medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Digital Health Tools for the Passive Monitoring of Depression: A Systematic Review of Methods

de Angel, V.,[1,2] Lewis, S. Y.,[1,3] White, K.,[1] Oetzmann, C.,[1] Leightley, D.,[1] Oprea, E.,[1] Lavelle, G.,[1] Matcham, F.,[1] Pace, A.,[4] Mohr, D. C.,[5,6] Dobson, R.,[2,7] Hotopf, M.[1,2]

1. Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
2. NIHR Maudsley Biomedical Research Centre, South London and Maudsley NHS Foundation Trust, London, UK
3. Department of Psychology, University of Bath, Bath, UK
4. Chelsea And Westminster Hospital NHS Foundation Trust
5. Center for Behavioral Intervention Technologies, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
6. Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
7. Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, 16 De Crespigny Park, London, SE5 8AF, UK

Summary

Background:
The use of digital tools to measure physiological and behavioural variables of potential relevance to mental health is a growing field sitting at the intersection between computer science, engineering and clinical science. We aim to summarise the literature on remote measuring technologies, mapping methodological challenges and threats to reproducibility, and to identify leading digital signals for depression.

Methods:
Medical and computer science databases were searched between January 2007 – November 2019. Published studies linking depression and objective behavioural data obtained from smartphone and wearable device sensors in adults with unipolar depression and healthy subjects were included (PROSPERO registration: 2019 CRD42019159929). A descriptive approach was taken to synthesise study methodologies.

Results
We included 52 studies and found threats to reproducibility and transparency arising from failure to provide comprehensive descriptions of recruitment strategies, sample information, feature construction and the determination and handling of missing data. The literature is characterised by small sample sizes, short follow-up duration and great variability in quality of reporting, limiting the interpretability of pooled results. Bivariate analyses show some consistency in statistically significant associations between depression and digital features from sleep, physical activity, location, and phone use data. Regression and classification machine learning models found predictive value of aggregated features.

Interpretation:
Recommendations are put forward to improve aspects of generalisability and reproducibility, such as wider diversity of samples, thorough reporting methodology and the potential for reporting bias in studies with numerous features.

Funding:
National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London.

Keywords: Digital Mental Health; smartphone; wearable device; depression; mobile health; mHealth; remote measurement; remote monitoring; machine learning; digital phenotype; systematic review

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction:

Depression remains the leading cause of disability worldwide (1), with a largely chronic course and poor prognosis (2). Early recognition and access to treatment, as well as better trial methodology has been linked to improved treatment outcomes and prognosis (3).

The use of digital technology to track mood and behaviour brings enormous potential for the clinical management and the improvement of research in depression. By passively sensing motion, heart rate and other physiological variables, smartphone and wearable sensors provide continuous data on behaviours that are central to psychiatric assessment, such as sociability (4), sleep/wake cycles (5), cognition, activity, (6) and movement (7).

With the global trend toward increased smartphone ownership (44.9% worldwide, 83.3% in the UK) and wearable device usage forecast to reach one billion by 2022 (8), this new science of "remote sensing", sometimes referred to as digital phenotyping or personal sensing (9) presents a realistic avenue for the management and treatment of depression. When combined with the completion of questionnaires, remote sensing may generate more objective, and frequent, measures of mood and other core dimensions of mental disorders, instead of relying on retrospective accounts of patients or participants.

The first step in generating meaningful clinical information from data derived from digital sensors is to generate features, which are the smallest constructed building blocks, designed to explain the behaviours of interest (see Mohr et al. (10) for a detailed analytical framework). These low-level features are often aggregated to define high-level behavioural markers, which can be understood as symptoms. For example, GPS data (sensor), can be translated into 'location type' (low-level feature), 'increased time at home location' (high-level behaviour) derived from location data may indicate social withdrawal or lack of energy (symptom), and may therefore be associated with depression severity.

One of the main challenges arises from this emerging field is that it sits at the intersection between computer science, engineering, and clinical science. The advantages of a multidisciplinary approach are evident, but these domains are yet to be brought together efficiently (11,12), giving rise to large differences in reporting standards with the risk that reproducibility may be threatened (13).

Previous reviews in affective disorders cite the level of heterogeneity across studies as a barrier to carry out meta-analytic syntheses of the results. Additionally, these reviews have included non-validated measures of depression, and a mix of bipolar and unipolar samples, characteristics which not only show divergent results (11,12,14), but add study diversity. There is therefore a need for a comprehensive review of methodologies, with more specific inclusion criteria, to highlight the sources of heterogeneity and methodological shortcomings in the field.

Given the difficulty in extracting a clear message from the available literature, the current work aims to review studies linking passive data from smartphone and wearable devices with depression and summarise key methodological aspects, to: (a) identify sources of heterogeneity and threats to reproducibility, and (b) identify leading digital signals for depression . We will also assess the quality of the included studies and evaluate their reporting of feasibility of passive data collection methods, participant retention and missing data.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

METHODS:

Search Strategy and selection criteria:

We searched Pubmed, IEEE Xplore, ACM Digital library, Web of Science, and Embase and PsychInfo via OVID, for studies published between January 2007 until November 2019, and used a combination of terms related to the key concepts of (1) depression and (2) digital sensors and Remote Measurement Technologies (RMTs) (full search in appendix 1). We also conducted searches based on bibliographies of reviews and meta-analyses on the topic. The protocol was uploaded on PROSPERO 2019 CRD42019159929.

Inclusion and exclusion criteria:

Studies had to have measured depressive symptoms in either clinical or epidemiological samples, and to consist of samples with mean ages between 18-65 years, due to the differences in behavioural patterns for older adults and children. We limited studies to those which had extracted data for at least 3 consecutive days (to allow for intraday mood fluctuations) from smartphones and wrist-worn devices commonly used by the public. This excludes medical-grade wearables such as electrocardiogram. Studies had to link data between validated scales of depression severity or status (case/non-case) and digital sensor-based variables including measures of behaviour, e.g. activity, sleep, etc., gathered passively. Studies had to be written in English, German or Spanish because these are the languages spoken by the reviewers, be published, peer-reviewed and with accessible full text.

Studies were excluded if their primary focus related to a condition other than depression. Studies focusing specifically on bipolar depression were excluded, however mixed studies consisting of unipolar and bipolar were included provided unipolar cases comprised a substantial majority (at least 80%) of the sample. We excluded studies published before 2007 as this was when the first smartphones became available.

Procedure

Studies were checked for eligibility by two researchers screening titles and abstracts. Potentially eligible studies' full texts were reviewed by one researcher, with a second researcher evaluating a sample of 10% of all texts. Any discrepancies were resolved by a third reviewer. Agreement of >90% was reached for all reviewer pairs. The eligibility process was documented according to PRISMA guidelines (15).

Data Extraction

Data extraction included the following variables: sample characteristics (N, mean age, gender, ethnicity), comorbidities, study design, study setting (clinical, community, student), depression outcome measures, length of follow up , device type, features measured, sensors used, statistical analyses and significance levels.

Study quality assessment

No single quality assessment tool was suitable because of the diversity of study types. We therefore combined the Appraisal Tool for Cross-Sectional Studies (AXIS tool (16) and the Newcastle-Ottawa Scale (NOS) for longitudinal studies (17). Items were scored with two points for fully fulfilled items, one point for partially fulfilled item, and zero for a non-fulfilled item (see appendix 2 for a description of each criterion). We added an item regarding having a published protocol prior to publishing results (1 point for a published protocol).

Feasibility

We collected information on five measures of feasibility of using digital health tools, with the aim of identifying potential obstacles to their implementation: engagement with study devices, reasons for study

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

drop out, reported problems with technology, percentage of study tasks completed, attrition and missing data.

Data Synthesis

Eight categories of behavioural features were identified: sleep, physical activity, circadian rhythm (rest-activity patterns through a 24-hour period), sociability, location, physiological parameters, phone use and environmental features. Appendix 3 provides a description for each feature. Within each behavioural category there are lower-level features, which group together several individual features as reported by each study. It was therefore possible for a single study to present multiple associations for the same feature. Significant associations according to 0.05 p-value thresholds are presented. Due to the heterogeneity of feature types, study designs and data reporting we did not conduct a meta-analysis.

Role of the Funding Source

The funders of the study had no role in study design, data extraction, data analysis, data interpretation, or writing of the review.

The data used in this review is available by request to the corresponding author.

RESULTS:

Fifty-two studies were included in the review (see Figure 1). The majority of articles (n=46) were published in medical journals, and 33 (61%) were from North America. A summary of included studies are presented in Table 2.

Studies were evenly divided between community samples (n=19), student samples (n=18) and clinical populations (n=15). The median sample size was 54, the median age of participants was 34 years, and the median percentage of females was 58%. However, there was a striking lack of information on some key data – with 11% and 7% of studies failing to give data on age or gender, respectively, and 63% failing to include information on ethnicity. Computer science journals were less likely to report age and gender but more likely to report ethnicity (33% studies failing to report each demographic). Fifteen different measures of depression were used, the most commonly used scales being the Center for Epidemiological Studies Depression Scale (CES-D [18]; n = 12 studies), Hamilton Rating Scale for Depression (HAM-D [19]; n = 13 studies), and Patient Health Questionnaire-9 (PHQ-9 [20]; n = 9).

Most studies had a cohort design, meaning that depression was measured at least at two different time points (see Table 1). However, these time points tended to be shorter than 2 weeks (Figure 2). Two studies provided no information on length of follow-up, instead only mentioning that data was obtained from participants providing at least 72hrs of consecutive data (5,21).

Table 1. The breakdown of study designs within each sample type.

Study design	*Total*	*Student*	*Community*	*Clinical*
Cross-sectional	19	4	10	5
Case control	7	0	1	6
Cohort	25	14	6	3
RCT	3	0	2	1
Total	**52**	**18**	**19**	**15**

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

To understand the relationships between depression and objective features, studies either looked at group differences (including classification analyses) or correlation and regression. Most studies presented direct bivariate relationships (n= 46), allowing for a closer evaluation of which features are promising markers of depressive symptomatology. Ten studies presented the result of a combination of features and their association with depressive state (n= 7), or depression severity (n= 8), using machine learning methods (Tables 3 and 4). Bivariate Pearson correlation coefficients were the most used analytical method (n= 32).

Table 2. Summary characteristics of included studies.

First Author	Year	Country	Field	N (RMT)*	% female	mean age (range/ SD)	RMT * follow up (days)	Sample Type	Depression Measure	Passive Feature Type: Sleep	Physical Activity	Circadian Rhythm	Sociability	Location	Phone use	Physiological	Environmental	Total feature types
Avila-Moraes (22)	2013	Brazil	M	30	100.0	44 (18 - 60)	7	Clinical	BDI, HAMD, MADRS		X	X				X		**3**
Ben-Zeev (23)	2015	USA	M	37	21.0	22.5 (19-30)	70	Student	PHQ-9	X	X							**3**
Boukhechba (24)	2018	USA	M	72	51.4	19.8 (2.4)	14	Student	DASS-21		X		X	X				**3**
Burns (25)	2011	USA	M	7	87.5	37.4 (19 - 51)	56	Community	PHQ-9	X	X		X	X	X			**5**
Byrne (26)	2019	Australia	M	42	0.0	(18-29)	7	Community	SCRAM - dep	X	X	X						**3**
Caldwell (27)	2019	USA	M	115	100.0	27.5 (6.1)	3	Community	BDI-II	X								**1**
Cho (4)	2016	South Korea	M	532	56.0	57	720	Community	BDI-II				X					**1**
David (28)	2018	USA	M	132	60.0	20.68 (18-21)	7	Student	PHQ-4				X		X			**2**
Difrancesco (29)	2019	Netherlands	M	359	62.4	50.1 (11.1)	7	Community	BDI-II	X	X	X						**3**
Dillon (30)	2018	Ireland	M	396	50.8	nr	7	Clinical	CES-D		X							**1**
Doane (31)	2015	USA	M	76	76.0	18.1 (0.4)	3	Student	CES-D	X								**1**
Doryab (32)	2014	USA	M	6	33.3	nr	120	Student	CES-D				X	X			X	**3**
Ghandeharioun (6)	2017	USA	CS	12	75.0	37 (20 - 73)	56	Clinical	HAM-D	X	X		X	X	X	X		**6**
Haeffel (33)	2017	USA	M	47	55.3	20.9	7	Student	BDI-II	X								**1**
Hori (34)	2016	Japan	M	40	52.5	39.8	7	Clinical	HAM-D	X								**1**
Jacobson (35)	2019	Brazil	M	15	87.0	47.6 (10.5)	7	Clinical	BDI, HAMD	X	X							**2**
Kawada (36)	2007	Japan	M	105	29.5	24.1 (1.8)	4	Student	CES-D	X	X	X						**3**

First Author	Year	Country	Field	N (RMT)*	% female	mean age (range/SD)	RMT * follow up (days)	Sample Type	Depression Measure	Feature Type: Sleep	Physical Activity	Circadian rhythm	Sociability	Location	Phone use	Physiological	Environmental	Total feature types
Knight (37)	2018	Australia	M	23	77.0	20.7 (3.2)	3	Community	DASS-21		X							**1**
Li (38)	2018	Australia	M	375	53.9	59.5 (5.5)	7	Community	CES-D		X							**1**
Lu (5)	2018	USA	CS	103	76.7	(18 -25)	nr	Student	QIDS	X	X			X				**3**
Luik (39)	2013	Netherlands	M	1734	53.4	62.3 (9.4)	7	Community	CES-D	X								**1**
Luik (40)	2015	Netherlands	M	1714	53.6	62.2 (9.4)	7	Community	CES-D	X								**1**
McCall (41)	2015	USA	M	58	67.0	42.1 (12.4)	56	Clinical	HAM-D	X								**1**
Mendoza-Vasconez (42)	2019	USA	M	266	nr	40.6 (9.9)	7	Community	HAM-D		X							**1**
Moukaddam (43)	2019	USA	M	22	76.0	50.3 (10.1)	56	Clinical	PHQ-9		X		X					**2**
Naismith (44)	2011	Australia	M	44	43	62.3	14	Clinical	HAM-D	X								**1**
Park (45)	2007	USA	M	54	57.4	43 (21 - 76)	14	Community	CES-D	X	X							**2**
Pillai (46)	2014	USA	M	39	73.8	55 (3.2)	7	Student	BDI-II	X								**1**
Pratap (47)	2019	USA	M	271	77.8	33.4 (10.7)	90	Community	PHQ-2				X	X				**2**
Robillard (48)	2013	Australia	M	66	62.7	21.5	7	Clinical	clinician assessment	X								**1**
Robillard (49)	2014	Australia	M	238	64.3	40.4	10	Clinical	HAM-D	X		X						**2**
Robillard (50)	2015	Australia	M	342	55.1	22.3	14	Clinical	clinician assessment	X		X						**2**
Robillard (51)	2016	Australia	M	25	48.0	20.9 (4.6)	14	Clinical	clinician assessment	X		X						**2**
Robillard (52)	2018	USA	M	12	58.0	20.1 (18-31)	13	Clinical	clinician assessment	X								**1**

First Author	Year	Country	Field	N (RMT)*	% female	mean age (range /SD)	RMT * follow up (days)	Sample Type	Depression Measure	Feature Type: Sleep	Physical Activity	Circadian rhythm	Sociability	Location	Phone use	Physiological	Environmental	Total feature types
Saeb (7)	2015	USA	M	21	71.4	28.9 (19 - 58)	14	Student	PHQ-9			X		X	X			**3**
Saeb (53)	2016	USA	M	38	20.8	nr	70	Community	PHQ-9		X			X				**2**
Sano (54)	2018	USA	M	47	72.0	(18 - 25)	30	Student	MCSF-12	X	X	X	X	X	X	X		**7**
Slyepchenko (55)	2019	Canada	M	70	57.9	(18 - 65)	15	Clinical	MINI	X	X	X						**3**
Smagula (a) (56)	2018 a	USA	M	145	67.0	60 (36-82)	9	Community	HAM-D			X						**1**
Smagula (b) (57)	2018	USA	M	45	38.8	38.08	10	Community	HAM-D			X						**1**
Stremler (58)	2017	Canada	M	101	62.7	34.1	5	Community	CES-D	X								**1**
Tao (59)	2019	China	M	220	52.3	20.3 (2.4)	7	Student	PROMIS - dep		X							**1**
Todder (60)	2009	Australia	M	27	48.1	49.5 (12.8)	7	Clinical	HAM-D		X	X						**2**
Vallance (61)	2013	Canada	M	385	0.0	65.3 (7.5)	3	Community	CES-D		X							**1**
Vanderlind (62)	2014	USA	M	35	42.3	19.8 (18-23)	21	Student	CES-D	X		X						**2**
Wahle (63)	2016	Switzerland	M	36	64.3	(20-57)	14	Community	PHQ-9		X		X	X	X			**4**
Wang(64)	2014	USA	CS	48	20.8	nr	7	Student	PHQ-9	X			X	X				**3**
Wang (65)	2018	USA	CS	83	51.8	20.1 (2.3)	126	Student	PHQ-8	X	X		X	X	X	X		**6**
White (66)	2017	USA	M	418	60.3	57 (35-85)	7	Community	CES-D	X		X						**2**
Yang (67)	2017	China	CS	48	nr	nr	70	Student	PHQ-9				X					**2**
Yaugher (68)	2015	USA	M	100	58.3	18.6 (18 - 27)	7	Student	PAI - dep	X								**1**
Yue (21)	2018	USA	CS	54	nr	(18-25)	nr	Student	PHQ-9		X			X				**2**
N = 52				**Median:** 54	**Median:** 57.7	**Median:** 34.2	**Median:** 9.5		**Total N:** 18	**Total:** 31	25	15	13	13	7	5	1	

* Number of participants/length of follow up included in passive data collection samples; these may be lower than overall study sample sizes.

RMT = Remote Measurement Technologies, SD= standard deviation, M = Medical Field, CS = Computer Science Field

BDI= Beck's Depression Inventory, HAM-D=Hamilton Depression Rating Scale, MADRS=Montgomery–Åsberg Depression Rating Scale, PHQ = Patient Health Questionnaire, PAI – dep = Personality Assessment Inventory – depression subscale, CES-D = Center for Epidemiologic Studies Depression Scale, MINI = Mini International Neuropsychiatric Interview, PROMIS = Patient-Reported Outcomes Measurement Information System,, MCSF-12 = Mental Component of the Short Form Health Survey, QIDS = Quick Inventory of Depressive Symptomatology, DASS = Depression Anxiety Stress Scales, SCRAM = Sleep, Circadian Rhythms, and Mood Questionnaire.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Quality Assessment and Feasibility

Figure 4 shows a breakdown of quality scores for each item (see appendix 4 for a distribution of scores on all eight quality assessment items). Justification of sample size was rarely given, and sample representativeness was poor, possibly reflecting that many reports were pilot or feasibility studies. Recruitment strategies and non-participation rates were not reported in the majority of cases. Missing data and strategies for handling missing data were infrequently described. Only four studies referred to a previously published protocol (25,29,30,54).

Only five studies reported engagement rates at follow up, and they all measured engagement at different time points, making comparisons difficult. Additionally, sensor data was sometimes obtained for a subsample, whereas acceptability measures were reported for the wider sample. Eighteen studies (34%) reported, or provided enough information to calculate, how many participants completed the study – results ranging from 22% adherence to the study (63) at 4 weeks, to 100% (69), with a median of 86.6% completers.

Reasons for dropouts were provided in four studies and were due to equipment malfunction and technical problems using devices (25,26,43,64) . Six additional studies reported issues including; lack of data for consecutive days, software error, participants forgetting to charge phones or devices, server and network connectivity problems, sensors breaking, missing clinical data which impeded comparisons with sensor data, and mobile software updates, which can interfere with data integrity (7,40,54,70–72).

Associations between Objective Features and Depression:

The association between groups of features and depression is given in Figure 3, broken down by feature type. We give the number of studies which have reported the feature and the number of feature-depression associations which reached statistical significance as a proportion of the total such associations reported.

Sleep

Twenty-nine studies collected data on sleep, typically ascertained using accelerometer, light and heart rate sensors. Nine different features of sleep are reported in Figure 3 A. Sleep quality, encompassing features relating to sleep fragmentation (number of awakenings and wake after sleep onset [WASO]), was the most commonly reported feature. Sleep efficiency is presented as a separate feature given its prevalence in studies. For all significant results, lower sleep efficiency or quality was associated with higher depression scores. Features with higher proportions of significant findings are features of sleep stability, sleep offset, time in bed; longer time in bed and later sleep offset were associated with higher depression scores.

Across studies finding significant results, sleep variability was higher for those with depression compared to controls (27), and those with more severe symptoms (28). The average length of follow up for studies showing significant associations between sleep stability and depression was 24.7 days (range = 4 – 63), whereas that for studies showing no significant associations was 8.6 (range = 3 – 21).

Total sleep time showed mixed directionality of significance, with some studies finding negative correlations between total sleep time and higher depression (36,64), others finding the depressed group having longer sleep time than controls (48).

Physical Activity

Measures of physical activity were collected in 20 studies using a mixture of smartphone (n=8) and wearable devices (n=12).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Activity levels were predominantly measured as gross motor activity within a day, and showed that depression was negatively correlated with physical activity (24,29). Out of the seven studies extracting 'activity levels' as a feature within physical activity, both studies using smartphones found a significant difference in depression severity, compared to one out of the five that used wrist actigraphy. Higher depressive symptoms were associated with less time spent engaging in physical activity (5), movement speed (21) and step count (43). Evidence for intensity and sedentary time was mixed, with Lu (5) and Difrancesco (29) finding lower depression in those with higher intensity activity and lower sedentary time, but Tao (59) and Todder (60) finding no such associations.

Circadian Rhythm

A total of 14 studies assessed movement patterns within a 24-hour period. All used accelerometery data, except for Saeb (7) who used GPS data for circadian movement.

All significant associations indicated that disturbed rest-activity patterns were associated with depressive symptoms, however, in the majority of instances where circadian rhythm was reported, no significant association with mood was detected. Depression has been associated with lower daytime activity and higher night-time activity (hour-based activity levels (22,55,60)), low intra-daily stability, more fragmented intra-daily movement, e.g., leaving for work and coming back at less regular times (7), later acrophase, or later activity peaks (50,56,66); lower amplitude, less difference between the average levels of activity during the peaks vs the troughs of activity (29,56). Four studies calculated circadian rhythmicity as a measure of the extent to which a participant's pattern follows an expected Cosinor model, finding lower circadian rhythmicity more likely to be associated with being depressed (49,53,55,74).

Sociability

Eleven studies assessed sociability. The average number of ingoing and outgoing calls was found to be negatively correlated with depressive symptoms in one small study (n=6), and only in men (32). Yang et al. (67), with a combination of microphone, GPS and Bluetooth sensing as a proxy for social proximity, found that an interaction between environmental noise and proximity to others was informative of depressive state, e. g. being in a quiet place with few people around, compared to either spending time outside alone, or in a noisy environment with more than 3 people. Other studies found that higher frequency of conversations in the day and at night correlated with lower depression (64), as well as being around human speech for longer (23).

Location

Location was assessed in 11 studies, measured via GPS. In addition to traditional statistical analyses, Saeb et al. (7) estimated accuracy and mean normalized Residual Mean Square Difference (NRMSD) to assess the performance of prediction models. We therefore do not have levels of significance as expressed via p-values for all features. Entropy was reported in 24 cases in four different studies. High entropy, or spending more time in fewer, more consistent locations, was associated with depression, as compared to lower entropy, where people spend more time in a greater number of more varied locations. Features of location variance – how varied a participant's locations are – show a negative correlation with depression, where the more varied the locations, the lower the likelihood of being depressed. Homestay – the amount of time spent at home – shows one of the most consistent patterns across the field, with all 11 included studies reporting a significant association with depression.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Phone Use

Three studies associated individual phone use features with depression. All studies found that increased unlock duration and unlock frequency were associated with depression, non p-value tests reported a mean NRMSD of 0.268 and 0.249, and 74.2% and 68.6 % accuracy in classifying depressed vs non depressed participants, respectively. Increased use of specific apps, such as Instagram, iOS maps, and the use of photo and video apps was associated with greater depression, whereas book apps were associated with milder symptoms (28).

Physiological Features

Temperature was measured by Ávila-Moraes et al. (22), who extracted more than 5 skin temperature features from a wrist-worn device, and found depressed people to have a longer time of elevated temperature compared to controls. One study (73), reported no association between heart rate and depression scores.

Environmental features

Ávila-Moraes et al. (22) also used a wrist-worn actigraphy device to measure light exposure and extracted four features. She found depressed groups to have lower variance of light intensity than controls. Another study found humidity to have a significant positive correlation with depressed symptoms (r=0.4) in women, but a negative correlation in men, suggesting females, but not males might feel worsening in their condition during rainy weeks (32).

Combined Features

Table 3 and Table 4 show the ten studies combining digital features to predict symptom severity (regression models) or depressive state (classification models). Twenty-four models in total were presented by all studies, the majority of which (n=18) included features of physical activity, followed by location (n=14), phone use (n=11) and sleep (n=9). Both classification and regression models showed predictive value, however, many of them lacked information regarding handling of missing sensor data and calibration. Those that do, report simple imputation methods such as mean imputation, with two studies using multiple imputation methods (6,21).

Table 3. Details for studies analysing combined features using classification models.

Study ID	First Author, Year	Device	Groups	N	N features	Feature type	Algorithm/ Model	Performance measure	Discrimination value	Missing Data Handling	Validation method	Comparison Models
1	Sano, 2018	Q sensor, smartphone	MCS SF-12 Low v High	47	204	PA, SC, Li	SVM RBF	Accuracy	85.1	Interpolation	10-fold cross validation	LASSO, SVM Linear
					441	PA, L, PU, SC, ST	SVM RBF	Accuracy	86.1			
					700	S, PA, PU, SC, ST, HR, Cl	SVM RBF	Accuracy	77.2			
					296	S, PA, PU	SVM RBF	Accuracy	78.7			
					25	PU	SVM RBF	Accuracy	71.1			
					25	S	SVM RBF	Accuracy	65			
2	Yue, 2018	Android	Clinician MDD vs HC	25	8	PA, L	SVM RBF	F1	0.66	Multiple Imputation	LOOCV	l2-regularized (ridge) regression
		iPhone		54	8	PA, L	SVM RBF	F1	0.76			
3	Wahle, 2016	Smartphone	PHQ-9 Dep vs HC	36	120	PA, So, L, PU	Random Forest	Accuracy	60.1	Unclear	LOOCV	SVM
4	Pratap, 2019	Smartphone	PHQ-2 Dep vs HC	93	10	So, L	Random Forest	Median AUC	> 0.50 (for 80.6% sample)	Mean imputation	none	
5	Saeb, 2015	Android	PHQ-9 Dep vs HC	18	8	CR, L	Elastic Net Logistic Regression	Accuracy	78.8	Unclear	LOOCV	
6	Wang, 2018	Smartphone	PHQ 4 Dep vs HC	83	9	S, PA, L, PU, HR	Lasso Logistic Regression	AUC	0.809	Unclear	10-fold cross validation	
7	Lu, 2018	smartphone and Fitbit	QIDS	69	36	S, PA, So	Multi-Task Deep Learning	F1	0.77	Exclusion	LO(W)OCV	STL (Lasso) STL (Ridge), MTL Lasso and Ridge

MCS SF = Mental Component Survey Short Form, PHQ = Patient Health Questionnaire

MDD = Major Depressive Disorder, HC = Healthy Control

S = Sleep, PA = Physical activity, CR = Circadian Rhythm, So = Sociability, L = Location, PU = Phone Use, SC = Skin Conductance, ST = Skin Temperature, HR = Heart Rate, Li = Light, Cl = clinical data

SVM RBF = Support Vector Machine - Radial Basis Function, AUC = Area Under the Curve, LOOCV = Leave One Out Cross Validation, STL = Single Task Learning, MTL = Multi-Task Learning

Table 4. Details for studies analysing combined features using regression models.

Study ID	Author, Year	Device	Outcome	N	N features	Feature type	Algorithm	Performance measure	Exact statistic	Missing Data Handling	Validation method	Comparison
2	Yue, 2018	Android	PHQ9	25	8	PA, L	SVM RBF	r	0.46	Multiple Imputation	LOOCV	Support Vector Multivariate Linear Regression
		iPhone	PHQ9	54	8	PA, L	SVM RBF	r	0.41			Support Vector Multivariate Linear Regression
4	Pratap, 2019	Smartphone	PHQ2	93	10	PA, So, L	Random Forests	R2	≈ 0	Mean Imputation	None Reported	
5	Saeb, 2015	Smartphone	PHQ9	18	8	CR, L	Elastic net Linear Regression	Mean NRMSD	0.251	Unclear	LOOCV	
				21	2	PU	Elastic net linear regression	Mean NRMSD	0.273			
6	Wang, 2018	Smartphone	pre PHQ 8	83	10	S, PA, L, PU, HR	Lasso Linear Regression	MAE	2.4	Unclear	10-fold cross validation	
			post PHQ 8	83	5	S, PA, So, L, PU	Lasso Linear Regression	MAE	3.6			
7	Lu, 2018	Smartphone, Fitbit	QIDS	69	36	S, PA, So	Multi-Task deep Learning	R2	0.44	Exclusion	LO(W)OCV	STL (Lasso) STL (Ridge), MTL Lasso and Ridge
8	Burns, 2011	Smartphone	PHQ9	7	38	PA, So, L, PU, Li	Regression Trees	Accuracy	nr	Unclear	10-fold cross validation	
9	Jacobson, 2019	Actiwatch	BDI-II	15	nr	PA, Li	Xgboost	r	0.86	Unclear	LOOCV	
10	Ghandeharioun, 2017	Empatica, Smartphone	HRDS	12	700	S, PA, PU	combination of regularised regression, robust-to-outlier, boosting, Random Forest and Gaussian Process	RMSE	4.5	Multiple Imputation	10-fold cross validation	

PHQ = Patient Health Questionnaire, QIDS = Quick Inventory of Depressive Symptomatology, nr = not reported

S = Sleep, PA = Physical activity, CR = Circadian Rhythm, So = Sociability, L = Location, PU = Phone Use, SC = Skin Conductance, ST = Skin Temperature, HR = Heart Rate, Li = Light, Cl = clinical data

SVM RBF = Support Vector Machine-Radial Basis Function, NRMSD = Normalized root-mean-square deviation, RMSE = Root-mean-square error, MAE = Mean Absolute Error, STL = Single Task Learning, MTL = Multi-Task Learning

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

DISCUSSION

We sought to summarise the literature on passive sensing for depression, in order to map the methodological challenges and threats to reproducibility, in an effort to generate standards in the literature that allow for a quantitative synthesis of results. We also assessed the available evidence for a relationship between sensor data and mood to identify leading digital signals for depression.

The first methodological shortcoming stems from the recency of this field. Studies have mostly employed opportunistic study designs, with small sample sizes, short follow-up windows and many being conducted on students, which limits generalisability. Different features may reach peak predictability of mood with different sampling timeframes, so shorter follow-ups may harm the prediction abilities for some behaviours (54). This is presumably more likely in feature types such as sleep and circadian rhythm which benefit from having more aggregated baseline data (75). There is no consensus on the timeframe window for optimal phenotyping, different windows therefore need to be evaluated.

A critical source of heterogeneity comes from the multitude of methods to create any individual feature, often without providing reasonable details of the process. A feature of sleep quality, for instance, defined in different studies as “Nocturnal Awakenings”, may have been constructed by measuring counts of awakenings, total number of minutes awake, or a proportion of awake vs asleep in a sleep session. Additionally, there may be differences in how raw sensor data is used to classify an event as sleep or awake. This heterogeneity challenges the ability of investigators to reproduce findings and hampered our ability to summarise results in meta-analysis.

The exploratory nature of many of these studies means that many different versions of the same feature may have been generated but studies do not transparently describe and justify feature selection and its association with depression. Researchers should provide a description of the feature, in the paper or supplement materials, that is sufficiently clear to allow for appropriate reproducibility.

Additionally, due to the large number of variables obtained in sensing studies, it is likely that published papers are selective in their reporting, and typically emphasise “positive” findings over “negative” ones. Preregistering studies and analyses would be one way of handling this. As the field matures and more studies are published, issues of rigour and reproducibility become more salient and preregistration becomes more important to reduce reporting bias and cherry-picking in the field.

The sources of heterogeneity arise from varying data collection timespans, depression assessment measures, feature construction, and analytical methods. Whilst differences in these areas represent a healthy heterogeneity in an evolving field, it means that nuance is required in interpreting the presence or absence of a relationship between any specific signal and depressed mood. For example, many studies recruited students, who have different socialisation patterns and smartphone usage to older adults (12,76). Prediction models based on younger populations have been found not to transfer to older age groups (77). Further, a signal detected in a clinical sample consisting of people with relatively severe depression may not be reproduced in a population sample where the majority of the sample have few or no depressive symptoms and there may be less variability in key sensor data (e.g. sleep or activity data).

For any broad concept (e.g. sleep or circadian rhythm) different sensor types or operating systems were used, and component features were derived using different approaches. For example, both iPhone and Android smartphone operating systems were included, and sometimes showed differences in significance

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

levels for the same variables (5,21). This could be due to differences in sampling and data collections for both operating systems, or differences in the user profiles of these products (78).

We found significant shortcomings of the literature in terms of fundamentals of reporting, including the most basic descriptors of sample characteristics, recruitment, attrition, and missing data. Whilst many of these shortcomings would be resolved by authors and journals following established reporting conventions (e.g. STROBE guidelines), there are a number of issues which are specific to this field.

One of those issues is missing data. Our quality assessments reflect poor reporting of missing data at both the sample level (e.g. attrition and study non-completion) and individual level (e.g. missing sensor data from participants). Missing data can arise from issues with technology, such as device and system failures, or from user-related issues which may be associated with depressed mood. For missing data to be used informatively, these two types need to be identified and dealt with in different ways in terms of their exclusion or analysis. Additionally, researchers set different thresholds as to what counts as missing data. This varies between studies and generates an important threat to reproducibility, making it crucial that these thresholds are reported. Our recommendation is that papers should clearly state how much data were missing and how it was managed in the analysis.

Remote sensing is a relatively new technology which potentially places considerable burden on study participants - it was therefore surprising that few studies reported on acceptability of the study protocol to participants. Where this did happen the emphasis was more on evaluating active questionnaire data rather than passive data and device use, where arguably greater issues over privacy and acceptability arise (6,49).

Association between mood and digital features:

Notwithstanding the cautions raised above, we were able to detect some associations between mood and digital features where there is growing consensus between studies. Location features appears to be most consistently associated with depression, with homestay and entropy both associated with mood in 4 and 5 studies, respectively. However, these studies do not determine the direction of causality, i.e. whether changes in sensed features such as homestay are merely a reflection of behaviours which appear in depression, such as reduced physical activity and social withdrawal (79,80) or whether they are, in themselves, predictors of deterioration in mood.

Several sleep features appear also to be consistently associated with depressed mood, with sleep stability showing the highest proportion of significant associations. When measuring socialisation, proximity-related features using Bluetooth and microphone sensors seem more sensitive to mood than call and message frequency counts. However, many of these studies have small sample sizes (54), student samples with a low mean age (24) or report high degree of intra- and interindividual variance in daily phone usage (82). Recent studies with larger and more diverse samples using classification machine learning techniques have found that a low average number and duration of calls made daily predicted depression state (83).

Even though disruptions in circadian rhythms have been thought to affect depression (84), the majority of studied features did not have a significant association with mood. As previously mentioned, this may be due to short follow-up since median follow-up times for circadian features = 9.5 days.

Feature Combinations

The findings of this review highlight the array of potential predictors that sensor data generates. As such, machine learning methods have been the choice analytic approach to the digital phenotyping of depression from multiple features. In addition to helping account for important interactions between the objective

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

features, for example how the effect of being alone is mediated by location (being indoors vs outdoors; (67), analysing multimodal data in this way may help cover missing data from one source to another. However, machine learning methods have been criticised for lacking transparency in how the model is built and how individual variables contribute to the overall prediction (85). Some studies in the current review do report their top predictors and bivariate associations with depression, but the question of how well these models can be replicated remains, highlighting the importance of thorough reporting.

Strengths and Limitations:

Our attempt to summarise the literature is necessarily crude because the reporting of feature-depression associations was too opaque and diverse to allow any credible attempt at meta-analysis. We have therefore had to rely on simple counts of associations reported, and this comes with caveats that reports are not weighted by sample size, follow up duration or study quality. It is possible that the associations we have reported are due to reporting bias, as mentioned in the previous section, where investigators emphasise "significant" findings over "non-significant" ones.

To present low-level features in a clear and meaningful way in this review, we combined them into broader low-level features and therefore some of the nuances between them were lost. For example, if one study extracted two features such as total number of minutes spent in phone calls and the average length of a phone call, they would both load into Call Duration, within the "Sociability" Feature Type (appendix 3).

Several studies included in this review have overlapping samples as they come from existing datasets. For example, four papers (53,64,67,73) use the StudentLife open dataset, where there is some similarity in the analysis, meaning that some of the feature associations may be duplicated.

Recommendations and conclusions:

Whilst there have been attempts at reporting standards for actively collected questionnaire data on mood (86), and guidelines exist for the reporting of observational data (STROBE), there is a need to develop consensus over the manner in which such mobile health studies are conducted and reported. This should not come at the expense of stifling innovation and should acknowledge that a new field of study takes time to develop. However, approaches to reporting feature construction, management of missingness, pre-specification of analytic plans, transparency of hypothesis testing conducted, and reporting of participant acceptability of procedures could readily be standardised. Doryab (87), for example, provides a generic framework for data processing and feature extraction, which was developed from student data but may be a reasonable starting point for other populations.

A further recommendation is to improve the generalisability of results by testing these technologies in well characterized clinical and general population samples, with larger study designs, and collected over longer periods of time. The importance of recruiting and reporting the diversity of study samples is highlighted by the difference in validity of these devices in detecting the behaviours of interest. For example, some wearable devices may be as less accurate on darker skin tones (88), and on women (89).

The literature we identified derives from both clinical and computer science disciplines and some of the heterogeneity we report results from these disciplines having distinct conventions, with medical outputs putting more weight sample and clinical outcome characteristics but often overlooking feature extraction and analysis description.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Our most pressing recommendation, however, is that there is a need for consistency in reporting in this field. The failure to report basic demographic information found in many studies, particularly from the computer science field, and the limited description in feature extraction and analysis in medical papers, has important implications for the interpretation of findings. A common framework, with standardised assessment and analytical tools, robust feature extraction and missing data descriptions, tested in more representative populations would be an important step towards improving the ability of researchers to evaluate the strength of the evidence.

Declaration of interests:
M.H. is principal investigator of the RADAR-CNS programme, a precompetitive public–private partnership funded by the Innovative Medicines Initiative and European Federation of Pharmaceutical Industries and Associations. The programme receives support from Janssen, Biogen, MSD, UCB and Lundbeck.

D.C.M. has accepted honoraria and consulting fees from Apple, Inc., Otsuka Pharmaceuticals, Pear Therapeutics, and the One Mind Foundation, royalties from Oxford Press, and has an ownership interest in Adaptive Health, Inc.

All other authors declare that they have no competing interests.

Acknowledgments:
This study represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust; (3) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; (4) the National Institute for Health Research University College London Hospitals Biomedical Research Centre; (5) the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; (7) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES:

1. World Health Organisation. Depression [Internet]. 2020 [cited 2021 May 27]. Available from: https://www.who.int/news-room/fact-sheets/detail/depression
2. Verhoeven JE, Verduijn J, Schoevers RA, van Hemert AM, Beekman ATF, Penninx BWJH. [Complete recovery from depression is the exception rather than the rule: prognosis of depression beyond diagnostic boundaries]. Ned Tijdschr Geneeskd. 2018 Sep 6;162.
3. Kraus C, Kadriu B, Lanzenberger R, Jr CAZ, Kasper S. Prognosis and improved outcomes in major depression: a review. Transl Psychiatry. 2019 Apr 3;9(1):1–17.
4. Cho YM, Lim HJ, Jang H, Kim K, Choi JW, Shin C, et al. A cross-sectional study of the association between mobile phone use and symptoms of ill health. Environ Health Toxicol [Internet]. 2016 Oct 26 [cited 2019 Jun 11];31. Available from: http://www.e-eht.org/journal/view.php?doi=10.5620/eht.e2016022
5. Lu J, Shang C, Yue C, Morillo R, Ware S, Kamath J, et al. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018 Mar 26;2(1):21:1-21:21.
6. 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). 2017. p. 325–32. (International Conference on Affective Computing and Intelligent Interaction).
7. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res. 2015 Jul 15;17(7):e175.
8. Vailshery LS. Ownership of smartphones in the UK 2020 [Internet]. Statista. 2021 [cited 2021 May 26]. Available from: https://www.statista.com/statistics/956297/ownership-of-smartphones-uk/
9. Mohr DC, Shilton K, Hotopf M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. Npj Digit Med. 2020 Mar 25;3(1):1–2.
10. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. In: Widiger, T and Cannon, TD, editor. ANNUAL REVIEW OF CLINICAL PSYCHOLOGY, VOL 13. 2017. p. 23–47. (Annual Review of Clinical Psychology; vol. 13).
11. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. JMIR MHealth UHealth. 2018;6(8):e165.
12. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health. 2020 Sep 30;ebmental-2020-300180.
13. Faurholt-Jepsen M, Brage S, Vinberg M, Christensen EM, Knorr U, Jensen HM, et al. Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state. J Affect Disord. 2012 Dec 10;141(2–3):457–63.
14. Dogan E, Sander C, Wagner X, Hegerl U, Kohls E. Smartphone-Based Monitoring of Objective and Subjective Data in Affective Disorders: Where Are We and Where Are We Going? Systematic Review. J Med Internet Res. 2017;19(7):e262.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med. 2009 Jul 21;6(7):e1000097.

16. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open. 2016 Dec 1;6(12):e011458.

17. Wells G, Shea B, O'Connell D, Robertson J, Peterson J, Welch V, et al. The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta- Analysis. 2000;21.

18. Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Appl Psychol Meas. 1977 Jun 1;1(3):385–401.

19. Hamilton M. A RATING SCALE FOR DEPRESSION. J Neurol Neurosurg Psychiatry. 1960 Feb;23(1):56–62.

20. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. J Gen Intern Med. 2001;16(9):606–13.

21. Yue C, Ware S, Morillo R, Lu J, Shang C, Bi J, et al. Fusing Location Data for Depression Prediction. IEEE Trans Big Data. 2018;1–1.

22. Ávila Moraes C, Cambras T, Diez-Noguera A, Schimitt R, Dantas G, Levandovski R, et al. A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. BMC Psychiatry. 2013 Mar 9;13:77.

23. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health. Psychiatr Rehabil J. 2015 Sep;38(3):218–26.

24. Boukhechba M, Cai L, Chow PI, Fua K, Gerber MS, Teachman BA, et al. Contextual Analysis to Understand Compliance with Smartphone-based Ecological Momentary Assessment. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare [Internet]. New York, NY, USA: ACM; 2018. p. 232–8. (PervasiveHealth '18). Available from: http://doi.acm.org/10.1145/3240925.3240967

25. Burns MN, Begale M, Duffecy J, Gergle D, Karr CJ, Giangrande E, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res. 2011 Aug 12;13(3):e55.

26. Byrne JEM, Bullock B, Brydon A, Murray G. A psychometric investigation of the sleep, circadian rhythms, and mood (SCRAM) questionnaire. Chronobiol Int. 2019 Feb 1;36(2):265–75.

27. Caldwell BA, Redeker NS. Sleep patterns and psychological distress in women living in an inner city. Res Nurs Health. 2009 Apr;32(2):177–90.

28. David ME, Roberts JA, Christenson B. Too Much of a Good Thing: Investigating the Association between Actual Smartphone Use and Individual Well-Being. Int J Human–Computer Interact. 2018 Mar 4;34(3):265–75.

29. Difrancesco S, Lamers F, Riese H, Merikangas KR, Beekman ATF, van Hemert AM, et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. Depress Anxiety. 2019 Oct;36(10):975–86.

30. Dillon CB, McMahon E, O'Regan G, Perry IJ. Associations between physical behaviour patterns and levels of depressive symptoms, anxiety and well-being in middle-aged adults: a cross-sectional study using isotemporal substitution models. BMJ Open. 2018 21;8(1):e018978.

31. Doane LD, Gress-Smith JL, Breitenstein RS. Multi-method assessments of sleep over the transition to college and the associations with depression and anxiety symptoms. J Youth Adolesc. 2015 Feb;44(2):389–404.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

32. Doryab A, Min JK, Wiese J, Zimmerman J, Hong J. Detection of Behavior Change in People with Depression. Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014;5.

33. Haeffel GJ. Don't sleep on it: Less sleep reduces risk for depressive symptoms in cognitively vulnerable undergraduates. J Pers Soc Psychol. 2017 Dec;113(6):925–38.

34. Hori H, Koga N, Hidese S, Nagashima A, Kim Y, Higuchi T, et al. 24-h activity rhythm and sleep in depressed outpatients. J Psychiatr Res. 2016 Jun;77:27–34.

35. Jacobson NC, Weingarden H, Wilhelm S. Using Digital Phenotyping to Accurately Detect Depression Severity. J Nerv Ment Dis. 2019 Oct;207(10):893–6.

36. Kawada T, Katsumata M, Suzuki H, Shimizu T. Actigraphic predictors of the depressive state in students with no psychiatric disorders. J Affect Disord. 2007 Feb;98(1–2):117–20.

37. Knight A, Bidargaddi N. Commonly available activity tracker apps and wearables as a mental health outcome indicator: A prospective observational cohort study among young adults with psychological distress. J Affect Disord. 2018 15;236:31–6.

38. Li X, Kearney PM, Fitzgerald AP. Accelerometer-Based Physical Activity Patterns and Correlates of Depressive Symptoms. In: Siuly, S and Lee, I and Huang, Z and Zhou, R and Wang, H and Xiang, W, editor. HEALTH INFORMATION SCIENCE (HIS 2018). 2018. p. 37–47. (Lecture Notes in Computer Science; vol. 11148).

39. Luik AI, Zuurbier LA, Hofman A, Van Someren EJW, Tiemeier H. Stability and fragmentation of the activity rhythm across the sleep-wake cycle: the importance of age, lifestyle, and mental health. Chronobiol Int. 2013 Dec;30(10):1223–30.

40. Luik AI, Zuurbier LA, Direk N, Hofman A, Someren EJWV, Tiemeier H. 24-Hour Activity Rhythm and Sleep Disturbances in Depression and Anxiety: A Population-Based Study of Middle-Aged and Older Persons. Depress Anxiety. 2015;32(9):684–92.

41. McCall WV. A rest-activity biomarker to predict response to SSRIs in major depressive disorder. J Psychiatr Res. 2015 May 1;64:19–22.

42. Mendoza-Vasconez AS, Marquez B, Linke S, Arredondo EM, Marcus BH. Effect of Physical Activity on Depression Symptoms and Perceived Stress in Latinas: A Mediation Analysis. Ment Health Phys Act. 2019 Mar;16:31–7.

43. Moukaddam N, Truong A, Cao J, Shah A, Sabharwal A. Findings From a Trial of the Smartphone and OnLine Usage-based eValuation for Depression (SOLVD) Application: What Do Apps Really Tell Us About Patients with Depression? Concordance Between App-Generated Data and Standard Psychiatric Questionnaires for Depression and Anxiety. J Psychiatr Pract. 2019 Sep;25(5):365–73.

44. Naismith SL, Rogers NL, Lewis SJG, Terpening Z, Ip T, Diamond K, et al. Sleep disturbance relates to neuropsychological functioning in late-life depression. J Affect Disord. 2011 Jul;132(1–2):139–45.

45. Park D-H, Kripke DF, Cole RJ. More Prominent Reactivity in Mood Than Activity and Sleep Induced by Differential Light Exposure Due to Seasonal and Local Differences. Chronobiol Int. 2007 Jan 1;24(5):905–20.

46. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res. 2014 Jul;77(1):70–5.

47. Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SD, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety. 2019 Jan 1;36(1):72–81.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

48. Robillard R, Naismith SL, Rogers NL, Scott EM, Ip TKC, Hermens DF, et al. Sleep-wake cycle and melatonin rhythms in adolescents and young adults with mood disorders: comparison of unipolar and bipolar phenotypes. Eur Psychiatry J Assoc Eur Psychiatr. 2013 Sep;28(7):412–6.

49. Robillard R, Naismith SL, Smith KL, Rogers NL, White D, Terpening Z, et al. Sleep-wake cycle in young and older persons with a lifetime history of mood disorders. PloS One. 2014;9(2):e87763.

50. Robillard R, Hermens DF, Naismith SL, White D, Rogers NL, Ip TKC, et al. Ambulatory sleep-wake patterns and variability in young people with emerging mental disorders. J Psychiatry Neurosci JPN. 2015 Jan;40(1):28–37.

51. Robillard R, Hermens DF, Lee RSC, Jones A, Carpenter JS, White D, et al. Sleep-wake profiles predict longitudinal changes in manic symptoms and memory in young people with mood disorders. J Sleep Res. 2016;25(5):549–55.

52. Robillard R, Carpenter JS, Rogers NL, Fares S, Grierson AB, Hermens DF, et al. Circadian rhythms and psychiatric profiles in young adults with unipolar depressive disorders. Transl Psychiatry. 2018 Oct 9;8(1):213.

53. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ. 2016;4:e2537.

54. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, et al. Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. J Med Internet Res. 2018 Jun 8;20(6):e210.

55. Slyepchenko A, Allega OR, Leng X, Minuzzi L, Eltayebani MM, Skelly M, et al. Association of functioning and quality of life with objective and subjective measures of sleep and biological rhythms in major depressive and bipolar disorder. Aust N Z J Psychiatry. 2019 Jul;53(7):683–96.

56. Smagula SF, Krafty RT, Thayer JF, Buysse DJ, Hall MH. Rest-activity rhythm profiles associated with manic-hypomanic and depressive symptoms. J Psychiatr Res. 2018;102:238–44.

57. Smagula SF, DuPont CM, Miller MA, Krafty RT, Hasler BP, Franzen PL, et al. Rest-activity rhythms characteristics and seasonal changes in seasonal affective disorder. Chronobiol Int. 2018;35(11):1553–9.

58. Stremler R, Haddad S, Pullenayegum E, Parshuram C. Psychological Outcomes in Parents of Critically Ill Hospitalized Children. J Pediatr Nurs. 2017 Jun;34:36–43.

59. Tao K, Liu W, Xiong S, Ken L, Zeng N, Peng Q, et al. Associations between Self-Determined Motivation, Accelerometer-Determined Physical Activity, and Quality of Life in Chinese College Students. Int J Environ Res Public Health. 2019 Aug 16;16(16).

60. Todder D, Caliskan S, Baune BT. Longitudinal changes of day-time and night-time gross motor activity in clinical responders and non-responders of major depression. World J Biol Psychiatry. 2009 Jan 1;10(4):276–84.

61. Vallance JK, Eurich D, Lavallee C, Johnson ST. Daily Pedometer Steps among Older Men: Associations with Health-Related Quality of Life and Psychosocial Health. Am J Health Promot. 2013 May 1;27(5):294–8.

62. Vanderlind WM, Beevers CG, Sherman SM, Trujillo LT, McGeary JE, Matthews MD, et al. Sleep and sadness: exploring the relation among sleep, cognitive control, and depressive symptoms in young adults. Sleep Med. 2014 Jan;15(1):144–9.

63. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. JMIR MHealth UHealth. 2016 Sep 21;4(3):e111.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

64. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct [Internet]. Seattle, Washington: ACM Press; 2014 [cited 2019 Jun 7]. p. 3–14. Available from: http://dl.acm.org/citation.cfm?doid=2632048.2632054

65. Wang R, Wang W, daSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018;2(1):1–26.

66. White KH, Rumble ME, Benca RM. Sex Differences in the Relationship Between Depressive Symptoms and Actigraphic Assessments of Sleep and Rest-Activity Rhythms in a Population-Based Sample. Psychosom Med. 2017 May;79(4):479–84.

67. Yang Z, Mo X, Shi D, Wang R. Mining relationships between mental health, academic performance and human behaviour. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). 2017. p. 1–8.

68. Yaugher AC, Alexander GM. Internalizing and externalizing traits predict changes in sleep efficiency in emerging adulthood: an actigraphy study. Front Psychol. 2015;6:1495.

69. Naismith SL, Rogers NL, Lewis SJG, Terpening Z, Ip T, Diamond K, et al. Sleep disturbance relates to neuropsychological functioning in late-life depression. J Affect Disord. 2011 Jul 1;132(1):139–45.

70. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res. 2014 Jul 1;77(1):70–5.

71. Vanderlind WM, Beevers CG, Sherman SM, Trujillo LT, McGeary JE, Matthews MD, et al. Sleep and sadness: exploring the relation among sleep, cognitive control, and depressive symptoms in young adults. Sleep Med. 2014 Jan;15(1):144–9.

72. Takano K, Sakamoto S, Tanno Y. Repetitive Thought Impairs Sleep Quality: An Experience Sampling Study. Behav Ther. 2014 Jan 1;45(1):67–82.

73. Wang R, Wang W, daSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018 Mar;2(1):43:1-43:26.

74. Robillard R, Hermens DF, Lee RSC, Jones A, Carpenter JS, White D, et al. Sleep-wake profiles predict longitudinal changes in manic symptoms and memory in young people with mood disorders. J Sleep Res. 2016 Oct;25(5):549–55.

75. Littner M, Kushida CA, Anderson WM, Bailey D, Berry RB, Davila DG, et al. Practice Parameters for the Role of Actigraphy in the Study of Sleep and Circadian Rhythms: An Update for 2002. Sleep. 2003 May 1;26(3):337–41.

76. Xu X, Creswell JD, Mankoff J, Dey AK, Chikersal P, Doryab A, et al. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2019 Sep 9;3(3):1–33.

77. Liu T, Nicholas J, Theilig MM, Guntuku SC, Kording K, Mohr DC, et al. Machine Learning for Phone-Based Relationship Estimation: The Need to Consider Population Heterogeneity. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2019 Dec 11;3(4):145:1-145:23.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.19.21260786; this version posted July 22, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

78. Gerpott TJ, Thomas S, Weichert M. Characteristics and mobile Internet use intensity of consumers with different types of advanced handsets: An exploratory empirical study of iPhone, Android and other web-enabled mobile users in Germany. Telecommun Policy. 2013 May 1;37(4):357–71.

79. Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychol Aging. 2006 Mar;21(1):140–51.

80. Segel-Karpas D, Ayalon L, Lachman ME. Loneliness and depressive symptoms: the moderating role of the transition into retirement. Aging Ment Health. 2018 Jan 2;22(1):135–40.

81. Moukaddam N, Truong A, Cao J, Shah A, Sabharwal A. Findings From a Trial of the Smartphone and OnLine Usage-based eValuation for Depression (SOLVD) Application: What Do Apps Really Tell Us About Patients with Depression? Concordance Between App-Generated Data and Standard Psychiatric Questionnaires for Depression and Anxiety. J Psychiatr Pract. 2019 Sep;25(5):365–73.

82. Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SD, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety. 2019;36(1):72–81.

83. Razavi R, Gharipour A, Gharipour M. Depression screening using mobile phone usage metadata: a machine learning approach. J Am Med Inform Assoc. 2020 Apr 1;27(4):522–30.

84. Germain A, Kupfer DJ. Circadian rhythm disturbances in depression. Hum Psychopharmacol Clin Exp. 2008;23(7):571–85.

85. Wall R, Cunningham P, Walsh P, Byrne S. Explaining the output of ensembles in medical decision support on a case by case basis. Artif Intell Med. 2003 Jun 1;28(2):191–206.

86. Faurholt-Jepsen M, Geddes JR, Goodwin GM, Bauer M, Duffy A, Vedel Kessing L, et al. Reporting guidelines on remotely collected electronic mood data in mood disorder (eMOOD)—recommendations. Transl Psychiatry. 2019 Jun 7;9(1):1–10.

87. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of Behavioral Features from Smartphone and Wearable Data. ArXiv181210394 Cs Stat [Internet]. 2019 Jan 8 [cited 2021 May 27]; Available from: http://arxiv.org/abs/1812.10394

88. Colvonen PJ, DeYoung PN, Bosompra N-OA, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. Sleep [Internet]. 2020 Oct 13 [cited 2020 Dec 7];43(10). Available from: https://academic.oup.com/sleep/article/43/10/zsaa159/5902283

89. Nuss KJ, Thomson EA, Courtney JB, Comstock A, Reinwald S, Blake S, et al. Assessment of Accuracy of Overall Energy Expenditure Measurements for the Fitbit Charge HR 2 and Apple Watch. Am J Health Behav. 2019 May 1;43(3):498–505.

Digital Health Tools for the Passive Monitoring of Depression: A Systematic Review of Methods

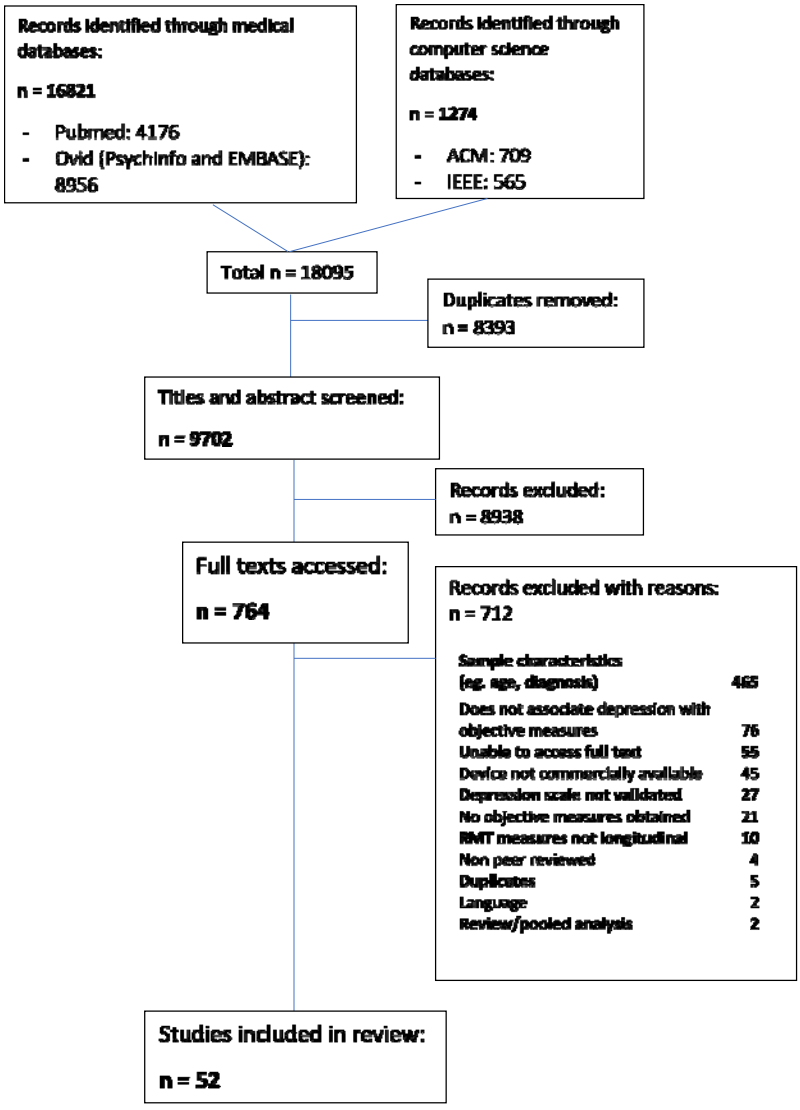


Figure 1: Study selection workflow

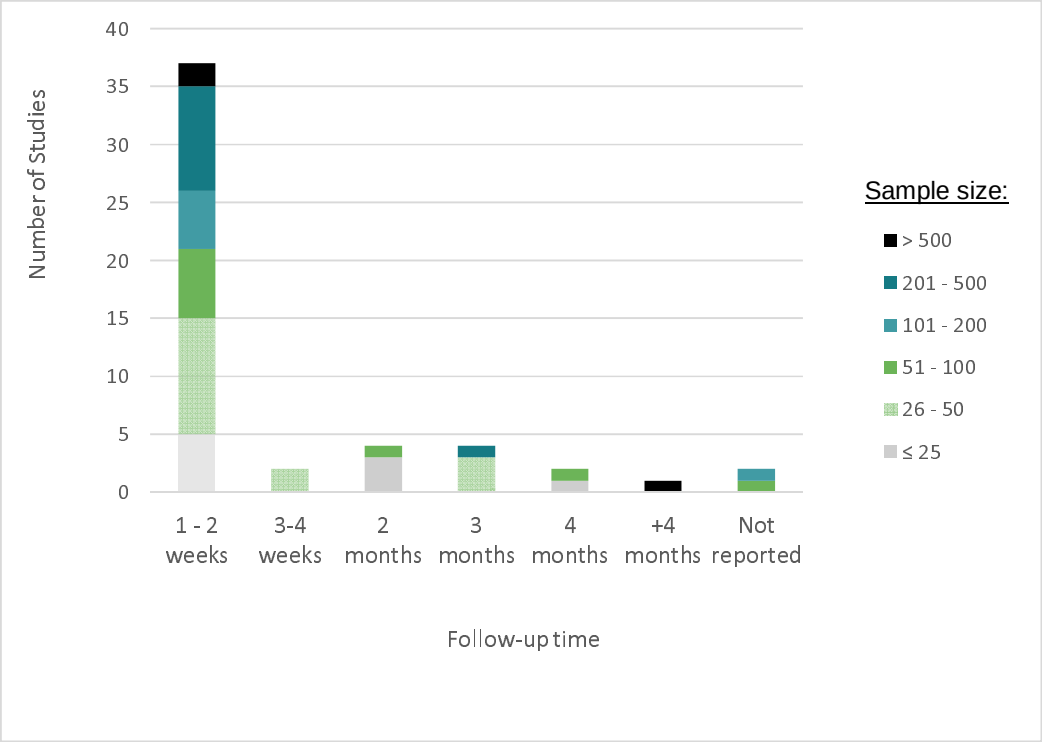


Figure 2. The sample sizes and follow-up times for all included studies

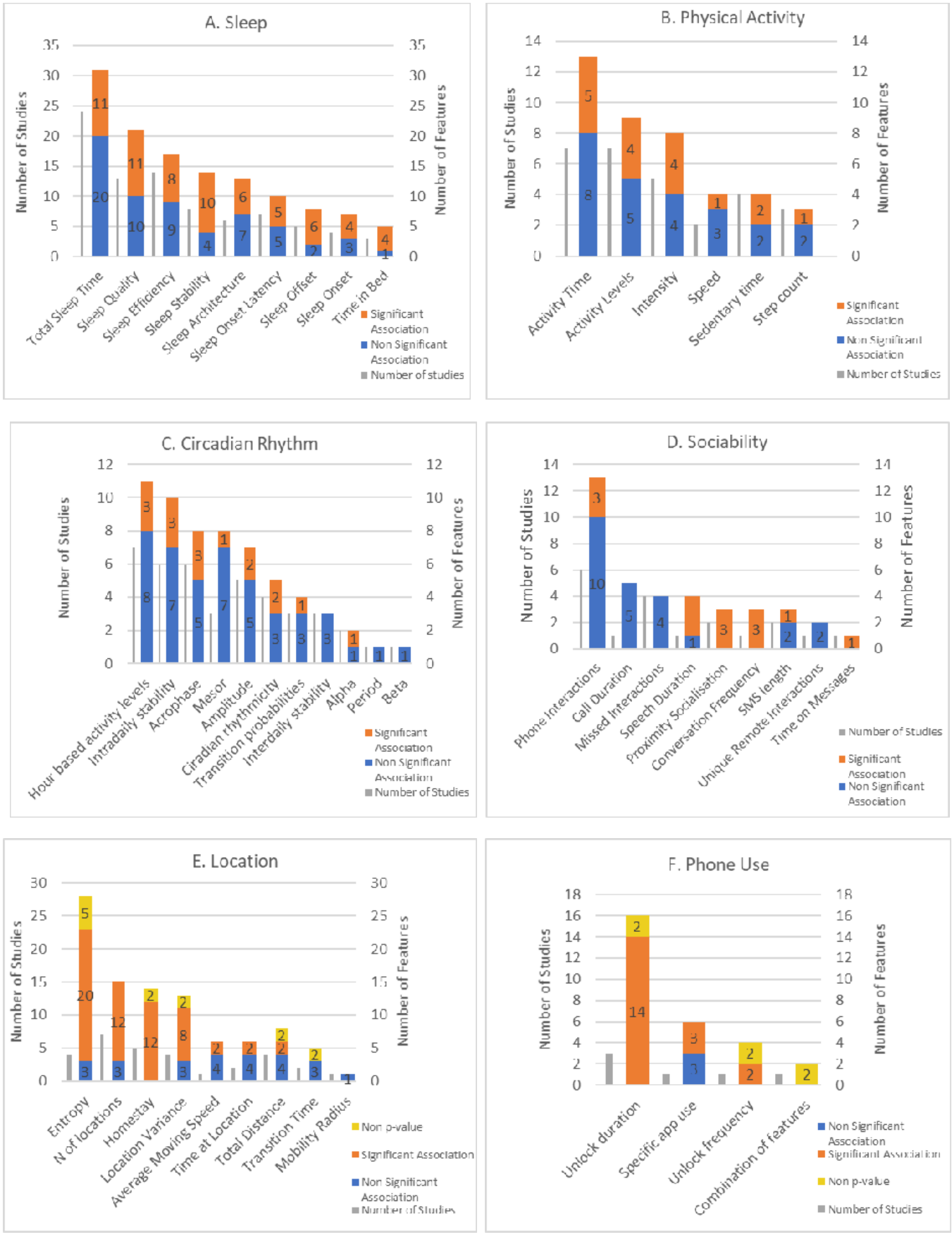


Figure 3. The number of times each feature has been reported in all included studies and their association with depression, where these associations are defined as having a below-threshold p-value ("Sig. Association"), above-threshold p-value ("Non Sig. Association"), and where statistical methods have been used that do not yield p-values ("Non p value"). The graphs show the number of studies assessing each feature.

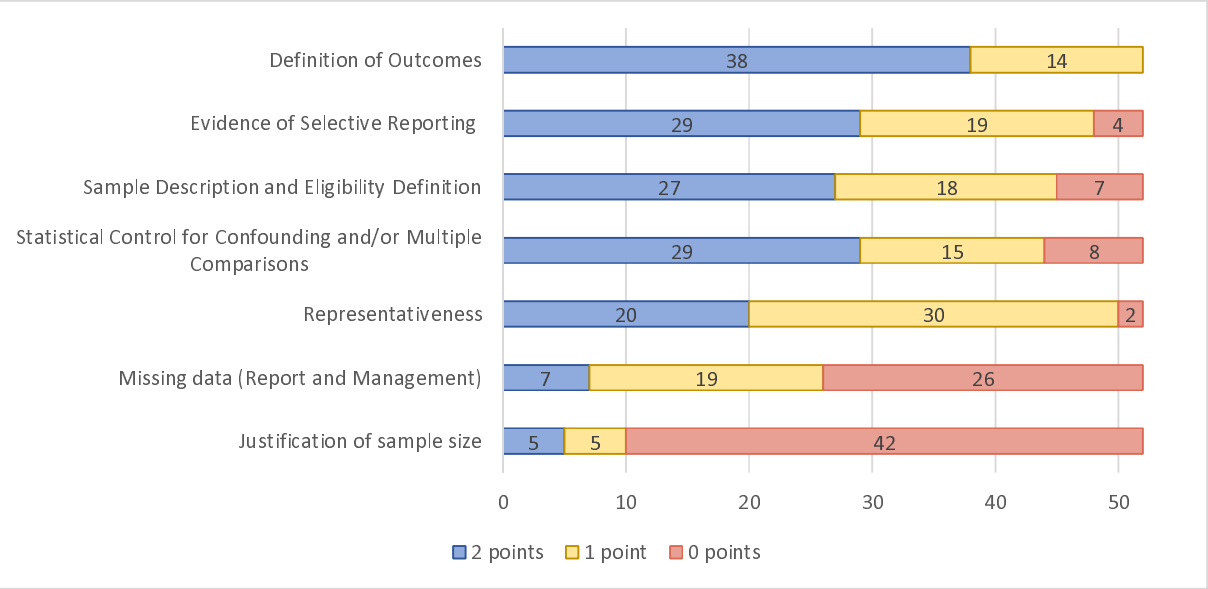


Figure 4 shows the number of studies scoring on each item.
2 points are given for fully addressing quality criteria, 1 point for partially addressing quality criteria, and 0 points for failing to address quality criteria.

Appendices

Appendix 1 – Search terms

Sources: 1) Pubmed, 2) Embase, PsychInfo via OVID, 3) IEEE Xplore, 4) ACM Digital library, 5) Web of science,

Searched for terms around the following key concepts:

1. Depression, depressive disorder
2. RMTs, sensors, technologies (Portable or wearable technology)

	Keywords	Ti/Ab
1	mood disorder; affective disorder; depression; depressive mood symptoms, mental health, depress*, Unipolar affective disorder, mental disorders	remote emotional health monitoring system, mood,
2	"Objective Behavioral Features", objective features; sensor data; "smart phone"; wearable devices; wearable, smartphone, app, apps, accelerometer, pedometer, actigraphy, motor activity, Psychomotor activity, Acceleration, Heart rate, heart rate and movement sensor, "digital biomarker", digital phenotype	activity measurement, wrist-worn, remote, Psychomotor activity, objectively measured activity parameters, Electronic monitoring, objective measure

1) **Pubmed search:**

	Depressive Disorder[Mesh] OR Major Depressive Disorder[Mesh] OR Depression[Mesh] OR depressi*[Title/Abstract] OR "affective disorder"[Title/Abstract] OR "mood disorder"[Title/Abstract]
AND	Remote Sensing Technology [MeSH] "digital"[Title/Abstract] OR smartphone[Title/Abstract] OR mobile[Title/Abstract] OR wearable[Title/Abstract] OR "objective measure"[Title/Abstract] OR "sensor data" OR "wearable devices" OR "smart phone" OR app OR apps OR "activity measure" OR acceleromet* OR pedomet* OR actigraph* OR "psychomotor activity" OR "remote monitoring" OR "GPS" OR "global positioning system" OR "mobile sensor" OR "RMT" OR "remote measurement technologies" OR mHealth OR "digital biomarker" OR "digital phenotype"

Limit to 2007

2) **OVID: PsychInfo and EMBASE**

exp Major Depression/

major depression/ or affective disorder/ or depressive disorder.mp.

("depressive mood" or "depressed mood" or "depressive symptoms" or "depressed symptoms" or "affective symptoms" or "mood disorder" or depression).ti,ab.

(digital or smartphone or mobile or wearable or objective measure).ti,ab.

("sensor data" or "wearable device" or "smart phone" or smartphone or accelerometer or pedometer or actigraphy or "psychomotor activity" or "remote monitoring" or "GPS" or "global positioning system" or "mobile

sensor" or "RMT" or "remote measurement technologies" or mhealth or "machine learning" or app or apps or "activity measure" or "digital biomarker").mp.

limit to yr="2007 -Current"

1	major depression/ or affective disorder/ or depressive disorder.mp.	262002
2	("depressive mood" or "depressed mood" or "depressive symptoms" or "depressed symptoms" or "affective symptoms" or "mood disorder" or depression).ti,ab.	716408
3	1 or 2	795172
4	(digital or mobile or wearable or objective measure).ti,ab.	318272
5	("sensor data" or "remote sensing technology" or "wearable device" or "smart phone" or smartphone or accelerometer or pedometer or actigraphy or "psychomotor activity" or "remote monitoring" or "GPS" or "global positioning system" or "mobile sensor" or "RMT" or "remote measurement technologies" or mhealth or "machine learning" or app or apps or "activity measure" or "digital biomarker").mp.	187037
6	4 or 5	485984
7	3 and 6	10688
8	limit 7 to yr="2016 -Current"	4635
9	limit 7 to yr="2007 - 2015"	4321
10	remove duplicates from 9	3526
11	remove duplicates from 8	3929
12	10 or 11	7455

3) **IEEE Xplore:**

("remote sensing technology" OR "psychomotor activity" OR "RMT" OR "mhealth" OR "accelerometer" OR "pedometer" OR "actigraphy" OR "sensor data" OR "sensing technology" OR "GPS" or "global positioning system" OR "mobile sensor" OR "smartphone" OR "mobile" OR "wearable" OR "smart phone" OR "app" OR "apps" OR "digital biomarker*" OR "digital phenotype") AND ("depression" OR "depressed mood" OR "depressive symptoms" OR "affective disorder" OR "mental health" OR "mood disorder" OR "mood")

Limit to 2007 onwards

4) ACM Digital library,

NOTE: adding “depressive mood” and/or “depressed symptoms” does not improve the search

"depressive mood" or "depressed mood" or "depressive symptoms" or "depressed symptoms" or "affective symptoms" or "mood disorder" or depression or "major depression" or "affective disorder" or "depressive disorder"

AND

"remote sensing technology" or "sensor data" or "wearable device" or "smart phone" or "smartphone" accelerometer or pedometer or actigraphy or "psychomotor activity" or "remote monitoring" or "GPS" or "global positioning system" or "mobile sensor" or "RMT" or "remote measurement technologies" or mhealth

"filter": {"publicationYear":{ "gte":2007 }},

{owners.owner=GUIDE}

5) Web of science,

# 9	3,689	#8 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=2007-2019*
# 8	4,305	#7 AND #4 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 7	327,872	#6 OR #5 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 6	147,948	TI=(wearable OR mobile) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 5	193,432	TS=("remote sensing technology" OR "sensor data" OR "wearable device" OR "smart phone" OR smartphone OR accelerometer OR pedometer or actigraphy OR "psychomotor activity" OR "remote monitoring" OR "GPS" OR "global positioning system" OR "mobile sensor*" OR "sensing technologies" OR "RMT" OR "remote measurement technologies" OR mhealth OR "digital biomarker*" OR "digital phenotype*") *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 4	542,297	#3 OR #2 OR #1 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 3	16,431	TS=("affective disorder" OR "mood disorder") *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 2	224,301	TI=(depress* OR "affective disorder" OR mood) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*
# 1	492,570	TS=(Depressive Disorder OR Major Depressive Disorder OR Depression) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=1900-2019*

Appendix 2 – Study Quality Assessment

1	**Protocol Published? [Y, N]**	Is a published protocol mentioned? 1= Yes, 0 = No
2	**Definition of Outcomes (Clinical outcomes and Objective Features)**	2. clear and appropriate definition of outcomes (depression and objective measures). 1. unclear or incomplete 0. none reported
3	**Evidence of Selective reporting (data measured but not reported)**	2. analysed data matches study objectives and post hoc analyses clearly defined as such 1. some variables measured not mentioned/ reported in results. 0. Significant results not defined at the outset nor in line with study objectives.
4	**Sample Description and Eligibility Definition**	2. Gives well-defined eligibility criteria, and the sources and methods of selection of participants. 1. Eligibility criteria incomplete or unclear, or clinical-based inclusion of depression defined by self report (and not assessed by clinician). 0. no mention of sampling strategy / eligibility
5	**Statistical Control for Confounding and/or Multiple Comparisons**	2. clear and appropriate . 1. unclear or incomplete 0. none reported
6	**Missing data (Report and management)**	2. clear and appropriate . 1. unclear or incomplete 0. none reported
7	**Representativeness**	2. sample representative of population of interest, 1. potential selection/sampling bias, 0. sample largely different to the populations it aims to study.
8	**Justification of sample size**	2. clear and appropriate justification of sample size. 1. unclear or incomplete explanation for sample size. 0. none reported

Appendix 3 – Feature Descriptions

Sleep

Low-Level Features	Description	Study-level feature examples
Total Sleep Time	The amount of actually sleep time in a sleep episode; this time is equal to the total sleep episode less the awake time	Total Sleep Time (TST) Total Sleep Duration Per Night Awake Duration
Sleep Quality	A combination of factors which relate to how much of the time that is intended for sleep is actually spent sleeping.	Intradaily Variability In Sleep Wake After Sleep Onset Number Of Nocturnal Awakenings Minutes After Wakeup
Sleep Efficiency	Sleep efficiency is another measure of sleep quality presented independently due to its popularity. It is the percentage of time spent asleep while in bed. It is calculated by dividing the amount of time spent asleep (in minutes) by the total amount of time in bed (in minutes).	Sleep Efficiency
Sleep Stability	Features of variability in sleep	Sleep Start Time Variability

		Interdaily Stability In Sleep Sleep Variability Standard Deviation Sleep Onset Variability
Sleep Architecture	The basic structure of sleep	Mean Mid-Sleep Time Acrophase Of Sleep Mid Sleep On Free Days Mean Mid-Sleep Time
Sleep Onset Latency	The amount of time it takes you to go from being fully awake to sleeping	Sleep Onset Latency (SOL) Sleep Onset Latency (Women) Sleep Onset Latency (Men)
Sleep Onset	The time at which sleep onset happens	Sleep Onset
Sleep Offset	The time at which the individual awakens.	Sleep Offset
Time In Bed	Total amount of time spent in bed	Time In Bed

Physical activity

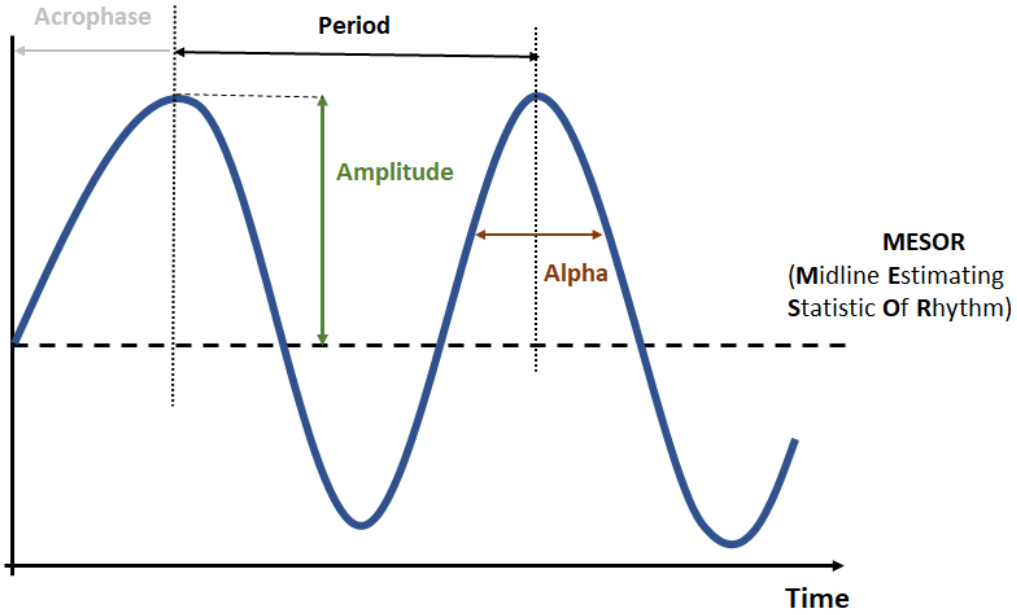
Low-Level Features	Description	Study-level feature examples
Activity Time	Time spent engaging in physical activity	Summation Of All Active Periods Fraction Of Time In Motion Minutes Per Week Of Physical Activity Average Wrist Activity/Min
Activity Levels	General levels of activity	Average 24-Hour Activity Gross Motor Activity Per Day Standard Deviation Motion Average Motion
Intensity	Activity features that differentiate between light, moderate and vigorous activity.	Minutes Of Heart Rate In Fat-Burn Zone Time In Moderate-To-Vigorous Physical Activity Time In Light Physical Activity
Speed	Movement speed	Speed Mean Speed Variance Average Moving Speed
Sedentary Time	Total time spent doing no activity	Sedentary Minutes Time In Sedentary Behaviours Stationary Time (Mean)
Step Count	Step count	Step Count

Circadian Rhythm

Low-Level Features	Description	Study-level feature examples
Hour Based Activity Levels	Activity levels at different times of day	Motor activity (diurnal) daytime activity levels(DALs) (am) daytime activity levels(DALs) (pm) mean activity during active period (day)
Intradaily Stability	The ratio of the hour-to-hour activity variability to the overall activity variability (higher values reflect more fragmented rhythms, e.g., due to frequent daytime napping or night-time awakenings.	Intradaily stability Intradaily variability change in intra-daily variability (IV),
Acrophase	peak of activity: a measure of the timing of overall high values recurring in each cycle, expressed in (negative) degrees in relation to a reference time set to 0°, with 360° equated to the period; and the period is the duration of one cycle	Acrophase
MESOR	MEAN activity levels: a rhythm-adjusted mean	MESOR

Amplitude	difference between peak and troughs of activity: difference between most active time and the least active time of the day, (higher values indicate a greater RAR amplitude) a measure of the extent of predictable change within a cycle	Relative amplitude
Circadian Rhythmicity	The coefficient of determination (or R2; not illustrated here), a measure reflecting the goodness of fit, was used as an indicator of circadian rhythmicity.	Circadian Rhythmicity
Transition Probabilities	The probability of transitioning from active to rest state or vice versa	Active to Rest - day Active to rest - night Rest to Active - night
Interdaily Stability	Ratio of variability within the mean 24-hour activity profile to the overall activity variability (higher values indicate greater stability of the mean 24-hour profile across days)	Interdaily stability
Alpha	Relative width of the curve at the middle of the peak. Higher alpha indicates relatively narrower active periods compared to rest periods.	Alpha
Period	The time in between activity peaks, usually 24 hours. Shorter periods lead to behaviour occurring at an earlier clock time and long periods to later timing	Period
Beta	Indicator of the steepness of the rise and fall of the curve, indicative of a faster transition from rest to active.	Beta

Circadian Rhythm Graph:



Sociability

Low-Level Features	Description	Study-level feature examples
Frequency of	Frequency of Phone Calls of Text	Daily Call count

Call Duration	Call duration	Call duration
Missed Interactions	Unreturned calls	Unreturned calls Missed interactions
Speech Duration	Length of detected speech	Speech duration Conversation duration during day Change in Conversation duration (slope)
Socialisation By Proximity	Detected proximity to others by nearby Bluetooth devices of speech.	Location/noise/voice Socialisation by proximity and noise
Conversation Frequency	Number of times conversation was detected nearby	Conversation frequency during day Conversation frequency during evening
SMS Length	SMS length	SMS length
Unique Remote Interactions	Total number of unique individuals with whom a participant interacted through phone calls or SMS messages on a particular day	Interaction diversity
Time Spent On Messages	Total time spent using messages	Total time spent using messages

Location

Low-Level Features	**Description**	**Study-level feature examples**
Entropy	The variability of time the participants spend at a certain location. High entropy translates to spending time more uniformly across different locations.	Entropy Normalised Entropy
N of Locations	The number of locations visited	Number of unique locations Total standard deviation of location Number of clusters
Home Stay	Amount of time spent at the location identified as Home	Homestay
Location Variance	The variability in a participant's location	Location Variance
Average Moving Speed	Average Moving Speed	Average Moving Speed
Time At Location	Average amount of time spent in a particular location.	Average staying time per visit across the study Cumulative staying time across the study Time at on-campus health facilities(mean)
Total Distance	Total Distance Travelled by a participant	Total Distance
Transition Time	The percentage of time during which a participant was in a non-stationary state. This was calculated by dividing the number of GPS location samples in transition states by the total number of samples.	Transition Time
Mobility Radius	The radius of the area within which a person moved.	Mobility Radius

Phone Use

Low-Level Features	**Description**	**Study-level feature examples**
Unlock Duration	The amount of time a person's phone is unlocked and therefore in use. Commonly referred as screen-time.	Total phone usage duration Mean phone usage duration at student accommodation
Specific App Use	The types of apps used.	total time spent using Instagram total time spent using maps total time spent using photo app
Unlock Frequency	The number of times a phone is unlocked.	phone usage frequency
Combination Of Phone Use	Combination of smartphone use features	combination of all smartphone use features

Speech

Low-Level Features	Description	Study-level feature examples
System Features	Articulation, formant features - changes in system speech components.	Formant 1 average acceleration (free response) Formant 1 average acceleration (read speech) Formant 1 variance (read speech)
Energy		Energy variability - velocity (speech) Energy variability - velocity (inversion)
Speech Rate		Phoneme rate speech (free response) Pseudosyllable rate speech (free response) Phoneme rate speech (read speech)
Pitch		Average pitch velocity (free response) Pitch variance (free response)
Shimmer		Shimmer
Aspiration		Aspiration
Jitter		Jitter

Physiology

Low-Level Features	Description	Study-level feature examples
Temperature	Temperature recorded from skin.	Diurnal Peripheral temperature Amplitude of temperature rhythm Mean elevated temperature time
Heart Rate	The number of heart beats per minute	Heart rate
Electrodermal Activity	Skin conductance	Electrodermal activity Difference in number of skin conductance level peaks

Environmental Features

Low-Level Features	Description	Study-level feature examples
Humidity	Environmental humidity	Humidity - males Humidity - females
Light	Ambient Light	Amplitude of light intensity Acrophase of light Mean elevated light time

Appendix 4 – Study Quality Assessment