Inferring mood disorder symptoms from multivariate time-series sensory data

Bryan M. Li[1]* **Filippo Corponi**[1]* **Gerard Anmella**[2] **Ariadna Mas**[2]
Miriam Sanabra[2] **Diego Hiadalgo-Mazzei**[2] **Antonio Vergari**[1]

[1]School of Informatics, University of Edinburgh
{bryan.li, filippo.corponi, avergari}@ed.ac.uk

[2]Hospital Clínic de Barcelona, University of Barcelona
{anmella, amasm, sanabra, dahidalg}@clinic.cat

Abstract

Mood disorders are increasingly recognized among the leading causes of disease burden worldwide. Depressive and manic episodes in mood disorders commonly involve altered mood, sleep, and motor activity. These translate to changes in sensory data that wearable devices can continuously and affordably monitor, thereby positioning themselves as a promising candidate to model mood disorders. Previous similar endeavors cast this problem in terms of binary classification (cases vs controls) or regress the total score of some commonly used psychometric scale. Nevertheless, these approaches fail to capture the variability within symptom domains described at the item level in psychometric scales. In this work, we attempt to infer mood disorder symptoms (e.g., depressed mood, insomnia, irritability) from time-series data collected with the medical grade Empatica E4 wristbands, as part of an exploratory, observational, and longitudinal study. We propose a multi-label framework to predict individual items from the two most widely used scales for assessing depression and mania. We experiment with two different approaches to preprocess the high-dimensional and noisy sensory data and attain results within a clinically acceptable level of error.

1 Introduction

Mood disorders, also referred to as affective disorders, are a group of diagnoses in the Diagnostic and Statistical Manual 5th edition [2] classification system, ranked among the top 25 leading causes of disease burden worldwide [18]. These are characterized by depressive and/or (hypo)manic episodes that have disturbances in mood, sleep patterns, and motor activity as predominant features. Such alterations correlate with changes in physiological parameters that wearable devices can continuously and affordably record in a patient's natural environment. Against the backdrop of a disappointingly limited clinical translation of psychiatric genetics and neuroscience research [9], the community has been looking to digital biomarkers from wearables as an alternative (complementary) paradigm [10].

The question we pursue is to which degree time-series data can be used to infer mood disorder symptoms pertaining to depression and mania, the two polarities of mood disorders, as described with Hamilton Depression Rating Scale-17 (HDRS) [8] and the Young Mania Rating Scale (YMRS) [24] respectively. These are 17-item and 11-item questionnaires used to measure depressive and manic symptoms respectively and take about 20 minutes each to administer. Questionnaire individual items

*Equal contributions.

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

are not scored on the same scale but different items have different rank numbers and some have a step-size of two between consecutive ranks to reflect that these should weigh more when computing the scale total score, i.e. the sum over individual items' score (ranging from 0 to 52 and from 0 to 60 for HDRS and YMRS respectively). A complete account of HDRS and YMRS items is given in Table A.1 and Table A.2 respectively. The nature of symptom domains mapped with HDRS and YMRS varies from mood, sleep, and psychomotor activity.

Our contribution is twofold. First, we cast depression and mania prediction in a multi-label learning framework, where we aim to infer each item in HDRS and YMRS. These fine-grained predictions in individual items in the psychometric scales are very desirable, as they provide a richer symptoms description, i.e. they clarify which symptom domains characterize the ongoing mood episode rather than a mere yes-or-no diagnosis. To the best of our knowledge, this is the first work to predict individual items in HDRS and YMRS for mood disorders. The majority of previous related works, adopted a binary (e.g., disease vs control, or acute vs remission) classification framework [5, 16] only one attempted to regress the HDRS total score [7]. However, patients with the same total score on YMRS and HDRS might have different clinical presentations. Take for example the case of two patients with the same depression severity level but one displaying psycho-motor agitation while the other psycho-motor retardation (two items from HDRS): this clinically obvious difference would go lost when reducing a psychometric scale to its total score and, as a result, it would not be used to personalize treatment. As an additional contribution, further to the standard approach of time-aligning multivariate time-series during preprocessing [1, 13], we experiment with learning an embedding representation of the raw recordings.

2 Multi-label learning of mood disorder symptoms

2.1 Study population & assessment

The analyses that follow are based on an original dataset 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. 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. All data were collected anonymously and stored encrypted in servers complying with all GDPR and HIPAA regulations.

Patients with an ongoing mood episode, patients with a historical mood disorder diagnosis clinically stable at present, and healthy controls are recruited. While healthy controls and euthymic patients are assessed only once, measurements on patients with a mood episode are taken at four time-points: acute phase (upon hospital admission or at the home treatment unit), response onset (usually mid-admission), remission (end of admission or soon after discharge), and recovery ($\sim$2 months into sustained remission). During each assessment, participants are interviewed by a specialized psychiatrist collecting socio-demographic and treatment info as well as YMRS and HDRS scores.

At the start of the clinical interview, participants are provided with an E4 Empatica wristband[2] and they are required to wear for about 48 hours. E4 devices have sensors collecting the following physiological data (sampling rates): 3D acceleration (ACC, 32Hz), inter-beat intervals (IBI, i.e. the time between two consecutive heart ventricular contractions), skin temperature (TEMP, 4Hz), blood volume pressure (BVP, 64Hz), electrodermal activity (EDA, 4Hz), and heart rate (HR, 1Hz). In total, data obtained from 25 healthy controls and 64 patients with mood disorders are used in this work, and an overview of the clinical-demographic variables of the population is available in Table 1. Figure A.1 shows the number of recording sessions by diagnosis.

2.2 Preprocessing

The raw data from an E4 Empatica recording session comes as a collection of recorded channels: ACC (3-dimensional), EDA, BVP, IBI, HR, and TEMP. IBI is computed from BVP signals and the reliability of its derivation depends upon the level of movement exhibited by the wearer[3]. Since IBI

[2] empatica.com/en-int/research/e4

[3] support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal

Table 1: Clinical-demographic summary of the study population (N = 89). HC (Healthy Controls); IQR (inter-quartile range); SD (standard deviation); N (number).

Age (years)	Mean (SD)	Median (IQR)
	41.34 (14.68)	39.5 (24.75)
Sex	Males - N (%)	Females - N (%)
	44 (49.44%)	45 (50.56%)
Diagnosis	HC - N (%)	Disease - N (%)
	25 (28.09%)	64 (71.91%)
YMRS (total score)	Mean (SD)	Median (IQR)
	7.54 (10.27)	3 (12)
HDRS (total score)	Mean (SD)	Median (IQR)
	6.86 (7.88)	4 (6)

is mostly used to study heart rate variability, which is not the focus of this study, we thus decided to exclude the channel from this work. Data from wearable devices (especially in a naturalistic setting) is inherently noisy, we, therefore, quality-controlled our data with the rules by Kleckner et al. [12] and the addition of a rule to remove HR values that exceed the physiologically plausible range (25-250 bpm). Each recording session was segmented using a sliding window of w and enforcing no overlap between bordering segments. After a grid search on $w \in (4, 2048)$, we found that a segment length of $w = 64$ seconds to be the optimal size. Since the sampling rate varies across different channels, the raw recordings were time-aligned to the level of a second in wall time. This method was used in a number of works that involve sensory data [1, 13], nevertheless, the down-sampling process (usually via max-pooling or averaging) can risk removing useful information in the raw recordings. As an alternative, we instead learned a latent representation for raw channel data, such dimension reduction approach has been proved effective in other tasks such as language and signal processing [15, 19]. To that end, we used either a fully-connected layer or a Gated Recurrent Unit (GRU) [4] layer to learn an embedding for each channel of the same dimensionality, i.e. 128. We could then concatenate the embeddings and feed them to the regression model in the same manner as the time-aligned data.

2.3 Experimental Design

We randomly divided each recording session from our dataset into train, validation, and test sets with a ratio of 70-15-15. We took a hard parameter-sharing approach to the multi-label problem, i.e. all 28 items shared the same model trunk, and thus the same base representation of the input data, but had one task-specific output layer. We experimented with a bidirectional Long Short-Term Memory (BiLSTM) model [20] with 256 hidden units and tanh activation followed by a dense layer to output 28 items. The model was trained to minimize the sum of mean squared error (MSE) from individual items weighted proportionally to the step size between consecutive ranks. The channel embeddings and the model were trained with Adam optimizer [11] end-to-end, with early stopping, for a maximum of 100 epochs. The codebase is available at github.com/INTREPIBD/TS4H2022.

3 Results

Towards comparability with previous studies and easier clinical interpretation, we computed the Root Mean Square Error (RMSE) on YMRS and HDRS in terms of total scores, i.e. RMSE over the sum of YMRS and HDRS, and item average, i.e. the average per-item RMSE over YMRS and HDRS. The results are shown in Table 2. Models trained with embedding representation of the raw recordings result in better test performance across both HDRS and YMRS, with GRU outperforming MLP embeddings (RMSE on YMRS total score = 5.6340; RMSE on HDRS total score = 4.4089).

An illustration of the per-item test performance (average and standard deviation of the residuals, i.e. signed difference between predicted and observed item values) is provided in Figure 1. Residuals are generally lower on the HDRS items. Some items stand out as having either very high (e.g. YMRS-6,

Table 2: Regression results on the test set where the RMSE are computed in terms of (1) total score, i.e. the sum of all items, and (2) item average, i.e. the average error of individual items.

			Total score		Item average	
Model	Time-aligned	Embeddings (dim)	YMRS	HDRS	YMRS	HDRS
1	True	N/A	6.0558	4.5957	1.3336	0.8599
2	False	MLP (128)	5.8560	4.5602	**1.3316**	0.8561
3	False	GRU (128)	**5.6340**	**4.4089**	1.3343	**0.8550**

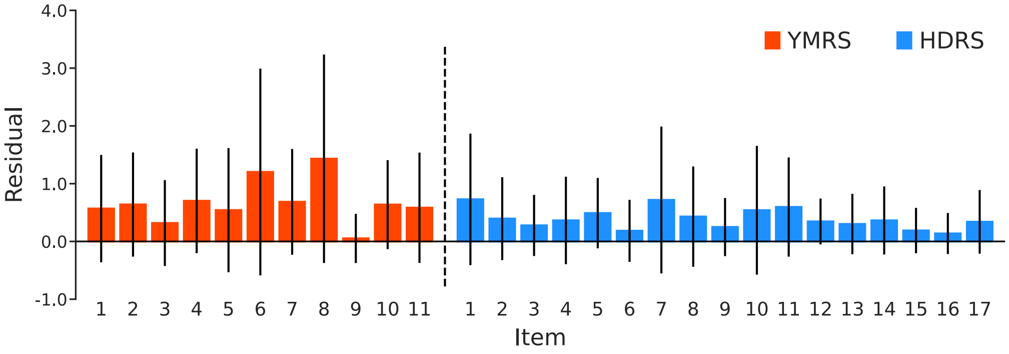

Figure 1: The per item regression differences in YMRS and HDRS using Model 3. Error bars report the standard deviation of the predictions. The residual is defined as the signed difference between the predicted and target values, therefore, a positive residual indicates an over-prediction and vice versa. Note that certain items (e.g. YMRS-6 and YMRS-8) have an interval of 2, which, naturally, can lead to larger prediction errors. Table A.1 and Table A.2 list the items in HDRS and YMRS.

YMRS-8, and HDRS-7) or very low (e.g. YMRS-9, HDRS-16) residuals. We should consider, however, that YMRS and HDRS are not designed to have the same range and have indeed as expected different distributions in our sample (Table 1). Furthermore, items admit different rank numbers (3, 4, or 5) and step-size (1 or 2) between consecutive ranks. Still, it is noteworthy that some comparable items have indeed quite different residuals distribution.

4 Discussion

It should be noted that acquiring data from a psychiatric population is a laborious process. On this note, previous studies trying to infer mental states from wearable device data in an actual psychiatric cohort had a sample size varying from about a dozen to only a few dozen subjects [1, 5, 6, 7, 21]. On the other hand, our analyses used a large sample size (N = 89). With reference to our research question, our results show that regressing YMRS and HDRS with a clinically acceptable error is viable. To provide some clinical perspective, it is common practice in psychiatry to discretize YMRS and HDRS total scores in five symptom severity bands. A five and three-point interval are the smallest bin widths for YMRS and HDRS respectively [3, 14], 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). Furthermore, our framework goes one step beyond merely providing YMRS and HDRS total scores since it outputs individual item ranks. This is of great clinical relevance since significant clinically meaningful information would be overlooked when reducing the fine-grained psychometric scales into a single score. On a more technical note, we noted that learning a latent representation of each channel can outperform time-aligning raw recordings, the preferred approach in previous works.

4.1 Future works

While it is common in the psychiatry literature to treat YMRS and HDRS items as continuous variables [17], each item is actually ordinal. For instance, YMRS-4 evaluates the sleep quality of the subject, from no decrease in sleep (rank 0) to denying the need for sleep (rank 4), which has a natural ordering. Therefore, instead of regressing on each item as nominal scales, ordinal regression can take advantage of the ordered nature of the psychiatric scales [22]. Secondly, linking the item residuals to other clinical-demographic variables available in our dataset is another future research direction. Thirdly, as psychometric measurements are noisy and prone to low rate of inter-specialists as well as intra-specialist agreement [23], unsupervised approaches would seem a promising pathway to data representations associated with clinically meaningful outcomes. Lastly, even though we ensured there were no overlapping segments, data from the same recording session could exist in the train, validation, and test sets depending on the random shuffle operation, where no-disease relevant features can potentially be learned by the model. A crucial direction of work is to assess out-of-sample performance, i.e. examine whether a model developed on a given population can generalize to new, previously unseen subjects.

Acknowledgments and Disclosure of Funding

This study was supported by a grant from Baszucki Brain Research Fund and the Instituto de Salud Carlos III (FIS PI21/00340, TIMEBASE STUDY). BML and FC 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 co-financed by the Fondo Social Europeo Plus (FSE+). Diego Hidalgo-Mazzei's research is supported by a Juan Rodés grant (JR18/00021) ISCIII.

References

[1] Adler, D. A., Wang, F., Mohr, D. C., and Choudhury, T. (2022). Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos one*, 17(4):e0266516.

[2] American Psychiatric Association, D., Association, A. P., et al. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

[3] Anderson, I., Pilling, S., Barnes, A., Bayliss, L., and Bird, V. (2010). The nice guideline on the treatment and management of depression in adults. *National Collaborating Centre for Mental Healt, National Institute for Health and Clinical Excellence. London: The British Psychological Society & The Royal College of Psychiatrists*.

[4] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[5] Côté-Allard, U., Jakobsen, P., Stautland, A., Nordgreen, T., Fasmer, O. B., Oedegaard, K. J., and Tørresen, J. (2022). Long–short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors. *IEEE Pervasive Computing*.

[6] Dalmeida, K. M. and Masala, G. L. (2021). Hrv features as viable physiological markers for stress detection using wearable devices. *Sensors*, 21(8):2873.

[7] Ghandeharioun, A., Fedor, S., Sangermano, L., Ionescu, D., Alpert, J., Dale, C., Sontag, D., and Picard, R. (2017). 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.

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

[9] Hidalgo-Mazzei, D., Young, A. H., Vieta, E., and Colom, F. (2018). Behavioural biomarkers and mobile mental health: a new paradigm. *International journal of bipolar disorders*, 6(1):1–4.

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

[11] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[12] Kleckner, I. R., Jones, R. M., Wilder-Smith, O., Wormwood, J. B., Akcakaya, M., Quigley, K. S., Lord, C., and Goodwin, M. S. (2017). Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. *IEEE Transactions on Biomedical Engineering*, 65(7):1460–1467.

[13] Li, B. M., Corponi, F., Anmella, G., Mas, A., Sanabra, M., Pacchiarotti, I., Valentí, M., Giménez-Palomo, A., Garriga, M., Agasi, I., et al. (2022). Can machine learning with data from wearable devices distinguish disease severity levels and generalise across patients? a pilot study in mania and depression. *medRxiv*.

[14] Lukasiewicz, M., Gerard, S., Besnard, A., Falissard, B., Perrin, E., Sapin, H., Tohen, M., Reed, C., Azorin, J.-M., and Group, E. S. (2013). 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. *International journal of methods in psychiatric research*, 22(1):46–58.

[15] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[16] Rykov, Y., Thach, T.-Q., Bojic, I., Christopoulos, G., Car, J., et al. (2021). Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR mHealth and uHealth*, 9(10):e24872.

[17] Salazar de Pablo, G., Moreno, D., Gonzalez-Pinto, A., Paya, B., Castro-Fonieles, J., Baeza, I., Graell, M., Arango, C., Rapado-Castro, M., and Moreno, C. (2021). Affective symptom dimensions in early-onset psychosis over time: a principal component factor analysis of the young mania rating scale and the hamilton depression rating scale. *European Child & Adolescent Psychiatry*, pages 1–14.

[18] Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., et al. (2021). 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.

[19] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[20] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

[21] Shah, R. V., Grennan, G., Zafar-Khan, M., Alim, F., Dey, S., Ramanathan, D., and Mishra, J. (2021). Personalized machine learning of depressed mood using wearables. *Translational psychiatry*, 11(1):1–18.

[22] Shi, X., Cao, W., and Raschka, S. (2021). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *arXiv preprint arXiv:2111.08851*.

[23] Way, B. B., Allen, M. H., Mumpower, J. L., Stewart, T. R., and Banks, S. M. (1998). Interrater agreement among psychiatrists in psychiatric emergency assessments. *American Journal of Psychiatry*, 155(10):1423–1428.

[24] Young, R. C., Biggs, J. T., Ziegler, V. E., and Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry*, 133(5):429–435.

A Appendix

Table A.1: Hamilton Depression Rating Scale (HDRS) by Hamilton [8]

Item	Description	Ranking
1	Depressed mood	0, 1, 2, 3, 4
2	Feelings of guilt	0, 1, 2, 3, 4
3	Suicide	0, 1, 2, 3, 4
4	Insomnia: early in the night	0, 1, 2
5	Insomnia: middle of the night	0, 1, 2
6	Insomnia: early hours of the morning	0, 1, 2
7	Work and activities	0, 1, 2, 3, 4
8	Retardation	0, 1, 2, 3, 4
9	Agitation	0, 1, 2, 3, 4
10	Anxiety psychic	0, 1, 2, 3, 4
11	Anxiety somatic	0, 1, 2, 3, 4
12	Somatic symptoms gastrointestinal	0, 1, 2
13	General somatic symptoms	0, 1, 2
14	Genital symptoms	0, 1, 2
15	Hypochondriasis	0, 1, 2, 3, 4
16	Loss of weight	0, 1, 2
17	Insight	0, 1, 2

Table A.2: Young Mania Rating Scale (YMRS) by Young et al. [24]

Item	Description	Ranking
1	Elevated mood	0, 1, 2, 3, 4
2	Increased motor activity-energy	0, 1, 2, 3, 4
3	Sexual interest	0, 1, 2, 3, 4
4	Sleep	0, 1, 2, 3, 4
5	Irritability	0, 2, 4, 6, 8
6	Speech (rate and amount)	0, 2, 4, 6, 8
7	Language-thought disorder	0, 1, 2, 3, 4
8	Content	0, 2, 4, 6, 8
9	Disruptive-aggressive behavior	0, 2, 4, 6, 8
10	Appearance	0, 1, 2, 3, 4
11	Insight	0, 1, 2, 3, 4

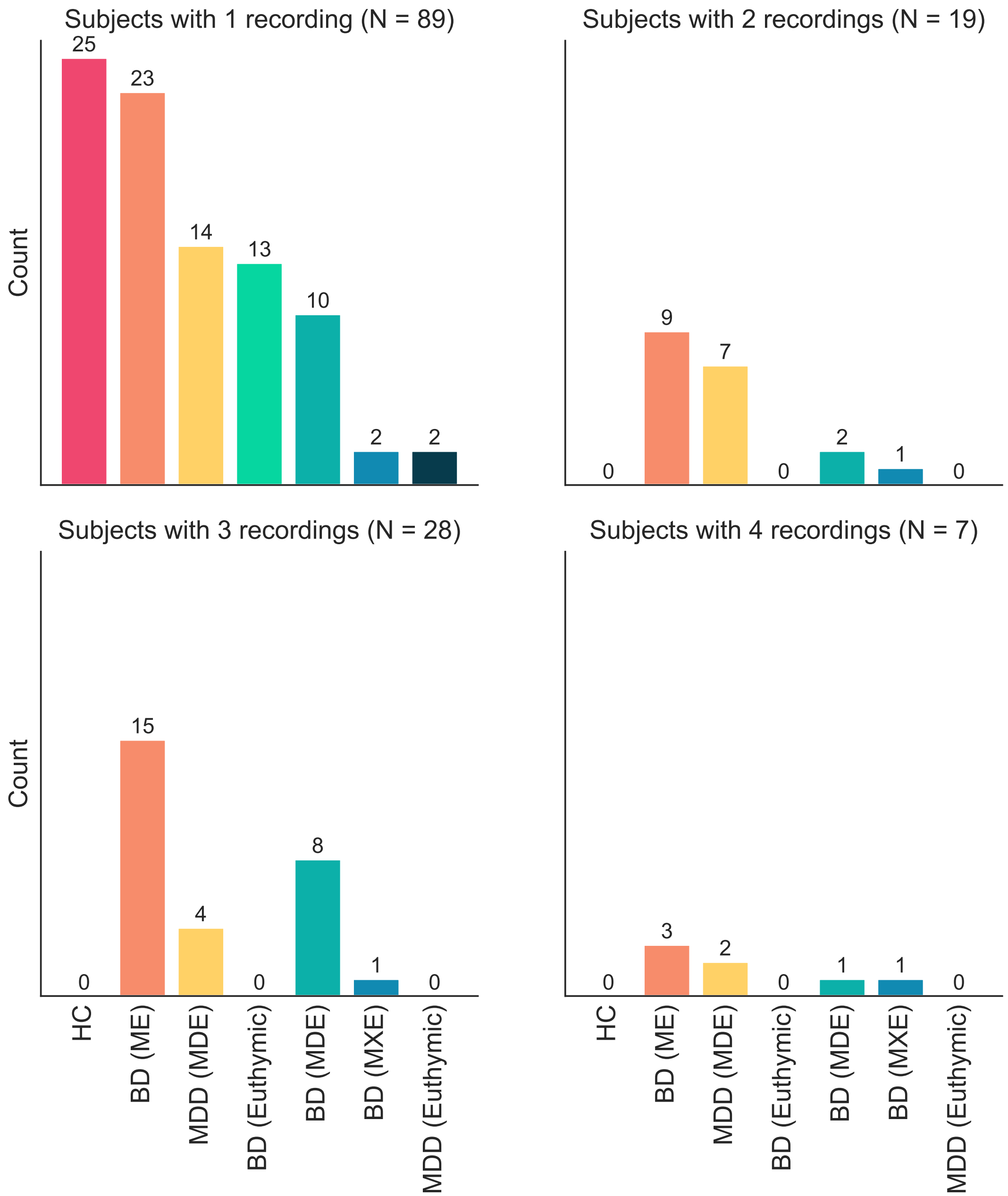

Figure A.1: Number of recordings session available at different time points by diagnosis. Some of the collected recordings and the corresponding clinical info have not yet been uploaded to our data storage and have therefore not been used for analyses. Mood disorders manifest in two polarities, mania, 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. 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. 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.