores above a certain threshold. This task can be pursued probabilistically, i.e. we try to learn $P(x)$ and identify data points that have low probability under our model. **Clustering** involves assigning a set of observations $\{(x_i)\}_{i=1}^{N}$ into K clusters $\{(C_k)\}_{k=1}^{K}$ such that observations within the same cluster are more similar to each other than to those in other clusters. For example, we may want to cluster patients based on some biological data and see to what extent the clusters align with medical diagnoses or are associated with clinical outcomes [106]. **Dimensionality reduction** aims to reduce the dimensionality of the data, such that $x_i \in \mathbb{R}^N$ is mapped to $z_i \in \mathbb{R}^M$ with $M << N$, while preserving some information about the original data. This is particularly important in high-dimensional data to improve computational efficiency, facilitate visualization, and reduce noise. Principal Component Analysis [107] is a staple linear dimensionality reduction technique in which we seek to project the data into a lower-dimensional space while preserving as much variance as possible.

2.3.3 Representation Learning vs Feature Engineering

Representation learning refers to the process of learning a representation $z = g(x)$ from an input object x towards a specific task [108]. This is in contrast to feature engineering, where a human expert designs features from the raw data [109]. Either way, the goal is to transform raw input data (e.g., acceleration values sampled from a wearable device) into a form that facilitates performing the task at hand (e.g., detecting whether the wearer is experiencing an acute mood episode). Traditional ML techniques, such as linear regression [110], support vector machines [111], and decision trees [112], rely on handcrafted features. In this approach, the transformation $g(x)$ is not learned from data but specified in advance by the researcher, often based on domain knowledge. For example, the minimum, maximum, or entropy of acceleration values recorded over a given period may be extracted and fed into a classifier [113]. While using handcrafted features can enhance interpretability since the features are human-designed, it also limits the model to a predefined set of features that may not be optimal for the task.

Deep Learning (DL), a collection of algorithms based on Artificial Neural Networks (ANN), have revolutionised the field of representation learning. DL algorithms are characterized by multiple stacked layers, each composing mathematical functions defined by the preceding layer. Research in ANN has shown that deeper networks, which stack more layers, outperform their shallower counterparts by learning hierarchical representations of data. The successive layers refine lower-level representations into higher-level abstractions. DL excels particularly in tasks involving unstructured data, such as images, text, and speech, where basis functions are automatically and adaptively learned from data. This approach eliminates the need for pre-specified feature extraction and allows the discovery of optimal features directly from data [114]. However, DL requires substantial computational resources and, more importantly, large datasets to perform well. This poses challenges in healthcare applications, where data collection is often a bottleneck (Section 2.5). Consequently, traditional ML techniques remain viable alternatives for small datasets [115].

In a supervised setting, DL models are trained to minimize a loss function $\mathcal{L}$ that quantifies the difference between the desired output y and the model's predictions $\hat{y} = h_\theta(x)$, where the form of h is an ANN with parameters θ. Crucially, $\mathcal{L}$ is differentiable with respect to the ANN's parameters θ. The training process involves back-propagation to compute the gradient of $\mathcal{L}$ with respect to θ and an optimization algorithm leveraging the gradient to minimize the loss. Back-propagation works by applying the chain

rule of calculus from the output layer backwards to the input layer. The optimization algorithm relies on the gradient, i.e. the vector pointing towards the direction of steepest increase of the loss as a function of the parameters θ, to navigate the loss function landscape towards a minimum, updating θ values in the oppositive direction of the gradient. Instead of computing the gradient over the entire dataset, which is typically computationally expensive, mini-batches of data are used to perform θ updates. For each mini-batch, the gradient of the loss function with respect to θ is computed, and θ values are updated in the direction that reduces the loss. This process is repeated over multiple iterations (epochs) over the dataset until a stopping criterion is met. As the θ parameters are updated, the representations (features) encoded by the layers of the network are refined [116, 114].

2.4 Which Machine Learning Task?

The field of personal sensing in MDs has yet to converge on what constitutes a fruitful and clinically relevant ML task. While many studies have adopted a supervised learning approach [94, 95, 96, 97, 98, 99, 100], there is considerable variability in the targets they aim to predict, even within the same diagnostic category (Chapter 5). These targets are typically scalar, such as the total scores on mood questionnaires [94] (regression) or the presence of an acute episode [95] (classification). Moreover, some studies rely on psychiatrist-led interviews, and others on patient-reported smartphone surveys. Additionally, retrospective labelling, where mood episodes are determined based on patient history preceding medical appointments, has been used [99]. A minority of studies have pursued unsupervised anomaly detection [105, 117, 118]. These studies too, however, depend on how ground truth labels are defined, as any anomaly in the data is validated against clinical outcomes.

Regardless of their specific approach, studies utilizing ML aim to determine whether and to what extent clinically relevant events can be inferred with personal sensing. However, the metrics employed to assess performance in these studies tend to overlook practical aspects of personal sensing integration into clinical practice. For example, how far in advance compared to ordinary clinical follow-up should an episode be recognized so that a significantly better outcome for the patient can be delivered? What is an acceptable trade-off between specificity and sensitivity in mood episode detection, considering costs, equally for the patient and the healthcare system, of a false alarm or,

on the other hand, of a missed exacerbation [119]?

A significant barrier to standardizing and accelerating research in ML solutions for personal sensing is the limited availability of benchmark datasets for MDs [120, 121]. While valid confidentiality concerns contribute to this issue, it is worth noting that the concept of benchmark datasets for ML research is relatively new in the clinical research community. Additionally, the vertical structure of clinical research, where access to data is often controlled by a few gatekeepers, may further hinder the sharing of datasets. Curating a collection of existing datasets would also allow for testing of the ML models on independent samples, which is not generally done at the moment.

2.5 Scarce and Noisy Labels

ML systems are typically trained on human-annotated data in a supervised manner, with performance heavily reliant on access to extensive annotated datasets, running into dozens of thousands (e.g. CIFAR-10 [122]) or even millions of data points (e.g. ImageNet [123]). In the realm of personal sensing for MDs, labelling entails the involvement of mental health specialists to assess patients from whom data is collected via wearables. However, this process is exceptionally resource-intensive, representing a significant bottleneck in curating large annotated datasets. Consequently, studies typically can afford to recruit only a few dozen patients [124, 29, 95, 98, 96, 94, 99, 125]. Since personal sensing is intended for longitudinal patient monitoring, multiple assessments from a single patient are required, exacerbating the issue of labelled data scarcity.

Data annotation poses a constraint across various healthcare applications of AI. Furthermore, more specifically to mental health, limited inter- and intra-rater reliability [126] jeopardize labels' quality. Inter-rater reliability refers to the consistency among different specialists when evaluating the same patient, such as scoring them on a psychometric scale or providing a diagnosis. On the other hand, intra-rater reliability pertains to the consistency of assessments made by the same specialist over time. These challenges stem from the limited pathophysiological insights into mental health, resulting in a diagnostic process heavily reliant on clinical observation. Consequently, labels are noisy and, ideally, assessments from multiple specialists would be required to train and gauge the performance of AI systems.

An alternative approach, albeit underexplored thus far, would involve utilizing "hard"

labels, eliminating the potential for inter- or intra-reliability issues. For instance, hospitalization could serve as a concrete outcome measure; however, further research, including engagement with individuals with lived experience, is warranted to delineate the most suitable outcomes. As regards the example of hospitalization, it is worth noting that only a fraction of acute mood episodes reaches a severity level warranting hospital admission [127, 128, 129, 130].

2.6 Heterogeneous Signal from Physiological Data

Based on the DSM-5 [10], an acute mood episode, be it MDE or ME, can be observed under different symptom combinations. For example, for a diagnosis of MDE, different combinations including a minimum of five out of nine criteria are allowed (where at least one of a) depressed mood or b) loss of interest is required). Additionally, disruptions within specific domains, such as sleep, can manifest in opposing extremes, e.g. insomnia or hypersomnia. Psychometric scales for symptom severity, such as HDRS [63] or YMRS [64], allow for diverse behavioural presentations for the same score on a given item. For instance, high agitation (HDRS item 9) may manifest through hand wringing, nail-biting, hair-pulling, or lip-biting. This shows that, labels in mental health, where DSM-5 and psychometric scales are staples of clinical research, allow for high heterogeneity in the patterns of physiological data related to mood episodes.

Recent studies [131] have confirmed that signals from wearables, indicative of MDs, vary across different populations. For instance, mobility may correlate positively with depression risk in one population but negatively in another. The features (or sensory modalities) most relevant for detecting MDs likely vary across different populations. Moreover, factors related to MDs are often entangled with factors influencing physiological variability, such as age, sex, or physical fitness. AI systems developed in the literature are typically trained on physiological data inputs only and do not account for clinical-demographic factors.

It has been shown that deviation from a person's mean rather than the absolute value of a feature (modality) is linked to MDs [132]. Consequently, AI systems, also considering the scarcity of labelled data mentioned above, struggle to generalize across different subjects or populations. This problem has been described in other healthcare applications of personal sensing, e.g. seizure detection [133], where it has been mitigated with domain adaptation solutions. An alternative strategy being explored [131] is the

development of **idiographic** models – personalized models tailored to each individual – contrasting with **nomothetic** models that assume a function mapping from wearable data to mood states generalizable across different individuals.

2.7 A Multitude of Different Devices

Numerous wrist-worn devices have been developed and utilized for personal sensing research in MDs. Fitbit emerges as the most commonly employed wearable device technology brand in the literature, closely followed by Empatica [73]. However, the presence of diverse devices contributes to fragmentation within the field of personal sensing, presenting a barrier to the aggregation of data from various studies. Notably, wearables exhibit disparities not only in the sensors or features they offer but also in the specifics of shared sensors, such as accelerometers, across different devices, resulting in disparate data distributions. Fitbit and Empatica exemplify the categories of **consumer-grade** and **research-grade** devices, respectively.

Consumer-grade devices usually provide only a limited selection of features, such as heart rate, steps, sleep duration, and activity intensity, extracted through proprietary algorithms, while access to the raw data is often limited. Furthermore, in order to maximize battery life, the sampling frequency is comparatively lower. They are generally cheaper and prioritize comfort and design [134]. It has been reported that manufacturers updated their preprocessing algorithms, without giving clinicians or researchers any notice, making lack of access to raw data even more of a problem [135]. Lastly, the reliability of the measurements with respect to gold standard laboratory equipment may be suboptimal [136, 137].

Research-grade devices, on the other hand, come equipped with a wider range of sensors and collect measurements at higher frequencies. This, however, comes at the expense of battery life. Studies using this kind of device, therefore, usually record shorter periods of time. Further to pre-processed features, extracted with black-box algorithm, raw measurements themselves are made available. Research-grade devices tend to be pricier and bulkier, but efforts are underway to enhance their comfort and aesthetics [77]. The reliability of their measurements is generally better than their consumer-grade counterparts [138, 136].

Compliance remains a crucial challenge toward real-world implementation of personal sensing, particularly given the intended use for ongoing remote monitoring [37]. Stud-

ies following up with patients over extended periods often incentivize participation monetarily to mitigate attrition, though such practices may not be feasible in a clinical setting [139]. Without immediate perceived benefits, patients may be less inclined to maintain device wear, especially if the device serves no purpose beyond data collection. Strategies such as gamification and equipping devices with additional functions relevant to patients' daily lives may play a role in enhancing compliance [140].

2.8 Unveiling the Black Box

Interpretability in personal sensing, like in other healthcare applications, is fundamental for ensuring the reliability and trustworthiness of ML models. It entails justifying the output of an ML model with an explanation that is meaningful for the user. Without some insight into the rationale behind an ML model's behaviour, both patients and clinicians may hesitate to trust it or act upon its recommendations [141].

Some models are considered *inherently* interpretable, at least to some extent. This is the case of simpler models inputting hand-crafted features, e.g. linear regression where the weight of each feature represents the mean change in the prediction given a one-unit increase of the feature. Some ANN architectures incorporate at least some degree of interpretability. A case in point is attention networks (e.g. Transformers [142]), essentially computing a weighted context vector as a conditional distribution over input sequences, or networks learning disentangled latent representations (e.g. β-Variational Autoencoders [143]), where each latent factor control a single, independent aspect of the original data, such as rotation or shape.

Another avenue for achieving interpretability, known as *post-hoc* interpretability [144], involves extracting interpretable information from a pre-trained model while keeping the model itself fixed. Examples of this method include backpropagation-based methods, computing the gradient of the model's output with respect to its input, thus highlighting the input features that most strongly influence the model's decision. Another intuitive paradigm is perturbation, which involves systematically perturbing input features and observing the resulting changes in the model's output, revealing the importance of each feature to the model's predictions. Lastly, approximation-based methods (e.g. Local Interpretable Model-agnostic Explanations [145]) generate locally faithful explanations by training interpretable surrogate models on perturbed versions of the input data, providing insights into the model's behaviour within specific regions of the input space.

An emerging and very useful tool for interpretability is causality. Causal interpretability delves deeper than simple correlations, aiming to explain the underlying causal relationships. Causal interpretability methods, drawing inspiration from frameworks like Judea Pearl's causal inference [146], aim to address this challenge. For instance, counterfactuals, or *what if* scenarios, can help explain the changes in input features necessary to observe a different output from an ML system [147]. In the realm of personal sensing, a unique challenge arises as, unlike other healthcare applications, specialists lack a causal model explaining how MDs affect physiological data. For example, say someone with pneumonia has a chest X-ray scan; a radiologist can pinpoint what aspects of the image are suggestive of a pulmonary infection and how the infective pathogen causes changes from a normal scan. This serves as a benchmark for any AI explanation. Conversely, with personal sensing, there is no established domain knowledge for telling whether the recording from a wrist-worn device is suggestive of a mood episode.

2.9 Trust and Actionability beyond Good Metrics

Beyond good test set metrics, ensuring calibration and quantification of uncertainty in ML model's predictions is crucial for reliable decision-making in personal sensing applications. Calibration refers to a model's ability to accurately represent the confidence it assigns to its predictions [148]. For example, if the model predicts a 70% probability of a new episode, a new episode should be observed 70% of the time when the model makes such a prediction. Imagine a model predicting an impending mood episode based on only partially suggestive variations in sleep patterns. If the model is not calibrated, it might express high confidence in its prediction, even when the evidence is weak. This could lead to unnecessary interventions or erode trust in the model.

Predictive uncertainty quantifies the level of confidence or reliability in the model's predictions on unseen data [149]. Uncertainty can be the result of noise inherent to the data (*aleatoric* uncertainty) or the model's lack of knowledge because of model misspecification or poor representation of the training data (*epistemic* uncertainty). While the latter can in principle be reduced with better models or better data, the former is irreducible. Without quantifying uncertainty, the model would not reveal the limitations of its knowledge. This is crucial in high-stakes environments, such as mental health. Quantifying uncertainty would enable users to improve the model or to tell when a model's output is unreliable so that closer human expert inspection is warranted.

Unfortunately, achieving calibration and uncertainty quantification with modern ANN, a popular choice for personal sensing and other healthcare applications, presents significant challenges. Neural networks often struggle to produce inherently well-calibrated outputs. Their complex architectures can lead to overconfidence, even when the predictions are inaccurate [150]. Additionally, quantifying uncertainty in neural networks is a computationally expensive and ongoing area of research. Addressing these challenges is crucial for building trustworthy and reliable ML models for personal sensing. Techniques like post-hoc calibration, like Isotonic Regression [151], and uncertainty-aware training methods, like Bayesian Neural Networks [152], are being explored to bridge the gap.

Chapter 3

The TIMEBASE/INTREPIBD study

We herewith introduce the iden**T**ifying d**I**gital bio**M**arkers of illn**E**ss activity in **B**ipolAr di**S**ord**E**r/**I**dentifying digital biomarkers of ill**N**ess activity and **T**reatment **RE**s**P**onse **I**n **B**ipolar **D**isorder (TIMEBASE/INTREPIBD) cohort. This is a longitudinal, observational, exploratory single-centre study with a fully pragmatic design integrated into existing real-world clinical practice, providing minimal disruption both for clinicians and patients. The original work presented in Chapters 4, 5, and 6 is based on this dataset. The study protocol along with some exploratory analyses was presented in the following publications I co-authored, which this chapter is based on: **Exploring Digital Biomarkers of Illness Activity in Mood Episodes: Hypotheses Generating and Model Development Study** published in JMIR uHealth mHealth and **Identifying digital biomarkers of illness activity and treatment response in bipolar disorder with a novel wearable device (TIMEBASE): protocol for a pragmatic observational clinical study** published in BJPsych Open .

3.1 Sample & Study Design

Participants are recruited by their psychiatrists, at outpatient clinics, acute inpatient units, or home hospitalization settings. HCs are drawn from a convenient sample of researchers and family members.

3.1.1 Inclusion criteria

Three groups of participants, designated A, B, and C, are recruited in the study. (Figure 3.1). The inclusion criteria for group A are as follows: (a) age 18–75 years; (b) current ME (subgroup A1), MDE in the context of either BD (subgroup A2) or MDD (subgroup A3), according to the DSM-5 criteria and confirmed by a semi-structured interview using the Structured Clinical Interview for DSM-5 (SCID-5-RV); (c) requiring admission to the acute in-patient or home hospitalisation units; and (d) willing and able to give consent (reconfirmed on clinical remission). Current acute DSM-5 affective episodes is measured at time point 0. As per DSM-5, a *mixed features* specifier is used to describe an acute mood episode where at least three symptoms of the opposite polarity are present. Symptom response is measured at time point 1, defined as a 30% improvement in the YMRS score or HDRS score. Time point 2 measures symptomatic remission, defined as a YMRS or HDRS score $\leq$7. Time point 3 assesses remission, defined as $\geq$8 weeks of sustained remission. If patients do not show clinical improvement (i.e. 30% improvement in clinical scores after time point 0), such as treatment-resistant patients, they are only recorded at time point 0 (acute phase). Likewise, patients attaining clinical remission (i.e. YMRS or HDRS $\leq$7) rapidly (e.g. in <1 week), skip time point 1 recording and are only recorded at time point 0 (acute) and time point 2 (remission phase). Patients in acute phases are assessed by the research team on a weekly basis to assess psychopathological changes corresponding to response or remission.

The inclusion criteria for group B are as follows: (a) age 18–75 years; (b) patients with a current diagnosis of BD or MDD, according to DSM-5 criteria and confirmed with SCID-5-RV; (c) sustained YMRS and HDRS scores of $\leq$7 for at least 8 weeks; and (d) willing and able to give consent.

The inclusion criteria for group C are as follows: (a) age 18–75 years; (b) no current or previous psychiatric disorder, according to the DSM-5 criteria and confirmed with SCID-5-RV, excluding nicotine substance use disorder; and (c) willing and able to give consent.

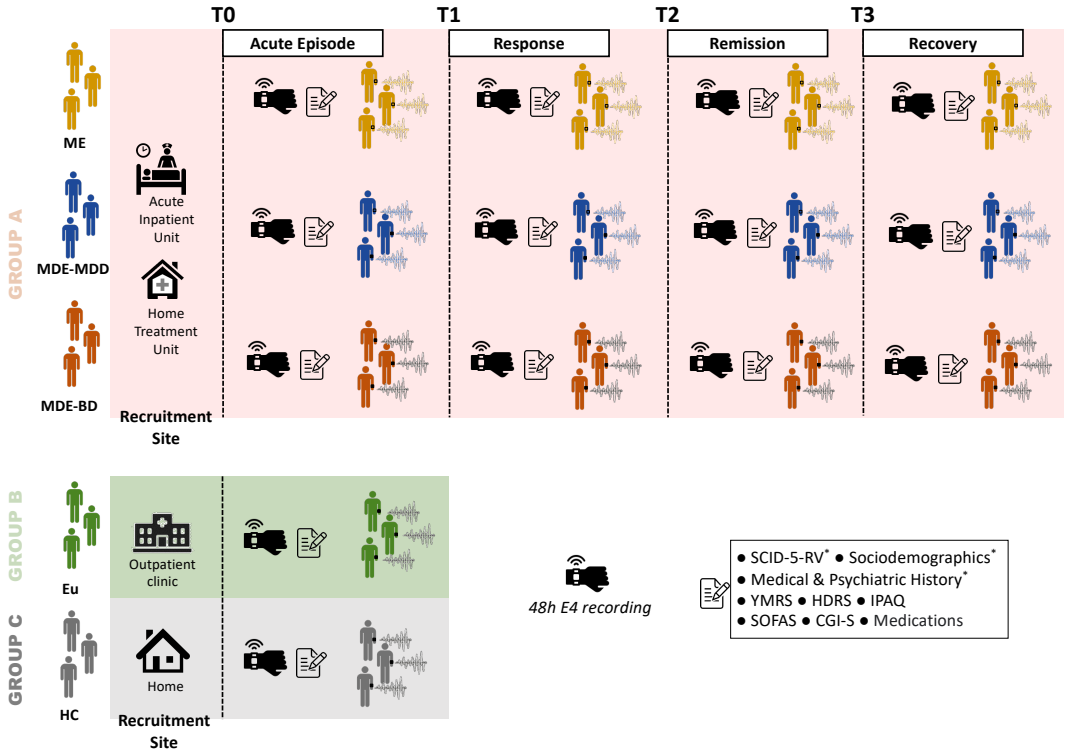


Figure 3.1: **Study Design**. Three sets of participants are recruited for the TIMEBASE/INTREPIBD study. Group **A** comprises individuals enrolled on an acute mood episode, that is a manic episode (ME) or a major depressive episode (MDE) in the context of either a bipolar disorder (BD) diagnosis or a major depressive disorder MDD diagnosis. Participants in group A have four longitudinal assessments: **T0** or acute episode; **T1** or response, defined as 30% improvement in the Young Mania Rating Scale (YMRS) or Hamilton Depression Rating Scale (HDRS) total score; **T2** or remission, that is YMRS or HDRS total score ≤ 7; **T3** or recovery, that is remission sustained for ≥ 8 weeks. Group **B** consists of individuals with an historical diagnosis of either BD or MDD, recruited into the study as they have YMRS or HDRS total score ≤ 7 for $\geq$ weeks. Group **C** comprises healthy controls. During each assessment, physiological data were recorded with an Empatica E4 wristband and the following clinical demographic features were recorded. Structured Clinical Interview for DSM-5 (SCID-5-RV); sociodemographics (e.g. age, sex); medical & psychiatric history; YMRS; HDRS; Social and Occupational Functioning Assessment Scale (SOFAS); Short Scale International Physical Activity Questionnaire (IPAQ); Clinical Global Impressions-Severity (CGI-S); medications. *denotes that assessment was not repeated during follow-up appointments.

3.1.2 Exclusion criteria

The exclusion criteria for all groups are as follows: (a) severe cardiovascular or neurological conditions that may cause autonomic dysfunction, ongoing cardiovascular arrhythmia or pacemaker use; (b) current comorbid substance use disorder as per DSM-5 criteria, excluding nicotine substance use disorder; (c) current comorbid psychiatric disorder causing significant interference; (d) current pharmacological treatment involving beta-blockers or other medications affecting the autonomic nervous system; and (e) ongoing pregnancy.

3.2 Assessments

3.2.1 Sociodemographic and clinical assessment

At baseline (time point 0), the study collects sociodemographic and clinical data including age, sex, psychiatric diagnoses according to DSM-5 criteria, medical and psychiatric comorbidities, illness duration, frequency of past manic and depressive episodes, longitudinal course specifiers (e.g., predominant polarity, seasonality, and rapid cycling), family history of mental illness, and prior substance misuse patterns.

3.2.2 Symptoms, severity, and functional assessment

Psychopathological assessments use the YMRS for manic symptoms and the 17-item HDRS for depressive symptoms. Disease severity is evaluated with the Clinical Global Impressions-Severity (CGI-S), where higher scores indicate greater severity. Functional evaluation uses the Social and Occupational Functioning Assessment Scale (SOFAS), which ranges from 1 to 100, with higher scores indicating better functionality. Clinical assessments are conducted cross-sectionally for groups B and C and at various time points (time point 0: acute, time point 1: response, time point 2: remission, time point 3: recovery) for group A.

3.2.3 Physical activity assessment

Physical activity is measured using the Short Scale International Physical Activity Questionnaire (IPAQ), which assesses different types of activities such as walking, moderate-intensity, and vigorous-intensity activities. Data on the frequency (days per week) and duration (time per day) of each activity type is collected. The continuous

score IPAQ results is expressed in metabolic equivalent of task minutes (MET-min) per week, calculated by multiplying the MET-min assigned to each activity (8 MET-min for vigorous, 4 MET-min for moderate, and 3.3 MET-min for walking) by the number of days it was performed in a week. MET-min corresponds to oxygen consumption during rest, equivalent to 3.5 ml of oxygen per kg of body mass per minute. Physical activity assessments are conducted cross-sectionally for groups B and C and at various time points (time point 0: acute, time point 1: response, time point 2: remission, time point 3: recovery) for group A.

3.2.4 Pharmacological treatments

All pharmacological treatments (both psychopharmacological and others) as recommended in international treatment guidelines are documented, including their generic name, starting day, and dose per day in each of the 48-hour records: cross-sectionally for groups B and C, and at different time points (time point 0: acute, time point 1: response, time point 2: remission, time point 3: recovery) for group A. Treatment decisions, including inpatient discharge, are made solely by the clinicians responsible for the case and with the patient's agreement, reflecting routine clinical care. Researchers involved in study recruitment do not participate in any clinical decisions regarding the patients included.

3.3 Recording of physiological data with wearables

During each assessment – cross-sectionally for groups B and C, and at different time points (time point 0: acute, time point 1: response, time point 2: remission, time point 3: recovery) for group A – participants are provided with an E4 Empatica [102] wristband, which they wear for approximately 48 hours, limited by the device's battery life. This non-interventional study ensures that individuals' behaviour remains unchanged apart from wearing the wristband. Patients from group A admitted to the psychiatric inpatient unit remain in the hospital until discharge, adhering to standard practices for acute patients. During inpatient admission, patients follow a structured routine, including set meal times (e.g., breakfast at 08:30 h) and sleep schedules (22:30 h to 08:30 h), along with daily activities such as psychiatrist consultations and therapy groups. Patients may nap during the day. Given the relatively uniform conditions during inpatient admissions, recordings at time points 0–2 are typically conducted in

this setting, minimizing variability between individuals. However, patients from group A admitted to home hospitalization units or outpatient settings (a minority of cases) are not subject to mobility restrictions. All participants are instructed to wear the wristband during their daily activities without altering their behavior. They are also instructed to put on the wristband themselves at the start of recording, and researchers ensure proper sensor contact with the wrist surface. Participants are advised to remove the device when showering to maintain its integrity.

Empatica E4 devices are equipped with sensors that collect physiological data at various sampling rates. During each recording session, the physiological data signals are obtained in either raw or processed formats. The raw data includes measurements from the following channels: three-axial acceleration (sampled at 32 Hz), electrodermal activity (EDA sampled at 4 Hz), skin temperature (TEMP sampled at 4 Hz), and blood volume pulse (BVP sampled at 64 Hz). The processed data includes inter-beat intervals (IBI), representing the time between consecutive heartbeats, and heart rate (HR), sampled at 1 Hz. The BVP signal is captured using PPG sensor, which measures blood volume changes. Empatica provides IBI as part of its output, calculated by detecting BVP peaks and determining intervals between adjacent beats. Similarly, heart rate is computed from IBI using a proprietary algorithm optimized to filter out artefacts.

Various research-grade wearable devices are available. For this project, the following factors were considered in selecting the Empatica E4: (a) signals of interest to be captured (e.g., actigraphy, EDA, HRV, TEMP); (b) data availability, as some wearable devices only provide general data (e.g., mean sleep time) and not fine-grained raw data, allowing for visual inspection, quality control, and detailed analysis; (c) study participants (BD, MDD, HCs); (d) study setting (inpatients, outpatients); (e) data confidentiality; and (f) previous literature supporting the device. The Empatica E4 was chosen because it meets all these criteria. It can measure all signals of interest (especially HRV and EDA) and provides raw fine-grained data, allowing for data processing according to study needs. Additionally, it is a device without direct internet connection and is durable in various physical situations, preserving confidentiality for patients experiencing acute affective episodes (who may lack insight and exhibit behavioural changes). The controlled inpatient environment limits the device's external communication. Previous literature also supports using this device in studies of bipolar disorder. Limited battery life could be a limitation if the study aimed to capture day-to-day variations in physiological signals. However, acute affective episodes show daily

Figure 3.2: **Empatica E4**. This wearable records (sampling rate) triaxial acceleration (32Hz), blood volume pressure (64Hz), electrodermal activity (4Hz), heart rate (1Hz), inter-beat intervals, i.e. the time between two consecutive heart ventricular contractions, and skin temperature (1Hz). Heart rate and inter-beat intervals are both features derived from blood volume pressure.

fluctuations, and response and recovery can only be assessed after sustained symptom reduction over several days or weeks. This need for sequential psychopathological status evaluation suggested that 48-hour records would adequately capture mood changes according to the study's requirements.

3.4 Ethics and confidentiality

All procedures in this study comply with the ethical standards outlined by national and institutional committees on human experimentation and adhere to the principles of the Helsinki Declaration of 1975, as revised in 2008. Approval for all procedures involving human patients was obtained from the Hospital Clinic Ethics and Research Board (approval numbers HCB/2021/104 and HCB/2021/1127). Before participating in the study, all participants provided written informed consent. Data collection is anonymous, and all data is securely stored in encrypted servers in compliance with the General Data Protection Regulation and Health Insurance Portability and Accountability Act regulations. The TIMEBASE/INTREPIBD study was designed before the start of my PhD, and my clinical practice as a consultant psychiatrist for the NHS has not been affected by any aspect of this study.

3.5 Strengths & Limitations

Annotating wearable recordings in populations with MDs is a resource-intensive task, particularly with multiple longitudinal assessments. This demand arises not only from the necessity of human expertise, such as psychiatrists, but also due to the nature of MDs, particularly MEs, which can hinder compliance. As of the writing of this thesis, the TIMEBASE/INTREPIBD study includes 98 patients recruited at the onset of an MD episode, 52 patients with a historical MD diagnosis recruited in euthymia, and 41 HCs, making it one of the largest sample sizes in this field. Figure 3.1 shows the number of assessments at different time points currently available in the TIMEBASE/INTREPIBD cohort. At the same time, Table 3.1 summarizes clinical-demographic features of the participant at the moment of study admittance. The figures herewith presented correspond to the dataset available in spring 2024. Recruitment for TIMEBASE/INTREPIBD is ongoing, and analyses in the following chapters are based on earlier dataset versions. Additionally, the study has recently expanded to include patients with psychotic disorders; these, along with participants facing technical issues during assessments (e.g., recording time $\leq$ 12 hours), are categorized as "Others" in the plot.

Comparatively, other studies in personal sensing for MDs have smaller sample sizes. For instance, Côté-Allard et al. 95 utilized a dataset of 47 patients (22 during an ME and 25 in euthymia) for a binary classification task. More recently, Jakobsen et al. 153 followed 37 patients with BD over six to twelve months. Earlier studies, like Ghandeharioun et al. 96, involved even smaller sample sizes, such as 12 patients with MDD.

In addition to the sample size, TIMEBASE/INTREPIBD distinguishes itself from other studies in the field by encompassing the entire spectrum of MDs, including both MDD and BD, spanning both polarities of BD. The study also longitudinally collects physiological data from clinically significant stages of acute episodes, allowing for the examination of associations between milestones during acute episodes and changes in physiological data. The inclusion of HC provides a baseline for "normal" physiological data. Additionally, HC and subjects categorized as "Others" (due to technical issues or new psychotic disorder recruitment) can be utilized in various unsupervised learning models.

The use of a state-of-the-art, research-grade device in TIMEBASE/INTREPIBD enables the collection of multiple physiological data modalities, unlike devices in other studies

Table 3.1: **Clinical demographic features at the point of study admittance** HC: Healthy Control; BD: Bipolar Disorder; ME: Manic episode; MDE: Major Depressive Episode; MDD: Major Depressive Disorder; YMRS: Young Mania Rating Scale (total score); HDRS: Hamilton Depression Rating Scale (total score).

	Age mean (sd)	Sex Males/Females	YMRS mean (sd)	HDRS mean (sd)
HC	38.08 (15.20)	14/27	1.1 (3.12)	1.93 (2.10)
BD (ME)	41.19 (15.21)	16/28	12.81 (10.61)	5.54 (3.85)
BD (MDE)	47.85 (15.20)	11/30	1.59 (2.68)	12.51 (7.29)
MDD (MDE)	49.62 (15.20)	7/6	1.36 (2.14)	17.62 (5.85)
BD (Euthymic)	49. 67 (12.24)	18/27	1.35 (2.14)	3.42 (2.38)
MDD (Euthymic)	45.14 (11.60)	4/3	0.29 (0.70)	3.57 (1.59)

that may only collect a single modality (e.g., acceleration in [98]). Access to raw data provides flexibility for custom analyses, as it is not limited by pre-specified features extracted with proprietary algorithms.

Despite its strengths, TIMEBASE/INTREPIBD is limited in its ability to explore the feasibility of personal sensing in clinical practice. The cohort's physiological data is sampled up to four times and only for 48 hours per session around clinically significant phases, creating a blind spot regarding within-patient baselines and transitions among clinical phases. Long-term data collection (e.g., over months), necessary for enabling early interventions, poses challenges of compliance and technology abandonment. The shorter battery life of research-grade relative to commercial devices requires frequent recharging, demanding an unrealistically high level of patient engagement.

Lastly, a single clinician collected psychometric scales for this study. As a result, inter-rater agreement issues cannot be addressed, potentially affecting the evaluation of machine learning models. Furthermore, no hard outcomes (e.g., hospitalization) were collected, limiting the robustness of outcome assessments.

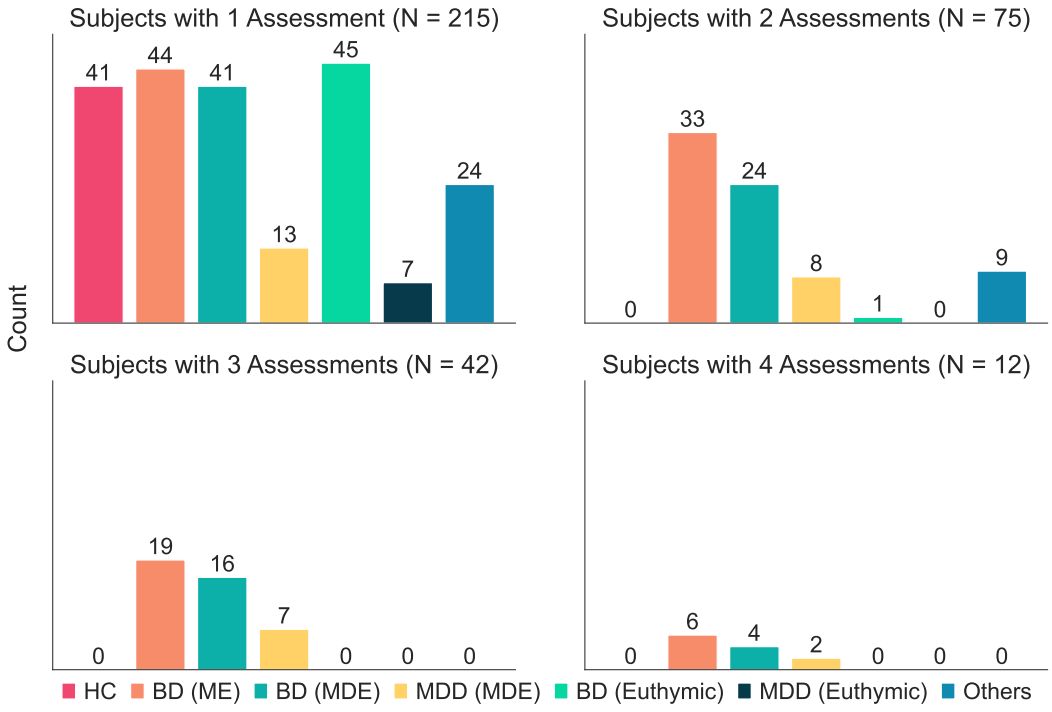


Figure 3.3: **Number of assessments available per longitudinal time point by diagnosis.** Mood disorders manifest in two polarities, mania, and depression. Major Depressive Disorder (MDD) is characterized by Major Depressive Episodes (MDEs) only, whereas Bipolar Disorder (BD) is characterized by the presence of (hypo)manic episodes (ME) that can alternate with MDEs. Patients with a former mood disorder diagnosis but currently clinically stable are said to be Euthymic. Except for Healthy Controls (HCs) and Euthymic Patients, other subjects are recruited at the onset of a disease episode. They are assessed at subsequent stages (maximum four) during their clinical course. The scope of TIMEBASE/INTREPIBD has been recently extended to include patients with a psychotic disorder. These, along with subjects recruited on the original study design but yielding recordings $\leq$ 24 hours due to compliance or other issues, are reported as "Others".

Chapter 4

Heart Rate Variability: a Promising Biomarker in Bipolar Disorder

4.1 Heart Rate Variability

HRV is a measure of the variation in time intervals between consecutive heartbeats, reflecting the dynamic interplay between the sympathetic (“fight-or-flight”) and parasympathetic ("rest-and-digest") branches of the ANS. ECG, whether ambulatory or performed with a portable Holter device, is considered the gold standard for HRV measurement [154]. Wrist-worn devices implementing ECG and PPG sensors, however, showed good reliability for HRV monitoring and are drawing interest given their widespread use and their non-invasive and continuous monitoring potential [155]. Some authors [156] observe that PPG measures blood volume changes and not ventricular heart contractions directly, unlike ECG. Thus, they suggest that PPG-derived HRV should be more appropriately indicated as pulse rate variability. We will follow the trend in the recent literature [157, 158] and use HRV, regardless of whether it is ECG- or PPG-derived.

Motion artefacts, resulting from physical movement, can significantly distort HRV readings from wrist-worn devices, complicating the accurate interpretation of autonomic activity. To mitigate this issue, researchers advocate for the collection of HRV data during night sleep time. This not only reduces the impact of motion artefacts but also provides a more standardized context for data comparison, as it reflects a more consistent state of autonomic function across different individuals and time points.

HRV is derived from the time intervals between consecutive heartbeats, i.e. IBI, upon cleaning the data from artefacts, which produces so-called Normal-to-Normal intervals (NN). However, a number of metrics exist to characterize HRV, grouped into time-domain, frequency-domain, and non-linear features, each offering unique insights into autonomic function [154]. Time-domain metrics, such as the standard deviation of NN intervals (SDNN) and the root-mean-square of successive differences (RMSSD), quantify variability over time, with SDNN reflecting the combined influence of all physiological factors contributing HRV and RMSSD highlighting parasympathetic activity specifically. Frequency-domain metrics analyse the distribution of power into different frequency bands, like low-frequency and high-frequency components, to understand the balance between autonomic branches. Non-linear metrics, including Poincaré plots and entropy measures, assess the unpredictability and complexity of a series of NN. The RMSSD is among the most commonly used HRV metric [159] and reliably captures parasympathetic activity. It is given as the reference HRV metric across both research and consumer-grade devices.

HRV has long been established as a critical indicator of cardiac stress, where lower HRV values are associated with increased risk of cardiovascular diseases, including myocardial infarction, hypertension, and heart failure [160, 161]. Moreover, HRV is linked to physical fitness, with higher HRV indicating better aerobic capacity and resilience to stress. Thus, HRV is being used in sports medicine and fitness to monitor recovery and guide training intensity [162].

More recently, HRV has gained traction in mental health where meta-analyses [163, 164, 165, 69] found a reduced HRV across a range of psychiatric conditions, with psychotic disorders featuring the greatest reduction. HRV can help monitor cardiac health in populations with MDs who experience higher cardiovascular co-morbidities. The literature consistently shows that individuals with MDD exhibit reduced HRV compared to healthy controls [166, 165]. MDD without any concurrent cardiovascular disease is also associated with reduced HRV, inversely correlated with depression severity. Critically, a variety of antidepressant treatments do not resolve these decreases despite the resolution of symptoms [167].

4.2 Studies into Bipolar Disorder

Research on HRV in BD is less extensive than that in MDD. A meta-analysis [69] suggested that HRV is reduced in BD compared to healthy controls, but heterogeneity and methodological issues limited the evidence. Most studies in BD to date analysed cross-sectional differences in BD relative to healthy controls. Evidence into longitudinal HRV changes across BD polarities (mania and depression) is limited. Investigating intra-individual HRV changes across affective states in BD is indeed a challenging and resource-intensive task. Longitudinal studies necessitate multiple follow-ups and assessments by mental health specialists, which is particularly demanding during MEs, often leading to poor patient compliance with study protocols. Consequently, recruiting large cohorts for HRV studies in BD has proven unfeasible, with previous studies involving only a few dozen participants [51, 168, 169, 27].

A key question is whether HRV improves as an acute episode of BD resolves, and whether different patterns across mania and depression exist. Results are mixed as two studies [51, 168] showed reduced HRV during mania relative to euthymia (i.e. sustained symptoms' remission) but another study reported an opposite pattern [27]. As for bipolar depression, no significant differences across depression and euthymia emerged in two studies[27, 169]. It should be noted that all these studies only sampled participants twice, that is on an acute episode and during euthymia, thus overlooking longitudinal trajectories of change in HRV as a function of symptoms' improvement. Furthermore, each relied on a small cohort, ranging from 15 [51] to 37 participants [169], and yet, as customary in psychiatric research, they all embraced frequentist null hypothesis significance testing (NHST).

4.3 Bayesian vs Frequentist Statistics

A core interest in medical research is parameter estimation, i.e. estimating the unknown value of some parameter(s) of a statistical model describing some phenomenon of interest in a population using a finite, random sample taken from that population. Two primary approaches exist to statistics, i.e. frequentist and Bayesian statistics, with distinct methodologies and philosophical underpinnings. Frequentist statistics, and in particular NHST, has long been the workhorse of statistical analysis in psychiatric, and medical research more broadly. In HRV studies, for example, NHST has been used to test a zero-mean difference in HRV values taken from the same subjects at two different

moments in their clinical course, i.e. mania and euthymia. Concretely, this was done with a two-tailed paired t-test [51]. Frequentist statistics, however, has recently become the object of a growing chorus of criticism and the reproducibility crisis in medical research has partly been blamed on its misuse.

Frequentist statistics operates on the premise that probability describes the long-run frequency of events; this philosophy shapes its approach to NHST. The frequentist statistician posits a null hypothesis (*H0*), usually representing the *status quo* or no effect, and an alternative hypothesis (*H1*), representing conversely the existence of an effect. Assuming the null to be true and under a given sampling intention, a test statistic, i.e. a function of the data with a known probability distribution (e.g. a t-distribution), is computed. The test statistic obtained from the data is checked against reference values from the known probability distribution and a *p*-value is born. Specifically, *p*-value is the probability of observing a result equal to or more extreme than that observed in the study sample, under the null and the posited sampling intention. A threshold of $p < 0.05$ is commonly (arbitrarily) used to reject *H0*. The dependence on sampling intention is usually overlooked and not adequately reported in clinical studies, especially in *post-hoc* analyses. For instance, the very same coin toss data (say $N = 24$ flips with $z = 7$ heads) can yield different *p*-values based on the form of the likelihood (binomial with N fixed or negative binomial with z fixed) [170].

Crucially, by its very definition, *p*-value is not the probability that the null hypothesis is true. It expresses, on the other hand, P(data | *H0*) but does not reverse this to P(*H0* | data); the null was indeed assumed to be true in the first place. It is not the probability that the alternative hypothesis is false, just because the data is unlikely under *H0* does not mean it is likely under *H0* without additional context. In summary, *p*-value cannot assess the extent to which the data supports H0 versus the alternative hypothesis H1. Moreover, it measures the existence of an effect but not its magnitude, and standardized measures of effect size, since grounded on frequentist statistics, inherit its limitations [170].

4.3.1 Bayesian Inference

The Bayesian framework mitigates some *p*-value shortcomings. The outputs of Bayesian inference are probability distributions over model parameters, which represent degrees of belief about the values of these parameters, given the data and underlying assumptions (the specified model and prior distribution over parameters). Inference is concerned with

gaining knowledge into unobserved parameters Θ of a given model $\mathcal{M}$ (left implicit in the notation if it is understood that only a single model is being considered), given a dataset D. The Bayesian rule, a consequence of the product rule of probability, dictates that the posterior distribution of Θ given D can be expressed as:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}$$

Here, $P(\Theta|D)$ is the posterior distribution, $P(D|\Theta)$ is the likelihood of the data under the model parameters, $P(\Theta)$ is the prior distribution, and $P(D)$ is the marginal likelihood or evidence, which acts as a normalizing constant, ensuring that $P(\Theta|D)$ is a valid probability distribution integrating to 1. The marginal likelihood $P(D)$ involves integration over all possible values of Θ: $\int P(D|\Theta)p(\Theta)d\Theta$.

This integral can be over a high-dimensional space. So, except for a few models where the prior and the posterior are conjugate, i.e. they belong to the same distribution family, an analytical closed-form solution for the posterior is not possible. A number of approximation methods have been developed to enable Bayesian inference, including Laplace approximation, Markov Chain Monte Carlo methods, and variational inference [171]. These are the inference workhorse in modern probabilistic programming languages, such as PyMC [172].

4.3.2 Assessing Evidence in a Bayesian Framework

Having the posterior distribution $P(\Theta|D)$ allows for directly interpretable statements about any model parameter of interest, providing insights into the evidence for both *H0* and the competing *H1*. This contrasts with frequentist p-values, which do not provide the probability that a parameter value is compatible with *H0*. Bayesian methods are particularly advantageous with small sample sizes, as is often the case with HRV studies in BD. They do not rely on the asymptotic properties of large samples and, due to their principled approach to handling uncertainty, they provide graded evidence that enables researchers to extract more information from small studies that might otherwise be underpowered to achieve statistical significance.

Testing a point (null) hypothesis in a Bayesian framework is commonly done in one of two ways. One is model comparison, where two models are instantiated, one with a prior distribution that allows only the value of interest and another that spreads probability over a wide range of values. The Bayesian factor, i.e. the ratio of two models' marginal likelihood, is then computed. This paradigm faces criticism for its sensitivity to the prior

specification, even when different priors lead to minor differences in the posterior [171]. The other approach encompasses a number of decision rules based on the posterior, among which the probability of direction (PD) and the region of practical equivalence (ROPE) are popular examples. PD is an index of effect existence, robust to the scale of both the response variable and the predictors. It ranges from 50% to 100%, representing the certainty with which an effect goes in a particular direction (i.e., is positive or negative), and is mathematically defined as the proportion of the posterior distribution that is of the median's sign. ROPE is a range of values, usually centred around the point (null) hypothesis value, which is considered negligible or too small to be of any practical relevance for the use case in question. The degree of overlap between the ROPE and the highest density interval 95 (HDI-95), i.e. the 95% most plausible values in a parameter's posterior, is then inspected.

4.4 The paper: Does heart rate variability change over acute episodes of bipolar disorder? A Bayesian analysis.

Below, we present our original contribution to the study of HRV in BD **A Bayesian analysis of heart rate variability changes over acute episodes of bipolar disorder**, published in npj Mental Health Research. Here we opted for RMSSD as a metric for HRV, motivated by its wide use and popularity; specifically, as customary in the field, we modelled its natural logarithm (lnRMSSD) since the log transformation achieves an easier to use, quasi Gaussian distribution [173, 174, 175, 176].

To our knowledge, we are the first to examine, within the same cohort, changes in lnRMSSD as acute episodes of both mania and depression resolve. We developed an interpretable probabilistic model, accounting for the hierarchical nature of the data, i.e. HRV measurements nested within individuals, and individuals experiencing acute BD episodes nested within manic and depressive states. This model considers the interactions of variables that influence lnRMSSD and attempts to explain the data-generating process for HRV, beyond a simple test statistic. We embrace the Bayesian paradigm and illustrate its advantages over NHST in our setting.

We applied our model to data from the TIMEBASE/INTREPIBD study, which includes at least three time points per person per affective episode. Unlike previous

studies that utilized only two time points (e.g., acute state versus euthymia), TIME-BASE/INTREPIBD enables capturing individual differences in lnRMSSD trajectories. Our findings show a positive rate of change of lnRMSSD as symptom severity lessens from acute episodes to euthymia. However, additional data is needed to determine whether the magnitude of this effect is clinically significant. Results do not support different HRV dynamics across the polarities of BD, i.e., mania and depression.

npj | mental health research

Article

https://doi.org/10.1038/s44184-024-00090-x

A Bayesian analysis of heart rate variability changes over acute episodes of bipolar disorder

Filippo Corponi[1] ✉, Bryan M. Li[1], Gerard Anmella[2,3,4,5], Clàudia Valenzuela-Pascual[2], Isabella Pacchiarotti[2], Marc Valentí[2], Iria Grande[2], Antonio Benabarre[2], Marina Garriga[2], Eduard Vieta[2], Stephen M. Lawrie[6], Heather C. Whalley[6,7], Diego Hidalgo-Mazzei[2] & Antonio Vergari[1]

Bipolar disorder (BD) involves autonomic nervous system dysfunction, detectable through heart rate variability (HRV). HRV is a promising biomarker, but its dynamics during acute mania or depression episodes are poorly understood. Using a Bayesian approach, we developed a probabilistic model of HRV changes in BD, measured by the natural logarithm of the Root Mean Square of Successive RR interval Differences (lnRMSSD). Patients were assessed three to four times from episode onset to euthymia. Unlike previous studies, which used only two assessments, our model allowed for more accurate tracking of changes. Results showed strong evidence for a positive lnRMSSD change during symptom resolution (95.175% probability of positive direction), though the sample size limited the precision of this effect (95% Highest Density Interval [−0.0366, 0.4706], with a Region of Practical Equivalence: [-0.05; 0.05]). Episode polarity did not significantly influence lnRMSSD changes.

Bipolar disorder (BD) is a severe mental health condition affecting > 1% of the global population[1]. With a population-level annual economic burden estimate of £6.43 billion in the UK alone[2] and an all-cause mortality rate 1.77 times higher than the general population[3], BD has huge personal and societal costs. Symptoms encompass disturbances in mood states, thought, energy, and vegetative functions manifesting during episodes of (hypo) mania and depression, the two polarities of BD.

Accumulating evidence[4] indicates autonomic nervous system dysregulation in BD, detectable through reduced vagally mediated heart rate variability (HRV). This is a measure of the variation in time between consecutive heartbeats and can be computed from interbeat interval (IBI) data collected via either electrocardiogram (ECG) or photoplethysmography (PPG). With the widespread adoption of wearable devices recording IBI data, HRV monitoring can be extended outside the doctor's office to the patient natural environment, in a near-continuous fashion, unlocking unprecedented opportunities for health monitoring[5]. A number of metrics have been developed to quantify HRV, grouped into time-domain, frequency-domain, and non-linear measures. Among these, the Root Mean Square of Successive RR interval Differences (RMSSD) has been suggested as a robust indicator of vagal tone and parasympathetic activity[6]. RMSSD is indeed the most commonly reported HRV output feature by a number of both commercial[7] and research-grade devices[8]. Modelling the natural logarithm of RMSSD (lnRMSSD) is common practice, as the log-transformation achieves an easier-to-use, quasi-Gaussian distribution[9–12].

Meta-analyses[13–16] found a reduced HRV across a range of psychiatric conditions, not just BD, with psychotic disorders featuring the greatest reduction. A reduced HRV is also a predictor of increased cardiovascular risk in the general population[17,18]. As of today it has not yet been fully investigated whether the resolution of symptoms over the course of a BD episode translates into changes in HRV and whether mania and depression, the two polarities of BD, display different HRV trajectories. In this study (Fig. 1) we fill this gap, leveraging the TIMEBASE/INTREPIBD study[19], a longitudinal cohort following up BD acute episodes.

Studying intra-individual HRV changes across affective states in BD is a challenging and resource-intensive endeavour, especially as longitudinal settings require patients to be followed up and assessed by a mental health specialist multiple times. This is particularly demanding with manic episodes, undermining patients' compliance to study instructions, such that recruiting large cohorts in HRV studies on BD proves unfeasible and all previous studies had only a couple dozen participants[20–23].

A case in points is Stautland et al.[20], limiting their analysis to a sample of 15 patients on a manic episode. A reduced RMSSD in mania relatively to

[1]School of Informatics, University of Edinburgh, Edinburgh, UK. [2]Bipolar and Depressive Disorders Unit, Hospital Clinic de Barcelona, Barcelona, Spain. [3]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [4]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain. [5]Departament de Medicina, Universitat de Barcelona, Barcelona, Spain. [6]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. [7]Generation Scotland, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, UK. ✉e-mail:

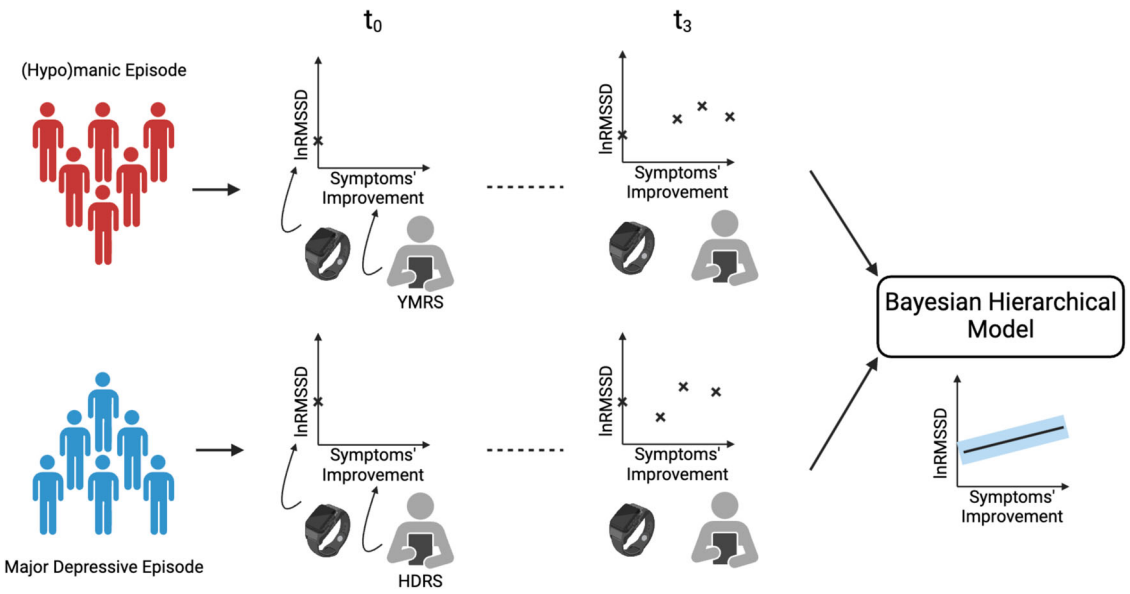


Fig. 1 | Longitudinal data from patients with bipolar disorder recruited at the onset of an acute episode is used to study the lnRMSSD trajectory as symptoms, as measured with clinician-administered rating scales, improve. Patients with bipolar disorder on either a manic (in red) or a depressive (in blue) episode are assessed up to four times, $t \in \{0, 1, 2, 3\}$, as their symptoms subside. During each assessment, lnRMSSD is collected with a smartwatch while symptoms'improvement is measured by a mental health specialist with a hetero-administered rating scales, the Young Mania Rating Scale[30] (YMRS) for mania and the Hamilton Depression Rating Scale-17[31] (HDRS) for depression. A Bayesian Hierarchical Model is fitted to the data to study the rate of change in lnRMSSD with respect to symptoms' improvement. Two models are developed and compared where the only difference is that in one the trajectory of lnRMSSD through symptoms' improvement is allowed to vary across polarities, to test whether a polarity-specific effect on lnRMSSD dynamics exists.

euthymia was found. Participants were assessed only twice – mania and euthymia – and paired two-tailed *t*-tests were used to test zero mean difference across manic and euthymic states. Similarly, Wazen et al.[21] recruited 19 patients with BD and showed a similar association between RMSSD and mania-to-euthymia transition. Again, only one acute state and one euthymia measurements were taken; a non-parametric (Wilcoxon's signed-rank) test was used, positing as null a zero median difference between paired observations. On the other hand, Hage et al.[22] found no significant HRV changes after 8 weeks in 37 patients with bipolar depression randomized to receive either escitalopram-celecoxib or escitalopram-placebo, regardless of treatment response status. The authors opted for a frequency-domain feature, i.e. high frequency (HF-HRV), as their HRV metric and employed repeated measures ANCOVA to evaluate differences between baseline and week 8. Lastly, Faurholt-Jepsen et al.[23] studied HRV changes in a sample of 16 patients with BD observed for a period of 12 weeks over as many different affective states (euthymia, depression, mania/mixed state) as possible, using a linear mixed-effect model. Investigators found an increased HRV during mania in comparison to both euthymia (in contradiction with[20,21]) and depression, but no significant difference across depression and euthymia. The difference between the second-shortest and the second-longest IBI collected during 30-second epochs was used a HRV measure.

All studies mentioned above[20–23] collected only one sample per patient per affective state (euthymia, mania/mixed state, depression) and thus did not consider HRV trajectories as a BD acute episode resolves. Moreover, while it is tempting to equate HRV increments/decrements between acute state and euthymia[20–22] to a process of positive/negative change in HRV, statistical literature[24,25] warns that two-time points are not sufficient to accurately capture individual differences in trajectories of change and are prone to confounding true change with measurement error. A minimum of three data points per subject is indeed recommended to investigate change over time. Furthermore, as customary in psychiatry research[20–23], all embraced frequentist null hypothesis significance testing (NHST), failing to propose a model explaining how HRV values are generated and which dependencies among variable govern HRV longitudinal dynamics. Despite its enduring popularity in psychiatry research, the NHST *p*-value has indeed been the object of a growing chorus of criticism[26,27]. The *p*-value serves solely for rejecting the null H_0 and lacks the capacity to assess the extent to which the data supports H_0 versus the alternative hypothesis H_1. Moreover, it measures the existence of an effect but not its magnitude; standardized measures of effect size, since premised on a frequentist framework, inherit its limitations. Further, by simply considering the distribution of a test statistic, previous studies relying on NHST did not elaborate a model trying to capture the data (HRV) generating process.

An alternative framework that has been gaining recognition and popularity in psychiatry research is Bayesian statistics, which mitigates some of the *p* values shortcomings[28,29]. The outputs of Bayesian methods are probability distributions over model parameters, representing the degree of beliefs about parameters' values, conditional on data and assumptions (the specified model and prior distribution over parameters). Posteriors can be used to make directly interpretable statements about any model parameter of interest, gaining insights into evidence equally for H_0 as for the competing H_1. This is in contrast to frequentist *p*-values, which do not give the probability that a parameter value is compatible with H_0. Bayesian methods are particularly useful with small sample sizes, as it is the case for HRV studies with BD. Indeed, they do not rely on the asymptotic properties of large samples and, thanks to their principled way of handling uncertainty, they yield graded evidence allowing us to gather more information from small studies that may be otherwise underpowered to reach statistical significance. As research into HRV (as well as other digital biomarkers) has the potential for delivering clinical decision support tools, interpretability, i.e. being able to clearly inspect and interrogate the data generating process, and a principled quantification of uncertainty in the model output, are key features of a Bayesian data analysis, that make it particularly appealing in clinical settings.

In this work, using data from the TIMEBASE/INTREPIBD study[19], we investigate lnRMSSD changes in patients with BD on either mania or depression as their symptoms' severity, measured with the total score on respectively Young Mania Rating Scale[30] (YMRS) and Hamilton Depression Rating Scale-17[31] (HDRS) respectively, wanes, from acute state up to euthymia, with at least three samples available per individual over the course of their episode. Our main contributions are as follows:

– We are the first to the best of our knowledge to study changes in lnRMSSD as an acute episode resolves across both mania and

depression within the same cohort.
- We develop an interpretable probabilistic model that captures the natural hierarchical structure in the data (HRV measurements are nested within subjects, subjects on an acute BD episode can be seen as themselves nested within mania and depression) and accounts for how variables interact in generating lnRMSSD. Relatedly, we illustrate the benefits of a Bayesian treatment over NHST, including a principled way to quantify uncertainty and better suitability to small samples than NSHT.
- We fit our model to the data from the TIMEBASE/INTREPIBD study where a minimum of three-time points per individual per affective episode is available. Unlike previous studies only using two-time points (e.g. acute state vs euthymia), this allows us to better capture individual differences in lnRMSSD trajectories. Data does not support the existence of different HRV dynamics across BD polarities, i.e. mania and depression. Results indicate a positive rate of change of lnRMSSD as symptoms' severity abates from acute episode up to euthymia; however, towards being able to claim that the magnitude of this effect has clinic significance, more data is needed.

Methods

The TIMEBASE/INTREPIBD cohort

Unlike other existing cohort, the TIMEBASE/INTREPIBD study[19] gathers multiple longitudinal assessments per patient over the course of an acute BD episode. This uniquely positions this cohort to investigate trajectories of change in lnRMSSD as an acute episode resolves. TIMEBASE/INTREPIBD is a prospective, exploratory, observational, single-center, longitudinal study with a fully pragmatic design embedded into current real-world clinical practice. A comprehensive description of the data collection campaign is detailed in Anmella et al.[19]. For the purpose of this work, subjects with a DSM-5 diagnosis of BD (equally type I and type II) were considered. Exclusion criteria comprised: concomitant severe cardiovascular or neurological medical conditions with a potential autonomic dysfunction, ongoing cardiovascular arrhythmia, or pacemaker; comorbid current substance use disorder according to the DSM-5 criteria, excluding nicotine substance use disorder; comorbid current psychiatric disorder with great interference of symptoms (e.g., obsessive-compulsive disorder with ritualized behaviours); ongoing pregnancy.

Patients were recruited at the onset of an acute BD episode, either mania or major depression, and were assessed up to four times over the course of their episode: acute phase, clinical response, remission, euthymia (score ≤7 on the HAMD and YMRS for at least 8 weeks[32]). During each assessment, patients were interviewed by a psychiatrist collecting clinical-demographics, including age, sex, medications being administered, and YMRS/HDRS. They were also required to wear the Empatica E4 device[33] on their non-dominant wrist until battery ran out (~48 hours). This wearable records (sampling rate) 3D acceleration (ACC, 32Hz), blood volume pressure (BVP, 64Hz), electrodermal activity (EDA, 4Hz), heart rate (HR, 1Hz), inter-beat intervals (IBI) and skin temperature (TEMP, 1Hz). Mixed BD episodes were not included in the present analyses in order to minimise diagnostic ambiguity and allow for an easier comparison between the two extreme polarities of BD, also considering that only two such episodes were available in the cohort at the time of this work. Hypomanic episode, on the other hand, were not collected in the TIMEBASE/INTREPIBD study[19].

HRV data extraction

During free-living wear, subjects might remove their device or contact to the wrist might be otherwise suboptimal; furthermore, PPG data is affected by motion artefacts, so wake HRV may be unreliable[34]. Thus, we first performed on-/off-body detection using discontinuity in EDA as a guide. In particular, similarly to[35,36], we considered measurements smaller than 0.05 μS as indicative of off-body status. Then, sleep/wake detection was carried out on on-body recording sequences using the algorithm by Van Hees et al.[37] which emerged as the best performing in a recent benchmark study on sleep-wake detection[38].

The RMSSD is arguably the most commonly used HRV metric[7,8] and reliably captures parasympathetic activity[6]. It is derived from RR intervals ($\boldsymbol{R}$) on either an ECG or a PPG reading and it is computed as follows:

$$\mathrm{RMSSD} = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N-1}\left(\boldsymbol{R}_{i+1}-\boldsymbol{R}_i\right)^2\right)} \quad (1)$$

where $\left(\boldsymbol{R}_{i+1}-\boldsymbol{R}_i\right)$ is difference between neighbouring RR intervals and N is the total number of RR intervals over which RMSSD is computed. Sleep occurring at nighttime between 10 pm and 5 am from each recording session was segmented with a sliding window of length and step size 5 and 1 minute, respectively, from which RMSSD was derived with FLIRT[39]. This is a popular open-access feature extractor toolkit compatible with E4 data, handling IBI pre-processing and RMSSD computation. The average of all valid 5-minute RMSSD values was taken as a measure for the full night's RMSSD. This approach to estimate RMSSD is implemented in commercial devices[40] and was used in previous research[41]. Five minutes is indeed a conventional length for RMSSD estimation[6]. Considering motion artefacts and circadian rhythms in HRV, nighttime sleep is a popular choice for HRV extraction; averaging over multiple 5-minute RMSSD is more robust than using just a random 5-minute RMSSD which would be susceptible to HRV variations across sleep stages[42]. Recording sessions from the TIMEBASE/INTREPIBD study stretched over 48 hours so, while two nights were available for HRV extraction, only the first one was considered, since closer to the time when HDRS/YMRS were taken. As standard practice[9–12], we modelled lnRMSSD, that is the natural logarithm of RMSSD, as this transformation results in an more convenient, quasi-Gaussian distribution. While wristbands today allow for collecting RMSSD, they do not provide a model explaining how features of the individual interact in generating RMSSD values. In the section that follows, we build a Bayesian model attempting to do just that.

Bayesian modeling

The goal of inference is to get to unobserved parameters (Θ), given the data. The Bayesian approach aims for a full distribution over Θ, not just a single value, which, especially when data is scarce, can be misleading, since it does not consider uncertainty and tells only a part of the story (e.g. the mean or the mode of the distribution). Our Bayesian analysis is particularly interested into the rate of change of lnRMSSD with respect to symptoms' severity, so this will be a key parameter of interest. The Bayesian paradigm commands to posit a process generating the data at hand governed by Θ, referred to as likelihood $P(\mathrm{Data}|\Theta)$, as well as a starting hypothesis as to what values Θ can credibly take, in advance of seeing any data, referred to as the prior $P(\Theta)$. The output of Bayesian inference is a posterior $P(\Theta|\mathrm{Data})$, where the prior beliefs about the values of Θ have been updated in light of the observed data.

As a running example to illustrate Bayesian methods, we temporarily assume here that the observed lnRMSSD values are sampled from a Gaussian distribution with mean μ and variance σ^2, the latter we assume given and equal to 1. As with ordinary regression, parameters can be modelled as a function of relevant covariates. For example, we might have reasons to believe that μ linearly depends on the symptoms' severity (V) of the individuals: $\mu = \theta_0 + \theta_1 V_i$, where i indexes the subjects in the study. The parameters of our model are thus θ_0 and θ_1 and our interest might be into θ_1, expressing the dependency of lnRMSSD on V. The likelihood $P(\mathrm{Data}|\Theta)$ is a function of the parameters, expressing the probability of observing the given data under particular values of Θ, in our example, how well different values of θ_0 and θ_1 explain the data.

The other key ingredient of a Bayesian model, further to the likelihood, is the prior probability over the parameters $P(\Theta)$, representing our beliefs about the parameters before seeing any data. The choice of prior can be informed by previous research. Alternatively, in case of lack of previous evidence or when the analyst does not want to favour one hypothesis over others, a non-committal prior can used, assigning equal credibility to

competing hypotheses. In the running example we might opt for $\theta_0 \sim \mathcal{N}(0,1)$ and $\theta_0 \sim \mathcal{U}(-1,1)$, i.e. a standard Gaussian for the intercept θ_0, favouring values around zero but not giving any preference to either positive or negative values, and a uniform distribution for the slope θ_1, assigning equal credibility to all values in the interval [-1,1].

Through Bayes' theorem, the prior is updated in light of the observed data to yield a posterior probability distribution $P(\Theta|\text{Data})$: this encapsulates the refined beliefs about the parameters, incorporating both prior knowledge and the information conveyed by the observed data. In our example, we might move from a flat prior over θ_1 to a distribution where the overwhelming majority of the probability density is concentrated on positive values. This posterior, $P(\theta_1|\text{Data})$, can be directly and naturally interpreted as our beliefs about values of θ_1, condition on the observed data and the posited model. This is arguably more intuitive for clinicians to use than a p-value, the probability of obtaining under the null hypothesis (H_0) and under the assumed sampling intention a result equal to or more extreme than the one observed from the data, and can be directly used to make statements about both the existence and the magnitude of an effect.

The only extra layer of complexity in Bayesian hierarchical models, on top of the vanilla Bayesian machinery we introduced above, is that parameters depend on other parameters too, referred to as hyperparameters, introducing dependencies between parameters at different hierarchical levels. This is particularly convenient as it allows us to model lnRMSSD observations as nested into subjects and subjects themselves as nested into episode polarity π. In our running example, we can modify the model to reflect that the relation between V and lnRMSSD might differ across polarities as follows: $\theta_1 \sim \mathcal{N}\left(\zeta_{\pi[i]}, 1\right)$ and $\zeta_\pi \sim \mathcal{U}(-1,1)$. This is now saying that the intercept θ_1 is sampled from a Gaussian whose mean is controlled by another parameter ζ_π with a uniform prior on [-1,1]. There are Π parameters ζ, one for each polarity and all sampled from the same uniform distribution. The notation $\pi[i]$ denotes the parameter ζ that corresponds to the polarity π to which the i^{th} individual's episode belongs to. It can be seen how hierarchical models provide a powerful framework for nested data: in our study, each patient (level-1) generates multiple lnRMSSD measures since patients are indeed assessed at multiple time points as their symptomatology improves; secondly, from each BD polarity (level-2) multiple patients are drawn. Hyperparameters enables sharing of information across level groups, while allowing for within-group variability. Conceptually, a hierarchical model provides a middle ground (*partial pooling*) between aggregating groups at a given level of the hierarchy (*complete pooling*), thus overlooking potential differences across groups, and treating them as completely independent (*no pooling*).

Variables preprocessing

We wanted to model how lnRMSSD changes as symptoms'severity, measured with the total score on either YMRS (manic episode) or HDRS (depressive episode), abates during the resolution of an acute BD episode. Each i^{th} individual of the N included in the analyses was sampled up to four times along their trajectory of symptoms' improvement, starting from episode onset $t = 0$. For the i^{th} individual, their improvement along this trajectory at time $t \in \{0, 1, 2, 3\}$ was expressed as $I_{i,t} = (\text{score}_{\pi[i],t=0} - \text{score}_{\pi[i],t})/(\text{score}_{\pi[i],t=0})$, where the notation $\pi[i]$ means that the total score on YMRS (HDRS) was used if the episode's polarity π of the i^{th} individual was manic (depressive). I therefore takes values in [0, 1], patients have a value of 0 at episode onset, i.e. study recruitment, and reach a value of 1 if their total score goes down to 0; intermediary values express fractional improvement with respect to episode's onset severity. For a given subject, successive recording sessions were required to have a strictly monotonic decrease in the relevant scale's total score.

A number of factors further to changes in symptoms' severity can influence HRV. We therefore controlled for relevant covariates available in our dataset, i.e. sex S (females = 1, males = 0), age A, and medications M. Age (in years) was standardized and treated as constant across different recording sessions for a given individual. Data for a number of drug classes known to affect HRV was recorded in the INTREPIBD/TIMEBASE dataset as boolean: lithium, selective serotonin reuptake inhibitors, serotonin and norepinephrine reuptake inhibitors, tricyclics, monoamine oxidase inhibitors, other antidepressants, typical antipsychotic, atypical antipsychotic, anticonvulsants, beta-blockers, opioids, amphetamines, antihistamines, antiarrhythmic agents, other anticholinergic medications, benzodiazepines. $M_{i,t}$ is simply the total number of such medications the i^{th} individual was on at time t. Lastly, as previous research in cross-sectional samples suggested that HRV is negatively correlated with symptoms' severity[43], we accounted for baseline severity $B_i = \text{score}_{\pi[i],t=0}/\max(q)$ where the denominator is the maximum value by design on either the YMRS or HDRS rating scale, depending on whether the episode's polarity of the i^{th} subject was mania or depression.

Regression models

We developed two hierarchical linear models, which we nicknamed *two-polarities-model* and *one-disease-model*, illustrated in Fig. 2, where the only difference is that the former allows the lnRMSSD rate of change with respect to symptoms' improvement to vary across polarities (manic and depressive), letting us test whether a specific polarity effect is supported by the data.

In the *two-polarities-model*, we assumed that lnRMSSD for the i^{th} subject at time t is drawn from a Gaussian $\mathcal{N}$ whose mean is a linear combination of the intercept $\beta_{0,i}$, symptoms' improvement $I_{i,t}$, and medications $M_{i,t}$:

$$\text{lnRMSSD}_{i,t} \sim \mathcal{N}\left(\beta_{0,i} + \beta_{1,i} I_{i,t} + \beta_2 M_{i,t}, \sigma_i\right) \quad (2)$$

The subscripts denote that while β_2 does not vary across either individuals or time, each individual has their own intercept term $\beta_{0,i}$ and coefficient $\beta_{1,i}$. This allows each individual to have their own intercept and rate of change with respect to I but crucially these parameters are drawn from a common distribution, as shown below. As regards $\beta_{0,i}$, i.e. the expected value lnRMSSD takes when $I_{i,t} = 0$ (episode onset) and $M_{i,t} = 0$ (no medications with a known effect on HRV), we modelled it as drawn from a Gaussian with a standard deviation fixed to 0.5 but whose mean linearly depends on sex S_i, age A_i, baseline severity B_i plus the intercept α_0:

$$\beta_{0,i} \sim \mathcal{N}\left(\alpha_0 + \alpha_1 A_i + \alpha_2 S_i + \alpha_3 B_i, 0.5\right) \quad (3)$$

As for $\beta_{1,i}$, i.e. the rate of change of lnRMSSD with respect to symptoms' improvement, subjects on different episode polarities draw their slope $\beta_{1,i}$ from Gaussian distributions centred at different values:

$$\beta_{1,i} \sim \mathcal{N}(\gamma_{\pi[i]}, 0.1) \quad (4)$$

Here $\pi[i]$ indeed signifies the mean γ corresponding to the group (polarity π) to which the i^{th} individual's ongoing episode belongs. We defined subject-specific lnRMSSD standard deviation σ_i as drawn from an inverse gamma distribution. The inverse gamma distribution is a convenient choice here, as it is the conjugate prior of a normal distribution with unknown mean and variance. Conjugacy speeds up inference by enabling a closed-form solution to (part of) the posterior:

$$\sigma_i \sim \mathcal{IG}(3, 0.5) \quad (5)$$

The prior for α_0 is a Gaussian centred at the sample average lnRMSSD, i.e. $\overline{\mu}_{lnRMSSD}$:

$$\alpha_0 \sim \mathcal{N}\left(\overline{\mu}_{lnRMSSD}, 0.1\right) \quad (6)$$

$\alpha 1, \alpha 2, \alpha 3$, and β_2 all had a Gaussian prior with mean -0.1 and standard deviation 0.1, informed by previous research showing that female sex, older age, greater symptoms' severity at onset, and the medications mentioned

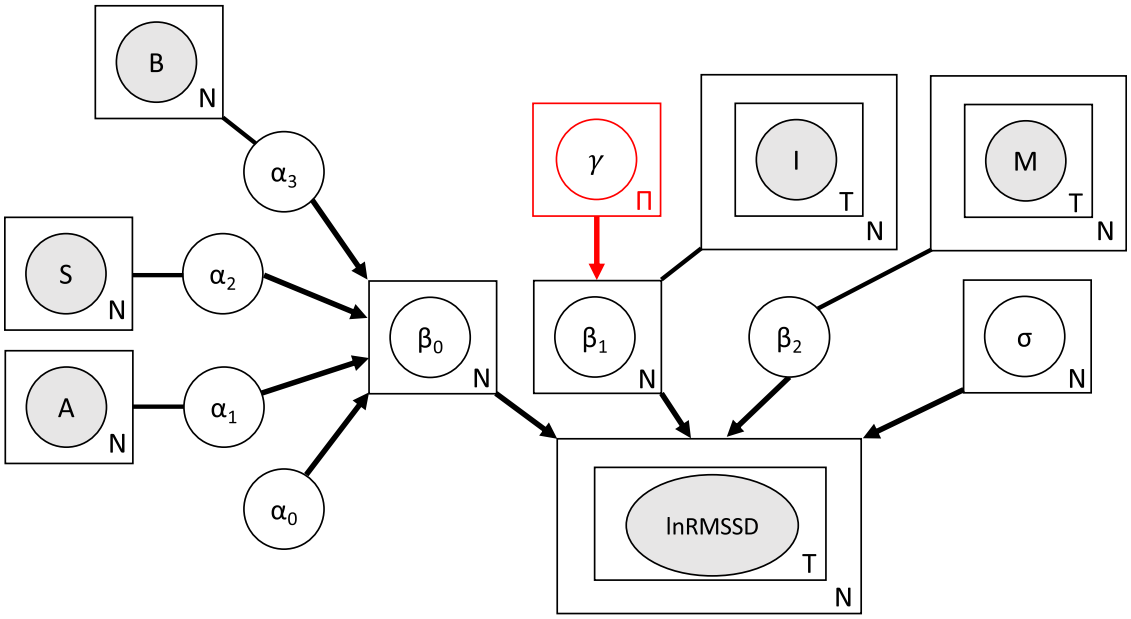


Fig. 2 | lnRMSSD data generating process assumed in the regression models. Grey-shaded nodes represent observed variables, while white nodes represent the model's parameters. Arrows define conditional dependencies in the model graph, while lines connecting parameters to their covariates do not define any probabilistic dependency but are shown simply to clarify which covariate a parameter refers to. The plate notation is used for observed variables and parameters that are repeated, where the letter indicates the number of repetitions; in other words, it indicates the nested structure in the data and in the model. For example, lnRMSSD is contained in two plates: the outer one indicating that samples are drawn at the subjects' level where N is the total number of subjects, the inner one indicating that within each of the N individuals, samples are taken at T times. The node for γ and its outgoing arrow are in red to mark that this node, and thus the dependency of its descendants on episode's polarity where there are $\Pi = 2$ polarities (mania and depression), is only present in the *two-polarities-model* into which the *one-disease-model*, differing only by the lack of this node, is nested. A: age; S: sex; B: baseline symptoms' severity; I: symptoms' improvement; M: medications.

above are associated with a lower HRV[43–45]:

$$\alpha_0, \alpha_1, \alpha_2, \beta_2 \sim \mathcal{N}(-0.1, 0.1) \quad (7)$$

On the other hand, we made a non-committal choice for the prior over γ_π, i.e. a uniform distribution assigning equal probability density to values in the zero-centered interval [-1, 1]:

$$\gamma_\pi \sim \mathcal{U}(-1, 1) \quad (8)$$

In other words, we start from a sceptical position and in advance of seeing any data we do not favour any value for the polarity-specific mean of the Gaussian from which $\beta_{1,i}$ is drawn.

The *one-disease-model* only differs by the lack of dependency of $\beta_{1,i}$ on the episode's polarity. Here, the prior on $\beta_{1,i}$ is a non-committal uniform:

$$\beta_{1,i} \sim \mathcal{U}(-1, 1) \quad (9)$$

Consequently, the *one-disease-model* pools subjects together regardless of polarity but, as with the *two-polarities-model*, $\beta_{0,i}$ and $\beta_{1,i}$ can still vary across subjects while being sampled from the same distribution.

There are different approaches to Bayesian inference. For example, simple models relying on exponential family distributions and conjugacy admit analytical solutions. Often times, however, with more complex models, as it is the case with our hierarchical models, different approaches are required, e.g. sampling-based solutions or variational inference. We adopted the Hamiltonian Monte Carlo (HMC) No-U-Turn Sampler (NUTS)[46], as state of the art inference algorithm and default choice across a number of probabilistic programming libraries[47,48]. In particular, we ran four parallel chains of 2000 tuning steps, 2000 samples, and a target acceptance probability of 0.99 was used for Bayesian inference in both models.

As explained above, the *two-polarities-model* and *one-disease-model* encapsulate different assumptions about the data-generating process. In particular, the former allows the rate of change of lnRMSSD with respect to symptoms' severity to vary across episode's polarity, while the latter does not account for episode polarity. Towards model comparison, i.e. to assess which of the two models betters explains our data, we used the Widely Applicable Bayesian Information Criterion (WAIC)[49]. WAIC calculates an estimate of the out-of-sample log-likelihood and adjusts for the effective number of parameters, providing a more accurate measure of a model's fit and predictive ability. The value of WAIC lacks inherent meaning and only becomes meaningful when comparing it across different models fitted to the same data. Lower WAIC values suggest a better fit of the model to the data. We chose WAIC over other criteria for its Bayesian consistency, effectiveness with complex models, incorporation of uncertainty, focus on predictive accuracy, applicability to hierarchical structures, and bias correction, offering a robust approach. The Bayesian factor, comparing model likelihoods based on observed data, is another tool for selecting between models but faces criticism for its sensitivity to the prior specification, even when different priors lead to minor differences in the posterior[50].

We plotted samples from the posterior distributions over the parameter(s) relevant to our investigation into RMSSD changes with respect to symptoms' improvement (potentially varying across polarities). Towards summarizing the posterior, we computed the Probability of Direction (PD)[51]. This is an index of effect existence, robust to the scale of both the response variable and the predictors. It ranges from 50% to 100%, representing the certainty with which an effect goes in a particular direction (i.e., is positive or negative), and is mathematically defined as the proportion of the posterior distribution that is of the median's sign. We also computed the 95% highest density interval (HDI-95), i.e. the 95% most plausible values in a parameter's posterior. This is more suited than the PD to measure the magnitude of an effect by comparing its overlap with a Regional of Practical Equivalence (ROPE); this is a range of values considered negligible or too small to be of any practical relevance for the use case in question[51,52]. Unlike PD, HDI and ROPE are sensitive to the parameter's scale. For the posterior over β_1, obtained by pooling together samples from all individuals' $\beta_{1,i}$ to

study the overall effect across individuals, we set a ROPE of [-0.05, 0.05]. As we are modelling lnRMSSD, for a given sample $\hat{\beta}_{1,i}$ of $\beta_{1,i}$ a unit change in $I_{i,t}$ (i.e., 100% improvement in symptoms over baseline severity) translates into a change of $\hat{\beta}_{1,i}$ in lnRMSSD for fixed values of other predictors in Equation (2). This is the standard interpretation of regression coefficients. When mapping back onto the original scale of RMSSD, if $\hat{\beta}_{1,i}$ equals an arbitrary value c, RMSSD changes with respect to its baseline value by a multiplicate factor of e^c, where e is the base of the natural logarithm. In fact, by the properties of logarithms, if $ln(Y_{t=T}) - ln(Y_{t=0}) = c$, then $Y_{t=T} = Y_{t=0} \times e^c$ for any arbitrary c. Thus, the ROPE of our choice considers negligible any multiplicate effect of a complete resolution of symptoms on RMSSD between $e^{-0.05} = 0.951$ and $e^{0.05} = 1.051$, in other words, a decrease (increase) of 4.9% (5.1%).

Ethical approval statement

The TIMEBASE/INTREPIBD study was conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice and the Hospital Clinic Ethics and Research Board (HCB/2021/104). All participants provided written informed consent prior to their inclusion in the study. All data were collected anonymously and stored encrypted in servers complying with all General Data Protection Regulation regulations.

Results

Study sample

At the time of this study, a total of 67 patients with BD had been recruited at the onset of a mood episode (29 depression, 38 mania) in the TIMEBASE/INTREPIBD study. Ultimately, a sample of 23 patients were available for this study: 41 dropped out before providing a minimum of three assessments, while 3 did not have a strictly monotonic decrease in their symptoms' severity, thus preventing the use of improvement on symptoms' severity to clock time in our model of change. 9 (resp. 14) individuals were recruited at the onset of a major depressive (resp. manic) episode. 17 (resp. 6) subjects had 3 (resp. 4) follow-up assessments. The median (resp. interquartile range) time (in years) since illness onset was 5 (resp. 17.5). Clinical-demographics are given in Table 1. Figures for the sleep time during the 10 pm to 5 am interval from which RMSSD was extracted are given in Supplementary Table 1. The median percentage of 5-minute sliding windows over sleep time not passing quality control with FLIRT, thus outputting a nan value, was 9.05 (interquartile range 1.95-25.32). Such segments were discarded from analyses and thus not considered in the computation of the night RMSSD.

Prior predictive checks

As customary in a Bayesian data analysis, before model fitting, we ran a series of checks, referred to as prior predictive checks, whose purpose is to assess the soundness of the model assumptions. This is particularly useful in hierarchical models, where the effect of hyperparameters might propagate downstream in the data-generating process in hard-to-predict ways. Specifically, we verified that, as desirable, in advance of seeing any data the implied distribution over lnRMSSD, i.e. the distribution obtained sampling from the model prior and generating synthetic lnRMSSD values, covered the sample distribution of lnRMSSD and had the bulk of the density lying within physiologically plausible values. Secondly, we verified that, before seeing the data, the model did not favour either positive or negative values for the lnRMSSD rate of change with respect to symptoms' improvement.

The top row of Fig. 3 shows the prior distribution over lnRMSSD across both the *two-polarities-model* (left) and the *one-disease-model* (right) against the one observed in the data. The two models have similar prior lnRMSSD distributions, which contain the observed data. However, probability is spread over a range of lnRMSSD values slightly broader than the one in the data, whist still keeping within physiologically plausible values. The 0.05, 0.5, and 0.95 quantiles ($q_{0.05}$, $q_{0.50}$, $q_{0.95}$) were respectively 1.88, 3.21, and 4.51 (1.89, 3.21, and 4.50) for the *two-polarities-model* (*one-disease-model*). The Kullback-Leibler divergence for the prior distribution over lnRMSSD from the *two-polarities-model* to the *one-disease-model*, a measure of "distance" between distributions taking values in $[0, +\infty]$, was 0.00006. On the other hand, $q_{0.05}$, $q_{0.50}$, $q_{0.95}$ were respectively 3.08, 3.60, and 4.18 for the sample lnRMSSD.

The bottom row of Fig. 3 shows the implied distribution over lines within a subject (shown as a way of example), each line representing a hypothesis, i.e. a sample from the prior, about the expected lnRMSSD value as a function of symptoms' improvement upon onset severity. In both models the subject's true values lie with the array of lines in both model, the lines' origin is centred roughly around the sample average lnRMSSD and, as a result of the non-informative prior, a broad range of slopes is credible under the prior with no preference for either positive or negative values (positive or negative rate of change of lnRMSSD with respect to symptoms' improvement).

Model convergence and comparison

In order to infer the posterior distribution over the model parameters, we resorted to Markov Chain Monte Carlo (MCMC) methods, in particular NUTS[46], as our models did not admit an exact, closed-form solution. MCMC involves generating a sequence of random samples, known as chains, which approximate the posterior distribution. However, convergence to the true posterior distribution is not guaranteed, so it's crucial to assess the convergence and mixing properties of the chains. This is typically done using diagnostics such as the Effective Sample Size (ESS), Gelman-Rubin convergence diagnostic ($\hat{R}$), and Bayesian Fractions of Missing Information (BFMIs). In both the *two-polarities-model* and the *one-disease-model* the chains mixed well with all ESS > 1000, all $\hat{R} = 1$, and all BFMIs ≥0.75.

The WAIS for the *two-polarities-model* and the *one-disease-model* was respectively -92.94 and -98.90, indicating that, conditional on our data, the latter model, not positing the lnRMSSD rate of change with respect to symptoms' improvement as dependent on the episode's polarity, is a better fit.

lnRMSSD rate of change with respect to symptoms' improvement

Further to investigating possible differences across the episode's polarities, a central question in our investigation was how lnRMSSD changed across the

Table 1 | Clinical-demographic features of the study sample

	AGE MEAN (STD)	FEMALES N (PERCENTAGE)	MEDICATIONS # MEAN (STD)	BASELINE SYMPTOMS' SEVERITY MEAN (STD)
MANIA	42.14 (12.81)	5 (35.71%)	2.86 (1.30)	YMRS
N=14				25.64 (5.09)
DEPRESSION	44.34.56 (13.03)	6 (66.67%)	3.78 (0.63)	HDRS
N=9				19.11 (3.21)

"Medications #" refers to the number of drugs recorded in our cohort with a known influence on HRV, which subjects were taking at the moment of study admittance; further details on medications are given in Supplementary Table 2. We report clinical-demographic features for the 44 patients not included in the present analyses as not providing a minimum of three HRV samples in Supplementary Table 3. Total score on Young Mania Rating Scale (YMRS) and Hamilton Depression Rating Scale-17 (HDRS) was used to track symptoms' severity in manic and depressive episodes, respectively. The figures herewith shown refer to the first assessment (acute episode onset). Note that, as YMRS and HDRS do not share the same range ([0-60] and [0-52], respectively), the percentage of improvement with respect to onset total score was used to clock time across polarities in the regression model.

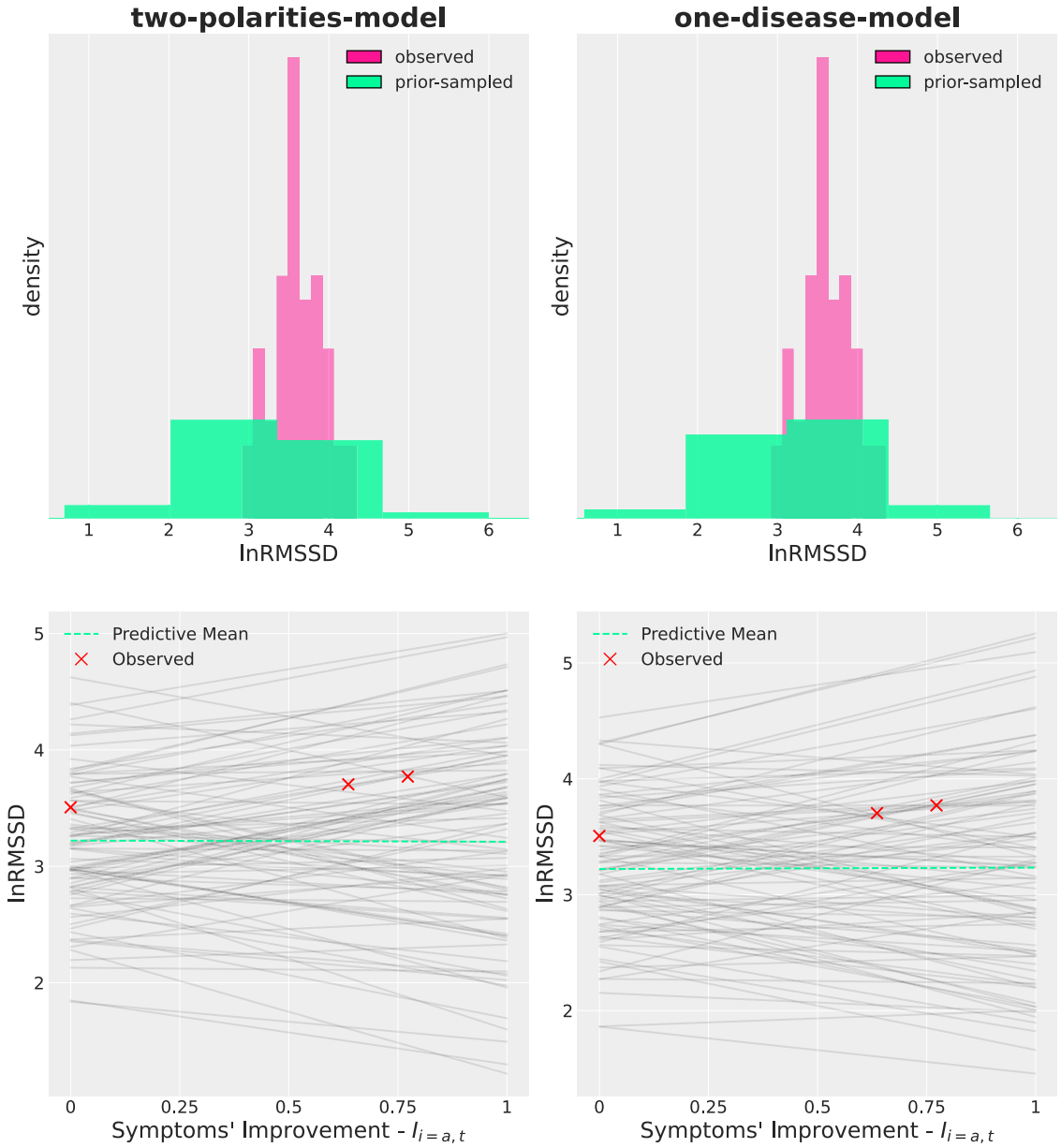


Fig. 3 | Prior predictive checks across the two regression models. The left column refers to the *two-polarities-model* while the right column to the *one-disease-model*. The (normalized) histograms in the top row show the observed lnRMSSD distribution against the lnRMSSD distribution implied by the prior. It can be seen that the observed lnRMSSD (pink) is tightly concentrated over a narrow range in comparison to the prior lnRMSSD (green), which puts some probability density on values at the boundaries of the physiologically plausible range. However, the bulk of the prior lnRMSSD contains the observed lnRMSSD. The three red crosses in each bottom row plot shows lnRMSSD measures at different stages of symptoms' improvement for a subject from our dataset, chosen as a way of example and assigned the dummy subject-id *a*. Superimposed are one hundred lines, each showing the expected lnRMSSD value for different draws from the prior. As a result of a vague and non-committal prior, lines can have a variety of slopes with no preference for either positive or negative values. The dashed green line represents the average across the one hundred black lines.

trajectory of symptoms' improvement, from episode onset up to euthymia. As the *one-disease-model* came out on top in model comparison, we collected and pulled together posterior samples from $\beta_{1,i}$ across the $N = 23$ individuals in our analyses, in order to study the overall effect β_1 regardless of the specific subject.

Figure 4 a illustrates the prior distribution, defined in Equation (9), for β_1. It can be seen how the prior is non-committal and vague, as it does not favour any value in the interval [-1,1] and admits a broad variability in the effect that $I_{i,t}$ can have on lnRMSSD, from -1 to 1 (the scale is logarithmic).

Figure 4b illustrates the posterior distribution over β_1. Bayesian inference reassigned credibility so that relatively strong effects of β_1 on lnRMSSD have very little probability densities, i.e. values below (above) -0.5 (0.5), while hypotheses compatible with the data now have higher density. Contrast how, upon conditioning on the data, the distribution on β_1 changed from Fig. 4a to b. We calculated commonly used statistics and decision rules on the posterior. The median (dashed red line) lies at 0.208. The PD indicates that β_1 is strictly positive with high probability, i.e. 95.175%. It can indeed be seen that samples from the posterior overwhelmingly favour positive values. The HDI-95, i.e. the narrowest interval containing 95% of the posterior probability density, spans [-0.03662–0.47061], thus overlapping but not containing the rope [-0.05, 0.05]. As per[52] recommendations, the HDI-95-based decision rule is therefore to withhold decision and collect more data to increase the precision of the estimates.

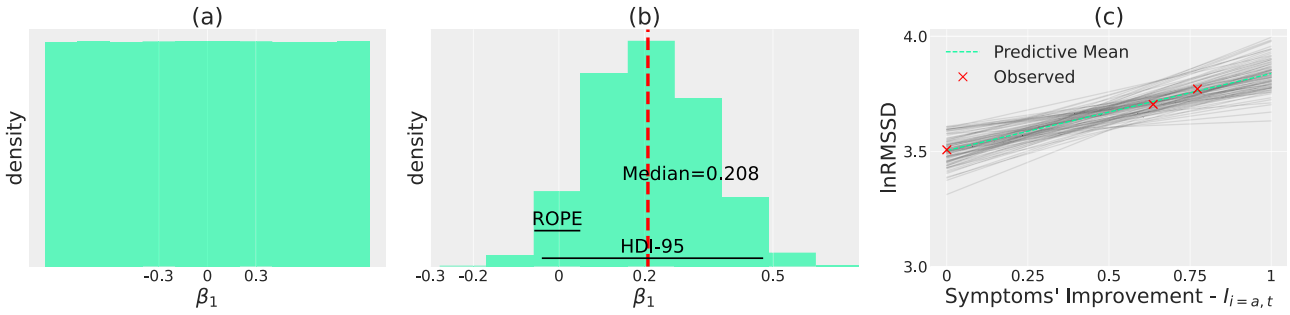


Fig. 4 | Prior and Posterior distributions over β_1. **a**: prior distribution over β_1. **b**: posterior distribution over β_1 along with median (red, dashed line), 95% Highest Density Interval (HDI-95) spanning [-0.03662-0.47061] and Region of Practical Equivalence (ROPE) at [-0.05, 0.05]. **c**: posterior distribution over expected lnRMSSD values as a function of symptoms' improvement for a subject recruited at the onset of a manic episode, identified with the dummy subject-id *a*. The posterior for the other subjects is available in Supplementary Fig. 2. Each black line (a total of one hundred is herewith displayed to avoid clutter) represents a single draw from the posterior, while the dashed green line is the average across all black lines sampled from the posterior. This illustrates how the Bayesian framework naturally incorporates uncertainty in its outputs, as in this plot we indeed have a distribution over lines and not just a single line. This notion of uncertainty enables better-informed decisions in a clinical setting, e.g. the confidence in a given positive trend in lnRMSSD is higher when lines are tightly packed around the average value.

Figure 4 c lastly shows the posterior for the same individual reported in prior predictive checks, bottom-right of Fig. 3, to whom the dummy identifier *a* was assigned. The distribution over lines now span only a narrow range of possible values, with a tendency for positive values. The posterior distribution for other subjects in the study can be seen in Supplementary Fig. 2 and overall confirms the positive trend in β_1 values.

The posterior over the co-variates' coefficients, i.e. age, sex, onset symptoms' severity, and number of medications with an influence on HRV, can also been seen in Supplementary Fig. 5. In general, the posterior did not differ much from the prior distribution in either shape or direction; however, for β_2, i.e. the coefficient associated with the number of medications known to affect HRV, the posterior sharpened and its HDI-95 excluded the 0 value.

Discussion

In this work, we studied how lnRMSSD changes as the symptoms' severity subsides over the course of an acute BD episode. Our findings do not support a specific effect of polarity, i.e. mania or depression, on the dynamics of change in lnRMSSD. To the best of our knowledge, only the work by Faurholt-Jepsen et al.[23] considered HRV across the full BD spectrum but only took one HRV sample per episode across patients, thus not investigating within-episode dynamics and limiting comparability with this study. The lack of a polarity-specific component to HRV trajectories in our study suggests that within-episode HRV changes may not be useful to distinguish between manic and depressive phases. On the other hand, our findings support with high confidence the existence of a positive rate of change of lnRMSSD with respect to symptoms' improvement over the course of an acute BD episode. However, our data did not show that the HDI-95 completely excludes the ROPE. This is likely related to the sample size, as sensitivity analyses (Supplementary Note 1) showed that increasing either the number of recruited subjects or the number of observations per subject led to a higher chance of a model fit where the HDI-95 completely excludes the ROPE, assuming a data generating process where the HDI-95 on the distribution for the lnRMSSD slope (β_1) does exclude the ROPE. While the Bayesian approach commands to consider the entire distribution, the HDI-95 summary and the ROPE-partial-overlap rule[52] suggests withholding decision and collect more data before developing an intervention that might depend on the parameter of interest completely excluding the ROPE.

Sample size is indeed a limitation of this and previous studies into intraindividual HRV changes in BD, since collecting longitudinal data from patients with BD, especially when on a manic episode, is a resource-intensive endeavour. The inherent limitation of sample size hinders the frequentist approach[53] used in previous studies. We thus opted for a Bayesian approach in our work, as it is more suitable to small samples and capable of quantifying uncertainty in a principled manner, a desirably property when data is used to inform decision-making in potentially high-risk environments such as healthcare. Furthermore, we went beyond simply assessing the distribution of a test statistic and proposed an explainable probabilistic model that attempts to explain how lnRMSSD values are generated across successive observations within-subjects and how different clinical-demographic covariates interact in this process.

Consistently with our results, the majority of previous studies investigating intra-individual HRV changes from mania to euthymia, while only collecting two samples per patient, found a positive difference[20,21]. Previous cross-sectional studies comparing patients on a manic episode to healthy controls also found a reduced HRV in mania[54]. Of importance, HRV in euthymic BD remains lower than in healthy controls despite full clinical remission, even though at least part of this difference is likely due to medications[55]. As regards studies into bipolar depression, one[22], taking only a sample from acute state and one from euthymia, did not find any significant difference in HRV across acute state and euthymia. However, a cross-sectional study[43] found a negative association between symptoms' severity and HRV. The inconsistency of findings in the literature may in part be a result of the sample size used in this type of studies and the frequentist approach. The Bayesian approach we herewith adopted is arguably better suited as it yields graded evidence, suggesting when collecting more data is likely to be fruitful. Secondly, we note that studies differ in the HRV metrics they employed and, more importantly, the device used for IBI data collection and the algorithms for IBI pre-processing. This could also explain inconsistency in findings. For the sake of transparency and reproducibility, we release the codebase we developed for these analyses.

The results of this study need to be balanced against some limitations. 1) We could not include BMI, alcohol, and nicotine intake as covariates in our models since these HRV confounders were not collected in the TIMEBASE/INTREPIBD study. Similarly, while unlike some previous studies (e.g.[16]) we included medications, we did not account for their plasma concentration, receptor profile, or interactions but only considered the total number of known interfering drugs. 2) We took one step beyond previous studies and fitted a model of change with at least three samples available per subject per episode, however the lack of a higher number of intra-individual observations constrained us to fit a linear model since non-linear patterns may not be identifiable with only three data points. However, we do not have reasons to exclude a non-linear trajectory. 3) The limited sample size likely prevented us from asserting the magnitude of the rate of change in lnRMSSD with respect to symptoms' improvement in a way to exclude a region of practical equivalence, and further research in this sense is needed.

In conclusion, previous converging evidence indicated an HRV reduction in BD relatively to healthy controls, pointing to an impairment in the autonomous nervous system. This study, the first to the best of our knowledge to include a minimum of three observations per patient per episode across both polarities of BD, suggests that an improvement in

symptoms' severity upon an acute episode is paralleled by a positive change in HRV. However, the pattern of HRV change does differ across mania and depression, the two polarities of BD. Thus, our findings suggest that HRV, thanks to an increasing adoption of wearable devices, may have a role in monitoring the course of an episode in clinical settings, acting as a measurable biological signal, which can complement clinical assessments; however, it may be not useful towards distinguishing polarities in BD. Studies of HRV in BD have been dogged by limited sample size, a limitation inherent to this type of studies. Crucially, unlike frequentist statitics, the Bayesian framework we herewith adopted, allowed for a fine-grained appreciation of the evidence, inspecting posterior distributions conditioned on the data (and the posited model), and the formulation of a generative, interpretable probabilistic model accounting for how different variables interact in generating HRV values within patients over the course of a BD episode.

Data availability

The data used for the present study can be made available through reasonable requests to the corresponding author due to data sharing restrictions

Code availability

The codebase developed for this work is available at https://github.com/april-tools/bayesian-hrv. Python 3.10 programming language was used, with Bayesian statistical modelling implemented in PyMC[48] and ArviZ[56].

Received: 19 March 2024; Accepted: 22 September 2024;
Published online: 03 October 2024

References

1. Merikangas, K. R. et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Arch. Gen. Psychiatry* **68**, 241–251 (2011).
2. Simon, J. et al. The costs of bipolar disorder in the united kingdom. *Brain Behav.* **11**, e2351 (2021).
3. Hayes, J. F., Marston, L., Walters, K., King, M. B. & Osborn, D. P. Mortality gap for people with bipolar disorder and schizophrenia: Uk-based cohort study 2000–2014. *Br. J. Psychiatry* **211**, 175–181 (2017).
4. Ramesh, A., Nayak, T., Beestrum, M., Quer, G. & Pandit, J. A. Heart rate variability in psychiatric disorders: A systematic review. *Neuropsychiat. Disease Treat.* 2217–2239 (2023).
5. Ronca, V. et al. Wearable technologies for electrodermal and cardiac activity measurements: A comparison between fitbit sense, empatica e4 and shimmer gsr3+. *Sensors* **23**, 5847 (2023).
6. Shaffer, F. & Ginsberg, J. P. An overview of heart rate variability metrics and norms. *Front. Public Health* 258 (2017).
7. Stone, J. D. et al. Assessing the accuracy of popular commercial technologies that measure resting heart rate and heart rate variability. *Front. Sports Active Living* 37 (2021).
8. Empatica EmbracePlus. Embrace plus user manual https://www.empatica.com/en-eu/embraceplus/ (2021). Accessed December 18 2023.
9. Plews, D. J., Laursen, P. B., Stanley, J., Kilding, A. E. & Buchheit, M. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. *Sports Med.* **43**, 773–781 (2013).
10. Plews, D. J. et al. Monitoring training with heart-rate variability: How much compliance is needed for valid assessment? *Int. J. sports Physiol. Perform.* **9**, 783–790 (2014).
11. Tarvainen, M., Lipponen, J., Niskanen, J. & Ranta-Aho, P. Kubios hrv version 3–user's guide. *Kuopio: University of Eastern Finland* (2017).
12. Nuuttila, O.-P., Nummela, A., Korhonen, E., Häkkinen, K. & Kyröläinen, H. Individualized endurance training based on recovery and training status in recreational runners. *Med. Sci. Sports Exercise.* **54** (2022).
13. Alvares, G. A., Quintana, D. S., Hickie, I. B. & Guastella, A. J. Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis. *J. Psychiatry Neurosci.* **41**, 89–104 (2016).
14. Chalmers, J. A., Quintana, D. S., Abbott, M. J.-A. & Kemp, A. H. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. *Front. psychiatry* **5**, 80 (2014).
15. Koch, C., Wilhelm, M., Salzmann, S., Rief, W. & Euteneuer, F. A meta-analysis of heart rate variability in major depression. *Psychol. Med.* **49**, 1948–1957 (2019).
16. Faurholt-Jepsen, M., Kessing, L. V. & Munkholm, K. Heart rate variability in bipolar disorder: a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **73**, 68–80 (2017).
17. Hillebrand, S. et al. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose–response meta-regression. *Europace* **15**, 742–749 (2013).
18. Sessa, F. et al. Heart rate variability as predictive factor for sudden cardiac death. *Aging (Albany NY)* **10**, 166 (2018).
19. Anmella, G. et al. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. *JMIR Mhealth Uhealth* (2023).
20. Stautland, A. et al. Reduced heart rate variability during mania in a repeated naturalistic observational study. *Front. Psychiat.* **14** (2023).
21. Wazen, G. L. L., Gregório, M. L., Kemp, A. H. & de Godoy, M. F. Heart rate variability in patients with bipolar disorder: from mania to euthymia. *J. Psychiatr. Res.* **99**, 33–38 (2018).
22. Hage, B. et al. Diminution of heart rate variability in bipolar depression. *Front. Public health* **5**, 312 (2017).
23. Faurholt-Jepsen, M., Brage, S., Kessing, L. V. & Munkholm, K. State-related differences in heart rate variability in bipolar disorder. *J. Psychiatr. Res.* **84**, 169–173 (2017).
24. Singer, J. D. & Willett, J. B.*Applied longitudinal data analysis: Modeling change and event occurrence* (Oxford university press, 2003).
25. Parsons, S. & McCormick, E. M. Two timepoints poorly capture trajectories of change: A warning for longitudinal neuroscience. *Available at SSRN 4415029* (2023).
26. Quintana, D. S. & Williams, D. R. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using jasp. *BMC Psychiatry* **18**, 1–8 (2018).
27. Colling, L. J. & Szűcs, D. Statistical inference and the replication crisis. *Rev. Philos. Psychol.* **12**, 121–147 (2021).
28. Wagenmakers, E.-J. et al. Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bull. Rev.* **25**, 35–57 (2018).
29. Rognli, E. W., Zahl-Olsen, R., Rekdal, S. S., Hoffart, A. & Bertelsen, T. B. Editorial perspective: Bayesian statistical methods are useful for researchers in child and adolescent mental health (2023).
30. Young, R. C., Biggs, J. T., Ziegler, V. E. & Meyer, D. A. A rating scale for mania: reliability, validity and sensitivity. *Br. J. psychiatry* **133**, 429–435 (1978).
31. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. psychiatry* **23**, 56 (1960).
32. Tohen, M. et al. The international society for bipolar disorders (isbd) task force report on the nomenclature of course and outcome in bipolar disorders. *Bipolar Disord.* **11**, 453–473 (2009).
33. Empatica. E4 wristband technical specifications - empatica support https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications (2020).
34. Li, K., Cardoso, C., Moctezuma-Ramirez, A., Elgalad, A. & Perin, E. Heart rate variability measurement through a smart wearable device: Another breakthrough for personal health monitoring? *Int. J. Environ. Res. Public Health* **20**, 7146 (2023).
35. Vieluf, S. et al. Twenty-four-hour patterns in electrodermal activity recordings of patients with and without epileptic seizures. *Epilepsia* **62**, 960–972 (2021).
36. Nasseri, M. et al. Signal quality and patient experience with wearable devices for epilepsy management. *Epilepsia* **61**, S25–S35 (2020).

37. Van Hees, V. T. et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PloS one* **10**, e0142533 (2015).
38. Patterson, M. R. et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. *NPJ Dig. Med.* **6**, 51 (2023).
39. Föll, S. et al. Flirt: A feature generation toolkit for wearable data. *Comput. Methods Prog. Biomed.* **212**, 106461 (2021).
40. Cao, R. et al. Accuracy assessment of oura ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: comprehensive analysis. *J. Med. Internet Res.* **24**, e27487 (2022).
41. de Vries, H., Kamphuis, W., van der Schans, C., Sanderman, R. & Oldenhuis, H. Trends in daily heart rate variability fluctuations are associated with longitudinal changes in stress and somatisation in police officers. In *Healthcare*, **10**, 144 (MDPI, 2022).
42. Boudreau, P., Yeh, W.-H., Dumont, G. A. & Boivin, D. B. Circadian variation of heart rate variability across sleep stages. *Sleep* **36**, 1919–1928 (2013).
43. Ortiz, A. et al. Reduced heart rate variability is associated with higher illness burden in bipolar disorder. *J. Psychosom. Res.* **145**, 110478 (2021).
44. O'Regan, C., Kenny, R., Cronin, H., Finucane, C. & Kearney, P. Antidepressants strongly influence the relationship between depression and heart rate variability: findings from the irish longitudinal study on ageing (tilda). *Psychol. Med.* **45**, 623–636 (2015).
45. Sammito, S. & Böckelmann, I. New reference values of heart rate variability during ordinary daily activity. *Heart Rhythm* **14**, 304–307 (2017).
46. Hoffman, M. D. et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).
47. Phan, D., Pradhan, N. & Jankowiak, M. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554* (2019).
48. Abril-Pla, O. et al. Pymc: a modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Sci.* **9**, e1516 (2023).
49. Watanabe, S. A widely applicable bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897 (2013).
50. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B.*Bayesian data analysis* (Chapman and Hall/CRC, 1995).
51. Makowski, D., Ben-Shachar, M. S., Chen, S. A. & Lüdecke, D. Indices of effect existence and significance in the bayesian framework. *Front. Psychol.* **10**, 2767 (2019).
52. Kruschke, J. K. Bayesian data analysis. *Wiley Interdiscip. Rev.: Cogn. Sci.* **1**, 658–676 (2010).
53. De Prisco, M. & Vieta, E. The never-ending problem: Sample size matters. *Eur. Neuropsychopharmacol.: J. Eur. Coll. Neuropsychopharmacol.* **79**, 17–18 (2023).
54. Bassett, D. A literature review of heart rate variability in depressive and bipolar disorders. *Aust. N.Z. J. Psychiatry* **50**, 511–519 (2016).
55. Bassett, D. et al. Reduced heart rate variability in remitted bipolar disorder and recurrent depression. *Aust. N.Z. J. Psychiatry* **50**, 793–804 (2016).
56. Kumar, R., Carroll, C., Hartikainen, A. & Martín, O. A. Arviz a unified library for exploratory analysis of bayesian models in python (2019).

Acknowledgements

We acknowledge the contribution of all the participants to the study.This project was funded by the Instituto de Salud Carlos III (ISCIII) (PI21/00340, TIMEBASE Study), cofunded by the European Union, as well as a Baszucki Brain Research Fund grant (PI046998) from the Milken Foundation. The ISCIII or the Milken Foundation had no further role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.F.C. and B.M.L. are supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.G.A. is supported by a Rio Hortega 2021 grant (CM21/00017) and M-AES mobility fellowship (MV22/00058), from the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and co-financed by the Fondo Social Europeo Plus (FSE+).C.V.P. is supported by a contract funded by MCIN/AEI/TED2021-131999BI00 Strategic Projects Oriented to the Ecological Transition and the Digital Transition 2021 and by the "European Union NextGenerationEU/PRTR".I.G. thanks the support of the Spanish Ministry of Science and Innovation (MCIN) (PI23/00822) integrated into the Plan Nacional de I+D+I and cofinanced by the ISCIII-Subdirección General de Evaluación y cofinanciado por la Unión Europea (FEDER, FSE, Next Generation EU/Plan de Recuperación Transformación y Resiliencia PRTR); the Instituto de Salud Carlos III; the CIBER of Mental Health (CIBERSAM); and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (2021 SGR 01358), CERCA Programme / Generalitat de Catalunya as well as the Fundació Clínic per la Recerca Biomèdica (Pons Bartran 2022-FRCB PB1 2022). A.V. is supported by the "UNREAL" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC.

Author contributions

F.C. conceived of the study, proposed the methodology, developed the software codebase for the analyses, and prepared the manuscript. B.M.L. contributed to the manuscript writing. G.A. contributed to manuscript writing and data collection. C.V.P., I.P., M.V., I.G.F., A.B., and M.G. collected the data for the TIMEBASE/INTREPIBD study. E.V., S.L., and H.W. critically reviewed the manuscript and provided feedback on the clinical side. D.H.M. is the principal investigator and the co-ordinator of the TIMEBASE/INTREPIBD study and critically reviewed the manuscript. A.V. contributed to the study design, methodology development, and manuscript writing.

Competing interests

G.A. has received CME-related honoraria, or consulting fees from Angelini, Casen Recordati, Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, Rovi, and Viatris, with no financial or other relationship relevant to the subject of this article. I.G. has received grants and served as consultant, advisor or CME speaker for the following identities: ADAMED, Angelini, Casen Recordati, Esteve, Ferrer, Gedeon Richter, Janssen Cilag, Lundbeck, Lundbeck-Otsuka, Luye, SEI Healthcare, Viatris outside the submitted work. She also receives royalties from Oxford University Press, Elsevier, Editorial Médica Panamericana. M.V. has received research grants from Eli Lilly & Company and has served as a speaker for Abbott, Bristol-Myers Squibb, GlaxoSmithKline, Janssen-Cilag, and Lundbeck. E.V. has received grants and served as consultant, advisor or CME speaker for the following entities: AB-Biotics, AbbVie, Adamed, Angelini, Biogen, Beckley-Psytech, Biohaven, Boehringer-Ingelheim, Celon Pharma, Compass, Dainippon Sumitomo Pharma, Ethypharm, Ferrer, Gedeon Richter, GH Research, Glaxo-Smith Kline, HMNC, Idorsia, Johnson & Johnson, Lundbeck, Luye Pharma, Medincell, Merck, Newron, Novartis, Orion Corporation, Organon, Otsuka, Roche, Rovi, Sage, Sanofi-Aventis, Sunovion, Takeda, Teva, and Viatris, outside the submitted work. All authors report no financial or other relationship relevant to the subject of this article.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44184-024-00090-x.

Correspondence and requests for materials should be addressed to Filippo Corponi.

Supplementary information

Supplementary Note 1 Sensitivity analyses

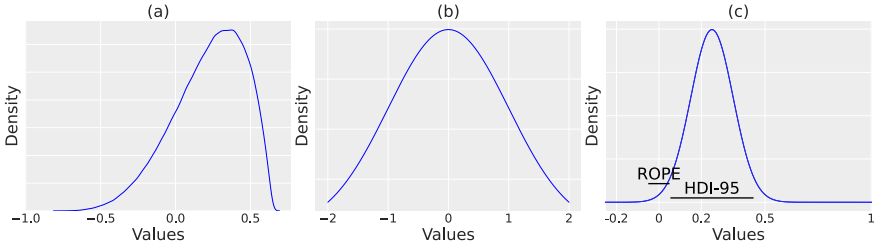
In sensitivity analyses, first (**i**), we assessed to what extent the results from the *one-disease-model* were influenced by the choice of prior for $\beta_{1,i}$. Second (**ii**), we examined with what frequency the HDI-95 would completely exclude the ROPE if multiple synthetic datasets were generated according to the *one-disease-model*, sampling $\beta_{1,i}$ from $\mathcal{N}(0.247, 0.1)$ – in other words positing for the data generating process a distribution on $\beta_{1,i}$ whose HDI-95 does exclude the ROPE [-0.05, 0.05] – and the *one-disease-model* described in was fit to each such dataset. This experiment indicates how likely our pipeline is to recover a distribution whose HDI-95 does not overlap with the ROPE, assuming the true data generating process is indeed governed by such a distribution, under the constraints of a limited sample size (N=23) with up to four follow-up measurements per patient. We also investigated how increasing either the number of follow-up observations per patients or the sample size would affect the chances of recovering a posterior whose HDI-95 excludes the ROPE.

i We experimented with two alternative choices of priors for $\beta_{1,i}$ in the *one-disease-model*: (a) a beta distribution with parameters a and b of 5 and 2, scaled by 1.5 and shifted by -0.85, and (b) a normal distribution with parameters μ and σ of 0 and 0.1. The probability density function (PDF) of the two distributions is displayed in Supplementary Figure 1. (a) is a distribution favouring positive values for $\beta_{1,i}$, the area under the curve (AUC) to the right of 0 is indeed 81.50%. On the other hand, (c) has its mode at zero and does not favour positive over negative values or vice versa, the AUC between -0.1 and 0.1 is 68.26%. (a) and (b) have a positive probability of direction of 96.677% and 95.152% respectively, but neither led to an HDI-95 excluding the ROPE ([-0.0120, 0.4658] and [-0.0952, 0.2500] respectively). The WAIC was -92.623 and -91.992 respectively.

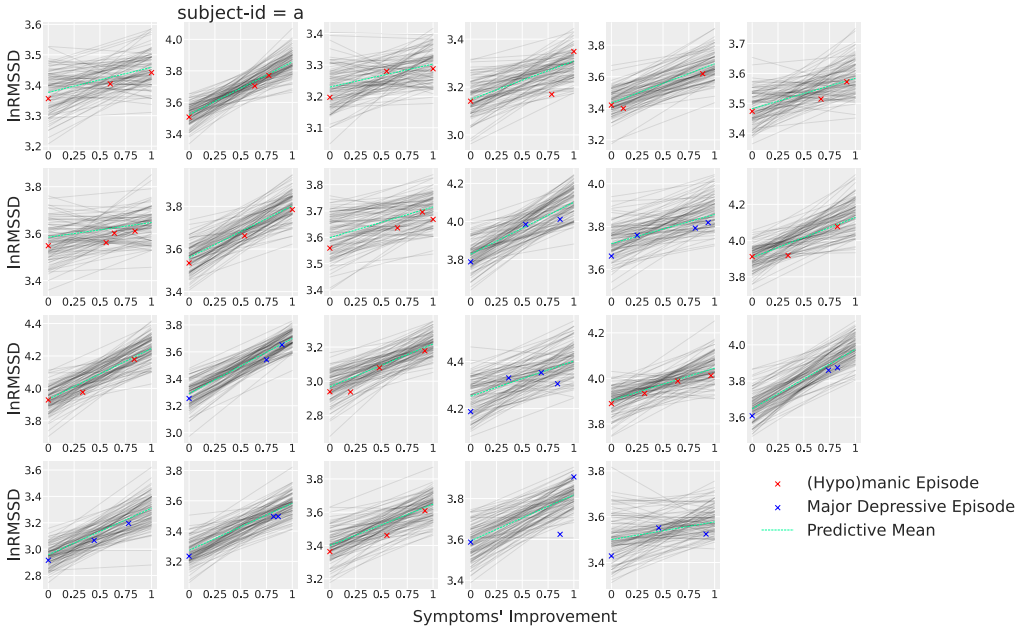
ii Synthetic data for age was generated from a Gaussian whose mean and standard deviation were set to the sample mean and standard deviation. Sex was sampled from a Bernoulli with mean set to the sample proportion of female participants. Baseline severity was sampled from a uniform, with support going from the sample minimum baseline severity to 1. Medications's number was sampled from a discrete uniform distribution over [2, 3, 4, 5, 6] where the probability for a patient of remaining on the same number of medications was 97.5%. This value reflects the clinical practice tendency to use continue the starting treatment regime throughout the episode. 73.91% (26.09%) of the subjects were sampled three (four) times over their trajectory of symptoms' improvement, where the first sampled was collected at 0, i.e. acute episode onset, as per the study design. Other observations were sampled uniformly at random at a symptoms' improvement position between 0.2 and 1. 73.91% (26.09%) corresponded to the fraction of patients in our sample with three (four) observations. All parameters but $\beta_{1,i}$ were sampled according to prior specified in . For $\beta_{1,i}$ we assumed a normal prior $\mathcal{N} \sim (0.25, 0.1)$. As shown in Supplementary Figure 1 (c), the HDI-95 [0.0540, 0.4460] is just to the right of the ROPE [-0.05, 0.05].

Synthetic datasets were sampled as described above. The model specified in , i.e. using a non-committal uniform prior on $\beta_{1,i}$, was subsequently fit to each synthetic dataset. We computed the proportion of times out of 100 simulations (i.e. 100 synthetic datasets) the HDI-95 from the posterior over $\beta_{1,i}$ excluded the ROPE. With 23 patients, of whom 73.91% (26.09%) have three (four) longitudinal observation, the HDI-95 completely lay to the right of the ROPE 62% of the time. Keeping the sample size N to 23 but generating synthetic datasets where all subjects are sampled 5 times over the trajectory of their symptoms' improvement, the fraction of time the HDI-95 excluded the ROPE rose to 75%. Lastly, keeping the number of longitudinal samples across subjects unvaried from that observed in our cohort but increasing N to 50, the HDI-95 excluded the ROPE 71% of the times.

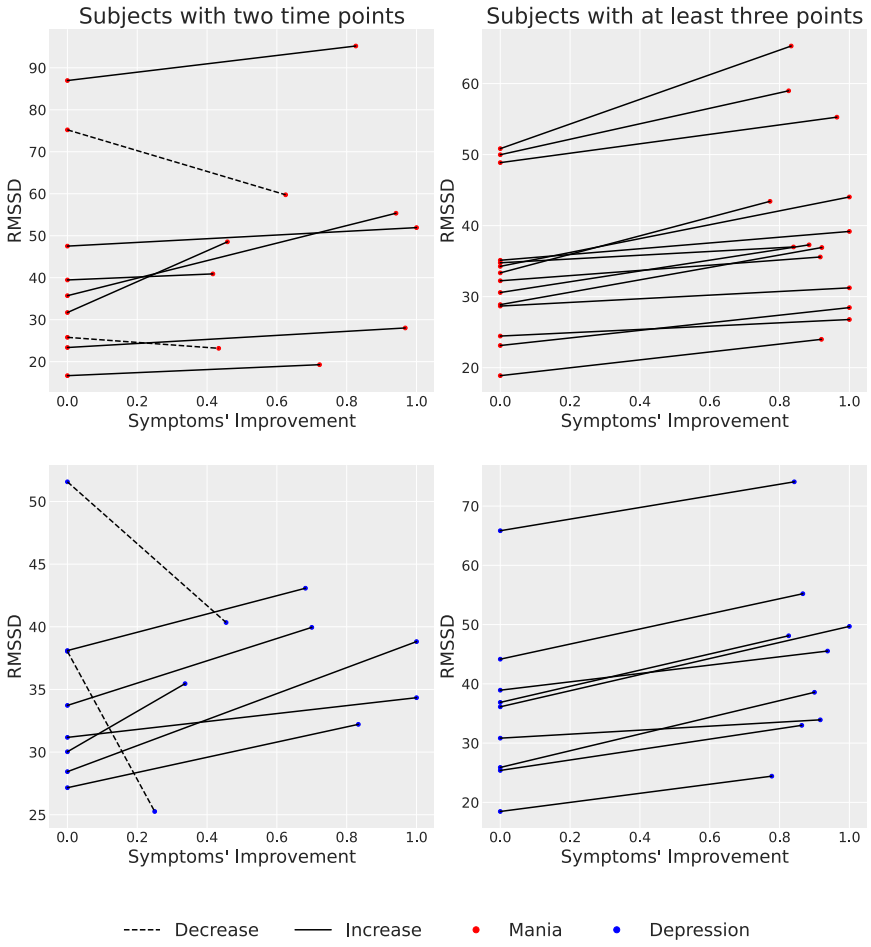
Supplementary figures



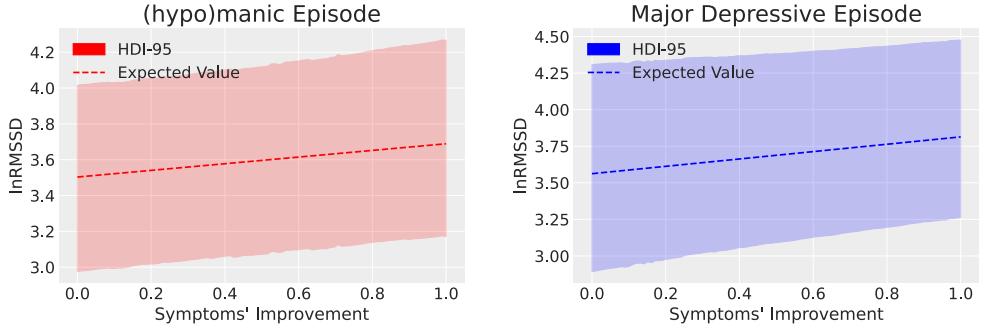
Supplementary Figure 1: **Priors for lnRMSSD rate of change with respect to symptoms' improvement** Probability density function of the priors on $\beta_{1,i}$ used in sensitivity analyses.



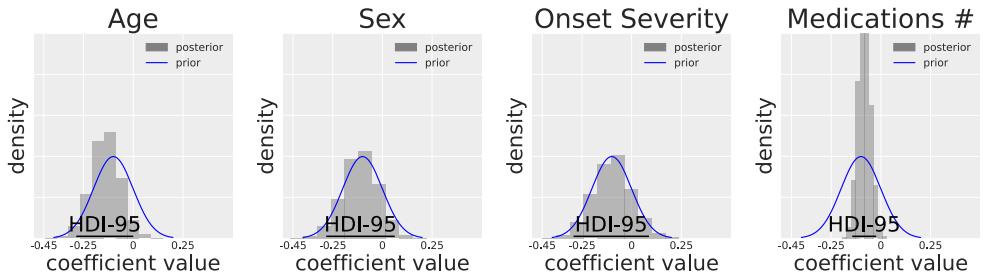
Supplementary Figure 2: **Posterior distribution from the *one-disease-model* over expected lnRMSSD values as a function of symptoms' improvement**. All subjects included in our are analyses are herewith shown. Red (blue) crosses indicates observed lnRMSSD values in patients on a mania (major depression) episode. The subplot with heading "subject-i=a" is the same as the one shown in Figure 4c. Within each subplot corresponding to a given subject in the cohort, each black line (a total of one hundred is herewith displayed to avoid clutter) represents a single draw from the posterior, while the dashed green line is the average across all black lines sampled from the posterior.



Supplementary Figure 3: **Trends in RMSSD when considering subjects from the TIME-BASE/INTREPIBD cohort with at least two measurements**. Plots on the left hand-side shows subjects with only two measurements available, while those on the right hand-side subjects with a minimum of three measurements, i.e. the very same subjects used for the main analyses and depicted in Supplementary Figure 2. For this latter group of subjects, we just retained the first and the last measurement available to avoid clutter and aid direct comparison. A positive trend in RMSSD as symptoms improve can be seen in subjects with just two RMSSD measurements available.



Supplementary Figure 4: **Posterior distribution from the *one-disease-model* over expected lnRMSSD values as a function of symptoms' improvement aggregated by episode's polarity**. The plot on the left (right) is obtained aggregating posterior draws from subjects recruited at the onset of a manic (major depressive) episode. The dashed line corresponds to the average of lnRMSSD expectations, i.e. the average across samples for the mean μ of the Gaussian in Equation (2), shown as a function of symptoms' improvement; the area shaded in red (blue) indicated the HDI-95 for the lnRMSSD expectation. Note that the *one-disease-model* is blind to information regarding episode polarity and was preferred, based on the WAIC, to the *two-polarities-model* which explicitly encodes the episode polarity.



Supplementary Figure 5: **Posterior distribution from the *one-disease-model* over co-variate coefficients**. The histogram, obtained from posterior samples, is normalized so that the area under the curve sums to one. Superimposed is the density of the prior, i.e. a Gaussian with mean and standard deviation of -0.1 and 0.1 respectively. Notice that the posterior for the coefficient associated with the number of medications, sharpened around a narrower range of values relative to the posterior for other co-variates and its HDI-95 excluded the 0 value, suggesting a significant effect of medications number on lnRMSSD changes.

Supplementary tables

Supplementary Table 1: **Nighttime sleep (hours) across episode polarities and follow-up assessments.** We herewith report the mean (sd) sleep time, in hours, dected with the algorithm by Van Hees et al. 37 during the 10 pm and 5 window on the first day of the recording.

	t_0	t_1	t_2	t_3
	MEAN (STD)	MEAN (STD)	MEAN (STD)	MEAN (STD)
MANIA	5.65 (0.69)	5.12 (0.34)	5.45 (0.76)	5.89 (0.61)
DEPRESSION	5.34 (0.45)	5.59 (0.58)	5.12 (0.63)	5.71 (0.24)

Supplementary Table 2: **Medications by class across manic and depressive episodes for each longitudinal assessment.** Each cell shows the average (standard deviation) number of medications under a given drug class for patients across manic episode (ME) and major depressive episode (MDE) during each of the follow-up assessments, $t \in \{t_0, t_1, t_2, t_3\}$. Li : lithium; SSRI: selective serotonin reuptake inhibitors; SNRI: serotonin and norepinephrine reuptake inhibitors; TCA: tricyclics; MAOI: monoamine oxidase inhibitors; OAD: other antidepressants; AP1: first generation antipsychotic; AP2 second generation antipsychotic; AED: antiepileptic drug; AMP: amphetamines; AH: antihistamines; AAD: antiarrhythmic drug; AC: other anticholinergic medications; BDZ: benzodiazepines

		LI	**SSRI**	**SNRI**	**TCA**	**MAOI**	**OAD**	**AP1**	**AP2**
t_0	ME	0.64 (0.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.21 (0.43)	1 (0)
	MDE	1 (0)	0.22 (0.44)	0.33 (0.5)	0.11 (0.33)	0 (0)	0.22 (0.44)	0.11 (0.33)	0.67 (0.5)
t_1	ME	0.85 (0.36)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.21 (0.43)	1 (0)
	MDE	0.88 (0.33)	0.44 (0.52)	0.11 (0.33)	0 (0)	0 (0)	0.11 (0.33)	0.11 (0.33)	0.78 (0.44)
t_2	ME	0.85 (0.36)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.14 (0.36)	0.86 (0.36)
	MDE	0.88 (0.33)	0.33 (0.11)	0.11 (0.33)	0 (0)	0 (0)	0.44 (0.22)	0 (0)	0.67 (0.5)
t_3	ME	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.25 (0.5)	0.5 (0.57)
	MDE	0.5 (0.71)	0 (0)	0 (0)	0 (0)	0 (0)	0.5 (0.71)	0 (0)	0.5 (0.71)

		AED	**β-BLOCKER**	**OPIOD**	**AMP**	**AH**	**AAD**	**AC**	**BDZ**
t_0	ME	0.36 (0.50)	0 (0)	0 (0)	0 (0)	0 (0)	0	0.21 (0.42)	0.43 (0.51)
	MDE	0.44 (0.52)	0.11 (0.33)	0 (0)	0 (0)	0 (0)	0 (0)	0.11 (0.33)	0.44 (0.52)
t_1	ME	0.36 (0.50)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.07 (0.27)	0.64 (0.50)
	MDE	0.55 (0.53)	0.11 (0.33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.55 (0.52)
t_2	ME	0.21 (0.43)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.07 (0.27)	0.43 (0.51)
	MDE	0. 55 (0.53)	0.11 (0.33)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.44 (0.53)
t_3	ME	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.5 (0.57)
	MDE	0.5 (0.71)	0.5 (0.71)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Supplementary Table 3: **Clinical-demographic features of patients on an acute manic or depressive episodes excluded from analysis as not providing a minimum of three samples** "Medications #" refers to the number of drugs recorded in our cohort with a known influence on HRV which subjects were taking at the moment of study admittance. The figures herewith shown refer to the first assessment (acute episode onset), when patients were surveyed twice.

	AGE	FEMALES	MEDICATIONS #	BASELINE SYMPTOMS' SEVERITY
	MEAN (STD)	N (PERCENTAGE)	MEAN (STD)	MEAN (STD)
MANIA	40.30 (14.87)	17 (73.91%)	3.04 (1.08)	YMRS
N=23				19.61 (8.27)
DEPRESSION	52.38 (11.34)	15 (71.42%)	3.43 (1.56)	HDRS
N=21				18.85 (5.42)

Chapter 5

Aligning mood states detection to psychiatry *modus operandi*

5.1 Inferring Mood States with Wearables and AI

In parallel to advancing knowledge into how physiological data are affected by MDs (Chapter 4), the advent of personal sensing promises to usher in a new clinical paradigm enabling early detection and early interventions. Researchers have therefore been investigating whether abnormal mood states can be inferred from physiological data with the use of AI. The majority of studies [124, 29, 96, 94, 99, 95, 98] cast personal sensing for MDs as a problem of time-series classification (Section 2.3.1) outputting a single scalar, i.e. the total score on a psychometric scale, such as the HDRS (Appendix A), or the disease state, for example disease exacerbation vs remission.

A minority of studies used anomaly detection (Section 2.3.2) for detecting acute episodes [105, 117, 118], viewed as departures from a patient's baseline mood, sleep patterns, and energy levels. Anomaly detection systems can be trained in an unsupervised fashion. This is an advantage over supervised learning as labelling in psychiatry (e.g. item scores on a psychometric scale) is resource-intensive, subject to limited inter-rater agreement [177], and imprecise regarding specific recording segments. Additionally, acute episodes can manifest differently across and within patients over time, leading to diverse data patterns. Anomaly detection might be more adaptable to these variations and less affected by data imbalance, as acute episodes represent a minority of a patient's life. However, anomaly detection requires long observation periods, which are difficult

to acquire, and still depend on human labelling for results validation.

5.2 A Clinically Meaningful Label

Overall, most studies look like exercises in AI as they do not try to align the AI systems' outputs with clinical psychiatry *modus operandi*. Supervised-learning works [124, 29, 96, 94, 99, 95, 98] predicting a single scalar have only limited clinical actionability, as the same label may underlie different symptom combinations, with different treatment needs. For example, an MDE with marked agitation may benefit from different medications than one with prominent retardation. Secondly, knowledge of what symptoms support a given single-label diagnosis would help build trust in otherwise opaque systems.

Similar considerations apply to studies using anomaly detection [105, 117, 118]. First, unlike what has been done in previous studies, anomalies should be detected online, i.e. as new data comes in, the clinical translatability is otherwise very limited. Knowing in hindsight that an episode might have occurred does put patients or clinicians in a position where better clinical outcomes can be attained. Secondly, for an anomaly detection system to be clinically actionable, and thus usable, it should provide explanations of its output, in a way that is clinically accessible. Anomaly detection in personal sensing requires long recording sessions, unavailable in TIMEBASE/INTREPIBD, where an anomaly (i.e. an acute episode) is seen within the context of "normality" (Section 7.1.1). I will herewith focus on supervised approaches.

5.3 Time-Series Classification with Machine Learning

In what follows, we propose a clinically meaningful supervised-learning task in personal sensing in MDs. This is in the context of time-series classification, which we herewith briefly review with a focus on modern neural architectures. A univariate time-series $\mathbf{x} = [x_1, x_2, \ldots, x_T]$ is an ordered set of real values of length T and can thus be represented with a vector $\mathbf{x} \in \mathbb{R}^T$. A multi-variate time-series $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D]$ consists of D univariate time-series of length T, with $d = 1, 2, \ldots, D$ indexing the features in the time series, and can therefore be represented with a matrix $\mathbf{X} \in \mathbb{R}^{T \times D}$. Multi-variate time series are common in personal sensing as wearables come equipped with various sensors, corresponding to the different features (channels or modalities) in the multi-

variate time series. Note however that, as wearable sensors may have different sampling frequencies, the $\mathbf{x}_d$ vectors in the multi-variate time-series may have different lengths T, while typically mapping to the same wall-time.

A dataset $D = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$ is a collection of $(\mathbf{X}_i, \mathbf{Y}_i)$ pairs, where $\mathbf{X}_i$ is a multivariate time-series (in the univariate case $\mathbf{X}_i \in \mathbb{R}^{T\times 1}$) and $\mathbf{Y}_i \in \mathbb{R}^{T\times C}$ is its corresponding one-hot label vector, i.e. at each time stamp t, $\mathbf{Y}_i$ takes value 1 at the index c corresponding to the correct class and 0 otherwise. In personal sensing, human labelling is not performed in a *point-wise* fashion, i.e. time-step by time-step, on the recording itself. On the other hand, a given label, e.g. a certain mental state or the score on a psychometric scale, acquired during an assessment, is posited to hold true for a given time interval, as suggested by domain knowledge.

The objective of time-series classification is to learn a function $f : \mathbf{X}_i \to \hat{\mathbf{Y}}_i$ where $\hat{\mathbf{Y}}_i$ is the predicted label. Some loss $\mathcal{L}(\mathbf{Y}_i, \hat{\mathbf{Y}}_i)$ measures the "compatibility" between the prediction $\hat{\mathbf{Y}}_i$ and the ground truth $\mathbf{Y}_i$. A widely-used option is the cross-entropy loss, which, given a vector of predicted probabilities over the c classes $\hat{\mathbf{y}}$ and the corresponding one-hot-encoding ground truth vector $\mathbf{y}$, is defined as:

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \mathbf{y}_c \log\left(\hat{\mathbf{y}}_c\right).$$

The majority of studies [124, 29, 96, 94, 99] in personal sensing used classical ML algorithms for the map $f : \mathbf{X}_i \mapsto \hat{\mathbf{Y}}_i$, while two studies [95, 98] experimented with DL. Regardless of the approach, segments obtained with a sliding window, rather than the whole time series itself, typically constitute the units of analysis inputted to ML models. Specifically, given a window length l and a step size s, the number of segments resulting from segmentation is equal to $\left\lfloor \frac{T-l}{s} \right\rfloor + 1$ where $\lfloor x \rfloor$ denotes the floor function, returning the greatest integer less than or equal to x. Segmentation is usually done as a form of data augmentation, i.e. to increase the number of data points for training ML models.

Capturing long-range dependencies, e.g. for a time-series sampled at 64 Hz stretching over weeks, is challenging. Segments are effectively treated as independent, identically distributed data points by most modern ML systems, including those used in personal sensing studies. Where appropriate, predictions over segments from the same recording, or parts thereof, can be aggregated with a majority vote. Classical ML models rely on hand-crafted features extracted from segments, i.e. human-engineered features, obtained

with pre-set functions, are derived from segments. On the other hand, DL models can operate on real-valued raw segments, whence they can adaptively and automatically extract features.

5.3.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are powerful ANN architectures, explicitly incorporating temporal dependencies [178]. They accomplish this by processing a time-series $\mathbf{X}$ one time step at a time, where $\mathbf{X} \in \mathbb{R}^{T \times D}$, T is the number of time steps and D is the number of features. While processing each time step, the RNN simultaneously considers information from previous time steps contained in the *hidden state*, which is updated continually as new steps are processed. In a standard RNN, the hidden state $\mathbf{h}$ at time step t is computed as follows:

$$\mathbf{h}_t = g\left(\mathbf{W}_{hh} \cdot \mathbf{h}_{t-1} + \mathbf{W}_{hx} \cdot \mathbf{x}_t + \mathbf{b}_h\right)$$

Here, $\mathbf{W}_{hh}$ is the weight matrix for the connections between the hidden states, $\mathbf{W}_{hx}$ is the weight matrix for the connections between input $\mathbf{x}_t$ (the t^{th} time step of $\mathbf{X}$) and hidden state, $\mathbf{b}_h$ is the bias, and g is an activation function, typically a non-linearity such as tanh or sigmoid. In classification tasks, the hidden state from the last time step T is generally used as a summary of the entire time series to determine the output label, for example inputting it into a Multilayer Perceptron (MLP) [179].

Some architectural improvements have been suggested to mitigate the vanishing (exploding) gradient problem, a limitation in standard RNNs where gradients become too small (large) during backpropagation, thus hindering the learning of long-term dependencies. The Long Short-Term Memory Network (LSTM) [180] is one such architecture that utilizes special gating mechanisms – including input, output, and forget gates – that control the flow of information through the network. This allows the network to maintain information over extended time lags while being less affected by the gradient instabilities. However, LSTMs still face challenges in handling very long sequences. Furthermore, processing information sequentially can lead to computational inefficiency, limiting parallelization during training.

5.3.2 Transformers

Transformers, a more recent architecture originally proposed for neural machine translation as an encoder-decoder sequence-to-sequence model [142], offer an alternative approach to modelling sequential data. Unlike RNNs, Transformers do not process data sequentially, therefore enabling parallelization. They rely on a mechanism called self-attention, which allows the model to attend to all parts – *tokens*, as borrowed from the NLP literature – of the input sequence simultaneously. This also enables Transformers to capture long-range dependencies more efficiently than RNNs.

Self-attention assigns a weight to each token, reflecting its importance for understanding the current token. These weights are used to create a context vector that summarizes the relevant information from the entire sequence. An analogy with the retrieval of a value from a database of key-value pairs, where the "similarity" between a query and the keys determines the retrieval, helps to build intuition and motivates the notation used in Transformers. With self-attention, rather than retrieving a single value for a given query, all values are returned but, crucially, each value is scaled by the computed similarity (using the scaled dot product) between the query and the keys.

Concretely, for a time-series $\mathbf{X} \in \mathbb{R}^{T \times D}$, T being the number of time steps and D being the number of features per time step, each token, i.e. each time-step t, is first transformed into a query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$ vectors using trainable weight matrices, in parallel, thus producing the following query, key, and value matrices:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^{V}$$

Where $\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, and $\mathbf{W}^{V}$ are the parameter matrices for queries, keys, and values, respectively. Self-attention is computed by taking a dot product of the query with all the keys, followed by a softmax operation to obtain the weights:

$$\text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V}$$

Here, d_k refers to the dimensionality of the key vectors, and the division by $\sqrt{d_k}$ is a scaling factor that helps in stabilizing gradients during training. The softmax function ensures that the weights sum to one, providing a probabilistic interpretation of relevance.

To enhance the ability of the Transformer to handle various aspects of the data simultaneously, multi-head self-attention is utilized. This approach involves splitting the query, key, and value matrices into multiple heads, and performing the self-attention process independently on each head. The outputs of these multiple attention processes are then concatenated and linearly transformed into the desired dimension:

$$\text{MultiHeadSelfAttenion} = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)\mathbf{W}^O$$

where each $\text{head}_i = \text{SelfAttention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$ and $\mathbf{W}^O$ is the output weight matrix that combines the contributions from all heads. Multi-head self-attention allows the Transformer model to simultaneously process information from different representational spaces, improving the model's ability to focus on different positions and features of the input sequence at once.

The output of self-attention is a set of context vectors, each weighted by its relevance to other tokens. This output is then typically passed through more layers in the network, including position-wise feedforward neural networks, normalization, and dropout layers, which jointly constitute a *Transformer block* [142]. The output of the final Transformer block can then be inputted into an MLP for classification tasks.

Transformers in time-series modelling originally used point-wise input tokens, in other words, they treated each time step of the input segment as a token. Recently, however, it has been shown [181] that aggregating time steps into subseries-level patches enhances the locality and captures comprehensive semantic information that is otherwise not available at point-level. This approach also reduces space and time complexity by a factor of the stride S; in other words, the number of tokens is reduced from the number of time steps T to the number of patches $P \approx T/S$. With multi-variate time-series, patches can be obtained from individual channels (*channel-independence*) [181] or all of them (*channel-mixing*) [182].

Transformers have nowadays become the *de facto* standard ANN architecture in DL across different domains, including time series. A recent work [183] however questioned the effectiveness of Transformer-based solutions in some time-series tasks (long-term time-series forecasting), showing this architecture is outperformed by simpler linear models.

5.4 The paper: Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number

Below, we present our work **Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number**, published in Translational Psychiatry. As its main contribution, this study proposes a new task, i.e. inferring all items in the YMRS and HDRS, among the most widely used psychometric scales for assessing symptoms of depression and mania respectively, as scored by a clinician. This task aligns with everyday psychiatric practice where the specialist, when recommending a given intervention, considers the specific features of a patient, including their symptom patterns, beyond a reductionist disease label. Secondly, we explore ML challenges associated with the new task. These include multi-task learning for multiple target variables, modelling ordinal data, learning subject-invariant representations to improve generalization, and learning with imbalanced classes.

Inferring individual symptoms, which can then be used to support a given diagnosis, works similarly to a concept bottleneck model [184]. Such a model first predicts a set of human-understandable concepts from the input data and then uses these concepts to make the final prediction. This structure allows for better interpretability because the predicted concepts can be scrutinized, making it easier to understand how the model arrived at its conclusion.

Translational Psychiatry www.nature.com/tp

ARTICLE OPEN

Check for updates

Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number

Filippo Corponi[1,8✉], Bryan M. Li[1,8], Gerard Anmella[2,3,4,5], Ariadna Mas[2,3,4,5], Isabella Pacchiarotti[2,3,4,5], Marc Valentí[2,3,4,5], Iria Grande[2,3,4,5], Antoni Benabarre[2,3,4,5], Marina Garriga[2,3,4,5], Eduard Vieta[2,3,4,5], Stephen M. Lawrie[6], Heather C. Whalley[6,7], Diego Hidalgo-Mazzei[2,3,4,5,9] and Antonio Vergari[1,9]

Mood disorders (MDs) are among the leading causes of disease burden worldwide. Limited specialized care availability remains a major bottleneck thus hindering pre-emptive interventions. MDs manifest with changes in mood, sleep, and motor activity, observable in ecological physiological recordings thanks to recent advances in wearable technology. Therefore, near-continuous and passive collection of physiological data from wearables in daily life, analyzable with machine learning (ML), could mitigate this problem, bringing MDs monitoring outside the clinician's office. Previous works predict a single label, either the disease state or a psychometric scale total score. However, clinical practice suggests that the same label may underlie different symptom profiles, requiring specific treatments. Here we bridge this gap by proposing a new task: inferring all items in HDRS and YMRS, the two most widely used standardized scales for assessing MDs symptoms, using physiological data from wearables. To that end, we develop a deep learning pipeline to score the symptoms of a large cohort of MD patients and show that agreement between predictions and assessments by an expert clinician is clinically significant (quadratic Cohen's κ and macro-average F1 score both of 0.609). While doing so, we investigate several solutions to the ML challenges associated with this task, including multi-task learning, class imbalance, ordinal target variables, and subject-invariant representations. Lastly, we illustrate the importance of testing on out-of-distribution samples.

Translational Psychiatry (2024)14:161 ; https://doi.org/10.1038/s41398-024-02876-1

INTRODUCTION

Mood disorders (MDs) are a group of diagnoses in the Diagnostic and Statistical Manual 5th edition [1] (DSM-5) classification system. They are a leading cause of disability worldwide [2] with an estimated total economic cost greater than USD 326.2 billion in the United States alone [3]. They encompass a variety of symptom combinations affecting mood, motor activity, sleep, and cognition and manifest in episodes categorized as major depressive episodes (MDEs), featuring feelings of sadness and loss of interest, or, at the opposite extreme, (hypo)manic episodes (MEs), with increased activity and self-esteem, reduced need for sleep, expansive mood and behavior. As per the DSM-5 nosography, MDEs straddle two nosographic constructs, i.e., Major Depressive Disorder (MDD) and Bipolar Disorder (BD), whereas MEs are the earmark of BD only [4].

Clinical trials in psychiatry to this day entirely rely on clinician-administered standardized questionnaires for assessing symptoms' severity and, accordingly, setting outcome criteria. With reference to MDs, Hamilton Depression Rating Scale-17 [5] (HDRS) and Young Mania Rating Scale [6] (YMRS) are among the most widely used scales to assess depressive and manic symptoms [7], quantifying behavioral patterns such as disturbances in mood, sleep, and anomalous motor activity. The low availability of specialized care for MDs, with rising demand straining current capacity [8], is a major barrier to this classical approach to symptom monitoring. This results in long waits for appointments and reduced scope for pre-emptive interventions. Current advances in machine learning (ML) [9] and the widespread adoption of increasingly miniaturized and powerful wearable devices offer the opportunity for personal sensing, which could help mitigate the above problems [10]. This can involve a near-continuous and passive collection of data from sensors, with the aim of identifying digital biomarkers associated with mental health symptoms at the individual level, therefore backing up clinical evaluation with objective and measurable physiological data. Personal sensing holds great potential for being translated

[1]School of Informatics, University of Edinburgh, Edinburgh, UK. [2]Bipolar and Depressive Disorders Unit, Department of Psychiatry and Psychology, Hospital Clínic de Barcelona, c. Villarroel, 170, 08036 Barcelona, Spain. [3]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), c. Villarroel, 170, 08036 Barcelona, Spain. [4]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain. [5]Departament de Medicina, Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona (UB), c. Casanova, 143, 08036 Barcelona, Spain. [6]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. [7]Generation Scotland, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [8]These authors contributed equally: Filippo Corponi, Bryan M. Li. [9]These authors jointly supervised this work: Diego Hidalgo-Mazzei, Antonio Vergari. ✉email:

Received: 7 July 2023 Revised: 9 March 2024 Accepted: 13 March 2024
Published online: 26 March 2024

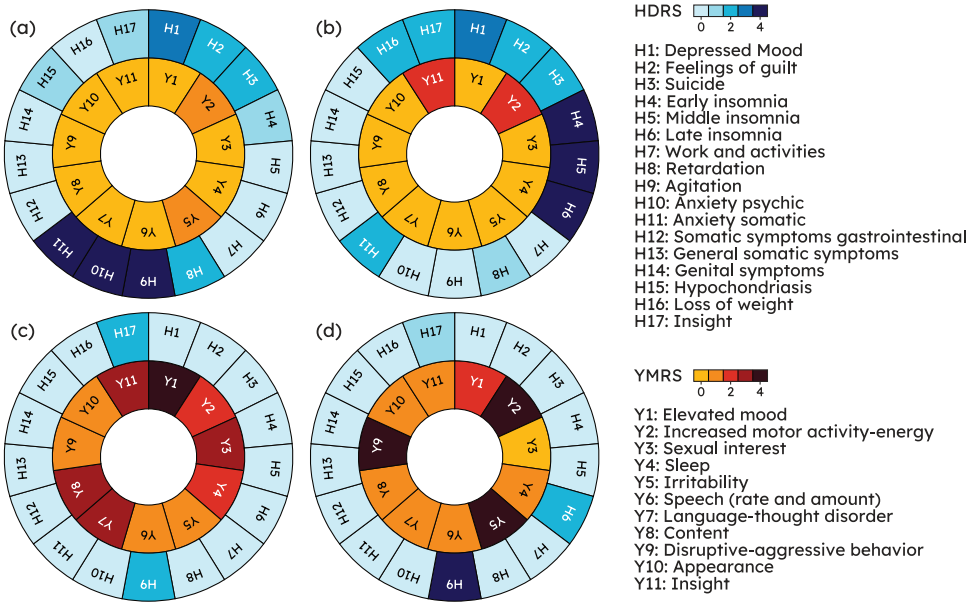


Fig. 1 The same severity level can be realized from different symptom combinations, underlying different treatment needs. Top row: a pair of patients with Major Depressive Disorder on a Major Depressive episode; while both share the same severity levels, total Hamilton Depression Rating Scale (HDRS) ≥ 23 [33]. Patient (**a**), with total HDRS = 24, exhibits high levels of anxiety (H9, H10, H11), whereas patient (**b**), with total HDRS = 26, displays a marked insomnia component (H4, H5, H6). Bottom row: a pair of patients with Bipolar Disorder on a Manic Episode with a total Young Mania Rating Scale (YMRS) ≥ 25. Patient (**c**), with total YMRS = 30, has an irritable/aggressive profile (Y2, Y5, Y9) whereas patient (**d**), with total YMRS = 30, has a prominently elated/expansive presentation (Y1, Y3, Y7, Y11). Knowing what specific symptoms underlie a given state may allow clinicians to tailor treatment accordingly: e.g., a molecule with a stronger anxiolytic profile such as paroxetine or a short course of a benzodiazepine as an antidepressant is introduced may be appropriate in patient (**a**) whereas patient (**b**) might benefit from a compound with marked hypnotic properties such as mirtazapine.

into clinical decision support systems [11] for the detection and monitoring of MDs. Specifically, it could be particularly appealing to automate the prediction of the items of the HDRS and YMRS scales as they correlate with changes in physiological parameters, conveniently measurable with wearable sensors [12–14].

However, so far, the typical approach has been to reduce MDs detection to the prediction of a single label, either the disease state or a psychometric scale total score [15, 16], which risks over-simplifying a much more complex clinical picture. Figure 1 illustrates this issue: patients with different symptoms and thus (potentially very) different scores on individual HDRS and YMRS items are "binned together" in the same category, leading to a loss of actionable clinical information. Predicting all items in these scales can instead align with everyday psychiatric practice where the specialist, when recommending a given intervention, considers the specific features of a patient, including their symptom patterns, beyond a reductionist disease label [17, 18]. Figure 1 illustrates a case in point where knowledge of the full symptom profile might enable bespoke treatment: on the face of it, patient (a) and (b) (top row) share the same diagnosis, i.e., MDE in the context of MDD; however, considering their specific symptom profile patient (a) might benefit from a molecule with stronger anxiolytic properties whereas patient (b) might require a compound with hypnotic properties. Furthermore, an item-wise analysis can lead to the identification of drug symptom specificity in clinical trials [19, 20].

Table 1 summarizes previous works in personal sensing for MDs and shows that all previous tasks collapsed the complexity of MDs to a single number. Côté-Allard et al. [21] explored a binary classification task, that is distinguishing subjects with BD on an ME from different subjects with BD recruited outside of a disease episode, when stable. The study experimented with different subsets of pre-designed features from wristband data and proposed a pipeline leveraging features extracted from both short and long segments taken within 20-hour sequences.

Pedrelli et al. [22], expanding on Ghandeharioun et al. [23], used pre-designed features from a wristband and a smartphone to infer HDRS residualized total score (that is total score at time t minus baseline total score) with traditional ML models. Tawaza et al. [24] employed gradient boosting with pre-designed features from wristband data and pursued case-control detection in MDD and, secondarily, HDRS total score prediction. Similarly, Jacobson et al. [25] predicted case-control status in MDD from actigraphy features with gradient boosting. Nguyen et al. [26] used a sample including patients with either schizophrenia (SCZ) or MDD wearing an actigraphic device and explored case-control detection where SCZ and MDD were either considered jointly (binary classification) or as separate classes (multi-class classification). Of notice, this was the first work to apply artificial neural networks (ANNs) directly on minimally processed data, showing that they outperformed traditional ML models. Lastly, the multi-center study of Lee et al. [27] investigated mood episode prediction with a random forest and pre-designed features from wearable and smartphone data. Further to proposing a new task, our work stands out for a sample size larger than all previous works by over 2 dozen patients, with the exception of a multi-center study by Lee et al. [27], where, however, clinical evaluation was carried out retrospectively, thereby inflating chances of recall bias [28] and missing out on the real-time clinical characterization of the acute phase. Indeed, collecting data from patients on an acute episode, using specialist assessments and research-grade wearables, is a challenging and expensive enterprise. Relatively to previous endeavors, the contribution of this work is two-fold: (1) Taking one step beyond the prediction of a single label, which misses actionable clinical information, we propose a new task in the context of MDs monitoring with physiological data from wearables: inferring all items in HDRS (17 items) and YMRS (11 items), as scored by a clinician, which enables a fine-grained appreciation of patients' psychopathology therefore creating opportunity for tailored

Table 1. This work is the first in personal sensing for MDs attempting to infer the full symptom profile, providing actionable clinical information beyond a single reductionist label, and it also stands out for the relatively large sample size (the largest among studies where MD acute phase clinical evaluation was not retrospective).

	Device(s)	Num. Patients	Patients Features	Task
This work	Empatica E4	75	MDD, BD; $M_{age} = 44.16$ $SD_{age} = 14.42$ $F_{\%} = 56$	HDRS and YMRS items multi-task regression
Côté-Allard et al. [21]	Empatica E4	47	BD; $M_{age} = 44$ $SD_{age} = 15$ $F_{\%} = 67.24$	Mania vs Euthymia binary classification
Ghandeharioun et al. [23]	Empatica E4 and Android Phone	12	MDD; $M_{age} = 37$ $SD_{age} = 17$ $F_{\%} = 75$	HDRS total score regression
Pedrelli et al. [22]	Empatica E4 and Smartphone	31	MDD; $M_{age} = 33.7$ $SD_{age} = 14$ $F_{\%} = 74$	HDRS total score regression
Jacobson et al. [25]	Actiwatch	23	MDD; $M_{age} = 48.2$ $SD_{age} = 11.0$ $F_{\%} = 43$	Depression detection binary classification
Tazawa et al. [24]	Silmee W20	45	MDD, BD; $M_{age} = 52.1$ $SD_{age} = 13.2$ $F_{\%} = 46.7$	Depression detection binary classification
Nguyen et al. [26]	Actiwatch	45	MDD, SCZ; $M_{age} = 44.70$ $SD_{age} = 11$ $F_{\%} = 73.33$	Disease detection binary/multi-class classification
Lee et al. [27]	Fitbit Charge Hr 2 or 3 and Smartphone	270	MDD, BD; $M_{age} = 23.3$ $SD_{age} = 3.63$ $F_{\%} = 54.5$	Mood episode prediction binary classification

Previous studies recruiting patients with either a DSM or an International Classification of Diseases (ICD) MD diagnosis and using passively collected wearable data are reported. $F_{\%}$: Percent Females; M_{age}: mean age; SD_{age}: standard deviation age.

treatment (Fig. 1). (2) We investigate some of the methodological *challenges* associated with the task at hand and explore possible ML solutions. **c1**: inferring multiple target variables (28 items from two psychometric scales), i.e., multi-task learning (MTL, see Section 3.4.1). **c2**: modeling ordinal data, such are HDRS and YMRS items (see Section 3.4.1). **c3**: learning subject-invariant representations, since, especially with noisy data and sample size in the order of dozens, models tend to exploit subject-idiosyncratic features rather than learning disease-specific features shared across subjects, leading to poor generalization [29] (see Section 3.4.2). **c4**: learning with imbalanced classes, as patients on an acute episode usually receive intensive treatment and acute states therefore tend to be relatively short periods in the overall disease course [30, 31] thereby tilting items towards lower ranks.

METHODS

Data collection and cohort statistics

The following analyses are based on an original dataset, TIMEBASE/INTREPIBD, being collected as part of a prospective, exploratory, observational, single-center, longitudinal study with a fully pragmatic design embedded into current real-world clinical practice. A detailed description of the cohort is provided in Anmella et al. [32]. In brief, subjects with a DSM-5 MD diagnosis (either MDD or BD) were eligible for enrollment. Those recruited on an acute episode had up to four assessments: **T0** acute phase (upon hospital admission or at the home treatment unit), **T1** response onset (50% reduction in total HDRS/YMRS), **T2** remission (total HDRS/YMRS ≤7), and **T3** recovery (total HDRS/YMRS continuously ≤7 for a period of ≥8 weeks) [33]. On the other hand, subjects with a historical diagnosis but clinically stable at the moment of study inclusion (euthymia, Eu) were interviewed only once. At the start of each assessment, a clinician collected clinical demographics, including HDRS and YMRS, and provided an Empatica E4 wristband [34] which participants were required to wear on their non-dominant wrist until the battery ran out (~48 h). A total of 75 subjects, amounting to a total of 149 recording sessions (i.e., over 7000 h), were available at the time of conducting this study. An overview of the cohort clinical-demographic characteristics is given in Table 2 and the number of recordings available per observation time (T0 to T3) by diagnosis is given in Supplementary Figure (SF) 1; observation times (T0 to T3) merely reflect how the data collection campaign was conducted and were not used (or implicitly assumed) as labels for any of the analysis herewith presented. Given the naturalistic study design, medications were prescribed as part of the regular clinical practice: subjects on at least one antidepressant, lithium, an anticonvulsant, or at least one antipsychotic were respectively 37.83%, 70.94%, 34.45%, 12.16% of the cohort. The median (interquartile range) time since disease onset was 6 (14) years.

Table 2. Clinical-demographic characteristics of the study population (N = 75).

	MEAN (SD)	MEDIAN (IQR)
AGE	44.66 (14.42)	45.00 (24.50)
HDRS (TOTAL)	7.27 (6.94)	4.00 (6.00)
YMRS (TOTAL)	7.21 (8.75)	3.00 (10.00)
NUMBER OF SUBJECTS (%)		
SEX	male: 33 (44) female: 42 (56)	
MOOD STATE	**MDE-MDD**: 9 (12) **EU-MDD**: 3 (4) **MDE-BD**: 12 (16) **ME**: 28 (37) **MX**: 7 (9) **EU-BD**: 16 (21)	
ASSESSMENT(S)	**1**: 75 (100) **2**: 44 (59) **3**: 22 (29) **4**: 8 (11)	

According to the DSM-5, an MD can be categorized as either a major depressive episode or a manic episode. As a bridge between these two, the DSM-5 admits a mixed symptoms specifier (MX) to cases where symptoms from both polarities are present.
EU-BD euthymia in bipolar disorder, *EU-MDD* euthymia in major depressive disorder, *HDRS* Hamilton Depression Rating Scale, *IQR* inter-quartile range, *MDE-BD* major depressive episode in bipolar disorder, *MDE-MDD* major depressive episode in major depressive disorder, *ME* manic episode, *MX* mixed symptoms episode, *SD* standard deviation, *YMRS* Young Mania Rating Scale.

The E4 records the following sensor modalities (we report their acronyms and sampling rates in parentheses): 3D acceleration (ACC, 32 Hz), blood volume pressure (BVP, 64 Hz), electrodermal activity (EDA, 4 Hz), heart rate (HR, 1 Hz), inter-beat interval (IBI, i.e., the time between

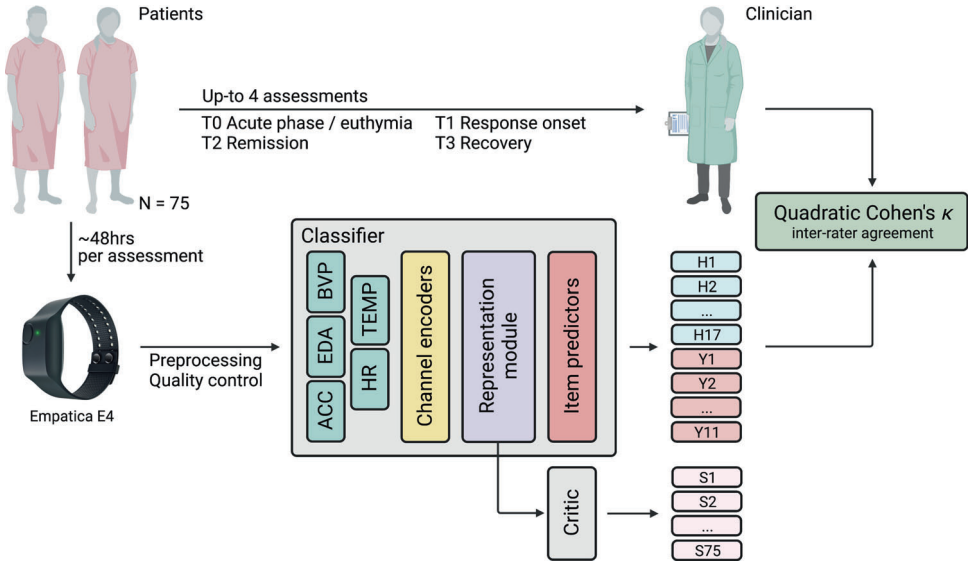


Fig. 2 Analysis workflow. Patients had up to four assessments. At the start of each assessment, a clinician scored the patient on the Hamilton Depression Rating Scale (H in the figure) and Young Mania Rating Scale (Y) and provided an Empatica E4 device asking the patient to wear it for ~48 h (i.e., average E4 battery life). An Artificial Neural Network (ANN) model is fed with recording segments and is tasked with recovering clinician scores. The quadratic Cohen's κ measures the degree to which the machine scores are in agreement with those of the clinician. The ANN model is made of Classifier (CF) and Critic (CR). The former comprises three main modules: (1) Encoder (EN), projecting input sensory channels onto a new space where all channels share the same dimensionality, regardless of the native E4 sampling frequency; (2) Representation Module (RM), extracting a representation h that is shared across all items; and (3) one Item Predictor IP_j for each item. CR is tasked with telling subjects (S in the figure) apart using *h* and is pitted in an adversarial game against RM(EN(·)), designed to encourage the latter to extract subject-invariant representations.

two consecutive heart ventricular contractions) and skin temperature (TEMP, 1 Hz). IBI was not considered due to extensive sequences of missing values across all recordings. This is likely due to high sensitivity to motion and motion artifacts, as observed previously [35].

Data pre-processing

An E4 recording session comes as a collection of 1D arrays of sensory modalities. We quality-controlled data to remove physiologically implausible values with the rules by Kleckner et al. [36] and the addition of a rule to remove HR values that exceeded the physiologically plausible range (25–250 bpm). The median percentage of data per recording session discarded from further analyses because of the rules above was 8.05 (range 1.95–32.10). Each quality-controlled recording session was then segmented using a sliding window, whose length (τ, in wall-time seconds) is a hyperparameter, enforcing no overlap between bordering segments (to prevent models from exploiting overlapping motifs between segments). These segments (x_i) and the corresponding 28 clinician-scored HDRS/YMRS items (y_i) from the subjects wearing the E4 formed our dataset, $\{(x_i, y_i)\}_{i=1}^{N}$. Note that all segments coming from a given recording session share the same labels, i.e., the HDRS/YMRS scores of the subject wearing the E4. HDRS/YMRS items map symptoms spanning mood, sleep, and psycho-motor activity. Some likely fluctuate over a 48-h session, especially in an ecological setting where treatments can be administered (e.g., Y9 disruptive-aggressive behavior may be sensitive to sedative drugs). To limit this, we isolated segments from the first five hours (*close-to-interview* samples) and used them for the main analysis, splitting them into train, validation, and test sets with a ratio of 70-15-15. Then, to study the effect of distribution shift, we tested the trained model on samples from each 30-min interval following the first five hours of each recording (*far-from-interview* samples). It should be noted that further to a shift in the target variables, a shift in the distribution of physiological data collected with the wearable device is to be expected [37], owing to different patterns of activity during the day, circadian cycles, and administered drugs. Details on the number of recording segments in train, validation, and test splits are given in Supplementary Table (ST) 1.

Evaluation metrics

HDRS and YMRS items are ordinal variables. For instance, *H11 anxiety somatic* has ranks 0-Absent, 1-Mild, 2-Moderate, 3-Severe, or 4-Incapacitating. The item distribution (see SF2) was imbalanced towards low scores due to patients on an acute episode usually receiving intensive treatment such that acute states tend to be relatively short-lived periods in the overall disease course [30, 31]. This can be quantified with the ratio between the cardinality of the majority rank and that of the minority rank ρ: e.g., say there are 100, 90, 50, 30, and 10 recording segments with an *H11* rank of respectively 0, 1, 2, 3, and 4, then ρ is 100/10 = 10 as 100 is the cardinality of the *H11* rank (0) with the highest number of segments and 10 is the cardinality of the *H11* rank (4) with the lowest number of segments. Metrics accounting for class imbalance should be used when evaluating a classification system in such a setting to penalize trivial solutions, e.g., systems always predicting the majority class in the training set regardless of the input features. We used Cohen's κ, in particular its quadratic version (QCK), since, further to its suitability to imbalanced ordinal data, it is familiar and easily interpretable to clinicians and psychometrists [38–41]. It expresses the degree to which the ANN learned to score segments in agreement with the clinician's assessments. This is similar to psychiatric clinical trials where prospective raters are trained to align with assessments made by an established specialist [42]. Cohen's κ takes values in $[-1,1]$, where 1 (−1) means perfect (dis)agreement. In a psychiatric context, 0.40–0.59 is considered a good range while 0.60–0.79 is a very good range [43]. Cohen's κ compares the observed agreement between raters to the agreement expected by chance taking into account the class distributions; the quadratic weightage in QCK penalizes disagreements proportionally to their squared distance. As individual HDRS/YMRS items have different distributions (see SF2), we checked whether item level performance was affected by sample Shannon entropy ($\mathcal{H}$). To this end, we computed a simple Pearson correlation coefficient (R) between item QCK and $\mathcal{H}$.

Model design

The task at hand is supervised, specifically, we sought to learn a function mapping recording segments to their HDRS and YMRS scores: $f : x_i \mapsto \hat{y}_i$. The model we developed to parametrize f comprised two independent sub-models (Fig. 2): (a) a **classifier** (*CF*), which learns to predict the HDRS/YMRS scores from patients' physiological data, and (b) a **patient critic** (*CR*), which penalize *CF* for learning subject-specific features (i.e., memorize the patient and their scores), rather than features related to the underlying disorder shared across patients. Both CF and CR are simply compositions of mathematical functions, that is layers of the neural network. The CF

module itself consisted of three sequential modules (or, equivalently, functions): (a.1) a *channel encoder* (EN) for projecting sensory modalities onto the same dimensionality regardless of the modality's native sampling rate so that they could be conveniently concatenated, (a.2) a *representation module* (RM) for extracting features, and lastly, to address **(c1)** multi-task learning, (a.3) 28 parallel (one for each item) *item predictors* (IP), each learning the probability distribution over item ranks conditional on the features extracted with RM. The critic module CR, instead, uses the representation from RM for telling subjects apart. CR competes in an adversarial game against EN and RM, designed to encourage subject-invariant representations. Details on the model's architecture, the mathematical form of CF and CR, and the model's loss are given in "Supplementary Methods – Model architecture and loss functions".

Learning from imbalanced data

We adapted to our use case the following three popular imbalance learning approaches. (i) Focal loss [44]: the categorical cross-entropy (CCE) loss from the item predictor IP_j was multiplied during training by a scaling factor correcting for rank frequency (such that under-represented ranks have a similar weight on the loss as over-represented ones) while at the same item focusing on instances where the model assigns a high probability to the wrong rank (these are instances the model is very confident about but its confidence it misplaced as it is outputting the wrong rank). (ii) Probability thresholding [45]: during inference, probabilistic predictions for each rank under the jth item were divided by the corresponding rank frequency (computed on the training set), plus a small term to avoid division by zero in case of zero frequency ranks. The new values were then normalized by the total sum. (iii) Re-sampling and loss re-weighting: HDRS/YMRS severity bins (defined in [33]) were used to derive a label which was then used to either random under-sampling (RUS) or random over-sampling (ROS) segments with, respectively, over-represented and under-represented labels. The loss of x_i was then re-scaled proportionally to the re-sampling ratio of its class.

Hyperparameter tuning

In order to find the hyperparameters that yield the best QCK in the validation set, we performed an exhaustive search using Hyperband Bayesian optimization [46]. ST2 shows the hyperparameters search space and the configuration of the best model after 300 iterations. We also computed which hyperparameters were the best predictors of the validation QCK. This was obtained by training a random forest with the hyperparameters as inputs and the metric as the target output and deriving the feature importance values for the random forest. Details on model training are given in "Supplementary Methods – Model training".

Baseline model using classical machine learning

Most previous works into personal sensing for MDs (as discussed in the Introduction) did not use deep learning for automatically learning features from minimally processed data but deployed classical ML models relying on hand-crafted features. Thus, we developed a baseline in the same spirit, in order to better contextualize our deep-learning pipeline performance on *close-to-interview* samples. Namely, from the same recording segments inputted to the ANN we extracted features (e.g., heart rate variability, entropy of movement) with a commonly used feature extractor for Empatica E4, named FLIRT [47], and developed random forest classifiers (28 in total, as many as there are HDRS and YMRS items), using random oversampling to handle class. We opted for random forest since it was a popular choice in previous relevant works [22, 27]. The hyperparameter space was explored with a random search of 300 iterations for each classifier. Details are given in ST3.

Prediction error examination

Towards gaining insights into the best-performing setting among those explored in the experiments detailed above we computed residuals on *close-to-interview* samples and illustrated their distribution across items. For the sake of better comparability, items with a rank step of two (e.g., *Y5 irritability*) were re-scaled to have a rank step of one like other items. Furthermore, towards investigating correlations between residuals, checking for any remarkable pattern in view of the natural correlation structure of HDRS and YMRS, we estimated a regularized partial correlation network, in particular a Gaussian graphical lasso (glasso [48]), over item residuals ("Supplementary Methods – Gaussian Graphical Lasso" for details). Lastly, towards having a subject-level perspective, we computed the item-average macro-averaged F1 score ($F1^M$) for each subject, checked for any pattern of cross-subjects variability in subject performance, and checked for association with available clinical-demographic variables (age, sex, HDRS/YMRS total score) using Pearson's R and independent samples t-test with Bonferroni correction.

Channels importance

In order to assess each sensory modality contribution to the HDRS-YMRS items prediction, we took a simple, model-agnostic approach to assess each individual channel contribution to the task at hand. That is to say, we selected the system performing best on the task and re-trained it including all channels (tri-axial ACC, EDA, BVP, HR, and TEMP) but one. For each left-out channel, we measured the difference in performance across items relative to the baseline model (the one trained on all channels).

RESULTS

Best model details – ANN

The loss type is the hyperparameter most predictive of validation QCK (ST4). The selected model employs the Cohen's κ loss with quadratic weightage [39] (**c2**). The best model uses a (small) critic penalty ($\lambda = 0.07$) added to the main objective, i.e., scoring HDRS/YMRS (**c3**). However, the training curve shows that the reduction in the multi-task loss (each item prediction can be thought of as a task) across epochs is paralleled by the reduction in the loss (cross-entropy) paid by CR, tasked with telling subjects apart. Resampling and loss re-weighting (**c4**) is the preferred strategy for class imbalance. We found that a segment length of 16 s yields the best result. The difference in QCK (Δ_{QCK}) for other choices of τ (in seconds) relative to the best configuration is −0.092 (8 s), −0.100 (32 s), −0.191 (64 s), −0.246 (128 s), 0.355 (256 s), −0.4431 (512 s), −0.577 (1024 s). Note that τ was explored among powers of 2 for computational convenience and that, when segmenting the first 5 hours of each recording, different τ values produced different sample numbers and lengths (the lower the τ values, the higher the number of samples, the shorter the sample). The predictive value of hyperparameter τ towards validation QCK is fairly low relative to other hyperparameters.

Main results

Our best ANN model achieves an average QCK across HDRS and YMRS items of 0.609 in *close-to-interview* samples, a value that can be semi-qualitatively interpreted as moderate agreement [49], confidently outperforming our baseline random forest model that only reached an average QCK of 0.214. Item level QCK correlates weakly ($R = 0.08$) with the degree of item class imbalance (ρ) but fairly ($R = 0.42$) with item $\mathcal{H}$. Table 3a shows QCK for each item in HDRS and YMRS. Briefly, QCK is highest for *H12 somatic symptoms gastrointestinal* (0.775) and lowest for *H10 anxiety psychic* (0.492). *H10* has also the highest $\mathcal{H}$ (1.370), however, *H7 work and activities*, despite having the second highest $\mathcal{H}$ (1.213), has a QCK of 0.629, ranking as the 9th best-predicted item.

Shift over time

When tested on *far-from-interview* samples, our system overall has a drop in performance (Table 3b and SF3). The average QCK is 0.498, 0.303, and 0.182 on segments taken respectively from the first, second, and third thirty-minute intervals. Thereafter, it fluctuates through the following thirty-minute intervals with 0.061 as the lowest value 15 h into the recording. The items with the biggest drop in QCK relative to their baseline value across the first three 30-min intervals are *H9 agitation*, *H10 anxiety somatic*, *Y4 sleep*, and *Y9 disruptive-aggressive behavior*. On the other hand, items that retain their original QCK value the most in the first three 30-min intervals are *H1 depressed mood*, *Y11 insight*, *H2 feelings of guilt*, and *H17 insight*. This pattern matches clinical intuition as items in the former group may be more volatile and reactive to environmental factors (including medications), whereas items in the latter group tend to change more slowly.

Table 3. (a) Quadratic Cohen's κ ranges from 0.775 on "somatic symptoms gastrointestinal" and to 0.492 on "anxiety psychic" (mean of 0.609). (b) Quadratic Cohen's κ deteriorated across both Hamilton Depression Rating Scale (HDRS) and Young Mania Rating Scale (YMRS) on segments taken further away from when the interview took place.

(a)

Item QCK	H1 0.642	H2 0.624	H3 0.694	H4 0.534	H5 0.595	H6 0.512	H7 0.629	H8 0.604	H9 0.508	H10 0.492
	H11 0.636	H12 0.775	H13 0.582	H14 0.594	H15 0.691	H16 0.637	H17 0.574	Y1 0.602	Y2 0.590	Y3 0.627
	Y4 0.629	Y5 0.591	Y6 0.572	Y7 0.582	Y8 0.588	Y9 0.755	Y10 0.602	Y11 0.566		

(b)

Item-average QCK	5:01–6:00	6:01–6:30	6:31–7:00	7:01–7:30	7:31–8:00	8:01–8:30	8:31–9:00
HDRS	0.483	0.301	0.183	0.178	0.180	0.177	0.178
YMRS	0.499	0.307	0.182	0.181	0.181	0.173	0.175

Notes for (a): Item level QCK across HDRS and YMRS items. See Supplementary Table 5 for macro-averaged F1 scores.
Notes for (b): Item average QCK is herewith shown,see Supplementary Fig. 3 for a zoom on individual items across all available 30-min intervals.

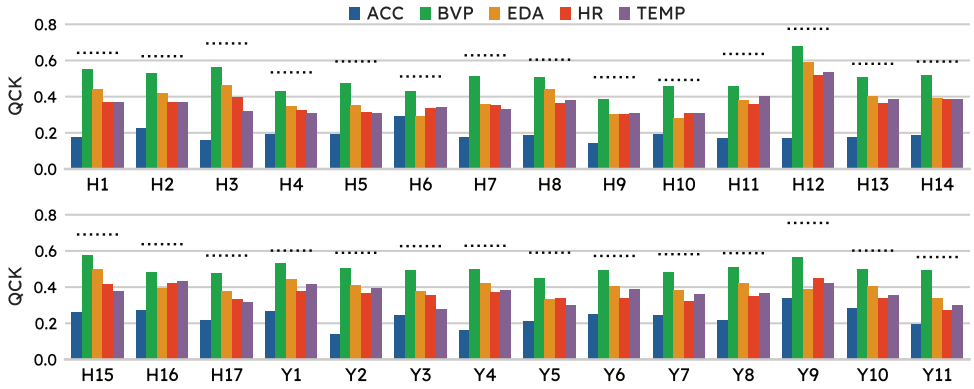


Fig. 3 All physiological modalities contributed to the test performance across items, however, this was particularly pronounced for Acceleration (ACC) and relatively modest for Blood Volume Pressure (BVP). Effect of dropping individual channels on item performance. The dotted line is at the level of baseline model performance while each bar indicates the performance upon re-training the best model including all channels but the one corresponding to the bar color code, as shown in the legend.

Post-hoc diagnostics

In order to gain further insights into the errors that our system made on *close-to-interview* holdout samples, we studied the distribution of residuals, i.e., the signed difference between prediction $\hat{y}$ and ground truth y. SF4 illustrates that the model is correct most of the time, residuals are in general evenly distributed around zero, and when wrong the model is most often off by just one item rank. Summing individual items' predictions, we could get predictions on HDRS/YMRS total sore (which is indeed simply the sum over the questionnaire items) which had a Root Mean Squared Error (RMSE) of 4.592 and 5.854, respectively.

Furthermore, we investigated the correlation structure among item residuals to check whether any meaningful pattern emerged. SF5 shows the undirected graphical model for the estimated probability distribution over HDRS and YMRS item residuals. The graph only has positive edges, that is, only positive partial correlations between item residuals and co-variates. HDRS and YMRS nodes tend to have weak interactions across the two scales, with the exception of nodes that map the same symptom, e.g., *Y11* and *H17* both query *insight*. Within each scale, partial correlations are stronger among nodes underlying a common symptom domain, e.g., *H1* and *H2* constitute "core symptoms of depression" [50], and speech (*Y6*) is highly related to mood (*Y1*) and thought (*Y7*, *Y8*) [51]. Average node predictability for HDRS and YMRS items, a measure of how well a node can be predicted by nodes it shares an edge with, akin to R^2, is 48.43%. Stability analyses showed that some edges are estimated reliably (i.e., they were included in all or nearly 500 bootstrapped samples), but there also is considerable variability in the edge parameters across the bootstrapped models. Subjects' item average F1 macro-averaged F1 score ($F1^M$) score had a mean value of 0.605 (std = 0.015) with no subjects standing out for a remarkable high (or low) performance. No associations with age, sex, HDRS/YMRS total score emerged (SF6).

Channels contribution

We were interested in whether physiological modalities contributed differently towards performance across items. This question, further to clinical interest, has also practical implications since other devices may not implement the same sensors as Empatica E4. Figure 3 shows that while all modalities seem to positively contribute to test performance across items, this is markedly the case with ACC as the model records the biggest drop in performance upon removal of this channel from input features. Specifically, upon zeroing out the contribution of ACC, the biggest deterioration in performance was observed for items mapping anxiety (e.g., *H11 anxiety somatic* $\Delta_{QCK} = -0.321$), *YMRS4 sleep*, and *YMRS9 disruptive- aggressive behavior* (with a Δ_{QCK} of −0.371 and −0.281 respectively), and core depression items (e.g., *H1 depressed mood* $\Delta_{QCK} = -0.276$). On the other hand, the contribution of BVP was relatively modest since, upon dropping this channel, items generally had only a marginal reduction in QCK.

DISCUSSION

In this work, we proposed a new treatment of MDs monitoring with personal sensing: inferring all 28 items from HDRS and YMRS, the most widely used clinician-administered scales for depression and mania respectively. Casting this problem as a single-label prediction, e.g., disease status or the total score on a psychometric scale, as done previously in the literature, dismisses the clinical complexity of MDs, thereby losing actionable clinical information, which is conversely preserved in the task we introduced here. Furthermore, the predicted total score on a psychometric scale can always be recovered if item-level predictions are available by simply summing them out, whereas the other direction, i.e., going from total score to individual item predictions, is not possible.

We developed and tested our framework with samples taken over five hours since the start of the clinical interview (*close-to-interview* samples), achieving moderate agreement [52] with expert clinician (average QCK of 0.609) on a holdout set and showing that our deep-learning pipeline vastly exceeded the performance (average QCK of 0.214) of traditional ML baseline relying on hand-crafted features. Item level performance showed a fair correlation with item $\mathcal{H}$, indicating that items with a higher "uncertainty" in their sample distribution tend to be harder to predict. The difference in $\mathcal{H}$ is partly inherent to the scale design, as different items admit a different number of ranks. HDRS/YMRS total scores, with the range of [0–52] and [0–60], were predicted with an RMSE of 4.592 and 5.854, respectively (note that item level error compounds across items when summing them out). A five and three-point interval are the smallest bin widths for YMRS and HDRS respectively [53, 54], e.g., a YMRS total score in the range of [20–25] is considered a mild mania and an HDRS total score in [19–22] is considered as severe depression. This shows that on average our model would be off by two score bands at most, in case of a true score falling on the edge of a tight severity bin (i.e., the ones reported above). We recommend caution in interpreting these results however as metrics suited for continuous target variables, unlike QCK and $F1^M$, are not robust in settings where the distribution is skewed (towards lower values in our case). Furthermore, while these results are comparable to previous ones (e.g., Ghandeharioun et al. [23] reported a RMSE of 4.5 on the HDRS total score), differences in the sample limit any direct comparison.

When used on samples collected from thirty-minute sequences following the first five hours of the recordings (*far-from-interview* samples), our model had a significantly lower performance with average QCK declining down to 0.182 in the third half-hour and then oscillating but never recovering to the original level. Consistently with clinical intuition, items suffering the sharpest decline relative to their baseline performance were those mapping symptoms that naturally have a higher degree of volatility (e.g., *H9 agitation*) while items corresponding to more stable symptoms (e.g., *H17 insight*) had a gentler drop in performance. Besides (some) symptoms plausibly changing over time, a shift in the physiological data distribution is very likely in a naturalist setting.

Residuals on holdout *close-to-interview* samples showed a symmetric distribution, centered around zero, thus the model was not systematically predicting either over- or under-predicting. The network of item residuals illustrated that our model erred along the correlation structure of the two symptom scales, as stronger connections were observed among items mapping the same symptom or a common domain. An ablation study over input channels showed that ACC was the most important modality, lending further support to the discriminative role of actigraphy with respect to different mood states [14]. Coherently, items whose QCK deteriorated the most upon removing this channel were those mapping symptom domains clinically observable through patterns of motor behavior.

In conclusion, we introduced a new task in personal sensing for MDs monitoring, overcoming limitations of previous endeavors which reduced MDs to a single number, with a loss of actionable clinical information. We indeed advocate for inferring symptoms' severity as scored by a clinician with the Hamilton Depression Rating Scale-17 [5] (HDRS) and the YMRS [6]. We developed a deep learning pipeline inputting physiological data recordings from a wearable device and outputting HDRS and YMRS scores in substantial agreement with those issued by a specialist (QCK = 0.609). This outperformed a competitive classical machine learning algorithm. We illustrated the main machine learning challenges associated with this new task and pointed to generalization across time as our key area of future research.

Limitations

We would like to highlight several limitations in our study. (a) All patients were scored on HDRS and YMRS by the same clinician. However, having scores from multiple (independent) clinicians on the same patients would help appreciate model performance in view of inter-rater agreement. (b) The lack of follow-up HDRS and YMRS scores within the same session did not allow us to estimate to what degree a shift in target variables might be at play. Relatedly, we acknowledge that the choice of five hours for our main analyses may be disputable and other choices may have been valid too. Five hours was an informed attempt to trade off a reasonably high number of samples with a minimal distribution shift over both target variables and physiological data; studying the effect of different cut-offs was not within the scope of this work. (c) Given the naturalist setting, medications were allowed, and their interference could not be ruled out. (d) As pointed out by Chekroud et al. [52], the generalizability of AI systems in healthcare remains a significant challenge. While we tested our method on out-of-distribution samples explicitly (close-to-interview vs far-from-interview), other aspects of generalization that are meaningful to personal sensing, such as inter-individual and intra-individual performance, have not yet been tested. For instance, we evaluated our methods on data obtained in a single centre, and it is unclear how well the model would perform in a cross-clinic setting.

Future work

(i) The decline in performance over future time points stands out as the main challenge towards real-world implementations and suggests that the model struggles to adapt to changes in background (latent) variables, e.g., changes in activity patterns. Research into domain adaptation should therefore be prioritized. We also speculate that MD symptoms and relevant physiological signals have slow- as well as fast-changing components. A segment length of 16 s would seem unsuitable for representing the former and an attempt should be made at capturing both. (ii) Generalization of unseen patients is a desirable property in real-world applications and something we consider exploring in the future. Another approach to tackle this point is to develop (or fine-tune) a model for each individual patient, as done in related fields [55]. (iii) Supervised learning systems notoriously require vast amounts of labeled data for training; as annotation (i.e., enlisting mental health specialists to assess individuals and assign them diagnoses and symptoms' severity scores) is a major bottleneck in mental healthcare [56], self-supervised learning [57] should be considered for applications using the E4 device. (iv) For an ML system to be trustworthy and actionable in a clinical setting, further research into model explainability and uncertainty quantification is warranted [58].

CODE AND DATA AVAILABILITY

The software codebase used is available at https://github.com/april-tools/wear-your-scales. Python 3.10 programming language was used for the symptoms scoring system, where deep learning models were implemented in PyTorch [59], hyperparameter tuning and visualization model performance were performed in Weights and Biases [60], and random forest classifiers were developed in scikit-learn [61]. Graphical modeling of the residuals and related analyses were performed in R

4.2.2 using packages *qgraph* [62] for network estimation and visualization, and *bootnet* [63] for bootstrapping. Data in de-identified form may be made available from the corresponding author upon reasonable request.

REFERENCES

1. American Psychiatric Association D, Association AP, others *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013
2. Santomauro DF, Herrera AMM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. 2021;398:1700–12.
3. Greenberg PE, Fournier A-A, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics. 2021;39:653–65.
4. Vieta E, Berk M, Schulze TG, Carvalho AF, Suppes T, Calabrese JR, et al. Bipolar disorders. Nat Rev Dis Prim. 2018;4:16.
5. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56.
6. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry. 1978;133:429–35.
7. Tohen M, Bowden C, Nierenberg AA, Geddes J. *Clinical trial design challenges in mood disorders*. Academic Press, 2015
8. Satiani A, Niedermier J, Satiani B, Svendsen DP. Projected workforce of psychiatrists in the United States: a population analysis. Psychiatr Serv. 2018;69:710–3.
9. Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry. 2020;10:1–26.
10. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol. 2017;13:23–47.
11. Jacobson NC, Feng B. Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. Transl Psychiatry. 2022;12:1–7.
12. Faurholt-Jepsen M, Brage S, Kessing LV, Munkholm K. State-related differences in heart rate variability in bipolar disorder. J Psychiatr Res. 2017;84:169–73.
13. Sarchiapone M, Gramaglia C, Iosue M, Carli V, Mandelli L, Serretti A, et al. The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis. BMC Psychiatry. 2018;18:1–27.
14. Tazawa Y, Wada M, Mitsukura Y, Takamiya A, Kitazawa M, Yoshimura M, et al. Actigraphy for evaluation of mood disorders: a systematic review and meta-analysis. J Affect Disord. 2019;253:257–69.
15. Culpepper L, Muskin PR, Stahl SM. Major depressive disorder: understanding the significance of residual symptoms and balancing efficacy with tolerability. Am J Med. 2015;128:S1–S15.
16. Earley W, Durgam S, Lu K, Ruth A, Németh G, Laszlovszky I, et al. Clinically relevant response and remission outcomes in cariprazine-treated patients with bipolar I disorder. J Affect Disord. 2018;226:239–44.
17. Salagre E, Vieta E. Precision psychiatry: complex problems require complex solutions. Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol. 2021;52:94–95.
18. Serretti A. Precision medicine in mood disorders. Psychiatry Clin Neurosci Rep. 2022;1:e1.
19. Vieta E, Durgam S, Lu K, Ruth A, Debelle M, Zukin S. Effect of cariprazine across the symptoms of mania in bipolar I disorder: analyses of pooled data from phase II/III trials. Eur Neuropsychopharmacol. 2015;25:1882–91.
20. Lisinski A, Hieronymus F, Näslund J, Nilsson S, Eriksson E. Item-based analysis of the effects of duloxetine in depression: a patient-level post hoc study. Neuropsychopharmacology. 2020;45:553–60.
21. Côté-Allard U, Jakobsen P, Stautland A, Nordgreen T, Fasmer OB, Oedegaard KJ. et al. Long-Short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors. IEEE Pervasive Comput. 2022;21:20–31.
22. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry. 2020;11:584711.
23. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C et al. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2017, pp 325–32.
24. Tazawa Y, Liang K, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon. 2020;6:e03274.
25. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. *NPJ*. Digit Med. 2019;2:3.
26. Nguyen D-K, Chan C-L, Li A-HA, Phan D-V, Lan C-H. Decision support system for the differentiation of schizophrenia and mood disorders using multiple deep learning models on wearable devices data. Health Inform J. 2022;28:14604582221137537.
27. Lee H-J, Cho C-H, Lee T, Jeong J, Yeom JW, Kim S. et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. Psychol Med. 2023;53:5636–44.
28. Hidalgo-Mazzei D, Young AH, Vieta E, Colom F. Behavioural biomarkers and mobile mental health: a new paradigm. Int J Bipolar Disord. 2018;6:1–4.
29. Özdenizci O, Wang Y, Koike-Akino T, Erdoğmuş D. Adversarial deep learning in EEG biometrics. IEEE Signal Process Lett. 2019;26:710–4.
30. De Dios C, Ezquiaga E, Garcia A, Soler B, Vieta E. Time spent with symptoms in a cohort of bipolar disorder outpatients in Spain: a prospective, 18-month follow-up study. J Affect Disord. 2010;125:74–81.
31. Verduijn J, Verhoeven JE, Milaneschi Y, Schoevers RA, van Hemert AM, Beekman AT, et al. Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: full recovery is the exception rather than the rule. BMC Med. 2017;15:1–9.
32. Anmella G, Corponi F, Li BM, Mas A, Sanabra M, Pacchiarotti I, et al. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. JMIR MHealth UHealth. 2023;11:e45405.
33. Tohen M, Frank E, Bowden CL, Colom F, Ghaemi SN, Yatham LN, et al. The International Society for Bipolar Disorders (ISBD) task force report on the nomenclature of course and outcome in bipolar disorders. Bipolar Disord. 2009;11:453–73.
34. Empatica. E4 wristband technical specifications – Empatica Support. E4 Wristband Tech. Specif. 2020. https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications. Accessed in June 2023.
35. Schuurmans AA, de Looff P, Nijhof KS, Rosada C, Scholte RH, Popma A, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst. 2020;44:1–11.
36. Kleckner IR, Jones RM, Wilder-Smith O, Wormwood JB, Akcakaya M, Quigley KS, et al. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. IEEE Trans Biomed Eng. 2017;65:1460–7.
37. Li X, Kane M, Zhang Y, Sun W, Song Y, Dong S, et al. Circadian rhythm analysis using wearable device data: novel penalized machine learning approach. J Med Internet Res. 2021;23:e18403.
38. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85:257–68.
39. de La Torre J, Puig D, Valls A. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. Pattern Recognit Lett. 2018;105:144–54.
40. Duran A, Dussert G, Rouvière O, Jaouen T, Jodoin P-M, Lartizien C. ProstAttentionNet: a deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. Med Image Anal. 2022;77:102347.
41. Czodrowski P. Count on kappa. J Comput Aided Mol Des. 2014;28:1049–55.
42. Alavi M, Biros E, Cleary M. A primer of inter-rater reliability in clinical measurement studies: pros and pitfalls. J Clin Nurs. 2022;31:e39–42.
43. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013;170:59–70.
44. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proc. IEEE international conference on computer vision*. 2017, pp 2980-8.
45. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249–59.
46. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: a novel bandit-based approach to hyperparameter optimization. J Mach Learn Res. 2017;18:6765–816.
47. Föll S, Maritsch M, Spinola F, Mishra V, Barata F, Kowatsch T, et al. FLIRT: a feature generation toolkit for wearable data. Comput Methods Prog Biomed. 2021;212:106461.
48. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–41.
49. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22:276–82.
50. Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. Dialogues Clin Neurosci. 2022;10:271–77.
51. Weiner L, Doignon-Camus N, Bertschy G, Giersch A. Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures. Sci Rep. 2019;9:1–10.
52. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science. 2024;383:164–7.
53. Lukasiewicz M, Gerard S, Besnard A, Falissard B, Perrin E, Sapin H, et al. Young Mania Rating Scale: how to interpret the numbers? Determination of a severity

threshold and of the minimal clinically significant difference in the EMBLEM cohort. Int J Methods Psychiatr Res. 2013;22:46–58.
54. Anderson I, Pilling S, Barnes A, Bayliss L, Bird V. The NICE guideline on the treatment and management of depression in adults. National Collaborating Centre for Mental Health, UK. Depression: the treatment and management of depression in adults (Updated Edition). British Psychological Society. 2010.
55. Saha S, Baumert M. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. Front Comput Neurosci. 2020;13:87.
56. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: *Proc. IEEE international conference on computer vision*. 2017, pp 843-52.
57. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. 2022;6:1346–52.
58. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. Npj Digit Med. 2023;6:6.
59. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G et al. PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc; 2019. pp. 8024−35.
60. Biewald L. Experiment tracking with weights and biases. 2020. https://www.wandb.com/.
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
62. Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: network visualizations of relationships in psychometric data. J Stat Softw. 2012;48:1–18.
63. Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: a tutorial paper. Behav Res Methods. 2018;50:195–212.

ACKNOWLEDGEMENTS

The authors thank Dr. Arno Onken for their comments and feedback on early versions of this work. The authors acknowledge the contribution of all the participants and collaborators of this study. This study has been funded by Instituto de Salud Carlos III (ISCIII) through the project "FIS PI21/00340, TIMEBASE Study" and co-funded by the European Union, as well as a Baszucki Brain Research Fund grant (PI046998) from the Milken Foundation. The ISCIII or the Milken Foundation had no further role in study design, collection, analysis, or interpretation of data, writing of the report, decision to submit the paper for publication. FC and BML are supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. GA is supported by a Rio Hortega 2021 grant (CM21/00017) from the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and cofinanced by Fondo Social Europeo Plus (FSE+). AM is supported by a contract funded by MCIN/AEI/TED2021- 131999BI00 Strategic Projects Oriented to the Ecological Transition and the Digital Transition 2021 and by the "European Union NextGenerationEU/PRTR". IP's research is supported by a FIS 2018 and 2021 grant (PI18/01001;PI21/00169) financed by the Instituto de Salud Carlos III (ISCIII). IG thanks the support of the Spanish Ministry of Science and Innovation (PI19/00954 and PI23/00822) integrated into the Plan Nacional de I+D+I and cofinanced by the ISCIII-Subdirección General de Evaluación y el Fondos Europeos de la Unión Europea (FEDER, FSE, Next Generation EU/Plan de Recuperación Transformación y Resiliencia_PRTR); the ISCIII; the CIBER of Mental Health (CIBERSAM); and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (2017 SGR 1365), Centres de Recerca de Catalunya (CERCA) Programme or Generalitat de Catalunya as well as the Fundació Clínic per la Recerca Biomèdica (Pons Bartran 2022-FRCB_PB1_2022). AB's research is supported by a FIS 2022 grant (PI122/01048) financed by the Instituto de Salud Carlos III (ISCIII), and grant Marató TV3 number 477/cC/2022. MG thanks the support of the Spanish Ministry of Science and Innovation (PI21/00340) integrated into the Plan Nacional de I+D+I and cofinanced by the ISCIII-Subdirección General de Evaluación y el Fondos Europeos de la Unión Europea (FEDER, FSE, Next Generation EU/Plan de Recuperación Transformación y Resiliencia_PRTR); the CIBER of Mental Health (CIBERSAM); and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (2017 SGR 1365), Centres de Recerca de Catalunya (CERCA) Programme or Generalitat de Catalunya. EV has received grants and served as consultant, advisor, or CME speaker for the following entities: AB-Biotics, AbbVie, Angelini, Biogen, Biohaven, Boehringer-Ingelheim, Celon Pharma, Compass, Dainippon Sumitomo Pharma, Ethypharm, Ferrer, Gedeon Richter, GH Research, Glaxo-Smith Kline, Idorsia, Janssen, Lundbeck, Medincell, Novartis, Orion Corporation, Organon, Otsuka, Rovi, Sage, Sanofi-Aventis, Sunovion, Takeda, and Viatris, outside the submitted work. DHM has received CME-related honoraria and served as consultant for Abbott, Angelini, Ethypharm Digital Therapy and Janssen-Cilag. AV is supported by the "UNREAL" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC.

AUTHOR CONTRIBUTIONS

FC conceived of the study, proposed the methodology, developed the software codebase for the analyses, and prepared the manuscript. BML contributed to codebase development and manuscript writing. GA, AM, IP, MV, IG, AB, and MG collected the data for the INTREPIBD study. EV, SML and HCW critically reviewed the manuscript and provided feedback on the clinical side. DHM is the coordinator and the principal investigator of the INTREPIBD/TIMEBASE study and critically reviewed the manuscript. AV supervised this study and contributed to the study design, methodology development, and manuscript writing.

COMPETING INTERESTS

All authors report no financial or other relationship relevant to the subject of this article. GA has received CME-related honoraria, or consulting fees from Angelini, Casen Recordati, Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, and Rovi, with no financial or other relationship relevant to the subject of this article. IP has received CME-related honoraria, or consulting fees from Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, CASEN Recordati and Angelini, with no financial or other relationship relevant to the subject of this article. MV has received research grants from Eli Lilly & Company and has served as a speaker for Abbott, Bristol–Myers Squibb, GlaxoSmithKline, Janssen–Cilag, and Lundbeck. MG has received CME-related honoraria, or consulting fees from Angelini, Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, and Ferrer, with no financial or other relationship relevant to the subject of this article. EV has received grants and served as consultant, advisor or CME speaker for the following entities: AB-Biotics, AbbVie, Adamed, Angelini, Biogen, Beckley-Psytech, Biohaven, Boehringer-Ingelheim, Celon Pharma, Compass, Dainippon Sumitomo Pharma, Ethypharm, Ferrer, Gedeon Richter, GH Research, Glaxo-Smith Kline, HMNC, Idorsia, Johnson & Johnson, Lundbeck, Luye Pharma, Medincell, Merck, Newron, Novartis, Orion Corporation, Organon, Otsuka, Roche, Rovi, Sage, Sanofi-Aventis, Sunovion, Takeda, Teva, and Viatris, outside the submitted work. DHM has received CME-related honoraria and served as consultant for Abbott, Angelini, Ethypharm Digital Therapy and Janssen-Cilag.

ETHICS APPROVAL

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice and the Hospital Clinic Ethics and Research Board (HCB/2021/104). All participants provided written informed consent prior to their inclusion in the study.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-024-02876-1.

Correspondence and requests for materials should be addressed to Filippo Corponi.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Table 1. **Number of recording segments across train, validation, and test splits by segment length.** While the same parts of the recording were used for train, validation, and test sets, regardless of the choice of segment length τ, different τ values (in seconds) resulted in different numbers of segments as shown below. While exploring different values of τ might add a layer of complexity to data splits, yet this was more principled than simply fixing segment length to an arbitrary number as done in previous works.

Segment length τ (sec)	**Train (N)**	**Validation (N)**	**Test (N)**
8	232894	49905	49905
32	58224	12476	12476
64	29112	6238	6238
128	14608	3130	3130
256	7252	1554	1554
512	3528	735	735
1024	1728	432	432

Supplementary Table 2. **Artificial Neural Network hyperparameter search space and final configuration after Hyper-band Bayesian optimization.**

Hyperparameter	**Search Space**	**Final Value**
Batch size	Uniform, min: 8, max 128, interval: 8	96
Critic λ	Uniform, min: 0, max 1	0.0723
Segment length τ	Uniform, 2^n, $3 \leq n \leq 10$	16
Focal loss γ	0, 1, 2, 3, 4, 5	5
Imbalance mode	N/A, Focal loss, prob. threshold, resample and reweight	Focal loss
Learning rate α_{lr}	Uniform, min: 0.0001, max: 0.01	0.0009
Preprocessing	N/A, Normalization, Standardization	Standardization
Loss function	Cross-entropy, Weighted κ, ONTRAM	Weighted κ
Weight decay	Uniform, min: 0, max: 1	0.1853
Channel encoders EN		
Embedding type	MLP, GRU, Time2Vec	MLP
Embedding Dim.	Uniform, min: 32, max: 64, interval: 8	312
Representation Module RM		
Num. Units	Uniform, min: 8, max: 2048, interval: 8	1568
Dropout	Uniform, min: 0, max: 1	0.4870

Supplementary Table 3. **Random Forest Classifiers search space.** 28 random forest classifiers were developed, one for each HDRS and YMRS item. Hyperparameters for each classifier were selected based on best validation set performance out of 300 randomly sampled configurations.

Hyperparameter	**Search Space**
Segment length τ	Uniform, 2^n, $3 \leq n \leq 10$
Criterion	Uniform: Gini, entropy log loss
Max features	Uniform, min: 8, max 184, interval: 8
Max depth	Uniform: 3, 9, 15, 30, None
Num. estimators	Uniform, 10^n, $2 \leq n \leq 5$

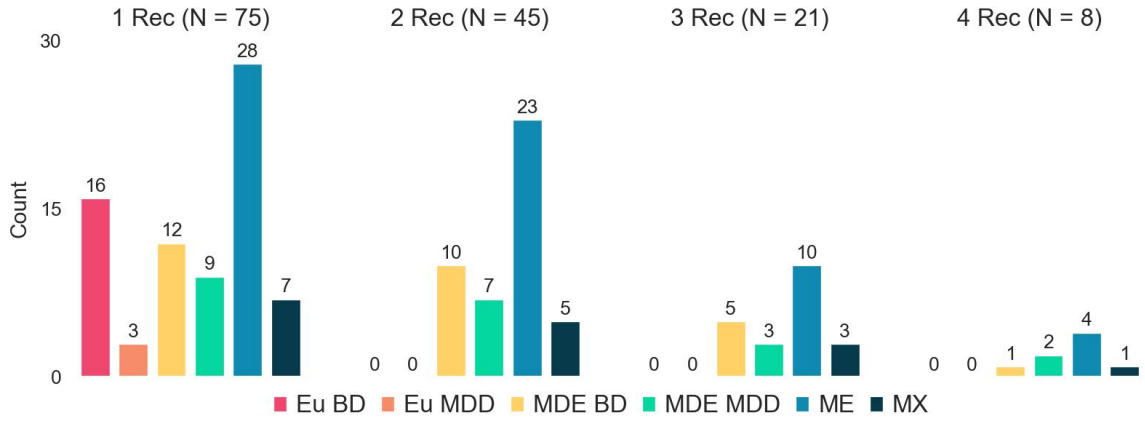
Supplementary Table 4. **Hyperparameters sorted by their importance towards predicting the monitored metric, validation Quadratic Cohen's κ (QCK).** QCK predictability from hyperparameters is derived by training a random forest with the hyperparameters as inputs and the metric as the target output and estimating feature importance values for the random forest. Details at docs.wandb.ai/parameter-importance.

Hyperparameter importance
loss function 0.205
dropout 0.073
batch size 0.070
weight decay 0.055
preprocessing 0.035
RM num. units 0.033
EN dim. 0.032
critic λ 0.026
learning rate α_{lr} 0.025
imbalance mode 0.022
focal loss γ 0.019
segment length τ 0.013

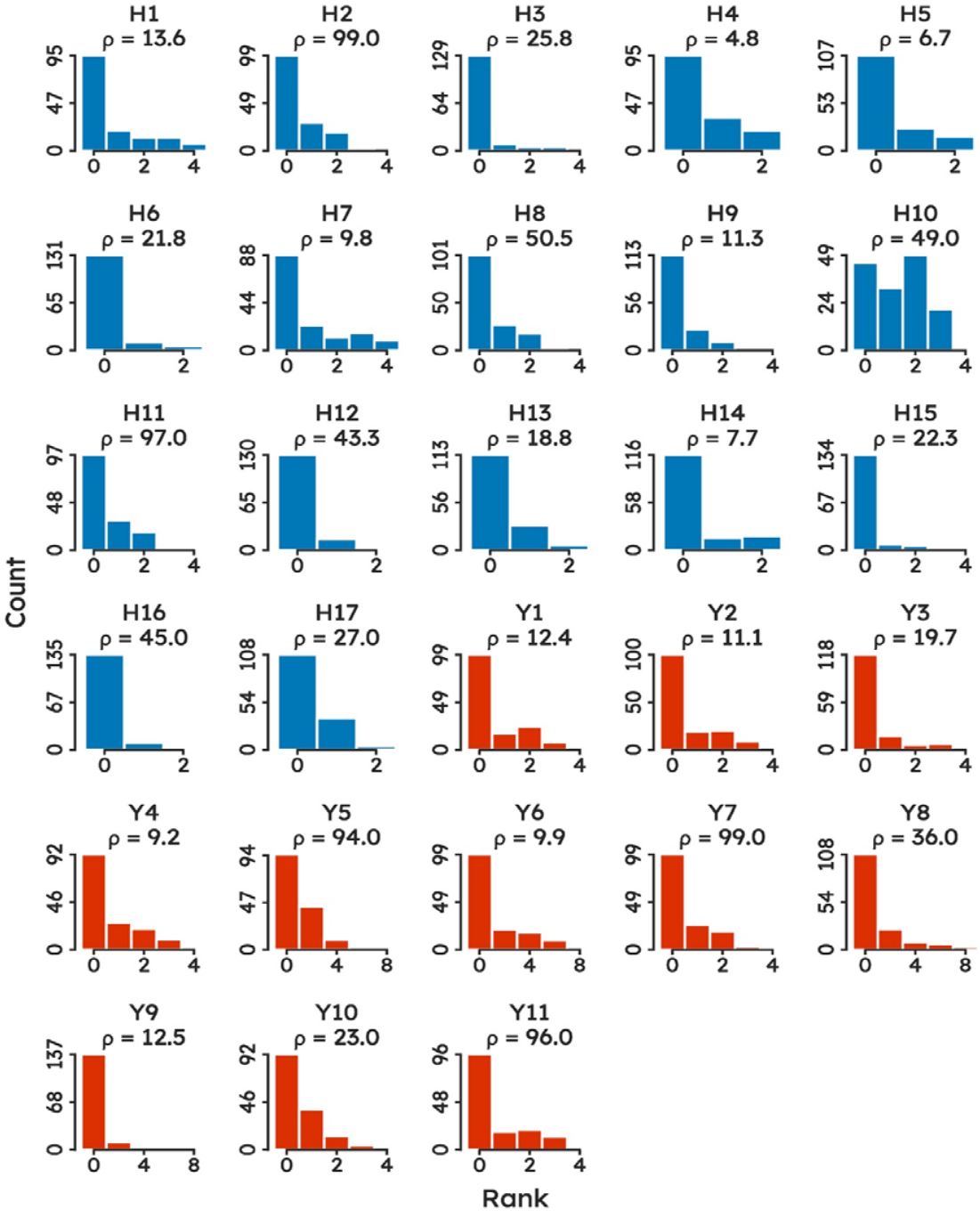
Supplementary Table 5. **F1 score ranges from 0.877 on "disruptive aggressive behavior" and to 0.379 on "anxiety psychic" with an average of 0.609**. Item level macro-averaged F1 score ($F1^M$) across HDRS and YMRS items.

Item $F1^M$	H1 0.512	H2 0.531	H3 0.445	H4 0.643	H5 0.702	H6 0.645	H7 0.610	H8 0.398	H9 0.562	H10 0.380
	H11 0.560	H12 0.860	H13 0.723	H14 0.641	H15 0.753	H16 0.736	H17 0.681	Y1 0.629	Y2 0.599	Y3 0.481
	Y4 0.647	Y5 0.710	Y6 0.581	Y7 0.501	Y8 0.562	Y9 0.877	Y10 0.653	Y11 0.430		

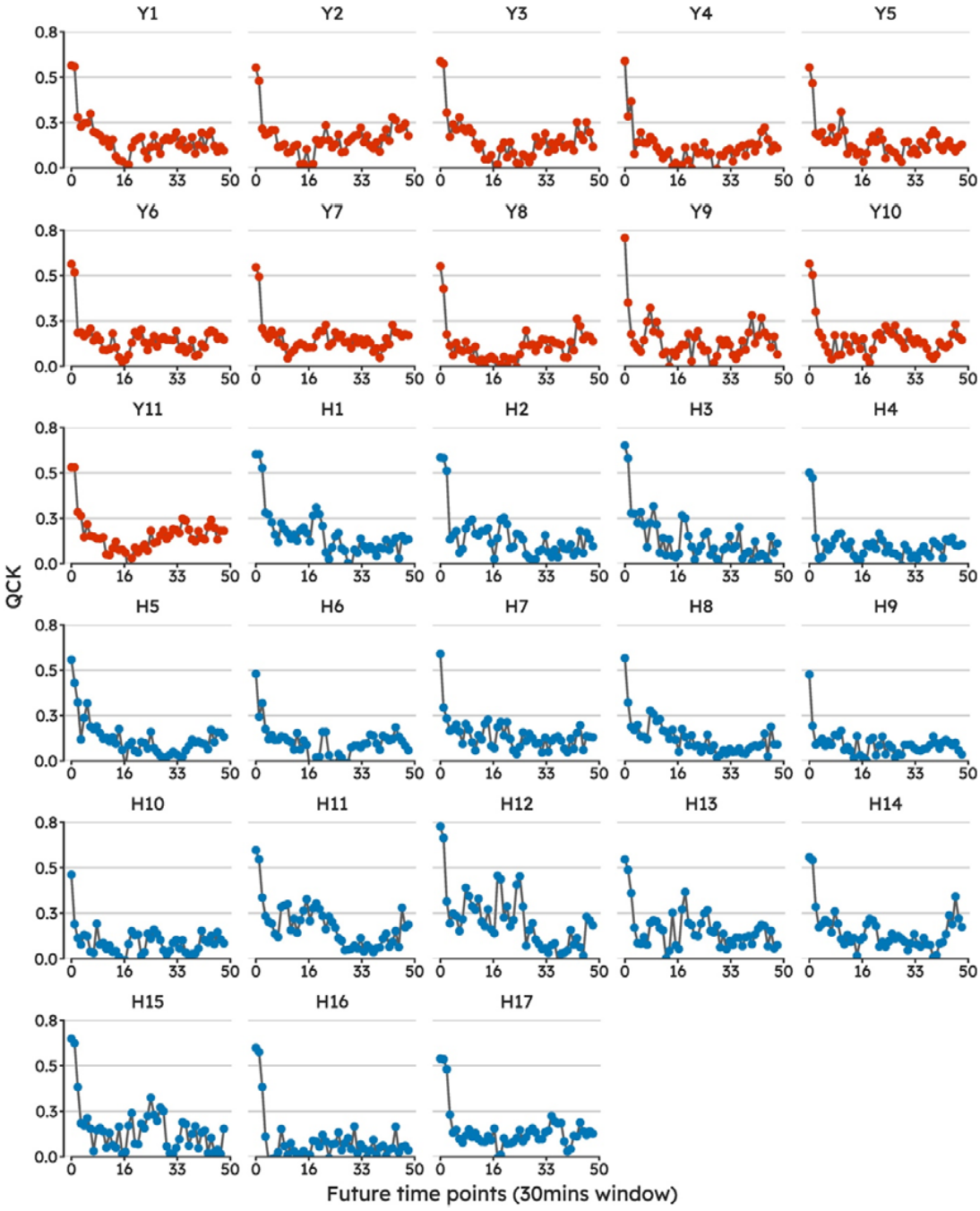
Supplementary Figure 1. **Number of recording sessions available at different time points (T0-T3) by diagnosis.** A total of 149 recordings were available from the data collection campaign at the moment of conducting this study. Mood disorders manifest in two polarities, mania, and depression. Major Depressive Disorder (MDD) is characterized by Major Depressive Episodes (MDEs) only, whereas Bipolar Disorder (BD) features (hypo)manic episodes (ME) that can alternate with MDEs. The presence of symptoms from both polarities within the same episode connotes a mixed episode (MX). Patients with a former mood disorder diagnosis, clinically stable at present are said to be Euthymic (Eu). With the exception of Healthy Controls (HCs) and Euthymic Patients, other subjects are recruited at the onset of a disease episode. They are assessed at subsequent stages (maximum four) during their clinical course. Exclusion criteria are co-morbidity with another psychiatric or neurological disorder or current drug abuse and pregnancy.



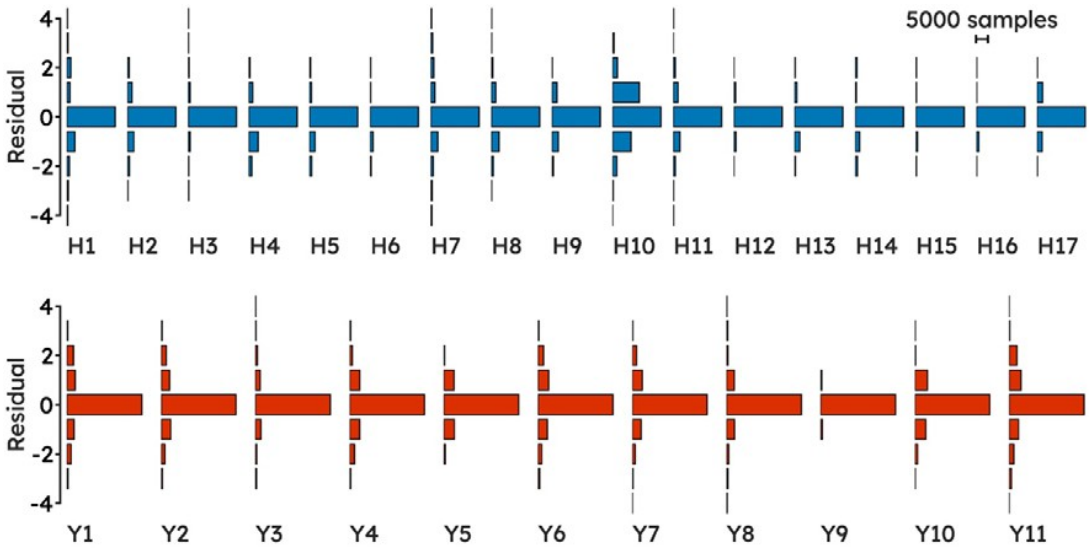
Supplementary Figure 2. **Distribution over Hamilton Depression Rating Scale (blue) and Young Mania Rating Scale (red) items across the recording sessions used in this study.** The number above each bar plot, ρ, is the cardinality of the majority class over that of the minority rank. Higher values of ρ thus indicate a more pronounced imbalance between the majority and the minority rank.



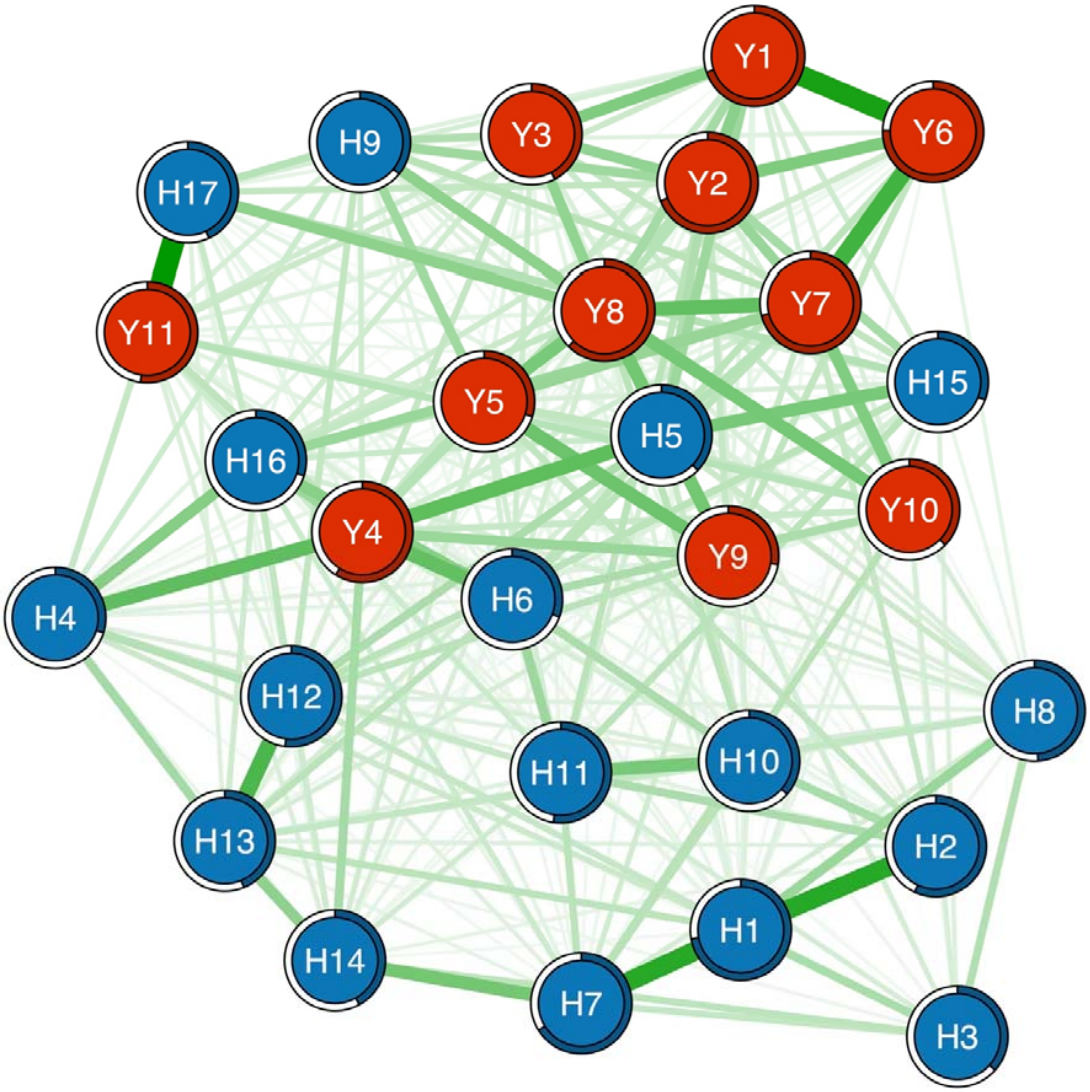
Supplementary Figure 3. **Quadratic Cohen's κ (QCK) deteriorated across all items when the model was tested on segments taken further away from when the interview took place.** The first point (0 on the x-axis) is the baseline performance, i.e. holdout segments from the first five hours of recordings (**close-to-interview**). The following points refer to the successive thirty-minute intervals (**close-to-interview**).



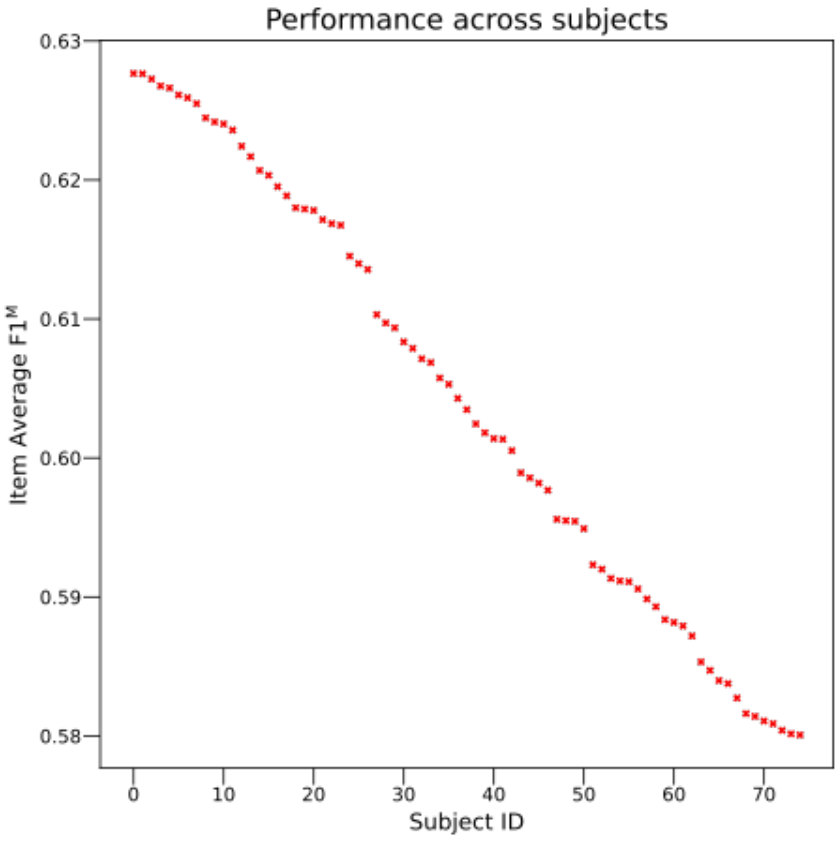
Supplementary Figure 4. **Item residuals overall show a symmetric distribution centered around zero, showing that the model is correct most of the time, it is not systematically either under- or over-predicting, and when wrong, it is usually off by only one.** Residuals, signed difference between prediction ($\hat{y}_i$) and ground truth (y), are shown across Hamilton Depression Rating Scale (Top) and Young Mania Rating Scale (Bottom) items.



Supplementary Figure 5. **Partial correlations between item residuals indicate that the model learned the scales' natural correlation structure.** Network displaying the relationship between HDRS (blue) and YMRS (red) item residuals. Green edges represent positive partial correlations between variables. Rings around nodes represent variance in a given variable with shadowed parts displaying the proportion of variance in that node that is explained by nodes that connect with it.



Supplementary Figure 6**. Performance was consistent across subjects, with no associations with age, sex, or total score on psychometric scales.** Each point on the scatter plot corresponds to a subject score as expressed with mean item macro-average F1 score ($F1^{M}$). Subjects are sorted in descending order by $F1^{M}$ and assigned a dummy ID for the sake of this plot. We could not reject the null hypothesis that the distributions of item-average $F1^{M}$ scores and, on the other hand, age/HDRS/YMRS underlying the samples are uncorrelated (Pearson R=0.016, p-val_B=3.762, Pearson R=0.117, p-val_B=3.964 for HDRS total score, Pearson R=0.152, p-val_B=3.114 for YMRS total score) or the null hypothesis that difference of item-average $F1^{M}$ mean across males and females is zero (t=1.422, p-val_B=0.631).



Supplementary Methods

1. Model architecture and loss functions

Our Artificial Neural Network (ANN) classifier (a), tasked with inferring Hamilton Depression Rating Scale (HDRS) and Young Mania Rating Scale (YMRS) items, consisted of three modules: channel encoders, representation modules, and item predictors. An auxiliary critic (b) aided the classifier in learning (**challenge c3**) subject invariant representations.

a) Classifier

The task of inferring HDRS and YMRS items is an instance of MTL (challenge c1), to which we adopted a hard parameter-sharing approach: all tasks shared the same model trunk $\mathsf{RM}(\mathsf{EN}(\cdot))$, and thus the same base representation of the input data $h_i = \mathsf{RM}\big(\mathsf{EN}(x_i)\big)$, which was then distributed across task-specific layers, IP [1]. As usually done with hard parameter-sharing[2], the multi-task loss was set equal to the average of task-specific losses, each weighted by the corresponding item rank step to account for different item weights on the scale total score:

$$\mathcal{L}_{\mathsf{MT}}(x_i; \xi, \phi, \psi) = -\frac{r_j}{\sum_j r_j} \ell_j\big(\mathsf{IP}_j(\mathsf{RM}(\mathsf{EN}(x_i))), y_i\big)$$

where r_j is the rank step size of the j^{th}item (e.g., Y5 irritability ranks have a step size of 2) and ξ, ϕ, and ψ are respectively EN, RM, and IP parameters. Details on the specific form of ℓ, the task-loss, are given below.

Channel encoders Since the sampling rate varies across the recorded signals within a segment, these are typically time-aligned, e.g. to the level of a second in wall-time usually via max-pooling or averaging[3,4], before being further

analysed. However, we took a different approach that did not pre-specify the function for time alignment and mapped each channel to the same dimensionality with the use of a channel encoder EN. We experimented with a simple Multilayer Perceptron (MLP), a Gated Recurrent Unit (GRU[5]) or, alternatively, the Time2Vec representation proposed in[6]. Before passing segments through EN, each channel was re-scaled. The optimal embedding dimensionality and re-scaling type (either standardization or normalization) were set during tuning.

Representation module We used a BiLSTM[7] as architecture for our RM, as it belongs to Recurrent Neural Networks (RNNs), a class of deep learning architectures specifically engineered to exploit dependencies in time series data, and, thanks to its bidirectionality in consuming an input sequence, it enjoys a richer time representation than a vanilla RNN.

Item predictors (challenge c2) The extracted representation $h = \mathsf{RM}\big(\mathsf{EN}(\cdot)\big)$ was then used as input to 28 IP, each dedicated to a specific HDRS/YMRS item. We experimented with three different treatments of the target variables, each translating to a different set-up of the task-specific layer and the task-specific loss. 1) Each item score prediction was simply treated as a multi-class classification problem. Accordingly, the j^{th} item predictor consisted of a fully connected layer, with as many output units as the number of ranks under that item, to which a SoftMax activation was applied, and the categorical cross-entropy (CCE) was used as loss function. 2) We used the same task-specific architecture as in 1) but adopted the QWK loss, as proposed in[8], which re-writes Cohen's κ in terms of probability distributions. 3) We implemented the ordinal neural network transformation model (ONTRAM[9]) which parameterizes the CCE loss to incorporate the order of the outcome, by deriving class probabilities from the conditional density function of a latent variable onto which observed classes are mapped.

b) Critic

We encouraged cross-subjects invariance in the representation extracted with $\mathsf{RM}(\mathsf{EN}(\cdot))$ by adding a critic CR, whose task was to correctly distinguish subjects apart from extracted representation, in an adversarial game, similarly to [10] and [11]. Concretely, CR, a simple MLP, inputs the extracted representation h and is trained to identify subjects from it. CR's task was therefore to minimize, with respect to CR's parameters θ, the following CCE loss:

$$\mathcal{L}_{\mathsf{CR}}\big(\mathsf{RM}\big(\mathsf{EN}(x_i)\big);\theta\big) = -\mathbb{1}_s \log\ \mathsf{CR}_s\left(\mathsf{RM}\big(\mathsf{EN}(x_i)\big)\right)$$

where $\mathbb{1}_s$ is an indicator taking value 1 when the i^{th}segments belong to the s^{th} subject and 0 otherwise, and $\mathsf{CR}_s(\cdot)$ is the critic output (i.e. probabilities from a softmax activation) for the s^{th} subject. On their part, EN and RM tried to trump the CR by filtering out from h information that could make CR 's task easy, while, at the same time retaining enough useful information for the item predictors IP_j. To achieve this, the following term was added to $\mathcal{L}_{MT}$ (the multi-task loss), which was minimized with respect to EN's and RM's parameters, ξ and φ :

$$\mathcal{L}_{\mathsf{R}}(x_i;\xi,\phi) = \lambda\left[-\mathbb{1}_s \log\left(1-\mathsf{CR}_s\left(\mathsf{RM}\big(\mathsf{EN}(x_i)\big)\right)\right)\right]$$

where $\lambda \in [0,1]$. The classifier's total loss was then $\mathcal{L}_{\mathsf{CF}} = \mathcal{L}_{\mathsf{MT}} + \mathcal{L}_{\mathsf{R}}$, where $\mathcal{L}_{\mathsf{R}}$ acted as a regulariser, a price CF paid for encoding subject-specific information in the representation h learned by $\mathsf{RM}(\mathsf{EN}(\cdot))$. Values of λ trade off learning cross-subjects invariant representations h against solving the main objective; for $\lambda = 0$, no incentive is given towards learning cross-subjects invariant representation.

2. Model training

All models were trained with AdamW[12] optimizer for a maximum of 400 epochs. Moreover, to speed up the training and search procedure, we employed an early stopping learning rate scheduler: we reduce the learning rate $\alpha_{\mathrm{lr}} = 0.3\alpha_{\mathrm{lr}}$ if the model has not improved in its validation performance after 10 consecutive epochs; we terminate the training procedure if the model has not improved after 2 learning rate reductions. Dropout[13] and weight decay were added to prevent overfitting.

3. Gaussian Graphical Lasso

Network edge sparsity, to avoid false positives, is enforced with the least absolute shrinkage and selection operator (LASSO[14]), which indeed shrinks all edge weights towards zero and sets small weights to exactly zero. The strength of the regularization is traded off by a hyperparameter λ, selected with the Extended Bayesian Information Criterion (EBIC[15]). The EBIC itself has a tuning parameter γ controlling the trade-off between sensitivity and precision, which we set to 0.25 as in [16]. We also estimated node predictability, measuring how well a node can be predicted by nodes it shares an edge with, which can be interpreted similarly to R^2[2,17]. Lastly, bootstrapping routines were used to gain insight into thestability of the estimated parameters.

References

1 Crawshaw M. Multi-task learning with deep neural networks: A survey. *ArXiv Prepr ArXiv200909796* 2020.

2 Ruder S. An overview of multi-task learning in deep neural networks. *ArXiv Prepr ArXiv170605098* 2017.

3 Adler DA, Wang F, Mohr DC, Choudhury T. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos One* 2022; **17**: e0266516.

4 Li BM, Corponi F, Anmella G, Mas A, Sanabra M, Hidalgo-Mazzei D *et al.* Inferring mood disorder symptoms from multivariate time-series sensory data. In: *NeurIPS 2022 Workshop on Learning from Time Series for Health*. 2022https://openreview.net/forum?id=awjU8fCDZjS.

5 Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv Prepr ArXiv14123555* 2014.

6 Kazemi SM, Goel R, Eghbali S, Ramanan J, Sahota J, Thakur S *et al.* Time2vec: Learning a vector representation of time. *ArXiv Prepr ArXiv190705321* 2019.

7 Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 1997; **45**: 2673–2681.

8 de La Torre J, Puig D, Valls A. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognit Lett* 2018; **105**: 144–154.

9 Kook L, Herzog L, Hothorn T, Dürr O, Sick B. Deep and interpretable regression models for ordinal outcomes. *Pattern Recognit* 2022; **122**: 108263.

10 Özdenizci O, Wang Y, Koike-Akino T, Erdoğmuş D. Learning invariant representations from EEG via adversarial inference. *IEEE Access* 2020; **8**: 27074–27085.

11 Cheng JY, Goh H, Dogrusoz K, Tuzel O, Azemi E. Subject-aware contrastive learning for biosignals. *ArXiv Prepr ArXiv200704871* 2020.

12 Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: *International Conference on Learning Representations*. 2019https://openreview.net/forum?id=Bkg6RiCqY7.

13 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; **15**: 1929–1958.

14 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 1996; **58**: 267–288.

15 Foygel R, Drton M. Extended Bayesian information criteria for Gaussian graphical models. *Adv Neural Inf Process Syst* 2010; **23**.

16 Haslbeck J, Waldorp LJ. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *ArXiv Prepr ArXiv151006871* 2015.

17 Haslbeck JM, Waldorp LJ. How well do network models predict observations? On the importance of predictability in network models. *Behav Res Methods* 2018; **50**: 853–861.

Chapter 6

Self-supervised Learning Mitigates the Annotation Bottleneck

6.1 The lack of labelled data cripples supervised and transfer learning

As mentioned in **Chapter 2**, labelling data in the context of personal sensing for MDs (as with other clinical applications) is expensive and time-consuming. Concretely, it involves enlisting mental health specialists to assess patients with structured, standardized questionnaires (Appendix A). Large annotated datasets are therefore extremely hard to curate (see Table 1 in Section 5.4). On the other hand, modern ML algorithms are notoriously data hungry, especially large ANNs, where bigger models have been shown to outperform smaller ones when data is not a constraint [185]. This means that task-specific supervised learning (Section 2.3.1) in personal sensing, in other words randomly initializing the parameters of some ANN and updating them by minimizing a task-specific loss (e.g. a cross-entropy classification loss), comes up against data size as a major limitation.

Naturally, researchers in AI have looked at ways to transfer knowledge across domains. Transfer learning [186] has been a first attempt in this direction and has been extensively investigated in CV. The general recipe is first training in a supervised way a model on some source domain, for which a large amount of labelled data is available, and then fine-tuning a sub-set of the model parameters using the target-domain data, for which only limited labelled data is available. As an example of a biomedical application

[187], researchers pre-trained convolutional ANNs on ImageNet [123] and then fine-tuned them on medical image tasks. This works because low-level features, typically learned in the initial layers, capturing basic concepts like edges, lines, or shapes, can be leveraged and transferred over to other domains in CV, such as medical imaging [188]. While more data efficient than end-to-end supervised learning, transfer learning still relies on some large source annotated datasets. This is problematic for personal sensing, as there is only limited data openly available for time series. For context, ImageNet is $\sim$150 GB whereas annotated datasets in the time-series domain are in the order of only a few GB [189].

6.2 Self-supervised learning

Self-supervised learning (SSL) has seen great success in the fields of CV and NLP by eliminating the need for human annotations, thereby leveraging the wealth of unlabelled data available. In this paradigm, the supervisory signal is derived from the data itself, training a model f_θ, an ANN parametrized by θ, on some *pre-text* (also referred to as *surrogate*) task [190]. There are two dominant paradigms for self-supervised pre-training based on the pre-text task design: generative and contrastive [191].

6.2.1 Generative

The generative approach involves training f_θ to generate or reconstruct parts of the data. A notable example in this category is BERT (Bidirectional Encoder Representations from Transformers) [192], where random words in a sentence are masked, and the model must predict these missing words. This forces the model to understand the context of the surrounding text, capturing deep semantic relationships. In the time-series domain, models learning to impute parts of the input set to missing with a Boolean mask follow the same principle. A mean squared error (MSE) objective can be used to guide training:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_{\text{masked},i} - \hat{\mathbf{x}}_{\text{masked},i})^2$$

where M is the number of masked inputs, $\mathbf{x}_{\text{masked},i}$ represents the actual values of these masked inputs, and $\hat{\mathbf{x}}_{\text{masked},i}$ denotes the model's predictions for these specific inputs. So, while the model generally inputs the entire vector $\mathbf{x}$ and reconstructs a full vector

$\hat{\mathbf{x}} = f_\theta(\mathbf{x})$, the MSE calculation is exclusively concerned with the reconstructed values corresponding to the originally masked components.

6.2.2 Contrastive

Contrastive methods, on the other hand, focus on learning representations by bringing similar samples closer together in the embedding space Z, with $\mathbf{z} = f_\theta(\mathbf{x})$, $\mathbf{z} \in \mathbb{R}^H$, $\mathbf{x} \in \mathbb{R}^D$ and $H << D$, while pushing representations of dissimilar samples apart. SimCLR (Simple Framework for Contrastive Learning of Visual Representations) [193] is a prime example of this approach. A critical component of contrastive learning is the use of data augmentation to create multiple views of the same data point. Transformations such as rotation and cropping create different views of the same image; the model learns representations invariant to these transformations. In the time-series domain, determining suitable transformations to create positive pairs is more challenging due to less intuitive human-level domain knowledge compared to vision tasks. Common transformations in this domain include time-warping, jittering, and permutation [194]. A popular choice for the contrastive loss is the Normalized Temperature-Scaled Cross-Entropy (NT-Xent) loss [195, 196], which for a positive pair of examples (i, j) is formulated as:

$$\mathcal{L}_{\text{NT-Xent}}(\mathbf{z}_i, \mathbf{z}_j) = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}$$

Here, each of the N samples in a mini-batch is augmented to produce a corresponding pair, creating N positive pairs in a batch of $2N$ augmented samples. The NT-Xent loss is computed for each sample against the augmented version of itself (forming a positive pair) and compared against the rest of the augmented mini-batch (as negative samples). The function $\text{sim}(\cdot,\cdot)$ denotes a similarity metric (e.g., cosine similarity), and τ is a temperature parameter controlling the sharpness of the softmax. This loss function effectively encourages the model to pull positive pairs' embeddings closer together while repelling negative pairs'. The final loss is computed across all positive pairs, both (i, j) and (j, i) in a mini-batch.

6.2.3 Fine-tuning vs Linear Readout

Two strategies are used to adapt self-supervised pre-trained models $f_{\tilde{\theta}}$, with $\tilde{\theta}$ indicating the model parameters learned during pre-training, to downstream tasks: fine-tuning and linear readout [190, 191].

Fine-tuning involves adjusting the entire set of model parameters $\tilde{\theta}$ on a labelled dataset for a specific target task. So, rather than taking on the target task with a randomly initialized model f_θ, like ordinary supervised learning (Section 2.3.1), $\tilde{\theta}$ are first copied onto θ; f_θ, i.e. $\theta \leftarrow \tilde{\theta}$, is then refined onto the target task for which labelled data is available:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{target}}(f_\theta(\mathbf{X}_{\text{target}}), \mathbf{Y}_{\text{target}})$$

The loss function $\mathcal{L}_{\text{target}}$ is defined with respect to some target task, where $\mathbf{X}_{\text{target}}$ and $\mathbf{Y}_{\text{target}}$ represent the input features and labels of the target task data, respectively. An optimizer adjusts θ iteratively to reduce this loss, culminating in the fine-tuned parameters θ^*.

In contrast, the linear readout approach freezes the pre-trained model parameters $\tilde{\theta}$, treating $f_{\tilde{\theta}}$ as a feature extractor, and introduces and trains on top of such featurizer a linear layer g parametrized by $\mathbf{w}$:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{target}}(g_{\mathbf{w}}(f_{\tilde{\theta}}(\mathbf{X}_{\text{target}})), \mathbf{Y}_{\text{target}})$$

Fine-tuning and linear readout have distinct advantages and limitations [190, 191]. The former allows model representations, captured by $\tilde{\theta}$, to adapt specifically to the nuances of the target task. This can be particularly beneficial when the downstream task differs significantly from the scenarios anticipated during pre-training, however it might come at the price of overfitting. So, fine-tuning might be better indicated when a substantial amount of data for the target task is available. Furthermore, as fine-tuning updates the entire model, it can be computationally expensive. On the other, linear readout is computationally more efficient and less prone to overfitting, since it recycles $f_{\tilde{\theta}}$ as a frozen featurizer and only trains a linear layer (or another simple model) on top of it. However, this method's simplicity may fall short of capturing the task-specific complexities that fine-tuning could address.

6.2.4 Foundation Models

Over the past few years, both the size of models and the corpus of unlabelled data for self-supervised learning have significantly expanded, leading to the development of foundation models, particularly in the natural language domain, such as BERT [192] and GPT-3 [197]. These are models trained on massive datasets which can generalize to new tasks with little to no task-specific training data, enabling few-shot and zero-shot learning. The creation of these foundation models has required substantial financial investments, making it a pursuit beyond the reach of most institutions. Additionally, the development and deployment of foundation models are associated with a significant carbon footprint, often disproportionately affecting disadvantaged economies [198].

Efforts to develop foundation models in time series are underway [199, 200] but are generally hampered by the relative data scarcity, as compared to the data size available for natural language, even when pulling together time series from different domains. Furthermore, while a token in natural language preserves a semantic across different domains, this is not the case with time series, which cannot be created from a fine number of discrete tokens. Motivated by these considerations, model reprogramming [201], a recent paradigm, tries to bootstrap the remarkable capabilities of foundation models from NLP, recycling them as a frozen backbone, while linear input and output transformations align the time-series and language modality and vice versa [202].

6.3 The paper: Wearable Data From Subjects Playing Super Mario, Taking University Exams, or Performing Physical Exercise Help Detect Acute Mood Disorder Episodes via Self-Supervised Learning: Prospective, Exploratory, Observational Study

Collecting and labelling data for personal sensing in MDs is extremely laborious. As a result, studies cannot afford to recruit but a few patients. This can severely limit the training of modern AI systems. Below, we present our original work **Wearable Data From Subjects Playing Super Mario, Taking University Exams, or Performing Physical Exercise Help Detect Acute Mood Disorder Episodes via Self-Supervised Learning: Prospective, Exploratory, Observational Study** published in JMIR mHealth uHealth

where using SSL I make progress towards overcoming the data annotation bottleneck.

I gathered eleven open-access datasets recording physiological data with an Empatica E4 device and developed a pre-processing pipeline for on-/off-body detection, sleep-wake detection, segmentation, and optional feature extraction. I have made both the pre-processing pipeline and the pre-processed data publicly available. This collection, named E4SelfLearning [203], includes 161 subjects and **is the largest open-access dataset of its kind to date**, aiming to stimulate future research into SSL with multivariate time-series sensory data by addressing barriers in pre-processing and data availability. E4SelfLearning can indeed be used for pre-training across a variety of personal sensing tasks, unrelated to MDs.

I proposed a novel Transformer-based architecture (E4mer) and demonstrated that SSL is a viable paradigm, outperforming both the fully-supervised E4mer and classical machine learning models using handcrafted features in distinguishing MD acute episodes from euthymia in a binary time-series classification task. I investigated the factors contributing to SSL's success, comparing two pretext task designs and conducting ablation analyses to study sensitivity to unlabelled data availability.

Original Paper

Wearable Data From Subjects Playing Super Mario, Taking University Exams, or Performing Physical Exercise Help Detect Acute Mood Disorder Episodes via Self-Supervised Learning: Prospective, Exploratory, Observational Study

Filippo Corponi[1], MSc, MD; Bryan M Li[1,2], MSc; Gerard Anmella[3,4,5,6], MD, PhD; Clàudia Valenzuela-Pascual[3,4,5,6], MSc; Ariadna Mas[3,4,5,6], MSc; Isabella Pacchiarotti[3,4,5,6], MD, PhD; Marc Valentí[3,4,5,6], MD, PhD; Iria Grande[3,4,5,6], MD, PhD; Antoni Benabarre[3,4,5,6], MD, PhD; Marina Garriga[3,4,5,6], MD, PhD; Eduard Vieta[3,4,5,6], MD, PhD; Allan H Young[7], MD, PhD; Stephen M Lawrie[8], MD; Heather C Whalley[8,9], PhD; Diego Hidalgo-Mazzei[3,4,5,6,7*], MD, PhD; Antonio Vergari[1*], PhD

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

[2]The Alan Turing Institute, London, United Kingdom

[3]Bipolar and Depressive Disorders Unit, Department of Psychiatry and Psychology, Hospital Clínic de Barcelona, Barcelona, Spain

[4]Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain

[5]Centro de Investigación Biomédica en Red de Salud Mental, Instituto de Salud Carlos III, Madrid, Spain

[6]Departament de Medicina, Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona, Barcelona, Spain

[7]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[8]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

[9]Generation Scotland, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Filippo Corponi, MSc, MD
School of Informatics
University of Edinburgh
Informatics Forum, 10 Crichton St, Newington
Edinburgh, EH89AB
United Kingdom
Phone: 44 131 651 5661
Email:

Abstract

Background: Personal sensing, leveraging data passively and near-continuously collected with wearables from patients in their ecological environment, is a promising paradigm to monitor mood disorders (MDs), a major determinant of the worldwide disease burden. However, collecting and annotating wearable data is resource intensive. Studies of this kind can thus typically afford to recruit only a few dozen patients. This constitutes one of the major obstacles to applying modern supervised machine learning techniques to MD detection.

Objective: In this paper, we overcame this data bottleneck and advanced the detection of acute MD episodes from wearables' data on the back of recent advances in self-supervised learning (SSL). This approach leverages unlabeled data to learn representations during pretraining, subsequently exploited for a supervised task.

Methods: We collected open access data sets recording with the Empatica E4 wristband spanning different, unrelated to MD monitoring, personal sensing tasks—from emotion recognition in Super Mario players to stress detection in undergraduates—and devised a preprocessing pipeline performing on-/off-body detection, sleep/wake detection, segmentation, and (optionally) feature extraction. With 161 E4-recorded subjects, we introduced E4SelfLearning, the largest-to-date open access collection, and its preprocessing pipeline. We developed a novel E4-tailored transformer (E4mer) architecture, serving as the blueprint for both SSL and fully supervised learning; we assessed whether and under which conditions self-supervised pretraining led to an

improvement over fully supervised baselines (ie, the fully supervised E4mer and pre–deep learning algorithms) in detecting acute MD episodes from recording segments taken in 64 (n=32, 50%, acute, n=32, 50%, stable) patients.

Results: SSL significantly outperformed fully supervised pipelines using either our novel E4mer or extreme gradient boosting (XGBoost): n=3353 (81.23%) against n=3110 (75.35%; E4mer) and n=2973 (72.02%; XGBoost) correctly classified recording segments from a total of 4128 segments. SSL performance was strongly associated with the specific surrogate task used for pretraining, as well as with unlabeled data availability.

Conclusions: We showed that SSL, a paradigm where a model is pretrained on unlabeled data with no need for human annotations before deployment on the supervised target task of interest, helps overcome the annotation bottleneck; the choice of the pretraining surrogate task and the size of unlabeled data for pretraining are key determinants of SSL success. We introduced E4mer, which can be used for SSL, and shared the E4SelfLearning collection, along with its preprocessing pipeline, which can foster and expedite future research into SSL for personal sensing.

(JMIR Mhealth Uhealth 2024;12:e55094) doi: 10.2196/55094

KEYWORDS

mood disorder; time-series classification; wearable; personal sensing; deep learning; self-supervised learning; transformer

Introduction

Mood disorders (MDs) are a group of mental health conditions in the *Diagnostic and Statistical Manual, Fifth Edition* (DSM-5) classification system [1]. They are chronic, recurrent disorders featuring disturbances in emotions, energy, and thought, standing out as a leading cause of worldwide disability [2,3] and suicidality [4]. Timely recognition of MD episodes is critical toward better outcomes [5]. However, this is challenging due to generally limited patient insight [6], compounded with the low availability of specialized care for MDs, with rising demand straining current capacity [7,8].

Personal sensing, involving the use of machine learning (ML) to harness data passively and near-continuously collected with wearable devices from patients in their ecological environment, has been attracting interest as a promising paradigm to address this gap [9]. Indeed, some of the core MD clinical features (eg, disturbance in mood and energy levels) translate into changes in physiological parameters measurable with wearable devices [10-12]. A major barrier to the development of clinical decision support systems featuring personal sensing has been the scarcity of labeled data, that is, data with annotations by clinicians about the MD state (eg, diagnosis, disease phase, symptom severity). Collecting and annotating data for personal sensing in MDs is, indeed, an expensive and time-consuming enterprise; thus, studies typically use samples running into only a few dozen patients [13-20].

In this work, we took a different perspective and leveraged *unlabeled* data collected with the Empatica E4 (hereafter E4) wristband [21], a popular research-grade device for personal sensing studies [22], as well as recent advancements in self-supervised learning (SSL) techniques that can learn meaningful representations from such unlabeled data. Specifically, we took advantage of open access data sets that record physiological data with the E4 across different settings but do not address MDs and therefore do not provide information about the mood state of the subjects involved. Although each such data set has only a limited number of subjects, our aggregated and preprocessed data set E4SelfLearning can break the labeled data bottleneck for personal sensing in MDs (Figure 1) [23-33].

Fully supervised systems require vast amounts of data to train, thus limiting their application in different fields, such as health care, where amassing large, high-quality data sets is demanding in terms of time and human resources [34]. Although previous studies on personal sensing for MDs have investigated different tasks, including acute MD episode detection [13-16], regression of a psychometric scale total score [17-19], and, more recently, multitask inference of all items in 2 commonly used psychometric scales [35], they all developed their models in a fully supervised fashion (ie, they were trained on samples for which ground-truth labels were available). As a result, considering that obtaining clinical annotations from patients, especially when on an acute MD episode, is a challenging and expensive enterprise, the sample size is generally modest (eg, N=52 in Côté-Allard et al [15], N=45 in Tazawa et al [13], and N=31 in Pedrelli et al [18]).

SSL, in contrast, is a framework where the model creates proxy supervisory signals within the data themselves, therefore alleviating the annotation bottleneck and allowing us to repurpose existing unlabeled data sets [36]. Specifically, SSL derives supervisory signals from the data themselves, thanks to pretext tasks, which are new supervised challenges, for example, imputing occluded parts of the input data. Through such preparatory pretext tasks, not requiring expert annotation, the model learns useful representations, partial solutions to the downstream target task of interest, for which only a comparatively small amount of annotated data are available [37]. On the back of the great success of SSL in computer vision (CV) [37] and natural language processing (NLP) [38], and with encouraging findings in other health care applications [39], we extended pioneering SSL works on multivariate time series [40-42] to personal sensing in MDs.

In this work, we made the following contributions:

- We gathered 11 open access data sets recording physiological data with an E4 wristband and developed a pipeline for preprocessing such data that performed on-/off-body detection, sleep/wake detection, segmentation, and (optionally) feature extraction. We made the

preprocessing pipeline and the preprocessed data publicly available. This collection (E4SelfLearning), with 161 subjects, is the biggest open access data set to date. We believe that this effort can stimulate future research into SSL with multivariate time-series sensory data by removing 2 barriers, preprocessing and data availability.

- We proposed a novel E4-tailored transformer (E4mer) architecture (Figure 2) [43] and showed that SSL is a viable paradigm, outperforming both fully supervised E4mer and classical machine learning (CML) models using handcrafted features in distinguishing acute MD episodes from clinical stability (euthymia in psychiatric parlance), that is, a time-series (binary) classification task.
- We investigated what makes SSL successful. Specifically, we compared 2 main pretext task designs (ie, masked prediction [MP] and transformation prediction [TP]) [44], and for the best-performing routine, we studied its sensitivity to the unlabeled data availability in ablation analyses. We inspected learned embeddings and showed that they capture meaningful semantics about the underlying context (ie, sleep/wake status) and symptom severity.

Figure 1. A total of ~6254 hours (261 days) of unlabeled recordings from 252 subjects while awake were used for self-supervised pretraining. Unlabeled data comprised a collection of 11 open access data sets, whose aggregation we make publicly available (E4SelfLearning), along with part of the TIMEBASE/INTREPIBD study that was not relevant for the target task under investigation (ie, acute episode vs euthymia classification). Unlabeled data were passed through a model consisting of an encoder and a transform head for self-supervised pretraining; the pretrained encoder block was then retained for the target task, while the transform head was replaced with a new, randomly initialized classification head. *The target task (labeled) training set from the TIMEBASE/INTREPIBD study was also used during self-supervised pretraining. Further details on the data sets used in this study are available in Table S1 in Multimedia Appendix 1. ADARP: Alcohol and Drug Abuse Research Program; PGG-DaLia: PPG Dataset for Motion Compensation and Heart Rate Estimation in Daily Life Activities; TIMEBASE/INTREPIBD: Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder; UE4W: Unlabeled Empatica E4 Wristband; WEEE: Wearable Human Energy Expenditure Estimation; WESAD: Wearable Stress and Affect Detection; WESD: Wearable Exam Stress Dataset.

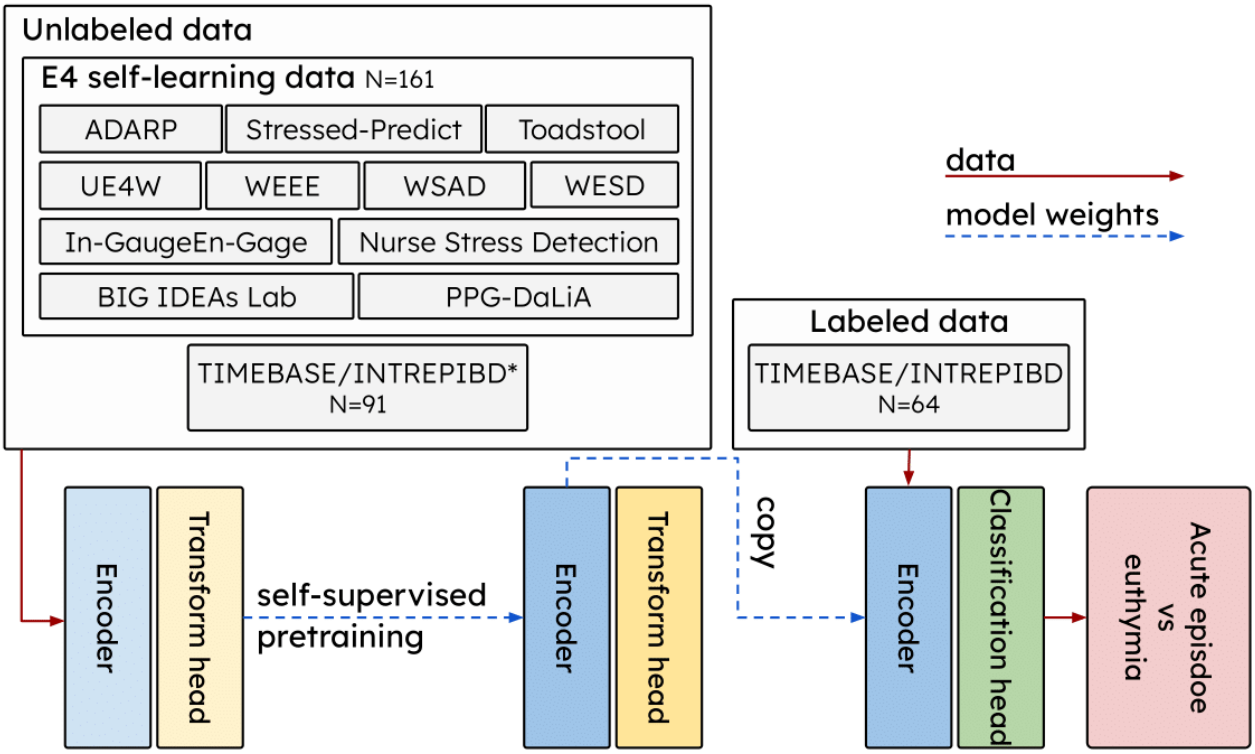
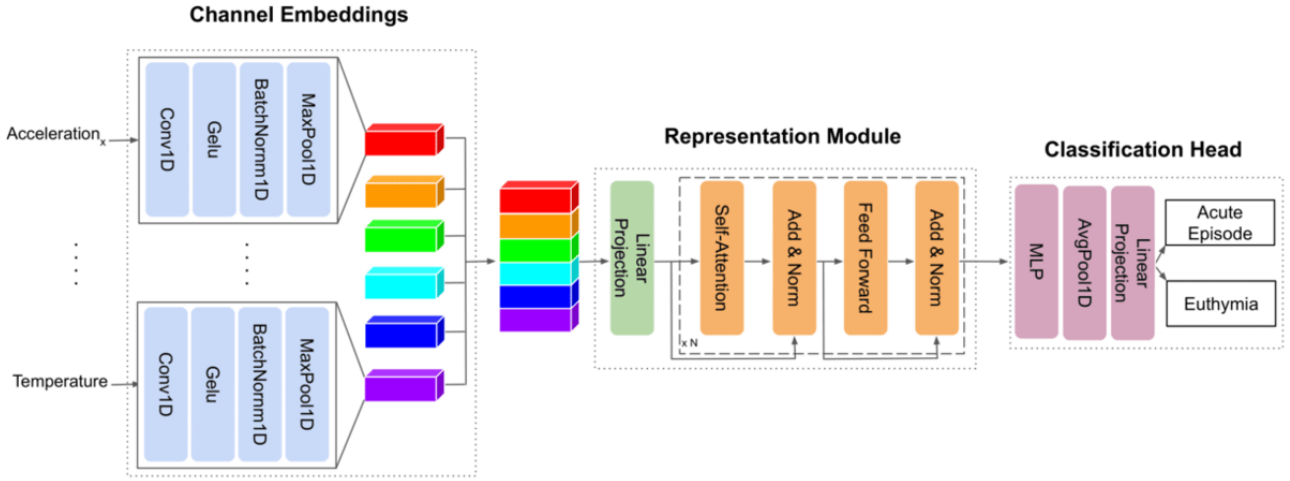


Figure 2. E4mer is a transformer model tailored to the Empatica E4 input data. E4mer consists of 3 sequential modules: (1) channel embeddings set in parallel, 1 for each Empatica E4 raw input channel (ie, acceleration$_x$, acceleration$_y$, acceleration$_z$, BVP, EDA, TEMP), extracting features and mapping channels to tensors of dimensionality (B=batch size, N=time steps, F=number of filters) so that they can be conveniently concatenated along dimension F; (2) RM learning contextual representations of the input time steps within the input segment, thanks to the multihead self-attention mechanism; (3) classification head outputting probabilities for the 2 target classes (ie, acute MD episode and euthymia). SSL models used in our experiments featured the same E4mer architecture described before, where, however, the classification head was replaced with a transform head projecting onto a label space compatible with the pretext task at hand. BVP: blood volume pressure; E4mer: E4-tailored transformer; EDA: electrodermal activity; MD: mood disorder; MLP: multilayer perceptron; RM: representation module; SSL: self-supervised learning; TEMP: temperature.



Methods

Study Sample

The TIMEBASE/INTREPIBD Cohort

Our target task was to distinguish acute MD episodes from euthymia using wearable data. We started from a data set for which we had labeled samples, the TIMEBASE/INTREPIBD (Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder) cohort [45]. A detailed description of the data collection campaign was given by Anmella et al [45]. In brief, this was a prospective, exploratory, observational study conducted at the Hospital Clìnic, Barcelona, Spain. Patients with a DSM-5 diagnosis of either major depressive disorder (MDD) or bipolar disorder (BD) were enrolled either in the acute affective episode group (defined according to the "Structured Clinical Interview" for DSM-5 disorder criteria) or in the euthymia group (score≤7 on the Hamilton Depression Rating Scale-17 [46] and the Young Mania Rating Scale [47] for at least 8 weeks [48], as confirmed with weekly ambulatory assessments). The former group had post–acute-phase follow-ups, which were, however, excluded from all analyses presented here. At the time of conducting this study, a total of 64 patients were available for the target task, half in the acute affective episode group and half in the euthymia group. Additionally, an extra 91 subjects (including healthy controls, subjects with schizophrenia, and subjects with a substance abuse disorder), whose status was not relevant to the target task, were available from the TIMEBASE/INTREPIBD cohort for self-supervised pretraining.

Patients were interviewed by a psychiatrist collecting clinical demographics (Table 1 and Table S2 in Multimedia Appendix 1) and were required to wear on their nondominant wrist an E4 wristband until the battery ran out (~48 hours). The E4 records 3D acceleration (sampling rate 32 Hz), blood volume pressure (BVP, sampling rate 64 Hz), electrodermal activity (EDA, sampling rate 4 Hz), heart rate (HR, sampling rate 1 Hz), interbeat interval (IBI, ie, the time between 2 consecutive heart ventricular contractions), and skin temperature (TEMP, sampling rate 1 Hz).

As shown in Table 1, MD episodes clinically lie on a spectrum, with depression on one end and mania on the other; mixed episodes, featuring symptoms from both polarities, are a bridge between the 2 spectrum extremes. In this study, we considered acute MD episodes of any polarity, and similarly, we considered euthymia as a unique class, whether in the context of a BD or an MDD diagnosis. Medication classes administered to the cohort are shown in Table S2 in Multimedia Appendix 1; Bonferroni-corrected chi-square tests found no significant association between treatment status (being on a given drug class or not) and target class (acute affective episode vs euthymia).

Table 1. Clinical-demographic features of the target task (acute affective episode vs euthymia classification) population (N=64).

Features	Acute affective episode group (n=32)	Euthymia group (n=32)
Age (years), means (SD)	50.56 (13.05)	47.22 (16.06)
Females, n (%)	15 (46.9%)	14 (43.8%)
MDE-BD[a]		
Patients, n (%)	9 (28.1)	—[b]
HDRS[c] score, mean (SD)	20.22 (6.34)	—
YMRS[d] score, mean (SD)	2.56 (3.94)	—
MDE-MDD[e]		
Patients, n (%)	7 (21.9)	—
HDRS score, mean (SD)	25.14 (4.78)	—
YMRS score, mean (SD)	1.86 (2.41)	—
ME[f]		
Patients, n (%)	14 (43.8)	—
HDRS score, mean (SD)	5.67 (4.37)	—
YMRS score, mean (SD)	20.13 (6.28)	—
MX[g]		
Patients, n (%)	2 (6.2)	—
HDRS score, mean (SD)	16 (4.24)	—
YMRS score, mean (SD)	13.5 (4.95)	—
BD[h]		
Patients, n (%)	—	26 (81.3)
HDRS score, mean (SD)	—	2.93 (1.73)
YMRS score, mean (SD)	—	1.3 (1.61)
MDD[i]		
Patients, n (%)	—	6 (18.7)
HDRS score, mean (SD)	—	3.14 (1.95)
YMRS score, mean (SD)	—	0.29 (0.76)

[a]MDE-BD: major depressive episode in bipolar disorder.

[b]Not applicable.

[c]HDRS: Hamilton Depression Rating Scale-17.

[d]YMRS: Young Mania Rating Scale.

[e]MDE-MDD: major depressive episode in major depressive disorder.

[f]ME: manic episode.

[g]MX: mixed episode.

[h]BD: bipolar disorder.

[i]HDRS: Hamilton Depression Rating Scale-17.

E4SelfLearning

For self-supervised pretraining, we gathered 11 open access data sets recording with an E4 [23-33]. Although they all used the same hardware, software, and firmware, such data sets could differ substantially for population, recording setting, and task: from students taking exams [29] or attending classes [31] to nurses carrying out their duty [30] and subjects performing different physical activities [28] or playing Super Mario [27]. Subjects that were not part of the target classes from the TIMEBASE/INTREPIBD study were also included in the unlabeled data for SSL.

Data Preprocessing

Our preprocessing encompassed the following sequential stages: on-/off-body detection, sleep/wake detection, segmentation, and (when preparing data for CML models) feature extraction.

During free-living wear, subjects might remove their device or contact with the wrist might be suboptimal. As a result, off-body periods can be erroneously mistaken for periods of sleep or sedentary behavior, due to the shared feature of an absence of movement. Signal discontinuity in biopotentials, such as EDA, due to a lack of skin contact can be reliably leveraged to detect nonwear periods. As shown by Vieluf et al [49] and Nasseri et al [50], we considered measurements less than 0.05 µS as indicative of off-body status. Furthermore, as we noticed occurrences of values greater than the EDA sensor range (ie, 100 µS [51]), as well as instances of TEMP values outside the physiological range (30°C-40°C), we set both to off-body.

As physiological data vary wildly across sleep and wake statuses, we used sleep/wake detection as a form of data cleaning to reduce the variance in the signal and considered only the wake time in our analyses, especially as most publicly available data sets are recorded in wake conditions. We opted for the algorithm developed by Van Hees et al (*Van Hees*) [52], which was reported as the best-performing algorithm in a recent benchmark study on sleep/wake detection (average F_1-score=79.1) [53]. Like most nonproprietary algorithms, Van Hees uses triaxial acceleration and, specifically, relies on a simple heuristic defining sleep with the absence of a change in the arm angle >5° for 5 minutes or more. To accommodate this rule, wherever on-body sampling cycles did not constitute unbroken sequences of at least a 5-minute duration, all the measurements in that period were considered as off-body and discarded from further analysis.

The wake time from each recording was then segmented with a sliding window, whose segment length (ω) and step size (Δω) were set to 512 and 128 seconds, respectively. This approach, also referred to as window slicing [54], is a common form of data augmentation in time-series classification as multiple segments are produced from a single recording, each one marked with the same label, and is common in personal sensing for MDs. Previous relevant works [15,18,55] have defined ω (Δω) based on clinical intuition and convenience concerning the available data. Another work [35] investigating the regression of HDRS and the YMRS items found the optimal ω through tuning, a computationally expensive approach in our setting; however, it showed that ω was not among the most important hyperparameters for the task at hand. Here, we opted for 512 seconds (~8.5 minutes, conveniently a power of 2 for computational efficiency in binary computers), similar to the 5-minute intervals used by Panagiotou et al [55] for training neural autoencoder architectures on anomaly detection by reconstruction error estimation. Our choice was a trade-off between clinical insight and technical constraints. Clinical intuition suggests that too small a value of ω may be ill suited to capture enough information toward acute affective episode versus euthymia discrimination. However, unlabeled data sets used for self-supervised pretraining recorded relatively short sessions (eg 1 hour [26]). As both CML and deep learning models are trained on individual segments and too long a segment length equates to fewer training data points, a 512-second-long segment allowed us to have enough data for developing ML models [55].

Recording segments constituted our basic unit of analysis, and for the target task, segments from the same recording all shared the same ground-truth label (ie, either acute affective episode or euthymia). When fed to deep learning models, segments were channel-wise standardized by subtracting the mean and dividing by the SD. Such statistics were learned from the target task training set or, in the case of SSL, its aggregation with unlabeled data. Acceleration, the BVP, EDA, and TEMP were considered in deep learning models, while the HR and the IBI, as features derived from the BVP through a proprietary algorithm, were excluded from the deep learning experiments shown here (see Multimedia Appendix 1). However, when using CML, handcrafted features were extracted from segments using *FLIRT* [56], a popular open access feature extraction toolkit for the E4. Note that a single row of features per segment was extracted; in other words, the window size parameter in FLIRT was set equal to ω. We used all features available through this package, derived with the *flirt.acc.get_acc_features* (eg, acceleration entropy), *flirt.eda.get_eda_features* (eg, tonic and phasic EDA components), and *flirt.hrv.get_hrv_features* (eg, HR and HR variability measures) functions. As FLIRT does provide built-in functions for TEMP, we also extracted the segment mean (SD) for this channel. Any missing value was handled with mean imputation. The percentage rate of missing values had a range of 0-37.31, with a mean of 10.44 (SD 16.78).

Data Splits and Metrics

In SSL experiments, we split unlabeled data in a ratio of 85:15 into train and validation sets, partitioning recordings across the 2 sets. For the target task, we investigated a time-split scenario, therefore splitting each recording into train, validation, and test sets again in a ratio of 70:15:15 along the recording time, thus testing generalization across future time points. We made sure that segments with overlapping motifs at the border between target task splits (resulting from using a sliding window with $\Delta\omega<\omega$) were confined to 1 split only, thus ultimately producing 18896, 3904, and 4128 segments for the train, validation, and test sets. The target task validation set doubled as a test set for estimating generalization performance on the SSL pretext task. The time-split scenario is common in personal sensing for MDs (eg, [18,35]), and indeed, despite efforts toward learning subject-invariant representations [57,58], cross-subject generalization remains an unsolved challenge, so personal sensing systems typically require access to each subject's physiological data distribution at training time [59].

The target task was a time-series binary classification. As expected in free-living wear, the total wear time and the off-body and wake times varies across subjects (and, as a result, so did the number of segments). Two-tailed *t* tests were performed to verify significant mean differences in off-body and wake times across individuals from the 2 target classes (acute affective episode and euthymia) but yielded a Bonferroni-corrected P value of >.05 (P=.56 for off-body time and P=.82 for wake time). An equal number of segments from each class was extracted for the target task. To that end, we found the pairing of euthymia and acute affective episode recordings that minimized the pairwise difference between the number of segments available per participant; next, within each pair, the first n segments were retained, where n is the number of

segments of the shortest recording in the pair. We optimized models on the target task for segment-level accuracy ($ACC_{segment}$). Second, to provide a subject-level perspective, we reported the subject ACC:

$$ACC_{subject} = \frac{1}{S}\sum_{s=1}^{S} 1(\hat{y}_s = y_s),$$

where y_s is the ground-truth mood state of the s-th subject, which is constant across all the s-th subject's recording segments, and $\hat{y}_s$ is a majority vote on the s-th subject, corresponding to the majority predicted class across the s-th subject's recording segments.

Machine Learning Models

We developed 2 types of baselines for the target task: (1) an E4-tailored deep learning pipeline inputting raw recording segments (E4mer) and (2) CML models using handcrafted features extracted with FLIRT from recording segments. We then assessed what boost in performance, if any, a self-supervised pretraining phase might deliver, where the SSL models shared the same building blocks as E4mer.

Baseline Models

E4-Tailored Transformer

E4mer is an artificial neural network discriminative classifier modeling the probability of an acute MD episode, given a recording segment. As shown in Figure 2, E4mer has 3 sequential blocks: (1) channel embeddings (CEs) set in parallel, consisting of the same 1D convolutions with a kernel size equal to the channel sampling frequency, followed by Gaussian error linear unit (Gelu) activation, 1D BatchNorm, and 1D MaxPooling using the channel sampling frequency as both kernel size and step size, so each CE output has the same dimensionality and can be conveniently concatenated with the others before being passed onto (2) a transformer [43] representation module (RM), and (3) a multilayer perceptron (MLP) classification head (H_{sl}). The CEs extract features from the input E4 channels and are designed to handle channels sampled at different frequencies; the RM, powered by multihead self-attention, learns contextual representations of the input tokens (timestamps in our case) within a recording segment; lastly, the H_{sl} maps such representations onto a label space appropriate for a binary classification. E4mer was trained to minimize the binary cross-entropy (BCE) loss between acute affective episode/euthymia predictions and the corresponding ground truth.

Classical Machine Learning

We experimented with the following algorithms, given their popularity and state-of-the-art performance in biomedical applications [60], including personal sensing [13,14]: elastic net logistic regression (ENET), K-nearest neighbor (KNN), support vector machine (SVM), and extreme gradient boosting (XGBoost).

Self-Supervised Learning Schemes

SSL schemes rely on devising a pretext task, for which a (relatively) large amount of unlabeled data is available, conducive to learning, during a pretraining phase, representations useful to solve the downstream target task [44]. What defines an SSL paradigm is thus its pretext task, consisting of a process, P, to generate pseudo labels and an objective to guide the pretraining. An SSL model typically consists of (1) an encoder EN(x;θ): X->V, learning a mapping from input views $x \in X$ to a representation vector $v \in R^d$, and (2) a transform head $H_{ssl}(\upsilon;\xi)$: V-> Z, projecting the feature embedding into a label space $z \in R^{d\prime}$ compatible with the pretext task at hand. When solving the target task, the pretrained encoder EN is retained as a partial solution to the target problem, whereas the pretrained transform head H_{ssl} is discarded and replaced with a new one, H_{sl}. Next, EN's parameter θ may be kept fixed and only H_{sl}'s parameters may be learned on the target task. This approach, often referred to as *linear readout* (LR), amounts to treating EN as a frozen feature extractor. Alternatively, instead of just training a new head, the entire network may be retrained on the target task, initializing EN's parameter θ to the values learned during self-supervised pretraining, a paradigm known as *fine-tuning* (FT). Our SSL models used the same architecture as E4mer, that is, an encoder EN, consisting of convolutional CEs, followed by a transformer RM, and an MLP for the transform head H_{ssl}. The success of SSL methods largely comes from designing appropriate pretext tasks that produce representations useful for the downstream target task. This usually involves domain knowledge of the target task. We investigated how different pretext tasks affected downstream performance, experimenting with 2 popular SSL routines that have shown success in other applications: MP and TP.

Masked Prediction

This family of SSL methods is characterized by training the model to impute data that have been removed or corrupted by P. It relies on the assumption that context can be used to infer some types of missing information in the data if the domain is well modeled. This strategy was popularized by the huge success of bidirectional encoder representations from transformers (BERT) [38] in NLP applications, and 1 of the first adaptations to multivariate time-series classification was proposed by Zerveas et al [41]. Similar to their implementation, for each segment channel, we sampled a Boolean mask where the sequences of 0s and 1s were sampled from geometric distributions with means of l_0 and l_1, respectively, with:

$$l_1 = \frac{1-r}{r} l_0,$$

where r is the masking ratio. As shown by Zerveas et al [41], the average length of the 0 sequences (l_m) and the proportion of masked values (r) were set to 3 seconds and 0.15, respectively. Each segment channel was then multiplied by its corresponding mask, effectively setting to 0 some of the channel-recorded measurements, and inputted to a model that was tasked to recover the original channel values. This was done by minimizing the root mean square error (RMSE) between the masked original value x(t, c) and its reconstruction outputted by the network $\hat{x}(t, c)$:

$$\mathcal{L}_{\text{RMSE}} = \sqrt{\frac{1}{|M|}\sum_{t\in M}\sum_{c\in M}(\hat{x}(t,c) - x(t,c))^2},$$

where c and t, respectively, index the channels, and the timestamps of the 0 values in the masks M and |M| are the total number of 0s sampled (ie, the masks' cardinality).

Transformation Prediction

We followed the implementation shown by Wu et al [42], which used SSL for a target task of emotion recognition with E4 recordings. In brief, for each channel, 1 of 6 transformations (ie, identity, Gaussian noise addition, magnitude warping, permutation, time warping, and cropping) was sampled uniformly at random and then applied. The transformed segment was then inputted into a model, which was tasked to guess, for each channel, which of the 6 transformations was applied. This amounted to a multitask, multiclass classification, where the model was trained to minimize channel average categorical cross-entropy (CCE):

$$\mathcal{L}_{\text{CCE}_{\text{Multitask}}} = \frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{T} -1_{c,j}\cdot\log(p_{c,j}),$$

where c indexes the channels and j the transformations, $1_{i,j}$ is an indicator taking value 1 when j is the correct transformation for channel c and 0 otherwise, and $p_{c,j}$ denotes the predicted probability that transformation j was applied to channel c. By solving this task, Wu et al [42] argued that the model learns representations robust to disturbances in the magnitude and time domains.

Tuning

A hyperparameter search for all models was carried out with hyperband Bayesian optimization [61]. For the target task, we selected the setting yielding the highest $ACC_{segment}$ in the validation set, whereas in self-supervised pretraining, we selected hyperparameters associated with the lowest relevant loss in the validation pretraining set. Multimedia Appendix 1 shows the hyperparameter search space and the best configuration across all models. Deep learning models were trained with the AdamW optimizer for a maximum of 300 epochs, with a batch size of 256. Moreover, to speed up the training and search procedure, we used an early stopping learning rate scheduler: we reduced the learning rate α_{LR} by a factor of 0.3 if the model did not improve in its validation performance after 10 consecutive epochs, and we terminated the training procedure if the model did not improve after 2 learning rate reductions. Dropout [62] and weight decay were added to prevent overfitting.

Post hoc Analyses

Toward elucidating key contributors to the viability of SSL, in addition to comparing different pretext task designs, we studied how (1) progressively downsampling unlabeled data sets or (2) removing each data set in turn from the unlabeled collection might impact the performance of our best SSL model. Thus, using the most performative self-supervised scheme, we retrained the SSL model from scratch under configurations (1) and (2) and then tested it on the target task. Note that in both settings, the entire target task training set was kept for pretraining; this is because pretraining on the training set can be always performed at no extra cost in terms of data acquisition. Lastly, we conducted statistical tests to better appreciate how the self-supervised E4mer compared against its fully supervised counterpart and the best-performing CML algorithm and how it was affected by different ablations. Based on whether we considered either (1) recording segments or (2) subjects as our basic analysis units, we had 2 different hypotheses. In (1), we used a linear mixed effects (LME) model to analyze the difference in correct class probabilities between the SSL model and each comparator, considering subjects as a random effect. This accounted for the nested structure of the data, where segments were sampled from individual subjects. A fixed effects intercept was included to test a 0 mean difference between the classifiers at the population level. Additionally, as the ML models we implemented, like most state-of-the-art algorithms [63], effectively treat segments as independent and identically distributed, we used a 2-tailed paired *t* test to assess whether a 0 mean difference in the probability assigned to the correct class was 0. In (2), we checked with a 2-tailed paired *t* test whether the between-classifiers mean difference in the $ACC_{segment}$ by subject was different from 0. To account for multiple testing, within both (1) and (2), a Bonferroni correction was applied. The number of tests was 19, that is, 17 different ablation settings plus 2 tests comparing the best baselines (fully supervised E4mer and the best CML) to SSL.

Code Used

Python 3.10 programming language was used where deep learning and CML models were implemented in PyTorch [64] and Scikit-learn [65]/XGBoost [66] respectively, while hyperparameter tuning was performed in both cases with weights and biases [67]. The best hyperparameter setting found during tuning for each model is reported in Multimedia Appendix 1. All deep learning models were trained on a single Nvidia A100 graphical processing unit (GPU).

Ethical Considerations

The TIMEBASE/INTREPIBD study was conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practice and the Hospital Clinic Ethics and Research Board (HCB/2021/104). All participants provided written informed consent prior to their inclusion in the study. All data were collected anonymously and stored encrypted in servers complying with the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Regarding other studies included in this work, we referred to relevant publications.

Results

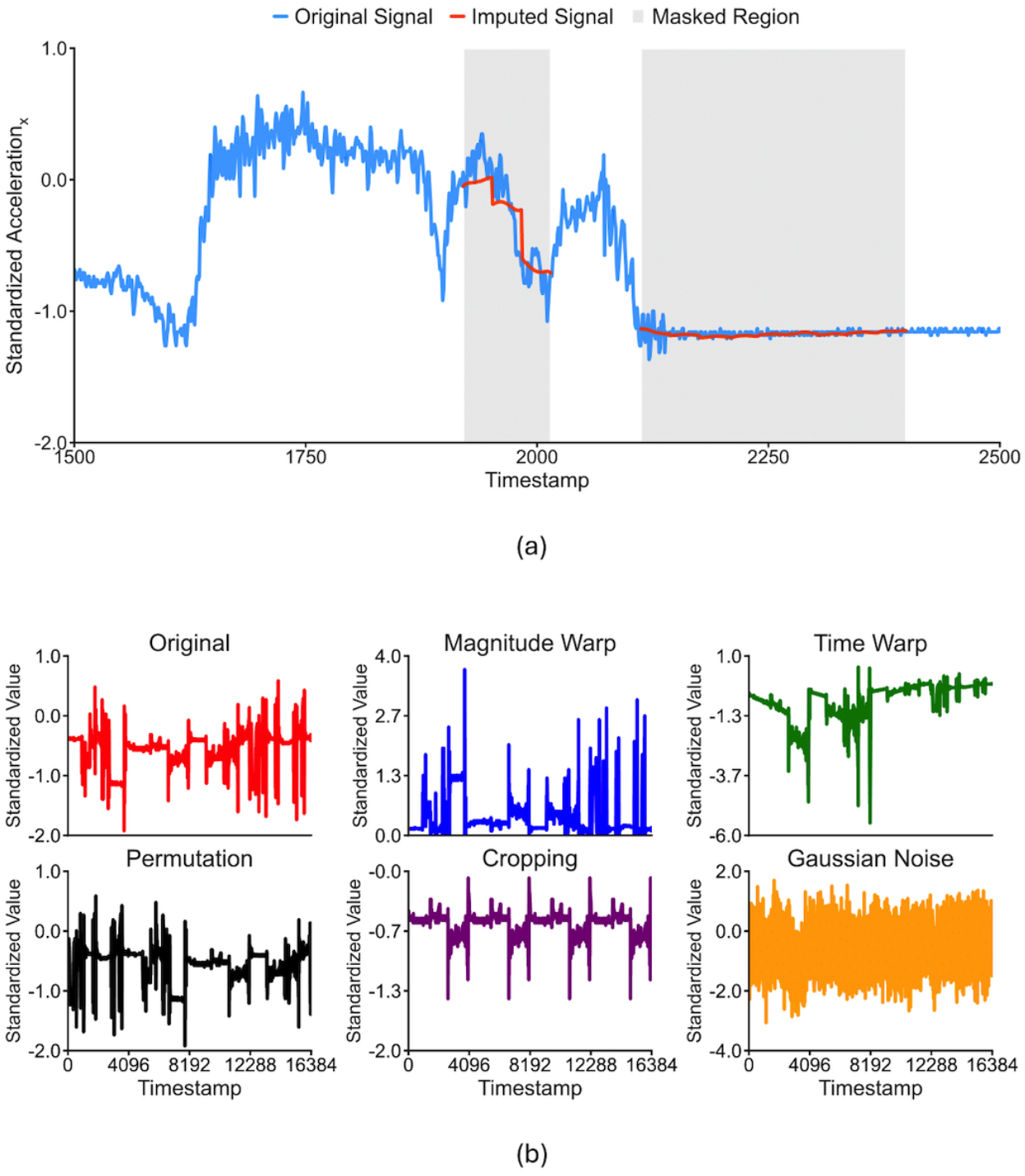
Surrogate Tasks Used in Self-Supervised Pretraining

The same model, using the E4mer architecture (Figure 2), was used across different pretext tasks. Figure 3 illustrates the surrogate tasks we experimented with. In MP (Figure 3a), parts of the input segments were zeroed out by multiplication with a Boolean mask sampled, as shown by Zerveas et al [41], and the

model was trained to recover the original input segments. Although the model output entire segments, only the masked values were considered toward the loss computation, that is, the RMSE. The assumption was that the model acquires good representations of the underlying structure of the data when learning to solve this task. Our best model had an error of 0.1347 on the test set (notice that input segments were channel-wise standardized).

In TP (Figure 3b), 1 transformation was sampled from a set and applied to each channel independently, and the model learned which transformation each channel underwent, minimizing the channel average CCE. We used the same transformations as Wu et al [42], who experimented with an E4 for a downstream task of emotion recognition. The rationale was to encourage robustness against signal disturbances introduced with the transformations. The test loss of the selected model was 0.5000.

Figure 3.



Target Task Performance Comparison

Table 2 illustrates the performance under each model we developed. Although they were all optimized for segment ACC, we also reported subject ACC since in a clinical scenario, a decision needs to be made at the subject level. Note that although ACC was a suitable metric in our use case as data were perfectly balanced, we also provided complementary metrics (precision, recall, F_1-score, and area under the receiver operating characteristic curve [AUROC]), both at the segment and at the patient level. At the subject level, the predicted class was the result of a majority vote over that subject's segments, while the predicted probabilities under each class were derived by summing segments' predicted probabilities for that subject and normalizing by the corresponding segment number. MP self-supervised pretraining comfortably outperformed end-to-end SSL, while also surpassing other self-supervised approaches.

Table 2. Performance in differentiating an acute MD[a] episode from euthymia across different models.

Model	ACC[b]		Precision		Recall		F_1 score		AUROC[c]	
	Segment	Subject	Segment	Subject	Segment	Subject	Segment	Subject	Segment	Subject
SL[d]										
ENET[e]	66.38	71.88	66.22	75	66.86	65.63	66.54	70	72.24	82.25
KNN[f]	70.37	82.81	69.09	80	73.74	81.2	71.34	80.6	73.27	83.26
SVM[g]	71.25	81.25	71.87	80	71.40	77.65	71.63	78.81	73.44	83.21
XGBoost[h]	72.02	82.81	71.33	83	72.11	81.1	71.72	82.03	72.44	83.17
E4mer[i]	75.35	81.25	73.46	80.55	75.34	82.14	74.39	81.33	75.68	82.22
SSL[j]										
MP[k] (LR[l])	77.53	87.5	78.34	88.6	77.41	88	77.87	88.3	78.02	89.2
MP (FT[m])	81.23[n]	90.63[n]	80.91[n]	90.11[n]	82.00[n]	92.87[n]	81.45[n]	91.47[n]	82.02[n]	93.11[n]
TP[o] (LR)	71.16	81.25	72.12	82.44	72.01	82.31	72.06	82.37	71.89	84.12
TP (FT)	75.69	84.38	75.41	82.11	74.79	83.9	75.1	83	75.21	84.23

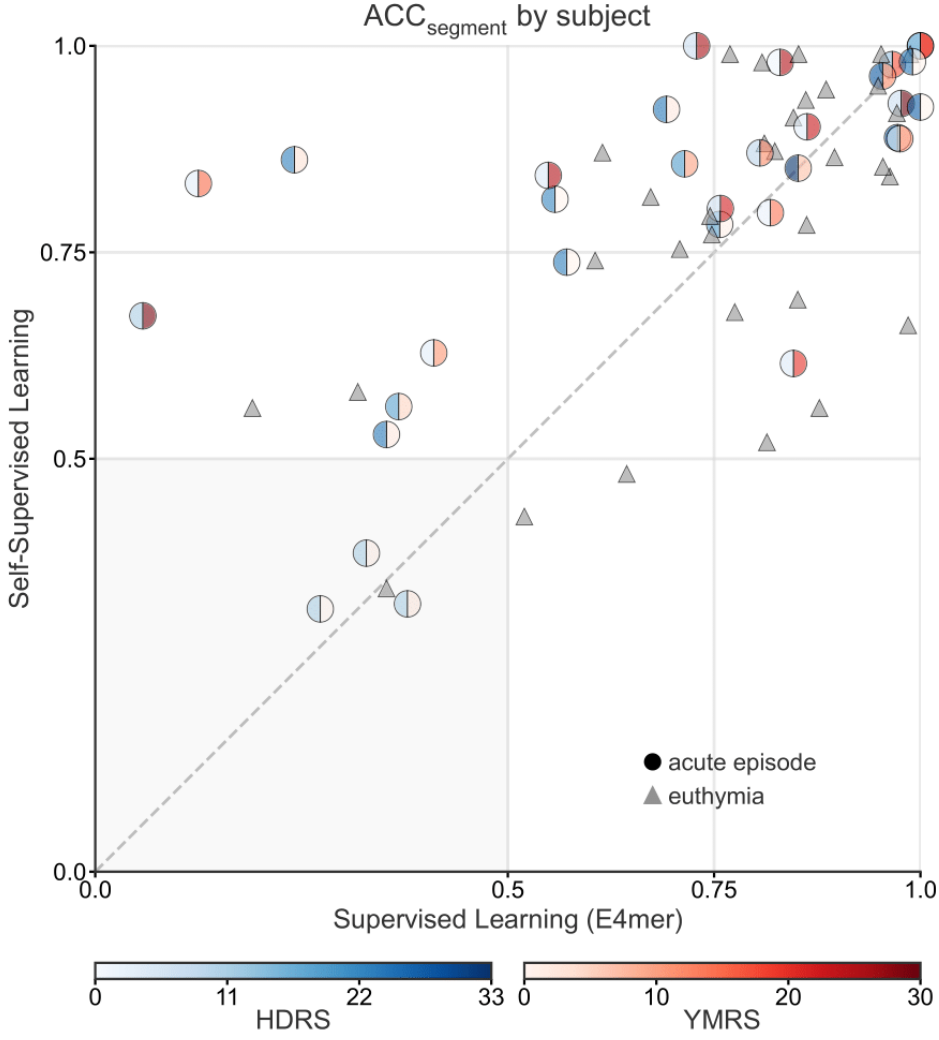
[a]MD: mood disorder.
[b]ACC: accuracy.
[c]AUROC: area under the receiver operating characteristic curve.
[d]SL: supervised learning.
[e]ENET: elastic net logistic regression.
[f]KNN: K-nearest neighbor.
[g]SVM: support vector machine.
[h]XGBoost: extreme gradient boosting.
[i]E4mer: E4-tailored transformer.
[j]SSL: self-supervised learning
[k]MP: masked prediction.
[l]LR: linear readout.
[m]FT: fine-tuning.
[n]The best results.
[o]TP: transformation prediction.

The E4mer and CML baselines performed to a similar level: although E4mer was superior to XGBoost in terms of $ACC_{segment}$ (75.35 vs 72.02), it was trumped by CML on $ACC_{subject}$ (82.81 vs 81.25). Other CML baselines fared worse than XGBoost. MP pretraining led to a target task performance, substantially higher than the baselines, under both metrics. Although both LR and FT dominated over supervised learning (SL), the latter scored the highest performance with $ACC_{segment}$ and $ACC_{subject}$ of 0.8123 and 0.9063, respectively. However, TP led to only modest improvement over E4mer. Statistical tests comparing the best SSL scheme (ie, MP with FT) against the fully supervised E4mer and XGBoost were significant at both the segment and the subject level. In particular, comparison with E4mer yielded $P_{Bonferroni}$ values of .03 for the LME model and <.001 and .02 for the *t* test at the segment and the subject level, respectively. For XGBoost, $P_{Bonferroni}$ values were .04 for the LME model and <.01 and .01 for the *t* test at the segment and the subject level, respectively.

Comparison of the best SSL with its SL counterpart in terms of $ACC_{segment}$ by subject (Figure 4) suggested that only 2 (3.1%)

patients with euthymia were misclassified by SSL but correctly classified by the supervised E4mer. However, SL mispredicted 8 (12.5%) individuals that SSL got right. Patients on an acute MD episode are shown as dots with a color gradient proportional to their total score on the HDRS [46] (left half) and the YMRS [47] (right half), 2 clinician-administered questionnaires tracking depression and mania severity, respectively. Subjects on an acute MD episode misclassified by SL included patients with severe depressive (or manic) symptomatology. Notably, both SSL and SL failed in the case of 4 (6.3%) subjects, including 3 (75%) patients on an acute MD episode with relatively moderate severity.

Figure 4. SSL beats SL by 4 (9.4%) more correctly classified subjects. ACCsegment under SSL and SL (E4mer) within each subject's test segments: subjects in the euthymia group are represented as triangles, while subjects on an acute affective episode are shown as circles with the left half colored in blue and the right half in red, with a gradient proportional to the total sum on the HDRS and the YMRS, respectively. Subjects' position on the x and y axes corresponds to their proportion of recording segments correctly classified by SL and SSL, respectively. Note that a subject's majority vote over their segments is in agreement with the subject's true mood state when the proportion of correctly classified segments from that subject is greater than 0.5. The HDRS and the YMRS range shown on the color bar refer to values scored in the TIMEBASE/INTREPIBD sample, while the total score, in general, range is 0-52 and 0-60, respectively. ACCsegment: segment accuracy; E4mer: E4-tailored transformer; HDRS: Hamilton Depression Rating Scale-17; SL: supervised learning; SSL: self-supervised learning; TIMEBASE/INTREPIBD: Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder.



Ablation Analyses and Learned Representations

Tables 3 and 4 show the difference in the target task $ACC_{segment}$ and $ACC_{subject}$ resulting from pretraining the best SSL on parts of the unlabeled data collection and then FT it onto the target task. Ablation analyses showed a positive trend between unlabeled data availability and target task performance, but data set–specific unobserved factors likely played a role. The difference in $ACC_{segment}$ and $ACC_{subject}$ from pretraining on just parts of the entire unlabeled data collection is shown in the tables. An LME model and a 2-tailed paired *t* test assessed whether the mean difference in predicted probabilities for the segment's correct class differed from 0, with the former correcting for subjects as a random effect. A 2-tailed paired *t* test assessed whether the mean difference in the number of correctly classified segments by subject differed from 0. In each test, the comparator was the best-performing self-supervised model. P values are corrected with Bonferroni's method. Note that a majority vote over a subject's segments was used to issue subject-level predictions, and $ACC_{subject}$ was simply the fraction of correct majority votes in the test set. $ACC_{subject}$, therefore, did not consider the proportion of votes over a subject's

segments in favor of the subject's correct class but just whether a majority, no matter how small or large, was reached in agreement with the correct class. However, the *t* test (subject) assessed a 0 mean difference in the proportion of votes, within subjects, for the correct class. As shown in Table 3, self-supervised pretraining, preceding FT on the target task, therefore used only a fraction of the total unlabeled collection. A resampling ratio of 0% meant that self-supervised pretraining was performed on the target training set only.

The Pearson correlation coefficient (PCC) between unlabeled data downsampling ratios and the difference in $ACC_{segment}$ and $ACC_{subject}$ was 0.9401 and 0.9449, respectively, indicating a strong dependence between performance and unlabeled data availability. Similarly, excluding individual data sets from pretraining impacted $ACC_{segment}$ and $ACC_{subject}$ proportionally to their relative size (PCC=–0.8185 and –0.4083, respectively). Notably, however, TIMEBASE/INTREPIBD, despite being collected at the same site as the target task data and making up the largest share of the unlabeled data collection, did not leave the largest dent in performance when excluded from training. Furthermore, excluding some data sets resulted in performance improvement. Differences in $ACC_{segment}$ and $ACC_{subject}$ did not always have the same sign because of the way they were defined. Indeed, it is, for example, possible that the absolute number of correctly classified segments decreased but enough previously misclassified segments within a subject were now correctly classified so that the majority vote for that subject flipped. Statistical analyses showed that the ablation of a single data set was associated with nonsignificantly different performance in terms of correctly classified segments within subjects. At the level of the probability assigned to the correct class for each segment, LME results were significant only for a data set, whereas results were mixed for *t* tests. Stratified resampling gave positive results, but the significance for LME was reached only at lower downsampling ratios.

Lastly, we visualized the representations learned by the encoder, EN, part of our best-performing models to gain further insights. As EN's output had dimensionality (B=number of segments, N=number of timestamps, D=transformer's model dimension), for visualization purposes, we averaged out the D axis and then used Uniform Manifold Approximation and Projection (UMAP) [68], a powerful nonlinear dimensionality reduction technique, to embed the resulting N-dimensional data points into 3 dimensions. The top-left plot of Figure 5 shows the representations learned during self-supervised pretraining with MP. The segments shown are the target task test segments, along with an equal number of segments belonging to the same sessions but taken from the sleep state, which the SSL model was never exposed to during training. Wake and sleep segments have different embeddings, suggesting that the model captured this structure in the physiological data: a Gaussian mixture model, indeed, recovered 2 clusters, one with predominantly sleep segments (n=4081, 82.66%) and the other with the majority of wake segments (n=3272, 95.58%). It should be noted that sleep and wake naturally have quite different semantics with respect to physiological data, and the algorithm we used for sleep/wake differentiation (Van Hees [52]) uses a simple heuristic defining sleep as a sustained lack of significant changes in the acceleration angle. The top-right and bottom plots of Figure 5 illustrate the representations from the SSL model upon FT on the target task. The top-right scatter plot displays the target task test segments, as well as pretraining validation set segments (except for the pretraining segments from the TIMEBASE/INTREPIBD collection). The latter group of segments we assumed as being taken from subjects without an acute MD episode and, arguably, most even without any historical MD diagnosis, since the open access data sets we found did not select for patients with an MD. The plot shows 3 clusters whose composition, as recovered with a Gaussian mixture model, was as follows: (1) n=1464 (79.26%) acute MD episode and n=383 (20.7%) euthymia; (2) n=1120 (74.16%) euthymia and n=390 (25.84%) acute MD episode; and (3) n=7801 (91.01%) unlabeled segments, n=683 (7.96%) euthymia, and n=88 (1.02%) acute MD episode. The bottom plots in Figure 5 show target task segments test segments only (no unlabeled segment), colored with a gradient proportional to symptoms' severity, as assessed with the HDRS [46] and the YMRS [47]. Embeddings would seem to suggest a progression in symptoms' severity across the 2 clusters of segments on the right of the scatter plot.

Table 3. Ablation analyses results: the unlabeled collection was downsampled, stratifying by data sets.

Resampling ratio	80%	60%	40%	20%	0%
$ACC^{a}_{segment}$ difference	–0.23[b]	–2.14[b]	–6.07[b]	–6.35[b]	–7.07[b]
$ACC_{subject}$ difference	–1.57[b]	–1.57[b]	–4.70[b]	–4.70[b]	–7.82[b]
LME[c] *P* value	.09	.07	.06	.05	.04
t Test (segment) *P* value	<.001	<.001	<.001	<.001	<.001
t Test (subject) *P* value	.001	.001	.001	.001	.001

[a]ACC: accuracy.

[b]Deterioration in performance upon retraining on the ablated unlabeled data collection.

[c]LME: linear mixed effects.

Table 4. Ablation analyses results: self-supervised pretraining was conducted, leaving out each data set in turn from the unlabeled collection.

Data set	Relative size	ACC[a]$_{segment}$ difference	ACC$_{subject}$ difference	LME[b] *P* value	*t* Test (segment) *P* value	*t* Test (subject) *P* value
Alcohol and Drug Abuse Research Program (ADARP)	12.34	–2.44[c]	–1.57[c]	.01	<.001	.99
Stress Predict	0.30	–0.21[c]	–1.57[c]	.23	.99	.99
Toadstool	0.04	0.52[d]	–3.13[c]	.99	.99	.99
Unlabeled Empatica E4 Wristband (UE4W)	2.32	–1.93[c]	–4.70[c]	.99	.003	.99
Wearable Human Energy Expenditure Estimation (WEEE)	0.18	1.19[d]	1.57[d]	.99	.90	.99
Wearable Stress and Affect Detection (WESAD)	0.42	–0.51[c]	–1.57[c]	.99	.99	.99
Wearable Exam Stress Dataset (WESD)	0.72	1.90[d]	1.57[d]	.06	.05	.99
In-GaugeEn-Gage	17.55	–4.44[c]	–4.70[c]	.99	<.001	.63
Nurse Stress Detection	11.82	–0.81[c]	–1.57[c]	.99	.99	.99
BIG IDEAs Lab	19.38	–2.09[c]	–1.57[c]	.99	.08	.99
PPG Dataset for Motion Compensation and Heart Rate Estimation in Daily Life Activities (PPG-DaLiA)	0.69	1.93[d]	4.70[d]	.53	<.001	.99
TIMEBASE/INTREPIBD[e]	34.24	–4.32[c]	–3.13[c]	.99	.99	.38

[a]ACC: accuracy.

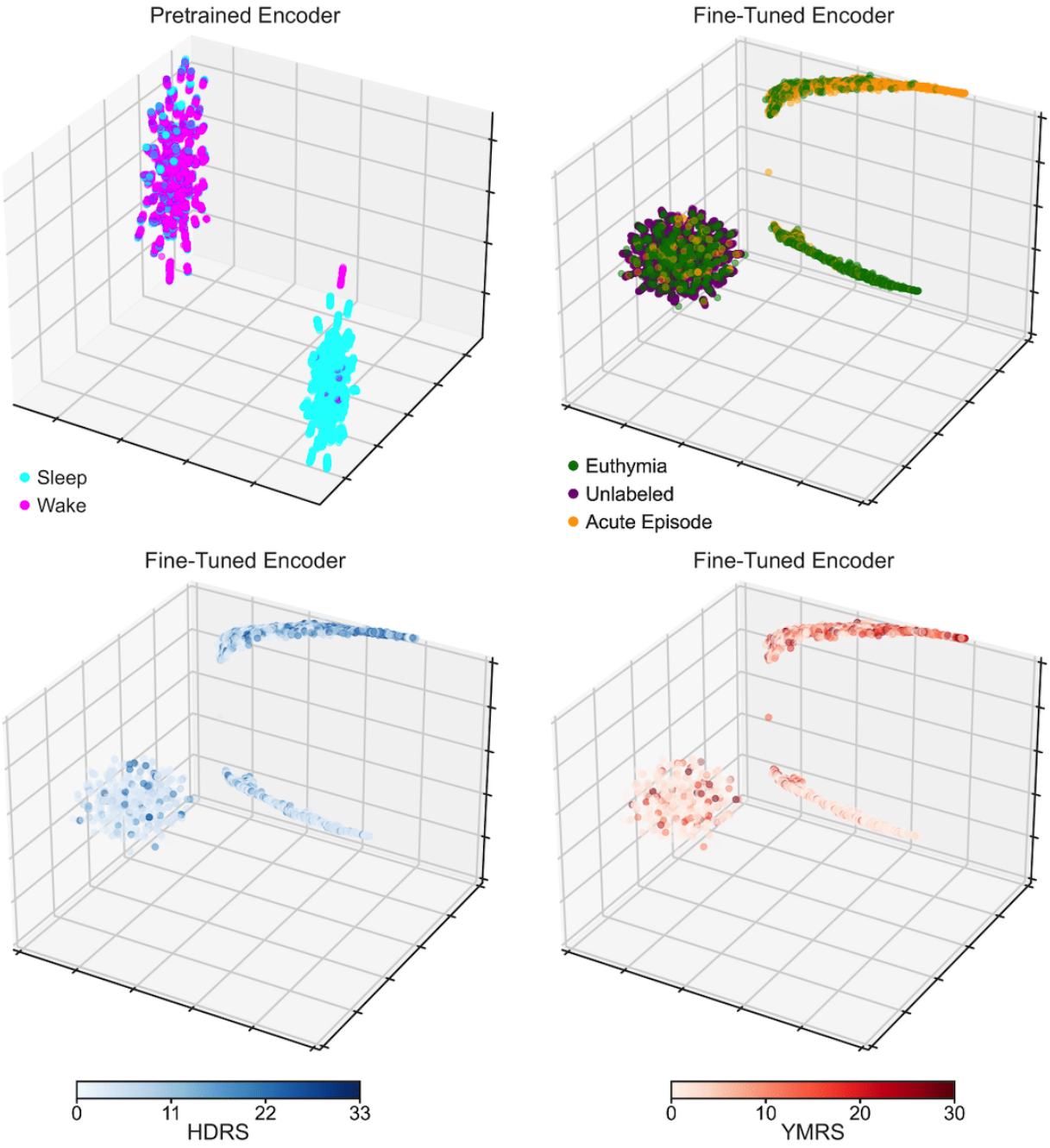
[b]LME: linear mixed effects.

[c]Deterioration in performance upon retraining on the ablated unlabeled data collection.

[d]Improvement in performance upon retraining on the ablated unlabeled data collection.

[e]TIMEBASE/INTREPIBD: Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder.

Figure 5. Reassuringly, the learned embeddings seem to have captured meaningful semantics about the underlying context. (Top left) Embeddings from the encoder pretrained on MP map sleep and wake segments to different parts of the latent space. (Top right) Embeddings from the encoder FT on the target task show that segments from the unlabeled open access data sets, which presumably do not contain subjects on an acute MD episode, tend to cluster with part of the segments from patients in euthymia. Embeddings from the fine-tuned encoder show a gradient in symptoms' severity across target task segments, as revealed by (bottom left) the HDRS and (bottom right) the YMRS total score. Note that unlabeled segments are not shown in the bottom left or right plot and that the HDRS and YMRS ranges shown on the color bar refer to values scored in the TIMEBASE/INTREPIBD sample, while the total score range, in general, can be 0-52 and 0-60, respectively. FT: fine-tuning; HDRS: Hamilton Depression Rating Scale-17; MD: mood disorder; MP: masked prediction; TIMEBASE/INTREPIBD: Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder; YMRS: Young Mania Rating Scale.



Discussion

Principal Findings

Personal sensing is likely to play a key role in health care supply, creating unprecedented opportunities for patient monitoring and just-in-time adaptive interventions [69]. Toward delivering on this promise, expert annotation is a major obstacle; this is especially the case with MDs, wherein data annotation is particularly challenging and time-consuming, considering the nature of the disorders.

To the best of our knowledge, we are the first to show that SSL is a viable paradigm in personal sensing for MDs, mitigating the annotation bottleneck, thanks to the repurposing of existing unlabeled data collected in settings as different as subjects playing Super Mario [27], taking university exams [29], or performing physical exercise [28].

We took on a straightforward yet fundamental task, that is, to distinguish acute MD episodes from euthymia. Timely recognition of an impending MD episode in someone with a historical MD diagnosis regardless of the episode polarity (depressive, manic, or mixed) may, indeed, enable preemptive interventions and better outcomes [5]. Our results suggest that with a sample size on the order of magnitude that is typical of studies into personal sensing for MDs, a modern deep learning fully supervised pipeline (E4mer) may offer no substantial improvements over simpler CML algorithms (eg, XGBoost), despite higher development and computational costs. However, the accumulation and repurposing of existing unlabeled data sets for an SSL pretraining phase leads to a confident margin of improvement: $ACC_{segment}$ and $ACC_{subject}$ improve by 7.8% and 11.54%, respectively, relative to the fully supervised E4mer, with 6 (9.4%) of 64 more subjects correctly classified.

Our findings further show that careful choice of the pretext task, as well documented in the literature on SSL [40], is key toward learning useful representations for the downstream target task. Unlike MP, improvement, if any at all, from TP was only modest. This is not to say that such a pretext task may in general fail to deliver on acute MD episode versus euthymia differentiation. Indeed, the specific transformations we implemented, borrowed from Wu et al [42], may have been suboptimal for our downstream task, pointing to the importance of domain knowledge (including clinical expertise) in pretext task design. Lastly, although SSL relaxes dependence on large, annotated data sets, our results indicate that its success relies on the size of unlabeled data. Ablation analyses, indeed, showed a positive correlation between target task performance and the size of the corpus available for pretraining. Data set–idiosyncratic factors accounting for the nonperfect correlation between the relative size and impact on target task performance may be present. Speculatively, these may include noise in the data, (dis)similarity of recording conditions, or (ir)relevance for the target task of the representations learned modeling the domain of the unlabeled data set.

Statistical analyses showed that excluding from pretraining any of the individual unlabeled data sets, while keeping all others, is not associated with a significant change in performance on the proportion of correctly classified segments within subjects. The lack of a significant effect in either direction (improvement or deterioration), along with a significantly superior performance of SSL over fully supervised schemes, indicate that pretraining on big data collections leads to higher performance than taking on the target task from scratch. Of importance, adding data sets for pretraining from domains not immediately related to the target task did not undermine the model. Pretraining under progressively lower downsampling ratios lent further support to the importance of data size. This is consistent with the deep learning recipe where the bigger the pretraining corpus, the better the results [70]. Results from tests at the level of segment-predicted probabilities are consistent with this view. Of the data sets comprising less than 1% of the entire unlabeled collection, only 1 reached statistical significance. LME has more flexibility to explain the data since rather than pooling all segments together in a unique (bigger) population, it treats them as embedded within subjects. This explains the lack of statistical significance relative to the *t* tests under various data ablation regimes.

Limitations

We acknowledge the following limitations of this study. We deliberately chose the simplest task that has clinical relevance in personal sensing for MDs since our focus was on SSL; however, we appreciate that a more fine-grained MD description, beyond a simple acute MD episode versus euthymia binary classification, may add further clinical value [35]. As the literature on SSL is expanding at a fast pace, a thorough search of different approaches was beyond the scope of this work. We acknowledge that other pretext tasks can be deployed, and although the architectural choice may have an impact on SSL, we settled for just 1 reasonable, modern model design with a transformer [43] as a workhorse for representation learning. Lastly, given the naturalist design of the study, reflective of the intended use of personal sensing in a clinical setting, we could not exclude the effect of confounders, including medications, on the physiological variables. However, we reported medication classes administered in the cohort and verified a lack of any significant association between target classes (euthymia vs acute MD episode) and being on a given medication class.

Future Directions

As our findings indicate that the choice of the pretext task has a significant impact on target task performance, further efforts should be put into pretext task design. Indeed, although MP is a general-purpose strategy inspired by the great success of BERT [38] in NLP, the literature on SSL [40] suggests that domain knowledge may help tailor the pretext task to the specific use case. A promising approach we did not explore is contrastive learning [71], which, indeed, relies on domain knowledge of how augmented views of the input are created, especially since most experience today is in computer vision and NLP, while physiological multivariate time series are relatively unexplored.

Conclusion

This work shows that SSL is a promising paradigm for mitigating the annotation bottleneck, 1 of the major barriers to the development of artificial intelligence–powered clinical decision support systems using personal sensing to help monitor MDs, thus enabling early intervention. The collection and preprocessing of open access unlabeled data sets that we curated (E4SelfLearning) can foster future research into SSL, therefore advancing the translation of personal sensing into clinical practice.

Acknowledgments

We acknowledge the contribution of all the participants of the study.

FC and BML were supported by the United Kingdom Research and Innovation (UKRI; grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising from this submission.

GA was supported by a Rio Hortega 2021 grant (CM21/00017) and an M-AES mobility fellowship (MV22/00058) from the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and cofinanced by the Fondo Social Europeo Plus (FSE+).

IG thanks the support of the Spanish Ministry of Science and Innovation (MCIN; PI23/00822) integrated into the Plan Nacional de I+D+I and cofinanced by the ISCIII-Subdirección General de Evaluación y confinanciado por la Unión Europea (FEDER, FSE, Next Generation EU/Plan de Recu- peración Transformación y Resiliencia PRTR); the Instituto de Salud Carlos III; the CIBER of Mental Health (CIBERSAM); the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (2021 SGR 01358), CERCA Programme/Generalitat de Catalunya; and the Fundació Clínic per la Recerca Biomèdica (Pons Bartran 2022-FRCB PB1 2022).

AHY's independent research was funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care. For the purposes of open access, the author applied a Creative Commons Attribution (CC BY) license to any accepted author manuscript version arising from this submission.

DHM was supported by a Juan Rodés grant (JR18/00021) by the ISCIII. AV was supported by the UNREAL project (EP/Y023838/1) selected by the European Research Council and funded by the UKRI Engineering and Physical Sciences Research Council.

Data Availability

The E4SelfLearning collection is available at Reference [72], and the codebase is available at Reference [73]. Data in deidentified form from the Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder study may be made available from the corresponding author upon reasonable request.

Authors' Contributions

FC conceived of the study, proposed the methodology, developed the software codebase for the analyses, prepared the manuscript, and curated data collection. BML contributed to codebase development and manuscript writing. GA, CVP, AM, IP, MV, IGF, AB, and MG collected the data for the TIMEBASE/INTREPIBD (Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder) study. EV, AHY, SL, and HW critically reviewed the manuscript and provided feedback on the clinical side. DHM is the coordinator of the TIMEBASE/INTREPIBD study and critically reviewed the manuscript. AV supervised this study and contributed to the study design, methodology development, and manuscript writing.

Conflicts of Interest

GA has received continuing medical education–related honoraria, or consulting fees from the Janssen-Cilag, Lundbeck, Lundbeck/Otsuka, and Angelini, with no financial or other relationship relevant to the subject of this paper. IG has received grants and served as a consultant, advisor, or CME speaker for the following identities: ADAMED, Angelini, Casen Recordati, Esteve, Ferrer, Gedeon Richter, Janssen-Cilag, Lundbeck, Lundbeck-Otsuka, Luye, SEI Healthcare, and Viatris outside the submitted work. She also receives royalties from Oxford University Press, Elsevier, and the Editorial Médica Panamericana. All authors report no financial or other relationship relevant to the subject of this paper.

Multimedia Appendix 1

Supplementary material.
[DOC File , 138 KB-Multimedia Appendix 1]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC. American Psychiatric Association; 2013.
2. Santomauro DF, Mantilla Herrera AM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. Nov 2021;398(10312):1700-1712. [doi: 10.1016/s0140-6736(21)02143-7]

3. Greenberg PE, Fournier A, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics. Jun 05, 2021;39(6):653-665. [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
4. Brådvik L. Suicide risk and mental disorders. Int J Environ Res Public Health. Sep 17, 2018;15(9):2028. [FREE Full text] [doi: 10.3390/ijerph15092028] [Medline: 30227658]
5. Joyce K, Thompson A, Marwaha S. Is treatment for bipolar disorder more effective earlier in illness course? A comprehensive literature review. Int J Bipolar Disord. Dec 9, 2016;4(1):19. [FREE Full text] [doi: 10.1186/s40345-016-0060-6] [Medline: 27613276]
6. Buchman-Wildbaum T, Váradi E, Schmelowszky Á, Griffiths MD, Demetrovics Z, Urbán R. The paradoxical role of insight in mental illness: the experience of stigma and shame in schizophrenia, mood disorders, and anxiety disorders. Arch Psychiatr Nurs. Dec 2020;34(6):449-457. [FREE Full text] [doi: 10.1016/j.apnu.2020.07.009] [Medline: 33280665]
7. Rimmer A. Mental health: staff shortages are causing distressingly long waits for treatment, college warns. BMJ. Oct 07, 2021;375:n2439. [doi: 10.1136/bmj.n2439] [Medline: 34620691]
8. Satiani A, Niedermier J, Satiani B, Svendsen DP. Projected workforce of psychiatrists in the United States: a population analysis. Psychiatr Serv. Jun 01, 2018;69(6):710-713. [doi: 10.1176/appi.ps.201700344] [Medline: 29540118]
9. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol. May 08, 2017;13(1):23-47. [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
10. Faurholt-Jepsen M, Brage S, Kessing LV, Munkholm K. State-related differences in heart rate variability in bipolar disorder. J Psychiatr Res. Jan 2017;84:169-173. [FREE Full text] [doi: 10.1016/j.jpsychires.2016.10.005] [Medline: 27743529]
11. Sarchiapone M, Gramaglia C, Iosue M, Carli V, Mandelli L, Serretti A, et al. The association between electrodermal activity (EDA), depression and suicidal behaviour: a systematic review and narrative synthesis. BMC Psychiatry. Jan 25, 2018;18(1):22. [FREE Full text] [doi: 10.1186/s12888-017-1551-4] [Medline: 29370787]
12. Tazawa Y, Wada M, Mitsukura Y, Takamiya A, Kitazawa M, Yoshimura M, et al. Actigraphy for evaluation of mood disorders: a systematic review and meta-analysis. J Affect Disord. Jun 15, 2019;253:257-269. [doi: 10.1016/j.jad.2019.04.087] [Medline: 31060012]
13. Tazawa Y, Liang K, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon. Feb 2020;6(2):e03274. [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03274] [Medline: 32055728]
14. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. NPJ Digit Med. Feb 01, 2019;2(1):3. [FREE Full text] [doi: 10.1038/s41746-019-0078-0] [Medline: 31304353]
15. Cote-Allard U, Jakobsen P, Stautland A, Nordgreen T, Fasmer OB, Oedegaard KJ, et al. Long–short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors. IEEE Pervasive Comput. Apr 1, 2022;21(2):20-31. [doi: 10.1109/mprv.2022.3155728]
16. Nguyen D, Chan C, Li AA, Phan D, Lan C. Decision support system for the differentiation of schizophrenia and mood disorders using multiple deep learning models on wearable devices data. Health Informatics J. Nov 01, 2022;28(4):14604582221137537. [FREE Full text] [doi: 10.1177/14604582221137537] [Medline: 36317536]
17. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C. Objective assessment of depressive symptoms with machine learning and wearable sensors data. 2017. Presented at: ACII 2017: Seventh International Conference on Affective Computing and Intelligent Interaction; October 23-26, 2017; San Antonio, TX. [doi: 10.1109/acii.2017.8273620]
18. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry. Dec 18, 2020;11:584711. [FREE Full text] [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
19. Lee H, Cho C, Lee T, Jeong J, Yeom JW, Kim S, et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. Psychol Med. Sep 2023;53(12):5636-5644. [doi: 10.1017/S0033291722002847] [Medline: 36146953]
20. Li BM, Corponi F, Anmella G, Mas A, Sanabra M, Hidalgo-Mazzei D, et al. Inferring mood disorder symptoms from multivariate time-series sensory data. 2024. Presented at: NeurIPS 2022 Workshop on Learning from Time Series for Health; December 2, 2022; New Orleans, LA. URL: https://openreview.net/forum?id=awjU8fCDZjS [doi: 10.1038/s41398-024-02876-1]
21. E4 wristband technical specifications. Empatica. URL: https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications [accessed 2024-06-24]
22. Ronca V, Martinez-Levy AC, Vozzi A, Giorgi A, Aricò P, Capotorto R, et al. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3. Sensors (Basel). Jun 23, 2023;23(13):5847. [FREE Full text] [doi: 10.3390/s23135847] [Medline: 37447697]
23. Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K. Deep PPG: large-scale heart rate estimation with convolutional neural networks. Sensors (Basel). Jul 12, 2019;19(14):3079. [FREE Full text] [doi: 10.3390/s19143079] [Medline: 31336894]
24. Sah RK, McDonell M, Pendry P, Parent S, Ghasemzadeh H, Cleveland M. ADARP: a multi modal dataset for stress and alcohol relapse quantification in real life setting. 2022. Presented at: 2022 IEEE-EMBS International Conference on Wearable

and Implantable Body Sensor Networks (BSN); September 27-30, 2022; Ioannina Greece. [doi: 10.1109/bsn56160.2022.9928495]
25. Schmidt P, Reiss A, Duerichen R, Marberger C, Van LK. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. 2018. Presented at: 20th ACM International Conference on Multimodal Interaction; October 16-20, 2018; Boulder, CO. [doi: 10.1145/3242969.3242985]
26. Iqbal T, Simpkin AJ, Roshan D, Glynn N, Killilea J, Walsh J, et al. Stress monitoring using wearable sensors: a pilot study and stress-predict dataset. Sensors (Basel). Oct 24, 2022;22(21):8135. [FREE Full text] [doi: 10.3390/s22218135] [Medline: 36365837]
27. Svoren H, Thambawita V, Halvorsen P, Jakobsen P, Garcia-Ceja E, Noori F. Toadstool: a dataset for training emotional intelligent machines playing Super Mario Bros. 2020. Presented at: MMSys '20: 11th ACM Multimedia Systems Conference; June 8-11, 2020; Istanbul Turkey. [doi: 10.1145/3339825.3394939]
28. Gashi S, Min C, Montanari A, Santini S, Kawsar F. A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices. Sci Data. Sep 01, 2022;9(1):537. [FREE Full text] [doi: 10.1038/s41597-022-01643-5] [Medline: 36050312]
29. Amin MR, Wickramasuriya D, Faghih R. A wearable exam stress dataset for predicting grades using physiological signals. 2022. Presented at: 2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT); March 10-11, 2022; Houston, TX. [doi: 10.1109/hi-poct54491.2022.9744065]
30. Hosseini S, Gottumukkala R, Katragadda S, Bhupatiraju RT, Ashkar Z, Borst CW, et al. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. Sci Data. Jun 01, 2022;9(1):255. [FREE Full text] [doi: 10.1038/s41597-022-01361-y] [Medline: 35650267]
31. Gao N, Marschall M, Burry J, Watkins S, Salim FD. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. Sci Data. Jun 02, 2022;9(1):261. [FREE Full text] [doi: 10.1038/s41597-022-01347-w] [Medline: 35654857]
32. Hinkle LB, Metsis V. Unlabeled Empatica E4 Wristband Data (UE4W) dataset. Zenodo. Jul 25, 2022. URL: https://zenodo.org/records/6898244 [accessed 2024-06-24]
33. Bent B, Cho PJ, Henriquez M, Wittmann A, Thacker C, Feinglos M, et al. Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. NPJ Digit Med. Jun 02, 2021;4(1):89. [FREE Full text] [doi: 10.1038/s41746-021-00465-w] [Medline: 34079049]
34. Shani C, Zarecki J, Shahaf D. The lean data scientist: recent advances toward overcoming the data bottleneck. Commun ACM. Jan 20, 2023;66(2):92-102. [doi: 10.1145/3551635]
35. Corponi F, Li BM, Anmella G, Mas A, Pacchiarotti I, Valentí M, et al. Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. Transl Psychiatry. Mar 26, 2024;14(1):161. [FREE Full text] [doi: 10.1038/s41398-024-02876-1] [Medline: 38531865]
36. Rani V, Nabi ST, Kumar M, Mittal A, Kumar K. Self-supervised learning: a succinct review. Arch Comput Methods Eng. Jan 20, 2023;30(4):2761-2775. [FREE Full text] [doi: 10.1007/s11831-023-09884-2] [Medline: 36713767]
37. Ohri K, Kumar M. Review on self-supervised image recognition using deep neural networks. Knowl-Based Syst. Jul 2021;224:107090. [doi: 10.1016/j.knosys.2021.107090]
38. Devlin DJ, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint posted online 2018. [doi: 10.48550/arXiv.1810.04805]. 2021. [doi: 10.48550/arXiv.1810.04805]
39. Huang S, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med. Apr 26, 2023;6(1):74. [FREE Full text] [doi: 10.1038/s41746-023-00811-0] [Medline: 37100953]
40. Zhang K, Wen Q, Zhang C, Cai R, Jin M, Liu Y, et al. Self-supervised learning for time series analysis: taxonomy, progress, and prospects. arXiv preprint posted online 2023. [doi: 10.48550/arXiv.2306.10125]. 2021. [doi: 10.48550/arXiv.2306.10125]
41. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A transformer-based framework for multivariate time series representation learning. 2021. Presented at: 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; August 14-18, 2021; Virtual. [doi: 10.1145/3447548.3467401]
42. Wu Y, Daoudi M, Amad A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Trans Affective Comput. Jan 2024;15(1):157-172. [doi: 10.1109/taffc.2023.3263907]
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention is all you need. 2017. Presented at: NIPS 2017: Advances in Neural Information Processing Systems 30; December 4-9, 2017; Long Beach, CA.
44. Ericsson L, Gouk H, Loy CC, Hospedales TM. Self-supervised representation learning: introduction, advances, and challenges. IEEE Signal Process Mag. May 2022;39(3):42-62. [doi: 10.1109/msp.2021.3134634]
45. Anmella G, Corponi F, Li BM, Mas A, Sanabra M, Pacchiarotti I, et al. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. JMIR Mhealth Uhealth. May 04, 2023;11:e45405. [FREE Full text] [doi: 10.2196/45405] [Medline: 36939345]
46. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. Feb 01, 1960;23(1):56-62. [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]

47. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry. Nov 29, 1978;133(5):429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
48. Tohen M, Frank E, Bowden C, Colom F, Ghaemi N, Yatham L. The International Society of Bipolar Disorders (ISBD) task force on the nomenclature of course and outcome in bipolar disorders. J Affect Disord. Apr 2010;122:S15. [doi: 10.1016/j.jad.2010.01.030]
49. Vieluf S, Amengual-Gual M, Zhang B, El Atrache R, Ufongene C, Jackson MC, et al. Twenty-four-hour patterns in electrodermal activity recordings of patients with and without epileptic seizures. Epilepsia. Apr 23, 2021;62(4):960-972. [doi: 10.1111/epi.16843] [Medline: 33619751]
50. Nasseri M, Nurse E, Glasstetter M, Böttcher S, Gregg NM, Laks Nandakumar A, et al. Signal quality and patient experience with wearable devices for epilepsy management. Epilepsia. Nov 04, 2020;61 Suppl 1(S1):S25-S35. [doi: 10.1111/epi.16527] [Medline: 32497269]
51. Emaptics user manual. ETH Zurich. URL: https://archive.arch.ethz.ch/esum/downloads/manuals/emaptics.pdf [accessed 2024-06-24]
52. van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. PLoS One. Nov 16, 2015;10(11):e0142533. [FREE Full text] [doi: 10.1371/journal.pone.0142533] [Medline: 26569414]
53. Patterson MR, Nunes AAS, Gerstel D, Pilkar R, Guthrie T, Neishabouri A, et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. NPJ Digit Med. Mar 24, 2023;6(1):51. [FREE Full text] [doi: 10.1038/s41746-023-00802-1] [Medline: 36964203]
54. Cui Z, Chen W, Chen Y. Multi-scale convolutional neural networks for time series classification. arXiv preprint posted online 2016. [doi: 10.48550/arXiv.1603.06995]. 2021. [doi: 10.48550/arXiv.1603.06995]
55. 55 PM, Zlatintsi A, Filntisis P, Roumeliotis A, Efthymiou N, Maragos P. A comparative study of autoencoder architectures for mental health analysis using wearable sensors data. 2022. Presented at: EUSIPCO 2022: 30th European Signal Processing Conference; August 29-September 2, 2022; Belgrade, Serbia. [doi: 10.23919/eusipco55093.2022.9909697]
56. Föll S, Maritsch M, Spinola F, Mishra V, Barata F, Kowatsch T, et al. FLIRT: a feature generation toolkit for wearable data. Comput Methods Programs Biomed. Nov 2021;212:106461. [FREE Full text] [doi: 10.1016/j.cmpb.2021.106461] [Medline: 34736174]
57. Özdenizci O, Wang Y, Koike-Akino T, ErdoĞmuŞ D. Learning invariant representations from EEG via adversarial inference. IEEE Access. 2020;8:27074-27085. [FREE Full text] [doi: 10.1109/access.2020.2971600] [Medline: 33747669]
58. Cheng JY, Goh H, Dogrusoz K, Tuzel O, Azemi E. Subject-aware contrastive learning for biosignals. arXiv preprint posted online 2020. [doi: 10.48550/arXiv.2007.04871]. 2021. [doi: 10.48550/arXiv.2007.04871]
59. Sabry F, Eltaras T, Labda W, Alzoubi K, Malluhi Q. Machine learning for healthcare wearable devices: the big picture. J Healthc Eng. 2022;2022:4653923. [doi: 10.1155/2022/4653923] [Medline: 35480146]
60. Strzelecki M, Badura P. Machine learning for biomedical application. Appl Sci. Feb 15, 2022;12(4):2022. [doi: 10.3390/app12042022]
61. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: a novel bandit-based approach to hyperparameter optimization. J Mach Learn Res. 2017;18:6765-6816. [doi: 10.1007/978-1-4899-7687-1_100200]
62. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(56):1929-1958.
63. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P. Deep learning for time series classification: a review. Data Min Knowl Disc. Mar 02, 2019;33(4):917-963. [doi: 10.1007/s10618-019-00619-1]
64. Paszke P, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. PyTorch: an imperative style, high-performance deep learning library. 2019. Presented at: NeurIPS 2019: 33rd Conference on Neural Information Processing Systems; December 8-14, 2019:8024-8035; Vancouver, Canada. [doi: 10.7551/mitpress/11474.003.0014]
65. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(85):2825-2830.
66. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
67. Experiment tracking: the system of record for your model training. Weights & Biases. URL: https://wandb.ai/site/experiment-tracking [accessed 2024-06-24]
68. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv oreprint posted online 2018. [doi: 10.48550/arXiv.1802.03426]. 2021. [doi: 10.48550/arXiv.1802.03426]
69. Birk RH, Samuel G. Digital phenotyping for mental health: reviewing the challenges of using data to monitor and predict mental health problems. Curr Psychiatry Rep. Oct 24, 2022;24(10):523-528. [doi: 10.1007/s11920-022-01358-9] [Medline: 36001220]
70. El-Nouby A, Izacard G, Touvron H, Laptev I, Jegou H, Grave E. Are large-scale datasets necessary for self-supervised pre-training. arXiv preprint posted online 2021. [doi: 10.48550/arXiv.2112.10740]. 2021. [doi: 10.48550/arXiv.2112.10740]

71. Kumar P, Rawat P, Chauhan S. Contrastive self-supervised learning: review, progress, challenges and future research directions. Int J Multimed Info Retr. Aug 05, 2022;11(4):461-488. [doi: 10.1007/s13735-022-00245-6]
72. FcmC / E4SelfLearning. Hugging Face. URL: https://huggingface.co/datasets/FcmC/E4SelfLearning [accessed 2024-07-02]
73. april-tools / e4selflearning. GitHub. URL: https://github.com/april-tools/e4selflearning [accessed 2024-07-02]

Abbreviations

ACC: accuracy
AUROC: area under the receiver operating characteristic curve
BD: bipolar disorder
BVP: blood volume pressure
CCE: categorical cross-entropy
CE: channel embedding
CML: classical machine learning
DSM-5: Diagnostic and Statistical Manual, Fifth Edition
E4mer: E4-tailored transformer
EDA: electrodermal activity
ENET: elastic net logistic regression
FT: fine-tuning
HDRS: Hamilton Depression Rating Scale-17
HR: heart rate
IBI: interbeat interval
KNN: K-nearest neighbor
LME: linear mixed effects
LR: linear readout
MD: mood disorder
MDD: major depressive disorder
ML: machine learning
MLP: multilayer perceptron
MP: masked prediction
NLP: natural language processing
PCC: Pearson correlation coefficient
RM: representation module
RMSE: root mean square error
SL: supervised learning
SSL: self-supervised learning
SVM: support vector machine
TEMP: skin temperature
TIMEBASE/INTREPIBD: Identifying Digital Biomarkers of Illness Activity in Bipolar Disorder/Identifying Digital Biomarkers of Illness Activity and Treatment Response in Bipolar Disorder
TP: transformation prediction
XGBoost: extreme gradient boosting
YMRS: Young Mania Rating Scale

Edited by L Buis; submitted 02.12.23; peer-reviewed by J Zulueta, B Montezano, W Speier; comments to author 29.02.24; revised version received 14.04.24; accepted 24.05.24; published 17.07.24

Please cite as:
Corponi F, Li BM, Anmella G, Valenzuela-Pascual C, Mas A, Pacchiarotti I, Valentí M, Grande I, Benabarre A, Garriga M, Vieta E, Young AH, Lawrie SM, Whalley HC, Hidalgo-Mazzei D, Vergari A
Wearable Data From Subjects Playing Super Mario, Taking University Exams, or Performing Physical Exercise Help Detect Acute Mood Disorder Episodes via Self-Supervised Learning: Prospective, Exploratory, Observational Study
JMIR Mhealth Uhealth 2024;12:e55094
URL: https://mhealth.jmir.org/2024/1/e55094
doi: 10.2196/55094
PMID:

Chapter 7

Conclusions

This dissertation explored the burgeoning field of personal sensing for MDs. These are prevalent and severe mental health conditions manifesting with recurring episodes involving disturbances in mood, rest-activity rhythms, and vegetative functions. Wearables, particularly wrist-worn devices, collect data reflecting such disturbances. This technology thus offers a unique opportunity to advance our understanding of MDs and improve symptom monitoring, enabling early detection and intervention.

Innovative data analytics approaches from AI can harness physiological data collected with wearables. This can be achieved with relatively simple models, such as those presented in Chapter 4, which are clinically interpretable. Such models enable researchers to test simple hypotheses and advance knowledge of the biological underpinnings of MDs. Alternatively, large(r) models can be used to mine wearable data to identify abnormal mental states, though this may come at the cost of lower interpretability. However, for a large model to be actionable and have clinical value beyond a mere exercise in AI, it must incorporate and reflect clinical knowledge, as shown in Chapter 5.

The road towards the clinical translatability of wearables for remote monitoring in MDs presents several challenges, as reviewed in Chapter 2. Chief among these is data scarcity since collecting and annotating data in personal sensing is resource-intensive. Chapter 6 explored how SSL can exploit unlabeled data, available in large(r) quantities, to learn useful representations without the need for human supervision, and then recycle these representations for supervised downstream tasks, such as detecting acute MD episodes. My original research contributions to the field of personal sensing in MD were presented in Chapter 4, Chapter 5, and Chapter 6.

Chapter 4 investigated longitudinal changes in HRV as acute episodes of BD subside. Motivated by the small sample size typical of this kind of study and seeking a principled way to quantify uncertainty in parameter values, I introduced a Bayesian framework to the study of HRV. I did not simply check the distribution of test statistics but proposed a probabilistic model, attempting to capture the HRV generating process. I showed a positive rate of change in HRV with respect to symptoms' improvement, albeit without polarity-specific (mania vs depression) patterns. HRV positions itself as a potential biomarker for recovery upon acute episodes of BD, underlying an improvement in the parasympathetic branch of the ANS.

Chapter 5 aligned time-series classification in personal sensing for MDs to everyday psychiatry practice, where patients sharing the same overall label (e.g. a diagnosis) may display different symptom combinations, requiring different management and therapeutic approaches. I therefore attempted to infer all items from two popular psychometric scales, assessing manic and depressive symptoms, respectively. These can be viewed as the concepts motivating a given diagnosis, therefore adding to the model's interpretability. This new task is hard and comes with a number of technical challenges, which I explored and proposed some solutions to.

Chapter 6 showed that SSL, leveraging unlabelled data collected for personal sensing tasks unrelated to MDs, can be deployed to mitigate the annotation bottleneck and outperforms fully-supervised learning. Labelling data for personal sensing in MDs is indeed very resource intensive, foiling attempts to develop ANN directly trained in a supervised way on the task at hand. In ablation *post-hoc* analyses, I showed which ingredients are key to the success of the SSL recipe. Furthermore, I release the curated unlabelled corpus of data, the largest open access to date, as well as the pre-processing pipeline doing, on-/off-body detection, sleep/wake detection, segmentation, and (optionally) feature extraction.

Overall, my work shows that personal sensing can contribute to a better understanding of the biological underpinnings of MDs, since it provides novel ways of measuring physiological data that correlate with various regulatory systems that are impaired in MDs. A case in point is HRV, a proxy for ANS activity, which my dissertation focused on. The case for personal sensing as a revolutionary paradigm enabling remote monitoring and therefore creating scope for timely interventions seems compelling. A solid basis is that core manifestations of MDs include changes in sleep, rest-activity patterns, and vegetative functions, which wearables can measure ecologically and near-

continuously. Towards delivering on the promises of personal sensing, novel solutions at the intersection between AI research and clinical science are needed. Despite enthusiasm around personal sensing, some fundamental challenges still stand along the way to clinical implementations.

7.1 Limitations

I hereby acknowledge some limitations across the works presented in this dissertation, in addition to what has already been discussed within each article. While TIMEBASE/INTREPIBD represents a commendable effort towards delivering personal sensing, certain aspects of its design, partly dictated by the choice of a research-grade device like the Empatica E4, limited the scope of our analyses.

7.1.1 Brief recording sessions

TIMEBASE/INTREPIBD records only brief, 48-hour sessions collected around clinically significant states, such as acute episodes, response, remission, and recovery (Chapter 3). Brief sessions, ranging from 48 hours to only a few weeks, are common in studies using research-grade devices [95, 94, 96]. The recording duration is constrained by battery life and the risk of technology abandonment in psychiatric cohorts. While wearables only work if patients are wearing them, compliance remains a major challenge in personal sensing. As it currently stands, it is impractical to expect patients with MDs to charge a wrist-worn device every few days and consistently wear it.

The TIMEBASE/INTREPIBD design implicitly assumes that the data-generating process remains the same within any 48-hour sample taken from a given mental state (*stationarity*) and disregards transitions across clinical stages. Such an assumption has not been verified in the literature, and it is likely that transitions across mental states are gradual rather than stepwise.

Sessions lasting 48 hours, or even a few weeks, are insufficient to assess the clinical potential of personal sensing. The intended use case is remote monitoring during daily life, which requires several months of monitoring to capture periods of disease exacerbation in the context of daily life. Studies extending over several months [139] use commercial devices with longer battery life but lack access to raw data and offer limited sensory modalities. Furthermore, no such cohort to date has included BD;

retention is likely even lower in this population, particularly during ME.

Lastly, the short recording time makes it difficult to model within-subject variability related to background characteristics, such as age, sex, and daily habits, separate from disease status. It is likely that a significant proportion of variability in wearable data is physiological and subject-specific rather than MD-related. Thus, with short recording sessions, it is challenging to determine the extent to which a model has learned MD-related patterns (possibly specific to subpopulations) versus memorizing subject-level idiosyncrasies in the signal.

7.1.2 Lack of hard outcomes

As explained in Chapter 2, labelling in mental health, i.e. scoring patients on psychometric scales (Appendix A) or issuing a diagnosis, is subject to significant inter- and intra-rater variability [126]. Clinical trials typically mitigate this issue by training raters until a minimum threshold of inter-rater agreement is reached. In TIMEBASE/INTREPIBD, only a single mental health specialist conducted the clinical assessments. Thus, when developing and validating our models, it was not possible to account for inter-rater reliability and noise in human annotation.

Furthermore, TIMEBASE/INTREPIBD did not record hard, objective outcomes, such as hospital admissions. While these labels may not have substantive biological meaning, as the literature does not support a specific and unique pathophysiology behind hospitalization in MDs, they represent hard outcomes, overcoming inter-rater agreement issues. Moreover, hospitalization accounts for a significant proportion of healthcare costs associated with MDs [127, 128, 129, 130]. However, only a minority of acute episodes in MDs result in hospital admission, making it challenging to collect a significant number of these events within a single cohort.

7.1.3 Generalization across subjects

Both in Chapter 5 and Chapter 6, data was partitioned into train and validation sets using a *time-split* (also referred to as *subject-dependent*) approach. This means that the same subjects, but different samples thereof, appeared in both the train and validation sets. This scenario potentially introduces 'information leakage', where the model may inadvertently learn and memorize subject-specific distributions rather than disease-related features. This contrasts with a *subject-split* (or *subject-independent*) approach,

where different subjects are used for training and validation. Developing a method that effectively captures common features among subjects while disregarding individual noise remains an unsolved problem [204]. For this reason, a *time-split* is a popular option, as it still retains value by testing the model's ability to generalize across different points in time (e.g., as a subject may engage in different activities).

7.2 Future Directions

A fundamental challenge towards clinical translation is patient compliance, i.e. ensuring that patients will wear the device. Studies over long observation periods [139] tend to monetarily reward patients to sustain engagement and still show non-negligible attrition rates. Most research in personal sensing in MDs focus on elucidating how physiological parameters are affected by illness states, or showing that data from wearables can reveal aberrant mood states with the help of AI. More effort should be given to the mundane problem of compliance, which is a requirement for clinical applications [205, 206, 205]. Gamification, text or email reminders, or equipping wrist-worn devices with functionalities beyond mere physiological data collection (e.g. checking emails or the news) may play a role. At the same time, digital navigators i.e. trusted individuals with experience in technology and mental health, whose role is to increase access to digital health, can be another impactful intervention, also towards maximizing access for people from marginalized communities or with less technology familiarity.

More efforts should go into collecting long longitudinal recordings, which are indeed better reflective of the intended setting for personal sensing. This could help establish how early an impending mood episode can be detected and, as a result, to what extent remote monitoring is conducive to better outcomes. Furthermore, having such long longitudinal observations would help contextualize MD-related changes in physiological data within a patient's baseline patterns. With long (months) recording, a *time-split* of the data and the development of *idiographic* (patient-tailored) models are better justified: the goal in this setting, in line with the envisioned clinical application, becomes to detect abnormal mental states within each subject.

In parallel to generating new datasets, sharing existing ones within the research community would increase data availability, which is a constraint in this field. Furthermore, the open-access framework would ensure transparency and reproducibility, as well as increase engagement of the ML community. AI-powered clinical decision support tools

notoriously suffer from poor generalization to independent samples [207]. Sharing datasets would make it possible to test the model on independent cohorts, as well as set up benchmark datasets.

Lastly, time series poses some unique challenges relative to other data modalities. State-of-the-art algorithms (e.g. Transformers [142]) were borrowed from other fields of AI, but recent works [183, 208] showed their shortcomings when modelling time-series. So, more methodological research in this data modality is probably needed.

Bibliography

[1] World Health Organization. WHO highlights urgent need to transform mental health and mental health care, 2022. Available online at: http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html, last accessed on 2024-04-19.

[2] Daniel Arias, Shekhar Saxena, and Stéphane Verguet. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine*, 54, 2022.

[3] GBD 2019 Mental Disorders Collaborators et al. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150, 2022.

[4] Albert Persaud, Geraint Day, Susham Gupta, Antonio Ventriglio, Roxanna Ruiz, Egor Chumakov, Geetha Desai, Joao Castaldelli-Maia, Julio Torales, Edgardo Juan Tolentino, et al. Geopolitical factors and mental health i. *International journal of social psychiatry*, 64(8):778–785, 2018.

[5] Anna Sri, Dinesh Bhugra, Albert Persaud, Rachel Tribe, Sam Gnanapragasam, João M Castaldelli-Maia, Julio Torales, and Antonio Ventriglio. Global mental health and climate change: A geo-psychiatry perspective. *Asian journal of Psychiatry*, page 103562, 2023.

[6] World Health Organization. Global burden of disease report, 2008. URL http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html. Available online at: http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html, last accessed on 2024-05-30.

[7] David McDaid and A-La Park. The economic case for investing in the prevention of mental health conditions in the uk, 2022. URL https://www.mentalhealth.org.uk/publications/mental-health-problems-cost-uk-economy-least-118-billion-year. Available online at: https://www.mentalhealth.org.uk/publications/mental-health-problems-cost-uk-economy-least-118-billion-year, last accessed on 2024-04-19.

[8] Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712, 2021.

[9] Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H Koenigsberg, and Ronald C Kessler. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *Pharmacoeconomics*, 39(6):653–665, 2021.

[10] DS American Psychiatric Association, American Psychiatric Association, et al. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC, 2013.

[11] Antonella Benvenuti, Mario Miniati, Antonio Callari, Michela Giorgi Mariani, Mauro Mauri, and Liliana Dell'Osso. Mood spectrum model: evidence reconsidered in the light of dsm-5. *World journal of psychiatry*, 5(1):126, 2015.

[12] Anand Satiani, Julie Niedermier, Bhagwan Satiani, and Dale P Svendsen. Projected workforce of psychiatrists in the united states: a population analysis. *Psychiatric Services*, 69(6):710–713, 2018.

[13] Heather Burrell Ward, Roscoe O Brady Jr, and Mark A Halko. Bridging the gap: strategies to make psychiatric neuroimaging clinically relevant. *Harvard review of psychiatry*, 29(3):185–187, 2021.

[14] Jessica L Bourdon, Rachel A Davies, and Elizabeth C Long. Four actionable bottlenecks and potential solutions to translating psychiatric genetics research: An expert review. *Public health genomics*, 23(5-6):171–183, 2021.

[15] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[16] Thomas R Insel. Digital phenotyping: technology for a new science of behavior. *Jama*, 318(13):1215–1216, 2017.

[17] James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. Icd-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21:1–10, 2021.

[18] Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493, 2014.

[19] Rutvik V Shah, Gillian Grennan, Mariam Zafar-Khan, Fahad Alim, Sujit Dey, Dhakshin Ramanathan, and Jyoti Mishra. Personalized machine learning of depressed mood using wearables. *Translational psychiatry*, 11(1):1–18, 2021.

[20] Eduard Vieta, Michael Berk, Thomas G Schulze, André F Carvalho, Trisha Suppes, Joseph R Calabrese, Keming Gao, Kamilla W Miskowiak, and Iria Grande. Bipolar disorders. *Nature reviews Disease primers*, 4(1):1–16, 2018.

[21] Wolfgang Marx, Brenda WJH Penninx, Marco Solmi, Toshi A Furukawa, Joseph Firth, Andre F Carvalho, and Michael Berk. Major depressive disorder. *Nature Reviews Disease Primers*, 9(1):44, 2023.

[22] Andrew A Nierenberg. Residual symptoms in depression: prevalence and impact. *The Journal of clinical psychiatry*, 76(11):26474, 2015.

[23] Sandeep Grover, Subho Chakrabarti, and Swapnajeet Sahoo. Prevalence and clinical correlates of residual symptoms in remitted patients with bipolar disorder: An exploratory study. *Indian Journal of Psychiatry*, 62(3):295–305, 2020.

[24] Christoph Kraus, Bashkim Kadriu, Rupert Lanzenberger, Carlos A Zarate Jr, and Siegfried Kasper. Prognosis and improved outcomes in major depression: a review. *Translational psychiatry*, 9(1):127, 2019.

[25] Antoine Oudin, Redwan Maatoug, Alexis Bourla, Florian Ferreri, Olivier Bonnot, Bruno Millet, Félix Schoeller, Stéphane Mouchabac, and Vladimir Adrien.

Digital phenotyping: Data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*, 25:e44502, 2023.

[26] Simone Schmidt and Simon D'Alfonso. Clinician perspectives on how digital phenotyping can inform client treatment. *Acta Psychologica*, 235:103886, 2023.

[27] Maria Faurholt-Jepsen, Søren Brage, Lars Vedel Kessing, and Klaus Munkholm. State-related differences in heart rate variability in bipolar disorder. *Journal of psychiatric research*, 84:169–173, 2017.

[28] Marco Sarchiapone, Carla Gramaglia, Miriam Iosue, Vladimir Carli, Laura Mandelli, Alessandro Serretti, Debora Marangon, and Patrizia Zeppegno. The association between electrodermal activity (eda), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC psychiatry*, 18(1): 1–27, 2018.

[29] Yuuki Tazawa, Masataka Wada, Yasue Mitsukura, Akihiro Takamiya, Momoko Kitazawa, Michitaka Yoshimura, Masaru Mimura, and Taishiro Kishimoto. Actigraphy for evaluation of mood disorders: A systematic review and meta-analysis. *Journal of affective disorders*, 253:257–269, 2019.

[30] Jukka-Pekka Onnela. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, 46(1):45–54, 2021.

[31] John Torous, Sandra Bucci, Imogen H Bell, Lars V Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3):318–335, 2021.

[32] Md Mobashir Hasan Shandhi, Karnika Singh, Natasha Janson, Perisa Ashar, Geetika Singh, Baiying Lu, D Sunshine Hillygus, Jennifer M Maddocks, and Jessilyn P Dunn. Assessment of ownership of smart devices and the acceptability of digital health data sharing. *npj Digital Medicine*, 7(1):44, 2024.

[33] Diego Hidalgo-Mazzei, Viktoriya L Nikolova, Simon Kitchen, and Allan H Young. Internet-connected devices ownership, use and interests in bipolar disorder: from desktop to mobile mental health. *Digital Psychiatry*, 2(1):1–7, 2019.

[34] World Health Organization. Recommendations on digital interventions for health system strengthening. *World Health Organization*, pages 2020–10, 2019.

[35] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[36] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[37] Kaela Van Til, Melvin G McInnis, and Amy Cochran. A comparative study of engagement in mobile and wearable health monitoring for bipolar disorder. *Bipolar disorders*, 22(2):182–190, 2020.

[38] Nicole Martinez-Martin, Henry T Greely, Mildred K Cho, et al. Ethical development of digital phenotyping tools for mental health applications: Delphi study. *JMIR mHealth and uHealth*, 9(7):e27343, 2021.

[39] Silvia Francesca Maria Pizzoli, Dario Monzani, Lorenzo Conti, Giulia Ferraris, Roberto Grasso, and Gabriella Pravettoni. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. *Frontiers in psychology*, 14:1103703, 2023.

[40] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics*, 20(4):1053–1060, 2015.

[41] Heng Zhang, Ahmed Ibrahim, Bijan Parsia, Ellen Poliakoff, and Simon Harper. Passive social sensing with smartphones: a systematic review. *Computing*, 105 (1):29–51, 2023.

[42] Auguste Vigouroux. *Study on electrical resistance in melancholic people*. PhD thesis, Faculte de medicine de Paris, 1890.

[43] Martin H Teicher. Actigraphy and motion analysis: new tools for psychiatry. *Harvard review of psychiatry*, 3(1):18–35, 1995.

[44] Pasquale Bufano, Marco Laurino, Sara Said, Alessandro Tognetti, and Danilo Menicucci. Digital phenotyping for monitoring mental disorders: Systematic review. *Journal of Medical Internet Research*, 25:e46778, 2023.

[45] Arfan Ahmed, Marco Agus, Mahmood Alzubaidi, Sarah Aziz, Alaa Abd-Alrazaq, Anna Giannicchi, and Mowafa Househ. Overview of the role of big data in mental health: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100076, 2022.

[46] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[47] Nicole R Karcher and Deanna M Barch. The abcd study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*, 46(1):131–142, 2021.

[48] Gerard Anmella, Filippo Corponi, Bryan M Li, Ariadna Mas, Miriam Sanabra, Isabella Pacchiarotti, Marc Valentí, Iria Grande, Antoni Benabarre, Anna Giménez-Palomo, et al. Exploring digital biomarkers of illness activity in mood episodes: hypotheses generating and model development study. *JMIR Mhealth and Uhealth*, 2023.

[49] Carsten Langholm, Scott Breitinger, Lucy Gray, Fernando Goes, Alex Walker, Ashley Xiong, Cindy Stopel, Peter Zandi, Mark A Frye, and John Torous. Classifying and clustering mood disorder patients using smartphone data from a feasibility study. *npj Digital Medicine*, 6(1):238, 2023.

[50] Shai Mulinari. Short-circuiting biology: Digital phenotypes, digital biomarkers, and shifting gazes in psychiatry. *Big Data & Society*, 10(1):20539517221145680, 2023.

[51] Andrea Stautland, Petter Jakobsen, Ole Bernt Fasmer, Berge Osnes, Jim Torresen, Tine Nordgreen, and Ketil J Oedegaard. Reduced heart rate variability during mania in a repeated naturalistic observational study. *Frontiers in Psychiatry*, 14: 1250925, 2023.

[52] Gerard Anmella, Ariadna Mas, Miriam Sanabra, Clàudia Valenzuela-Pascual, Marc Valentí, Isabella Pacchiarotti, Antoni Benabarre, Iria Grande, Michele De Prisco, Vincenzo Oliva, et al. Electrodermal activity in bipolar disorder: Differences between mood episodes and clinical remission using a wearable

device in a real-world clinical setting. *Journal of Affective Disorders*, 345:43–50, 2024.

[53] Emese Sükei, Agnes Norbury, M Mercedes Perez-Rodriguez, Pablo M Olmos, and Antonio Artés. Predicting emotional states using behavioral markers derived from passively sensed data: data-driven machine learning approach. *JMIR mHealth and uHealth*, 9(3):e24465, 2021.

[54] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–21, 2019.

[55] John Torous, Jessica Firth, and Samuel B Goldberg. Digital mental health's unstable dichotomy—wellness and health. *JAMA Psychiatry*, 2024. doi: 10.1001/jamapsychiatry.2024.0532.

[56] National Institute for Health and Care Excellence. Devices for remote monitoring of parkinson's disease, Jan 2023. Available online at: `https://www.nice.org.uk/guidance/dg51`, last accessed on 2024-04-12.

[57] National Institute for Health and Care Excellence. Nice recommends life changing technology is rolled out to people with type 1 diabetes, 2023. Available online at: https://www.nice.org.uk/news/article/nice-recommends-life-changing-technology-is-rolled-out-to-people-with-type-1-diabetes, last accessed on 2024-02-13.

[58] Empatica. E4 wristband technical specifications – empatica support, Jan 2020. URL `https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications`.

[59] Gerard Anmella, Filippo Corponi, Bryan M. Li, Ariadna Mas, Marina Garriga, Miriam Sanabra, Isabella Pacchiarotti, Marc Valentí, Iria Grande, Antoni Benabarre, and et al. Identifying digital biomarkers of illness activity and treatment response in bipolar disorder with a novel wearable device (timebase): protocol for a pragmatic observational clinical study. *BJPsych Open*, 10(5):e137, 2024. doi: 10.1192/bjo.2024.716.

[60] Bryan M. Li, Filippo Corponi, Gerard Anmella, Ariadna Mas, Miriam Sanabra, Diego Hidalgo-Mazzei, and Antonio Vergari. Inferring mood disorder symptoms

from multivariate time-series sensory data. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=awjU8fCDZjS.

[61] Clàudia Valenzuela-Pascual, Ariadna Mas, Roger Borràs, Gerard Anmella, Miriam Sanabra, Meritxell González-Campos, Marc Valentí, Isabella Pacchiarotti, Antoni Benabarre, Iria Grande, et al. Sleep–wake variations of electrodermal activity in bipolar disorder. *Acta Psychiatrica Scandinavica*, 2024.

[62] A John Rush Jr, Michael B First, and Deborah Blacker. *Handbook of psychiatric measures*. American Psychiatric Pub, 2009.

[63] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[64] Robert C Young, Jeffery T Biggs, Veronika E Ziegler, and Dolores A Meyer. A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry*, 133(5):429–435, 1978.

[65] Mauricio Tohen, Ellen Frank, Charles L Bowden, Francesc Colom, S Nassir Ghaemi, Lakshmi N Yatham, Gin S Malhi, Joseph R Calabrese, Willem A Nolen, Eduard Vieta, et al. The international society for bipolar disorders (isbd) task force report on the nomenclature of course and outcome in bipolar disorders. *Bipolar disorders*, 11(5):453–473, 2009.

[66] Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders. *American Journal of psychiatry*, 167(7):748–751, 2010.

[67] Ann Elizabeth Fowler La Berge. The history of science and the history of microscopy. *Perspectives on Science*, 7(1):111–142, 1999.

[68] YEUN-CHUNG Chang, Kou-Mou Huang, Jyh-Horng Chen, and Cheng-Tau Su. Impact of magnetic resonance imaging on the advancement of medicine. *Journal of the Formosan Medical Association= Taiwan yi zhi*, 98(11):740–748, 1999.

[69] Maria Faurholt-Jepsen, Lars Vedel Kessing, and Klaus Munkholm. Heart rate variability in bipolar disorder: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 73:68–80, 2017.

[70] Christian Otte, Stefan M Gold, Brenda W Penninx, Carmine M Pariante, Amit Etkin, Maurizio Fava, David C Mohr, and Alan F Schatzberg. Major depressive disorder. *Nature reviews Disease primers*, 2(1):1–20, 2016.

[71] Mujeeb Z Banday, Aga S Sameer, and Saniya Nissar. Pathophysiology of diabetes: An overview. *Avicenna journal of medicine*, 10(04):174–188, 2020.

[72] Katherine W Scangos, Matthew W State, Andrew H Miller, Justin T Baker, and Leanne M Williams. New and emerging approaches to treat psychiatric disorders. *Nature medicine*, 29(2):317–333, 2023.

[73] Arfan Ahmed, Sarah Aziz, Mahmood Alzubaidi, Jens Schneider, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, Barry Solaiman, and Mowafa Househ. Wearable devices for anxiety & depression: a scoping review. *Computer Methods and Programs in Biomedicine Update*, 3:100095, 2023.

[74] Nuno Gomes, Matilde Pato, André Ribeiro Lourenço, and Nuno Datia. A survey on wearable sensors for mental health monitoring. *Sensors*, 23(3):1330, 2023.

[75] Mijeong Kang and Kyunghwan Chai. Wearable sensing systems for monitoring mental health. *Sensors*, 22(3):994, 2022.

[76] Nannan Long, Yongxiang Lei, Lianhua Peng, Ping Xu, Ping Mao, et al. A scoping review on monitoring mental health using smart wearable devices. *Math. Biosci. Eng*, 19(8):7899–7919, 2022.

[77] Mohan Babu, Ziv Lautman, Xiangping Lin, Milan HB Sobota, and Michael P Snyder. Wearable devices: Implications for precision medicine and the future of health care. *Annual Review of Medicine*, 75:401–415, 2024.

[78] Pew Research Center. Mobile fact sheet, 2024. Available online at: `https://www.pewresearch.org/internet/fact-sheet/mobile/`, last accessed on 2024-05-09.

[79] Pew Research Center. About one-in-five Americans use a smart watch or fitness tracker, 2020. Available online at: `https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/`, last accessed on 2024-05-09.

[80] Ruba Fadul, Hessa Alfalahi, Aamna Al Shehhi, and Leontios Hadjileontiadis.

Depressive disorder remote detection through touchscreen typing behaviour. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[81] Rafail-Evangelos Mastoras, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Seada Kassie, Taoufik Alsaadi, Ahsan Khandoker, and Leontios J Hadjileontiadis. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific reports*, 9(1):13414, 2019.

[82] Jaeeun Shin and Sung Man Bae. A systematic review of location data for depression prediction. *International Journal of Environmental Research and Public Health*, 20(11):5984, 2023.

[83] Elicia Toon, Margot J Davey, Samantha L Hollis, Gillian M Nixon, Rosemary SC Horne, and Sarah N Biggs. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and psg in a clinical cohort of children and adolescents. *Journal of Clinical Sleep Medicine*, 12(3):343–350, 2016.

[84] Ju-Yu Wu, Congo Tak-Shing Ching, Hui-Min David Wang, and Lun-De Liao. Emerging wearable biosensor technologies for stress monitoring and their real-world applications, 2022.

[85] Abigail Ortiz, Kamil Bradler, and Arend Hintze. Episode forecasting in bipolar disorder: Is energy better than mood? *Bipolar disorders*, 20(5):470–476, 2018.

[86] Ian B Hickie, Jan Scott, Kathleen R Merikangas, and Elizabeth M Scott. Are depressive and other mood disorders best conceptualized as disorders of energy, and related motor activity, rather than mood? *Research Directions: Depression*, 1:e4, 2024.

[87] Guohua Lu, F Yang, J Andrew Taylor, and John F Stein. A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. *Journal of medical engineering & technology*, 33(8):634–641, 2009.

[88] Wayne S Fenton and Ellen S Stover. Mood disorders: cardiovascular and diabetes comorbidity. *Current Opinion in Psychiatry*, 19(4):421–427, 2006.

[89] Ashvita Ramesh, Tanvi Nayak, Molly Beestrum, Giorgio Quer, and Jay A Pandit.

Heart rate variability in psychiatric disorders: A systematic review. *Neuropsychiatric Disease and Treatment*, pages 2217–2239, 2023.

[90] Ashley E Mason, Patrick Kasl, Severine Soltani, Abigail Green, Wendy Hartogensis, Stephan Dilchert, Anoushka Chowdhary, Leena S Pandya, Chelsea J Siwik, Simmie L Foster, et al. Elevated body temperature is associated with depressive symptoms: results from the tempredict study. *Scientific Reports*, 14 (1):1884, 2024.

[91] Alan Brnabic and Lisa M Hess. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC medical informatics and decision making*, 21:1–19, 2021.

[92] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[93] Grand View Research, Inc. AI In Healthcare Market Size, & Share Trends Analysis Report By Component (Hardware, Services), By Application, By End-use, By Technology, By Region, And Segment Forecasts, 2024 - 2030, 2024. URL `https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market`.

[94] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry*, 11:584711, 2020.

[95] Ulysse Côté-Allard, Petter Jakobsen, Andrea Stautland, Tine Nordgreen, Ole Bernt Fasmer, Ketil Joachim Oedegaard, and Jim Tørresen. Long–short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors. *IEEE Pervasive Computing*, 2022.

[96] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag, and Rosalind Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, pages 325–332. IEEE, 2017.

[97] Yuuki Tazawa, Kuo-ching Liang, Michitaka Yoshimura, Momoko Kitazawa, Yuriko Kaise, Akihiro Takamiya, Aiko Kishi, Toshiro Horigome, Yasue Mit-

sukura, Masaru Mimura, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*, 6(2):e03274, 2020.

[98] Duc-Khanh Nguyen, Chien-Lung Chan, Ai-Hsien A Li, Dinh-Van Phan, and Chung-Hsien Lan. Decision support system for the differentiation of schizophrenia and mood disorders using multiple deep learning models on wearable devices data. *Health Informatics Journal*, 28(4):14604582221137537, 2022.

[99] Heon-Jeong Lee, Chul-Hyun Cho, Taek Lee, Jaegwon Jeong, Ji Won Yeom, Sojeong Kim, Sehyun Jeon, Ju Yeon Seo, Eunsoo Moon, Ji Hyun Baek, et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in south korea: a prospective nationwide cohort study. *Psychological Medicine*, pages 1–9, 2022.

[100] Alaa Abd-Alrazaq, Rawan AlSaad, Farag Shuweihdi, Arfan Ahmed, Sarah Aziz, and Javaid Sheikh. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *NPJ Digital Medicine*, 6(1):84, 2023.

[101] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006.

[102] ETH Zurich. Empatica user manual, 2023. URL `https://archive.arch.ethz.ch/esum/downloads/manuals/emaptics.pdf`. last accessed on 2023-09-13.

[103] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[104] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[105] Daniel A Adler, Dror Ben-Zeev, Vincent WS Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR mHealth and uHealth*, 8 (8):e19962, 2020.

[106] Luigi A Maglanoc, Nils Inge Landrø, Rune Jonassen, Tobias Kaufmann, Aldo Córdova-Palomera, Eva Hilland, and Lars T Westlye. Data-driven clustering reveals a link between symptoms and functional brain connectivity in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(1):16–26, 2019.

[107] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[108] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[109] Guozhu Dong and Huan Liu. *Feature engineering for machine learning and data analytics*. CRC press, 2018.

[110] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.

[111] Shan Suthaharan and Shan Suthaharan. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235, 2016.

[112] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[113] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann. Flirt: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine*, 212: 106461, 2021.

[114] Simon JD Prince. *Understanding Deep Learning*. MIT press, 2023.

[115] K Shailaja, Banoth Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018.

[116] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[117] M Panagiotou, Athanasia Zlatintsi, Panagiotis Paraskevas Filntisis, AJ Roumeli-

otis, Niki Efthymiou, and Petros Maragos. A comparative study of autoencoder architectures for mental health analysis using wearable sensors data. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1258–1262. IEEE, 2022.

[118] Kennedy Opoku Asare, Aku Visuri, Julio Vega, and Denzil Ferreira. Me in the wild: An exploratory study using smartphones to detect the onset of depression. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 121–145. Springer, 2021.

[119] Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[120] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[121] Md Atiqur Rahman Ahad, Anindya Das Antar, Masud Ahmed, Md Atiqur Rahman Ahad, Anindya Das Antar, and Masud Ahmed. Sensor-based benchmark datasets: comparison and analysis. *IoT Sensor-Based Activity Recognition: Human Activity Recognition*, pages 95–121, 2021.

[122] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research), 2010.

[123] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[124] Nicholas C Jacobson, Hilary Weingarden, and Sabine Wilhelm. Digital biomarkers of mood disorders and symptom change. *NPJ digital medicine*, 2(1):3, 2019.

[125] Filippo Corponi, Bryan M Li, Gerard Anmella, Ariadna Mas, Isabella Pacchiarotti, Marc Valentí, Iria Grande, Antoni Benabarre, Marina Garriga, Eduard Vieta, et al. Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. *Translational Psychiatry*, 14(1):161, 2024.

[126] Mousa Alavi, Erik Biros, and Michelle Cleary. A primer of inter-rater reliability in clinical measurement studies: Pros and pitfalls. *Journal of Clinical Nursing*,

31(23-24):e39–e42, 2022. doi: https://doi.org/10.1111/jocn.16514. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jocn.16514.

[127] A Matthew Prina, Theodore D Cosco, Tom Dening, Aartjan Beekman, Carol Brayne, and Martijn Huisman. The association between depressive symptoms in the community, non-psychiatric hospital admission and hospital outcomes: a systematic review. *Journal of psychosomatic research*, 78(1):25–33, 2015.

[128] Gernot Fugger, Thomas Waldhör, Barbara Hinterbuchinger, Nathalie Pruckner, Daniel König, Andrea Gmeiner, Sandra Vyssoki, Benjamin Vyssoki, and Matthäus Fellinger. Pattern of inpatient care for depression: an analysis of 232,289 admissions. *Bmc Psychiatry*, 20:1–7, 2020.

[129] U Ösby, A Tiainen, L Backlund, G Edman, M Adler, J Hällgren, K Sennfält, M van Baardewijk, and P Sparen. Psychiatric admissions and hospitalization costs in bipolar disorder in sweden. *Journal of affective disorders*, 115(3): 315–322, 2009.

[130] Hamish Innes, James Lewsey, and Daniel J Smith. Predictors of admission and readmission to hospital for major depression: a community cohort study of 52,990 individuals. *Journal of Affective Disorders*, 183:10–14, 2015.

[131] Daniel A Adler, Caitlin A Stamatis, Jonah Meyerhoff, David C Mohr, Fei Wang, Gabriel J Aranovich, Srijan Sen, and Tanzeem Choudhury. Measuring algorithmic bias to analyze the reliability of ai tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research*, 3(1):17, 2024.

[132] Jeremy F Huckins, Alex W DaSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K Nepal, Jialing Wu, Mikio Obuchi, Eilis I Murphy, Meghan L Meyer, et al. Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of medical Internet research*, 22(6): e20185, 2020.

[133] Bingzhao Zhu and Mahsa Shoaran. Unsupervised domain adaptation for cross-subject few-shot neurological symptom detection. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 181–184. IEEE, 2021.

[134] Rosie Dobson, Melanie Stowell, Jim Warren, Taria Tane, Lin Ni, Yulong Gu,

Judith McCool, and Robyn Whittaker. Use of consumer wearables in health research: Issues and considerations. *Journal of Medical Internet Research*, 25: e52444, 2023.

[135] Jukka-Pekka Onnela. Exporting the same data from a wearable twice doesn't give you the same data, Year. Available online at: http://www.beiwe.org/exporting-the-same-data-from-a-wearable-twice-doesnt-give-you-the-same-data/, last accessed on 2024-06-29.

[136] Vincenzo Ronca, Ana C Martinez-Levy, Alessia Vozzi, Andrea Giorgi, Pietro Aricò, Rossella Capotorto, Gianluca Borghini, Fabio Babiloni, and Gianluca Di Flumeri. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between fitbit sense, empatica e4 and shimmer gsr3+. *Sensors*, 23(13):5847, 2023.

[137] Simone Benedetto, Christian Caldato, Elia Bazzan, Darren C Greenwood, Virginia Pensabene, and Paolo Actis. Assessment of the fitbit charge 2 for monitoring heart rate. *PloS one*, 13(2):e0192691, 2018.

[138] Salvatore Tedesco, Marco Sica, Andrea Ancillao, Suzanne Timmons, John Barton, and Brendan O'Flynn. Accuracy of consumer-level and research-grade activity trackers in ambulatory settings in older adults. *PloS one*, 14(5):e0216891, 2019.

[139] Faith Matcham, C Barattieri di San Pietro, Viola Bulgari, G De Girolamo, R Dobson, Hans Eriksson, AA Folarin, Josep Maria Haro, Maximilian Kerz, Femke Lamers, et al. Remote assessment of disease and relapse in major depressive disorder (radar-mdd): a multi-centre prospective cohort study protocol. *BMC psychiatry*, 19:1–11, 2019.

[140] Nila Armelia Windasari and Fu-ren Lin. Why do people continue using fitness wearables? the effect of interactivity and gamification. *Sage Open*, 11(4): 21582440211056606, 2021.

[141] A Saranya and R Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230, 2023.

[142] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[143] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[144] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[145] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[146] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[147] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[148] Weijie Chen, Berkman Sahiner, Frank Samuelson, Aria Pezeshk, and Nicholas Petrick. Calibration of medical diagnostic classifier scores to the probability of disease. *Statistical methods in medical research*, 27(5):1394–1409, 2018.

[149] Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, page 107441, 2023.

[150] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[151] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

[152] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[153] Petter Jakobsen, Ulysse Côté-Allard, Michael Alexander Riegler, Lena Antonsen Stabell, Andrea Stautland, Tine Nordgreen, Jim Torresen, Ole Bernt Fasmer, and Ketil Joachim Oedegaard. Early warning signals observed in motor activity preceding mood state change in bipolar disorder. *Bipolar Disorders*, 2024.

[154] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

[155] Veronica Dudarev, Oswald Barral, Chuxuan Zhang, Guy Davis, and James T Enns. On the reliability of wearable technology: A tutorial on measuring heart rate and heart rate variability in the wild. *Sensors*, 23(13):5863, 2023.

[156] Elisa Mejía-Mejía, James M May, Robinson Torres, and Panayiotis A Kyriacou. Pulse rate variability in cardiovascular health: A review on its applications and relationship with heart rate variability. *Physiological Measurement*, 41(7): 07TR01, 2020.

[157] DS Quintana, Gail A Alvares, and JAJ Heathers. Guidelines for reporting articles on psychiatry and heart rate variability (graph): recommendations to advance research communication. *Translational psychiatry*, 6(5):e803–e803, 2016.

[158] Shohei Sato, Takuma Hiratsuka, Kenya Hasegawa, Keisuke Watanabe, Yusuke Obara, Nobutoshi Kariya, Toshikazu Shinba, and Takemi Matsui. Screening for major depressive disorder using a wearable ultra-short-term hrv monitor and signal quality indices. *Sensors*, 23(8):3867, 2023.

[159] Jason D Stone, Hana K Ulman, Kaylee Tran, Andrew G Thompson, Manuel D Halter, Jad H Ramadan, Mark Stephenson, Victor S Finomore Jr, Scott M Galster, Ali R Rezai, et al. Assessing the accuracy of popular commercial technologies that measure resting heart rate and heart rate variability. *Frontiers in Sports and Active Living*, page 37, 2021.

[160] Stefanie Hillebrand, Karin B Gast, Renée de Mutsert, Cees A Swenne, J Wouter Jukema, Saskia Middeldorp, Frits R Rosendaal, and Olaf M Dekkers. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose–response meta-regression. *Europace*, 15 (5):742–749, 2013.

[161] Francesco Sessa, Valenzano Anna, Giovanni Messina, Giuseppe Cibelli, Vincenzo Monda, Gabriella Marsala, Maria Ruberto, Antonio Biondi, Orazio Cascio,

Giuseppe Bertozzi, et al. Heart rate variability as predictive factor for sudden cardiac death. *Aging (Albany NY)*, 10(2):166, 2018.

[162] Scott Michael, Kenneth S Graham, and Glen M Davis. Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals—a review. *Frontiers in physiology*, 8:259883, 2017.

[163] Gail A Alvares, Daniel S Quintana, Ian B Hickie, and Adam J Guastella. Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis. *Journal of psychiatry and neuroscience*, 41(2):89–104, 2016.

[164] John A Chalmers, Daniel S Quintana, Maree J-Anne Abbott, and Andrew H Kemp. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. *Frontiers in psychiatry*, 5:80, 2014.

[165] Celine Koch, Marcel Wilhelm, Stefan Salzmann, Winfried Rief, and Frank Euteneuer. A meta-analysis of heart rate variability in major depression. *Psychological Medicine*, 49(12):1948–1957, 2019.

[166] Zuxing Wang, Yuanyuan Luo, Yuan Zhang, Lili Chen, Yazhu Zou, Jun Xiao, Wenjiao Min, Cui Yuan, Yu Ye, Mingmei Li, et al. Heart rate variability in generalized anxiety disorder, major depressive disorder and panic disorder: A network meta-analysis and systematic review. *Journal of Affective Disorders*, 2023.

[167] Andrew H Kemp, Daniel S Quintana, Marcus A Gray, Kim L Felmingham, Kerri Brown, and Justine M Gatt. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. *Biological psychiatry*, 67 (11):1067–1074, 2010.

[168] Guilherme Luiz Lopes Wazen, Michele Lima Gregório, Andrew Haddon Kemp, and Moacir Fernandes de Godoy. Heart rate variability in patients with bipolar disorder: from mania to euthymia. *Journal of psychiatric research*, 99:33–38, 2018.

[169] Brandon Hage, Briana Britton, David Daniels, Keri Heilman, Stephen W Porges, and Angelos Halaris. Diminution of heart rate variability in bipolar depression. *Frontiers in public health*, 5:312, 2017.

[170] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[171] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[172] Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C Luhmann, Osvaldo A Martin, et al. Pymc: a modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516, 2023.

[173] Daniel J Plews, Paul B Laursen, Jamie Stanley, Andrew E Kilding, and Martin Buchheit. Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring. *Sports medicine*, 43:773–781, 2013.

[174] Daniel J Plews, Paul B Laursen, Yann Le Meur, Christophe Hausswirth, Andrew E Kilding, and Martin Buchheit. Monitoring training with heart-rate variability: How much compliance is needed for valid assessment? *International journal of sports physiology and performance*, 9(5):783–790, 2014.

[175] MP Tarvainen, J Lipponen, JP Niskanen, and P Ranta-Aho. Kubios hrv version 3–user's guide. *Kuopio: University of Eastern Finland*, 2017.

[176] Olli-Pekka Nuuttila, Ari Nummela, Elisa Korhonen, Keijo Häkkinen, and Heikki Kyröläinen. Individualized endurance training based on recovery and training status in recreational runners. *Medicine and science in sports and exercise*, 54 (10), 2022.

[177] Bruce B Way, Michael H Allen, Jeryl L Mumpower, Thomas R Stewart, and Steven M Banks. Interrater agreement among psychiatrists in psychiatric emergency assessments. *American Journal of Psychiatry*, 155(10):1423–1428, 1998.

[178] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[179] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[180] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[181] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[182] Shiyi Qi, Liangjian Wen, Yiduo Li, Yuanhang Yang, Zhe Li, Zhongwen Rao, Lujia Pan, and Zenglin Xu. Enhancing multivariate time series forecasting with mutual information-driven cross-variable and temporal modeling. *arXiv preprint arXiv:2403.00869*, 2024.

[183] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[184] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[185] Amina Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):24, 2021.

[186] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[187] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128:104115, 2021.

[188] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[189] Yannik Hahn, Tristan Langer, Richard Meyes, and Tobias Meisen. Time series dataset survey for forecasting with deep learning. *Forecasting*, 5(1):315–335, 2023.

[190] Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30(4):2761–2775, 2023.

[191] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[192] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[193] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[194] Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised learning for time series: Contrastive or generative? *arXiv preprint arXiv:2403.09809*, 2024.

[195] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[196] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[197] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

[198] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023.

[199] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*, 2024.

[200] Sabera J Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: Tokenized time series embeddings for general time series analysis. In *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024. URL https://openreview.net/forum?id=jzIdR2TXlK.

[201] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22584–22591, 2024.

[202] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[203] Filippo Corponi. E4SelfLearning: Dataset hosted on Hugging Face datasets. https://huggingface.co/datasets/FcmC/E4SelfLearning, 2024. last accessed on 2024-07-28.

[204] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *arXiv preprint arXiv:2405.19363*, 2024.

[205] Yuezhou Zhang, Abhishek Pratap, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Faith Matcham, Srinivasan Vairavan, Judith Dineley, Yatharth Ranjan, Zulqarnain Rashid, et al. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. *NPJ digital medicine*, 6(1):25, 2023.

[206] Judith Borghouts, Elizabeth Eikey, Gloria Mark, Cinthia De Leon, Stephen M Schueller, Margaret Schneider, Nicole Stadnick, Kai Zheng, Dana Mukamel, and Dara H Sorkin. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. *Journal of medical Internet research*, 23 (3):e24387, 2021.

[207] Adam M Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R Corlett, Nikolaos Koutsouleris, Harlan M Krumholz, et al. Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167, 2024.

[208] Dongbin Kim, Jinseong Park, Jaewook Lee, and Hoki Kim. Are self-attentions effective for time series forecasting? *arXiv preprint arXiv:2405.16877*, 2024.

Appendix A

Hamilton Depression Rating Scale

Hamilton Depression Rating Scale (HDRS) [63] which consist of 17 items, each with a score to indicate severity of the symptom.

H1. Depressed Mood (sadness, hopeless, helpless, worthless)

0 - Absent.

1 - These feeling states indicated only on questioning.

2 - These feeling states spontaneously reported verbally.

3 - Communicates feeling states non-verbally, i.e. through facial expression, posture, voice and tendency to weep.

4 - Patient reports virtually only these feeling states in his/her spontaneous verbal and non-verbal communication

H2. Feelings of guilt

0 - Absent.

1 - Self reproach, feels he/she has let people down.

2 - Ideas of guilt or rumination over past errors or sinful deeds.

3 - Present illness is a punishment. Delusions of guilt.

4 - Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations.

H3. Suicide

0 - Absent.

1 - Feels life is not worth living.

2 - Wishes he/she were dead or any thoughts of possible death to self.

3 - Ideas or gestures of suicide.

4 - Attempts at suicide (any serious attempt rate 4).

H4. Insomnia: early in the night

0 - No difficulty falling asleep.

1 - Complains of occasional difficulty falling asleep, i.e. more than 1⁄2 hour.

2 - Complains of nightly difficulty falling asleep.

H5. Insomnia: middle of the night

0 - No difficulty.

1 - Patient complains of being restless and disturbed during the night.

2 - Waking during the night – any getting out of bed rates 2 (except for purposes of voiding).

H6. Insomnia: early hours of the morning

0 - No difficulty.

1 - Waking in early hours of the morning but goes back to sleep.

2 - Unable to fall asleep again if he/she gets out of bed.

H7. Work and Activities

0 - No difficulty.

1 - Thoughts and feelings of incapacity, fatigue or weakness related to activities, work or hobbies.

2 - Loss of interest in activity, hobbies or work – either directly reported by the patient or indirect in listlessness, indecision and vacillation (feels he/she has to push self to work or activities).

3 - Decrease in actual time spent in activities or decrease in productivity. Rate 3 if the patient does not spend at least three hours a day in activities (job or

hobbies) excluding routine chores.

4 - Stopped working because of present illness. Rate 4 if patient engages in no activities except routine chores, or if patient fails to perform routine chores unassisted.

H8. Retardation

0 - Normal speech and thought.

1 - Slight retardation during the interview.

2 - Obvious retardation during the interview.

3 - Interview difficult.

4 - Complete stupor.

H9. Agitation

0 - None.

1 - Fidgetiness.

2 - Playing with hands, hair, etc.

3 - Moving about, can't sit still.

4 - Hand wringing, nail biting, hair-pulling, biting of lips.

H10. Anxiety Psychic

0 - No difficulty.

1 - Subjective tension and irritability.

2 - Worrying about minor matters.

3 - Apprehensive attitude apparent in face or speech.

4 - Fears expressed without questioning.

H11. Anxiety Somatic (physiological concomitants of anxiety)

0 - Absent.

1 - Mild.

2 - Moderate.

3 - Severe.

4 - Incapacitating.

H12. Somatic Symptoms Gastro-Intestinal

0 - None.

1 - Loss of appetite but eating without staff encouragement. Heavy feelings in abdomen.

2 - Difficulty eating without staff urging. Requests or requires laxatives or medication for bowels or medication for gastro-intestinal symptoms.

H13. General Somatic Symptoms

0 - None.

1 - Heaviness in limbs, back or head. Backaches, headaches, muscle aches. Loss of energy and fatigability.

2 - Any clear-cut symptom rates 2.

H14. Genital Symptoms

0 - Absent.

1 - Mild.

2 - Severe.

H15. Hypocondriasis

0 - Not present.

1 - Self-absorption (bodily).

2 - Preoccupation with health.

3 - Frequent complaints, requests for help, etc.

4 - Hypochondriacal delusions.

H16. Loss of Weight

0 - Less than 1 lb weight loss in week.

1 - Greater than 1 lb weight loss in week.

2 - Greater than 2 lb weight loss in week.

H17. Insight

0 - Acknowledges being depressed and ill.

1 - Acknowledges illness but attributes cause to bad food, climate, overwork, virus, need for rest, etc.

2 - Denies being ill at all.

Young Mania Rating Scale

Young Mania Rating Scale (YMRS) [64] which consist of 11 items, each with a score to indicate severity of the symptom.

Y1. Elevated Mood

0 - Absent.

1 - Mildly or possibly increased on questioning.

2 - Definite subjective elevation; optimistic, self-confident; cheerful; appropriate to content.

3 - Elevated; inappropriate to content; humorous.

4 - Euphoric; inappropriate laughter; singing.

Y2. Increased Motor Activity-Energy

0 - Absent.

1 - Subjectively increased.

2 - Animated; gestures increased.

3 - Excessive energy; hyperactive at times; restless (can be calmed).

4 - Motor excitement; continuous hyperactivity (cannot be calmed).

Y3. Sexual Interest

0 - Normal; not increased.

1 - Mildly or possibly increased.

2 - Definite subjective increase on questioning.

3 - Spontaneous sexual content; elaborates on sexual matters; hypersexual by self-report.

4 - Overt sexual acts (toward patients, staff, or interviewer).

Y4. Sleep

0 - Reports no decrease in sleep.

1 - Sleeping less than normal amount by up to one hour.

2 - Sleeping less than normal by more than one hour.

3 - Reports decreased need for sleep.

4 - Denies need for sleep.

Y5. Irritability

0 - Absent.

2 - Subjectively increased.

4 - Irritable at times during interview; recent episodes of anger or annoyance on ward.

6 - Frequently irritable during interview; short, curt throughout.

8 - Hostile, uncooperative; interview impossible.

Y6. Speech (Rate and Amount)

0 - No increase.

2 - Feels talkative.

4 - Increased rate or amount at times, verbose at times.

6 - Push; consistently increased rate and amount; difficult to interrupt.

8 - Pressured; uninterruptible, continuous speech.

Y7. Language-Thought Disorder

0 - Absent.

1 - Circumstantial; mild distractibility; quick thoughts.

2 - Distractible, loses goal of thought; changes topics frequently; racing thoughts.

3 - Flight of ideas; tangentiality; difficult to follow; rhyming, echolalia.

4 - Incoherent; communication impossible.

Y8. Content

0 - Normal.

2 - Questionable plans, new interests.

4 - Special project(s); hyper-religious.

6 - Grandiose or paranoid ideas; ideas of reference.

8 - Delusions; hallucinations.

Y9. Disruptive-Aggressive Behavior

0 - Absent, cooperative.

2 - Sarcastic; loud at times, guarded.

4 - Demanding; threats on ward.

6 - Threatens interviewer; shouting; interview difficult.

8 - Assaultive; destructive; interview impossible.

Y10. Appearance

0 - Appropriate dress and grooming.

1 - Minimally unkempt.

2 - Poorly groomed; moderately disheveled; overdressed.

3 - Disheveled; partly clothed; garish make-up.

4 - Completely unkempt; decorated; bizarre garb.

Y11. Insight

0 - Present; admits illness; agrees with need for treatment.

1 - Possibly ill.

2 - Admits behavior change, but denies illness.

3 - Admits possible change in behavior, but denies illness.

4 - Denies any behavior change.