MSLife — Digital Behavioral Phenotyping of Multiple Sclerosis Symptoms in the Wild Using Wearables and Graph-Based Statistical Analysis

GABRIEL GUO, Columbia University
HANBIN ZHANG, SUNY Buffalo
LIUYI YAO, SUNY Buffalo
HUINING LI, SUNY Buffalo
CHENHAN XU, SUNY Buffalo
ZHENGXIONG LI, University of Colorado Denver
WENYAO XU, SUNY Buffalo

Treatment for multiple sclerosis (MS) focuses on managing its symptoms (*e.g.*, depression, fatigue, poor sleep quality), varying with specific symptoms experienced. Thus, for optimal treatment, there arises the need to track these symptoms. Towards this goal, there is great interest in finding their relevant phenotypes. Prior research suggests links between activities of daily living (ADLs) and MS symptoms; therefore, we hypothesize that the behavioral phenotype (revealed through ADLs) is closely related to MS symptoms. Traditional approaches to finding behavioral phenotypes which rely on human observation or controlled clinical settings are burdensome and cannot account for all genuine ADLs. Here, we present *MSLife*, an end-to-end, burden-free approach to digital behavioral phenotyping of MS symptoms in the wild using wearables and graph-based statistical analysis. *MSLife* is built upon (1) low-cost, unobtrusive wearables (*i.e.*, smartwatches) that can track and quantify ADLs among MS patients in the wild; (2) graph-based statistical analysis that can model the relationships between quantified ADLs (*i.e.*, digital behavioral phenotype) and MS symptoms. We design, implement, and deploy *MSLife* with 30 MS patients across a one-week home-based IRB-approved clinical pilot study. We use the GENEActiv smartwatch to monitor ADLs and clinical behavioral instruments to collect MS symptoms. Then we develop a graph-based statistical analysis framework to model phenotyping relationships between ADLs and MS symptoms, incorporating confounding demographic factors. We discover 102 significant phenotyping relationships (*e.g.*, later rise times are related to increased levels of depression, history of caffeine consumption is associated with lower fatigue levels, higher relative levels of moderate physical activity are linked with decreased sleep quality). We validate their healthcare implications, using them to track MS symptoms in retrospective analysis. To our best knowledge, this is one of the first practices to digital behavioral phenotyping of MS symptoms in the wild.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: Multiple Sclerosis, Digital Behavioral Phenotyping, Mobile Health, Wearables, Graph-Based Statistical Analysis

Authors' addresses: Gabriel Guo, gzg2104@columbia.edu, Columbia University; Hanbin Zhang, hanbinzh@buffalo.edu, SUNY Buffalo; Liuyi Yao, SUNY Buffalo; Huining Li, huiningl@buffalo.edu, SUNY Buffalo; Chenhan Xu, chenhanx@buffalo.edu, SUNY Buffalo; Zhengxiong Li, zhengxiong.li@ucdenver.edu, University of Colorado Denver; Wenyao Xu, wenyaoxu@buffalo.edu, SUNY Buffalo.

2474-9567/2021/12-ART158 $15.00
https://doi.org/10.1145/3494970

ACM Reference Format:
Gabriel Guo, Hanbin Zhang, Liuyi Yao, Huining Li, Chenhan Xu, Zhengxiong Li, and Wenyao Xu. 2021. MSLife — Digital Behavioral Phenotyping of Multiple Sclerosis Symptoms in the Wild Using Wearables and Graph-Based Statistical Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 158 (December 2021), 35 pages. https://doi.org/10.1145/3494970

1 INTRODUCTION

Multiple sclerosis (MS) is one of five common neurological disorders, affecting 2.5 million people worldwide [1, 2]. In the United States, approximately 900,000 people live with MS [3], and its annual healthcare expense is estimated between $8528-$54,244 per patient per year [4]. Besides the first signs of vision and mobility problems, MS patients over time develop a spectrum of MS symptom complications, including depression, fatigue, and impaired sleep quality [5, 6]. Currently, there is no cure for MS, and the treatment typically focuses on managing these symptoms through intervention and behavioral therapies [2, 7–9]. The specific intervention depends on the symptoms a patient experiences: for instance, a patient experiencing depression may be prescribed antidepressants like desipramine, while a patient experiencing fatigue may be prescribed antifatigue medications like Symmetrel [2, 9, 10]. Thus, to ensure optimal treatment, there arises the need to track these experienced symptoms. To meet this goal of tracking MS symptoms, a recently popular area of research interest is finding their relevant phenotypes [11–18] (defined as "an individual's observable traits" [19]). Accordingly, in this paper, our aim is to find relevant phenotypes of MS symptoms.

In particular, whether or not the behavioral phenotype — revealed through activities of daily life (ADLs) — is closely related to MS symptoms is an open scientific/medical question [20–22]. Prior research suggests some link between ADLs and MS symptoms; therefore, we hypothesize that the behavioral phenotype is closely related to MS symptoms [9, 20, 21, 23, 24]. Traditionally, the way to find behavioral phenotypes would be to conduct a trial in a controlled clinical setting. The shortcoming of such controlled clinical trials is that, by definition, they are unable to capture data in the wild (*i.e.*, genuine activities of daily life). Therefore, in order to identify the specific ADLs that make up the behavioral phenotypes, we are limited to observational data collected in the wild. Yet, in collecting in the wild observational data, it is costly, burdensome, and unreliable to expect MS patients or clinicians to be aware of and manually record every single detail of their every single ADL.

Thus, to verify our hypothesis, we must meet the following challenges: (1) Which specific ADLs should we monitor as potential behavioral phenotypes of MS symptoms? (2) How can we design a convenient, low-cost system to continuously monitor these ADLs in the wild? (3) Given an observational dataset, how can we design a framework which models and identifies the relationships among MS symptoms and their behavioral phenotypes, while accounting for the simultaneous interactions among all the different variables in the dataset? Addressing challenge (1), we must account for activities done while awake as well as sleep patterns, as prior research suggests that these are all linked to MS symptoms [6, 25–28]. We should capture both micro (*e.g.*, time spent at different intensities of physical activity, day-by-day rise time) and macro (*e.g.*, the total number of physically active periods, the total number of awakenings) characteristics of these ADLs. Regarding challenge (2), digital phenotyping, the "moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices", is a suitable approach [29]. Particularly, we can utilize low-cost, unobtrusive wearable devices with embedded sensors (*i.e.*, smartwatches) to passively, continuously monitor the (digital) behavioral phenotype (ADLs) in the wild [30]. Finally, as for challenge (3), graph-based causal discovery and inference algorithms allow us to discover and quantify the relationships between different variables (*i.e.*, ADLs, MS symptoms) from purely observational data (*i.e.*, from wearable devices), allowing us to see which ADLs are the behavioral phenotypes. The advantage of these algorithms is that they automatically control for all possible confounding variables (*i.e.*, covariates) from the observational dataset [31, 32].

With this in mind, we design, implement, and deploy the end-to-end *MSLife* approach with a cohort of 30 MS patients across a one-week home-based IRB-approved clinical pilot study. We leverage GENEActiv smartwatches (equipped with accelerometers and light sensors) to continuously, passively monitor ADLs; and clinical behavioral instruments to collect MS symptoms — thus yielding 5040 total hours of data. After extracting features from the raw sensor data, we design a graph-based statistical analysis framework that leverages causal discovery (Fast Greedy Equivalence Search) and causal inference (Propensity Score Matching) algorithms to discover and quantify the relationships between MS symptoms and ADLs. We also incorporate individuals' demographic information (*e.g.*, genders, ages, and medical conditions) as confounding factors in the graph-based statistical analysis.

Fulfilling the task of digitally behaviorally phenotyping MS symptoms, our pilot study discovers a graph including 102 relationships among MS symptoms, ADLs, and demographics. These include: later rise times are related to increased depression, history of caffeine consumption is associated with lower fatigue levels, and higher relative levels of moderate physical activity are connected to decreased sleep quality and time. We assess the validity of these relationships in two folds: (1) We corroborate our graph-based analysis with traditional statistical metrics, namely the Pearson correlation coefficient (r) and the p-value. This traditional analysis confirms that the 102 relationships from the graph are indeed statistically significant, with $|r|_{avg} = 0.461$ and $p_{avg} = 9.1 * 10^{-4}$. (2) Since the purpose of finding digital behavioral phenotypes is to eventually track symptom development (*i.e.*, disease progression), we leverage the digital behavioral phenotyping relationships to identify MS symptoms in a retrospective analysis. This is greatly effective, significantly outperforming baseline machine learning methods by over 10% in classifying whether or not a patient has a particular MS symptom (*e.g.*, depression, fatigue, poor sleep quality), according to the standard classification performance metrics commonly used for mobile health: accuracy (75.6% vs 64.0%), precision (71.8% vs 61.5%), and recall (76.6% vs 64.9%).

In summary, our contribution is three-fold:

- We design a novel, graph-based approach to digitally behaviorally phenotyping MS symptoms using unobtrusive wearable devices in daily life. The advantage of our graph-based approach is that it automatically controls for the effects of covariates. Furthermore, it is generalizable to studying the digital behavioral phenotypes of any chronic, multi-symptomatic disease.
- We discover a graph of digital behavioral phenotypes of MS symptoms including 102 relationships (*e.g.*, later rise times are associated with increased depression, history of caffeine consumption is connected with lower fatigue levels, and higher relative levels of moderate physical activity are related to decreased sleep quality). These results were derived from deploying our approach with a cohort of 30 MS patients across a one-week home-based IRB-approved clinical pilot study. Traditional statistical models confirm that they are strong and statistically significant.
- We find that our digital behavioral phenotypes greatly improve the tracking of major MS symptoms, based on our retrospective machine learning-based analysis. This paves the way for many important implications, including monitoring MS progression and facilitating precision medicine.

2 BACKGROUND

2.1 Multiple Sclerosis: Medical Perspective

Multiple sclerosis is a neurodegenerative disease that affects 2.5 million people across the world [1, 2]. On a biological level, MS occurs when a person's immune system attacks myelin, which is the fatty tissue that insulates nerve cells; this interrupts the communication between the brain and the rest of the body [1, 9, 33]. The annual MS-related healthcare expense is estimated to be between $8528-$54,244 per patient per year [4].

Symptoms are diverse and unpredictable, including muscle weakness, impaired coordination, paralysis, tremors, dizziness, impaired speech, hearing loss, depression, fatigue, mood swings, and sleep disorders [2, 6–9, 25, 34].

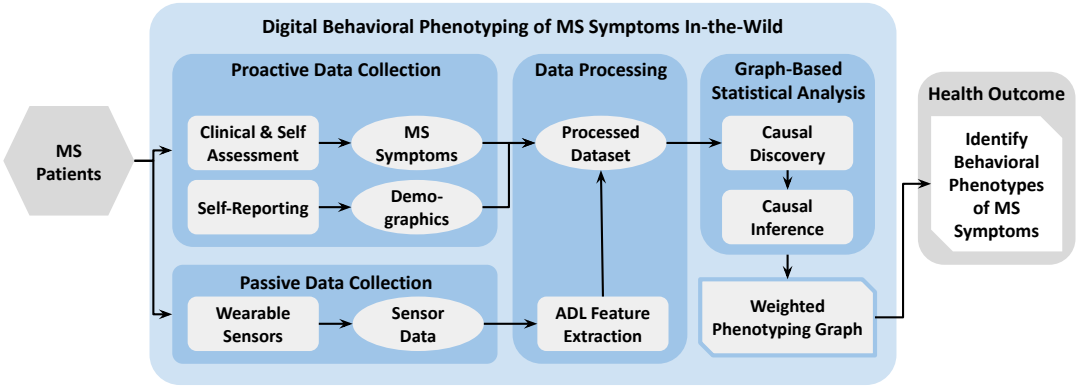

Fig. 1. End-to-end architecture of *MSLife*.

Although a cure for MS itself currently does not exist, symptoms can be managed through behavioral therapies and intervention [7–9]. The specific intervention depends on the symptoms a patient experiences: for instance, a patient experiencing depression may be prescribed antidepressants like desipramine, while a patient experiencing fatigue may be prescribed antifatigue medications like Symmetrel or even undergo magnetic therapy [2, 9, 10]. Thus, to ensure optimal treatment, there arises the need to track these symptoms.

2.2 Digital Phenotyping

Digital phenotyping is the "moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices" [29]. It has two parts: (1) using personal digital devices (*e.g.*, smartwatch) to continuously monitor the phenotype in daily life; (2) leveraging statistical algorithms on the collected data to identify which phenotype variables are actually relevant [30]. Digital phenotyping has been used in various contexts, from mental health [29, 35, 36] to Alzheimer's Disease [37]. For example, the StudentLife study found that certain features extracted from smartphone sensor data could be used to digitally phenotype mental health and educational performance variables in college students [38].

3 SYSTEM DESIGN: ADL MONITORING HARDWARE

3.1 Design Goals

In designing a system to monitor ADLs of MS patients in the wild, we have the following design goals:

- *Unobtrusive:* The system must not interfere with the authenticity of the ADLs, as that would defeat the purpose of monitoring them in the wild (vs. controlled clinical setting). Patients should perform ADLs as if no monitoring system was present.
- *Continuous:* Continuous monitoring allows us to have precise, highly detailed data; which gives us more insights regarding ADLs and their relationships to MS symptoms.
- *Long-Term:* The system must collect data for the entire 168-hour duration of the study (day and night).
- *Valid:* The system must get correct measurements.
- *Reliable:* The system must be consistent in its measurements (low noise/random variation).
- *Exportable Open-Format Data:* We must be able to freely access and analyze the collected data to find relevant digital behavioral phenotypes.
- *Widely Available:* To promote the scalability of our approach (in both this study and future ones).

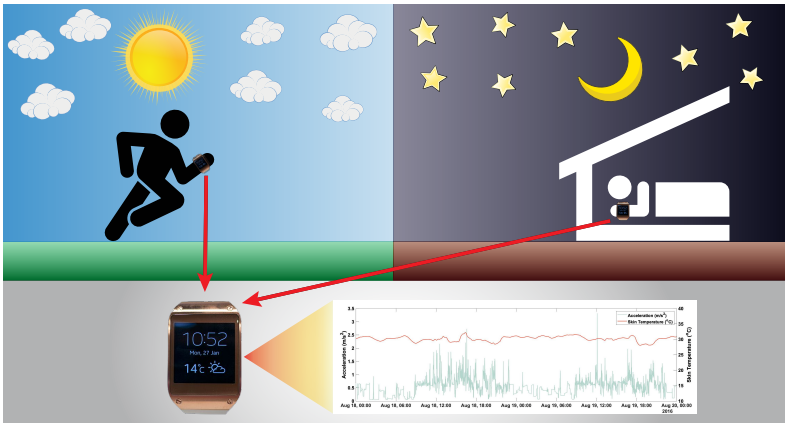

Fig. 2. We leverage wearable devices to continuously, passively monitor ADLs of MS patients in the wild 24/7.

3.2 Wearable Devices

To satisfy these goals, we leverage the GENEActiv smartwatch. It is lightweight (16 g) and compact (43 × 40 × 13 mm^3), thus minimizing interference with ADLs (*unobtrusive*). Its embedded tri-axial accelerometers can monitor ADLs at a sample rate of 100Hz (*continuous*) for 7 uninterrupted days (*long-term*), while its silicon photodiodes can simultaneously sense light exposure at wavelengths 400-1100nm [39, 40]. The accelerometers measure in a range of $+/-8g$, where $g = 9.81\frac{m}{s^2}$ with a resolution of $7.8mg$, where $mg = 10^{-3} \times g$; the light sensors measure in a range of 0-3000 Lux and resolution of 5 Lux [41]. GENEActiv has been shown to be: *valid*, with its acceleration measurements having a Pearson correlation coefficient of $r = 0.97$ ($P < 0.001$) with the true acceleration generated by the multi-axis shaking table in the study by [42]; and *reliable*, with the accelerometer's intra-device coefficient of variation being only 1.8% and the inter-device coefficient of variation among 47 GENEActiv accelerometers being only 2.4% [42]. Similarly, GENEActiv's light sensors can measure light intensity with accuracy $+/-10\%$ [41]. After data collection is finished, data can be exported to `.csv` files, which can be analyzed with tools like MatLab or R (*exportable open-format data*) [43]. Finally, it can be purchased online via a public website (*widely available*) [44].

Each participant wore one GENEActiv smartwatch on one wrist (the other wrist having no smartwatch). We chose this single-sensor, wrist-worn approach for its convenience and low obtrusiveness in the wild, as compared to multi-sensor approaches like [11]; additional smartwatches on other body parts (*e.g.*, ankle, elbow) could potentially interfere with the genuineness of the in the wild data. Furthermore, we believe that a wrist-worn sensor alone can accurately measure ADLs, based on the widespread commercialization and acceptance of wrist-worn sensors that do so (*e.g.*, Fitbit, Apple Watch) [44–46].

4 STUDY DESIGN

4.1 Participants

We collaborate with medical professionals in our Nursing School to carry out a clinical study, which is approved by our Institutional Review Board.

4.1.1 Recruitment: Recruiting of participants took place via physician referral, word of mouth, and the Gateway Chapter of the National Multiple Sclerosis Society.

Table 1. Demographics of participants (N = 30)

Characteristics	Values
Age (years), M (SD)	45.5 (10.4)
Gender (Female), n (%)	28 (88)
Body Mass Index, M (SD)	27.0 (7.5)
Ethnicity (Caucasian), n (%)	23 (72)
Years since diagnosis, M (SD)	11.0 (8.0)
Years since symptom onset, M (SD)	15.8 (8.9)

4.1.2 Inclusion Criteria: The criterion for participants is that they have to have been diagnosed with MS, be between 18 and 70 years of age, and be proficient in English. The lack of a healthy control group is intentional because we wish to track symptoms among people who already have MS. This is in line with similar studies that wish to measure outcomes among MS patients only [47–50].

4.1.3 Exclusion Criteria: People who had one or more of the following conditions are excluded from recruitment: taking interferon drug treatment, currently receiving cancer treatment, experiencing pregnancy or menopause, having severe chronic obstructive pulmonary disease, or having Parkinson's disease (as these conditions can be confounding factors).

4.1.4 Enrolled Cohort: We finally enroll 48 people who are clinically diagnosed as MS patients, with 30 patients ultimately finishing our study (some patients dropped out due to personal reasons). This is comparable to the cohort sizes in similar previous studies involving mobile health technologies (MyTraces [51] had 28 subjects, StudentLife [38] had 48 subjects, CrossCheck [52] had 21 subjects). Regarding the demographics of the cohort, we note that Caucasians and Females make up a majority of the participants; this reflects the fact that MS most commonly occurs among Caucasians and Females [53].

4.2 Study Procedure

4.2.1 Data Collection: Once participants were recruited, they had their symptoms clinically assessed via standard clinical questionnaires to establish the baseline level of symptom severity they normally experience. They also self-reported their demographic characteristics, to account for possible confounders in MS progression [54–57].

Following that, they were given the GENEActiv smartwatch. Over the next 168 hours, the wrist-worn smartwatch continuously, passively monitored their ADLs in regular non-clinical environments; via tri-axial accelerometers (to sense user kinematic motion at sample rate 100Hz) and silicon photodiodes (to sense light exposure at wavelengths 400-1100nm).[40, 44]. Furthermore, participants were also given a symptom severity diary to fill out on a daily basis (once a day, before going to sleep for the night) during the 168-hour data collection period. These daily self-assessments, when considered in tandem with smartwatch ADL data, allow us to better see the patterns in MS symptoms and how they are affected by daily activities since they correspond to the same time period.

In all, the study produced 5040 total hours of data. Our study duration is on par with that of previous studies based on passive mobile sensing for health (Wang et al.'s study lasted for 14 days [58], MyTraces lasted for 20 days [51], a plurality of subjects (43%) in SugarMate had 6-10 days of data collected [59]).

4.2.2 Privacy Considerations: We anonymized the data by giving each participant a randomized ID code. The mapping of ID code to participant identity is securely stored separately from the rest of the study data so that participants cannot be identified from the main dataset.

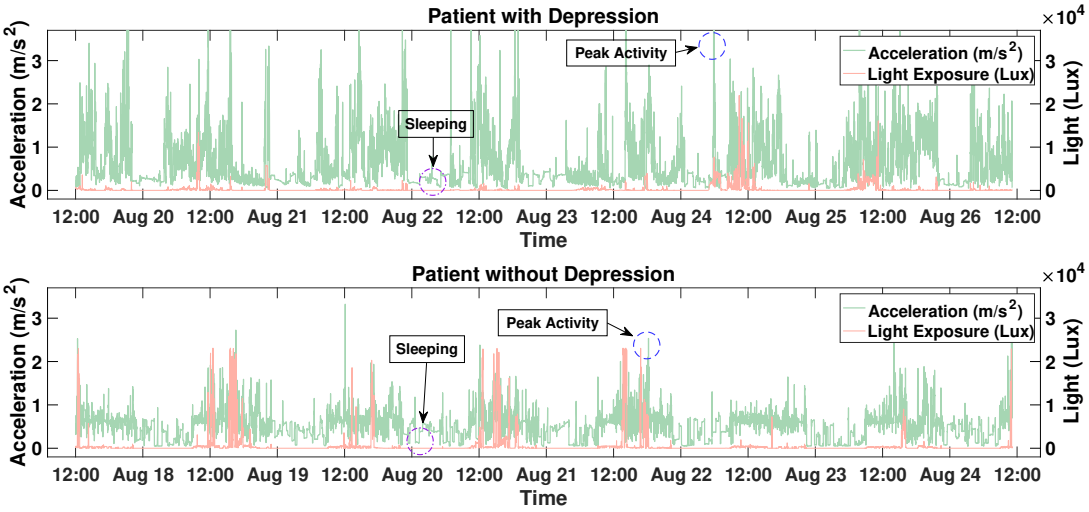

Fig. 3. Raw acceleration and light exposure data from our smartwatches, for a patient with depression and a patient without depression. By analyzing the patterns, we can infer information about ADLs. For instance, the cyclical relative maxima and minima show when MS patients are at peak physical activity and when they are sleeping.

4.2.3 Compliance: The GENEActiv smartwatch automatically monitors when it is being worn or not. On average, patients wore the smartwatches for 23.7 hours per day, indicating a high level of user compliance.

4.2.4 Data Quality: Data quality was high, with missing values only 2.7% of the dataset. We ultimately replaced the missing values with the average values for the corresponding variables, due to the graph-based statistical analysis algorithms' [60, 61] requirements that there be no missing values. We feel that average values are a reasonable choice to replace missing values, as they would have a minimal impact on the distribution of the dataset (as opposed to replacing missing values with, for example, the min or max, which would skew the dataset from its original distribution); this is especially true since there were so few missing values.

5 DATA PROCESSING: ADL FEATURE EXTRACTION

5.1 Feature Extraction Goals

Now that we have collected the raw ADL data via smartwatch sensors, we need to extract relevant features. In doing so, we make sure to account for ADLs from all times of day, *i.e.*, activities done while awake as well as sleep patterns, since prior research suggests that these are all linked to MS symptoms [6, 25–28]. Additionally, for ADLs, we want to measure both their micro characteristics (which are more granular and concern the different ADL measurements day-to-day) and macro characteristics (which are more concerned with summarizing the overall features of ADLs over the entire study period). Examples of micro characteristics are time spent at different intensities of physical activity, day-by-day rise time, day-by-day sleep duration. Examples of macro characteristics are the total number of activity periods, the total number of awakenings, median sleep duration.

5.2 Actigraphy Algorithms

Towards this goal, we leverage the power of actigraphy algorithms, whose ability to extract ADL features has been clinically verified [62–64]. Their input is the raw tri-axial accelerometer data. The output of the actigraphy

algorithms is the extracted ADL features, listed in the "Activities of Daily Life" category of Table 2. We use GENEActiv's implementations of actigraphy algorithms in R, run on a standard desktop CPU [44]; which have been shown to provide reliable extraction of the aforementioned features (*e.g.*, activity levels) [42, 65, 66]. Although the full-length implementation details are outside this paper's scope, we briefly describe the actigraphy algorithms:

5.2.1 Awake Calculations: The main consideration for actigraphic calculations while awake is determining activity intensity levels. These are determined from the raw accelerometer measurements. First, the gravity-adjusted acceleration magnitude is calculated as

$$|a|_{adj} = \sqrt{a_x^2 + a_y^2 + a_z^2} - g, \tag{1}$$

where $g = 9.81 \frac{m}{s^2}$ is the force of gravity [42, 67, 68]. We use the following standard thresholds [42] to classify the intensity of this gravity-adjusted acceleration magnitude, where $mg = 9.81 * 10^{-3} \frac{meters}{second^2}$ and MET is a standard unit of energy expenditure which stands for the metabolic equivalent of task (according to the CDC, "One MET is defined as the energy expenditure for sitting quietly", and is equivalent to "3.5 ml of oxygen uptake per kilogram of body weight per minute") [69, 70]:

- Sedentary: $|a|_{adj} \in [0, 62.8)$ *mg*, corresponding to $[0.0, 1.5)$ MET
- Light: $|a|_{adj} \in [62.8, 112.9)$ *mg*, corresponding to $[1.5, 4.0)$ MET
- Moderate: $|a|_{adj} \in [112.9, 407.1)$ *mg*, corresponding to $[4.0, 7.0)$ MET
- Vigorous: $|a|_{adj} \in [407.1, \infty)$ *mg*, corresponding to $[7.0, \infty)$ MET

In terms of the reliability of these thresholds, prior work uses the metrics of $sensitivity = \frac{TP}{TP+FN}$ (also known as recall) and $specificity = \frac{TN}{TN+FP}$ to quantify, where TP, TN, FP, FN respectively stand for true positive, true negative, false positive, false negative. For the sedentary-light threshold (62.8 mg), sensitivity and specificity are 98% and 96%; for the light-moderate threshold (112.9 mg), sensitivity and specificity are 98% and 64%; for the moderate-vigorous threshold, sensitivity and specificity are 78% and 98% [42].

5.2.2 Asleep Calculations: For detecting sleep, actigraphy algorithms use the (previously mentioned) accelerometer data to quantify a person's level of activity; the mean, standard deviation, duration, and the number of activity events are then considered as parameters to a threshold-based calculation, for which a value above 0 is considered as asleep, and a value below 0 is considered as awake [71, 72]. Sleep detection with GENEActiv has been experimentally determined to be within 16.9% of the true sleep duration as measured by clinical "gold-standard" methods [65].

6 DATASET

We combine the extracted ADL features with MS symptoms and demographics to make a three-category dataset. This dataset, encompassing 5040 hours, has 57 variables (Tab. 2).

6.1 MS Symptoms

We assess symptoms that significantly affect the lives of MS patients, including (1) depression, (2) fatigue, (3) sleep quality, (4) mood, and (5) functional disability [5, 6, 12, 25, 34, 73–79].

6.1.1 Depression: Overall level of *Depression* is reported on the Center for Epidemiological Studies Depression Scale (CES-D), ranging from 0 (lowest) to 60 (highest). It consists of 20 statements (*e.g.*, "I was bothered by things that don't usually bother me."); where the patient is asked to rate how often they "felt this way during the past week" on a scale of 0-3: "rarely or none of the time" (0), "some or a little of the time" (1), "occasionally or a moderate amount of time" (2), or "most or all of the time"(3). The final score is the sum of these measurements [80].

Table 2. Variables in *MSLife* dataset.

Information Category	Subcategory	Acquired Variables
MS Symptoms	Depression	Overall Depression
	Fatigue	Overall Fatigue
		Diary Fatigue
	Sleep Quality	Overall Sleep Quality
		Diary Sleep Quality
		Restorative Sleep Levels
		Circadian Rhythm
	Mood	Diary Mood
	Functional Disability	Daytime Function
		Disability
Activities of Daily Life (ADLs)	Daytime ADLs (Awake)	Time spent in (sedentary, light, moderate, vigorous) activity
		Energy used in (sedentary, light, moderate, vigorous) activity
		% of activity at (sedentary, light, moderate, vigorous) intensity
		Number of activity periods
		Activity duration
	Nighttime ADLs (Asleep)	Bedtime
		Rise time
		Wake duration
		Sleep duration
		Time spent in bed
		Number of awakenings
		Sleep efficiency
		Light exposure
	Miscellaneous	Time without watch
		Day of week
Demographics	Basic	Age
		Gender
		Race
	Medical	BMI
		Year MS start
		Year MS diagnosed
		General health
		Smoking history
		Alcohol history
		Caffeine history
		Exercise history
	Social	Marital status
		Health insurance
		Employment
		Annual income

6.1.2 Fatigue: We measure *Overall fatigue* using the clinical standard Fatigue Severity Scale (FSS), which ranges from 9 (lowest) to 63 (highest). It consists of 9 statements, such as "I am easily fatigued", where the patient ranks their agreement on a scale of 1-7, where 1 is strong disagreement, and 7 is strong agreement. The total score is the sum of these ratings [81]. We also have patients record daily self-assessed fatigue on a 0-5 scale, giving

us *Diary fatigue*. Thus, *Overall fatigue* is the baseline level of fatigue, while *Diary fatigue* allows us to see how variations in ADLs affect fatigue from day to day.

Note: The reason why we do not take daily measurements of depression (while we do for other symptoms like fatigue), is that depression is defined as a persistent, long-term state lasting for at least one week [80]. It is distinct from feeling sad for one day due to a specific temporary event [82]. On the other hand, it is common to have, for example, fatigue one day, then not have it the next; it is also common to have fatigue as a persistent symptom.

6.1.3 Sleep: Similarly, we measure *Overall sleep quality* with the Pittsburgh Sleep Quality Index (PSQI) on a scale of 0 (worst) to 21 (best) to establish a baseline sleep quality. (Technically, PSQI was originally on a decreasing scale of 0 (best) - 21 (worst); we invert it to be on an increasing scale of 0 (worst) - 21 (best) so that higher ratings correspond to higher sleep quality when we report results.) PSQI is calculated by summing 7 component scores, each of which ranks a specific aspect of sleep quality (*e.g.*, "Sleep latency", "Use of sleeping medication") on a scale of 0 (worst) - 3 (best) [83]. To capture day-to-day changes in sleep quality that may be associated with ADLs, we have patients record subjectively self-assessed sleep quality on a scale of 0 (worst) to 5 (best) in diaries on a daily basis, giving *Diary sleep quality*. Similarly, we wish to know how restorative (restful, relaxing) sleep is from day-to-day, so patients record this on a 0 (not restorative) to 5 (very restorative) scale; thus, we obtain *Restorative sleep levels*. Next, we assess *Circadian rhythm* on a scale of 1 to 3, with 1 meaning "morning person", 2 meaning neither morning nor night person, and 3 meaning "night person".

6.1.4 Mood: Regarding mood, patients self-record *Diary mood* on a scale of 0 (worst, very ill-tempered, very irritable) to 5 (best, very well-tempered, very easygoing). This is done once every night, before they go to sleep, as a reflection on their mood throughout that day. We also take the median of these measurements, to get an impression of the patient's overall mood, in addition to seeing how the changes in daily activities and events affect mood.

6.1.5 Functional Disability: We assess *Daytime functionality* via the Functional Outcomes of Sleep Questionnaire-10 (FOSQ-10), which ranges from 5-20, where higher ratings reflect superior functionality [84]. It has 10 questions (*e.g.*, "Do you have difficulty performing employed or volunteer work because you are sleepy or tired?") ranked on a scale of 1-4, where 1 indicates extreme difficulty and 4 indicates no difficulty in completing an activity due to sleepiness. The total score is the sum of the 10 question scores, divided by 2 [85]. Also, we measure *Disability* on the clinical standard Kurtzke Expanded Disability Status Scale (EDSS), on a scale of 0 (no disability, neurologically normal) to 10 (death from MS) [76].

6.2 Digital Behavioral Phenotype (Activities of Daily Living)

6.2.1 Daytime ADLs (Awake): We take a plethora of metrics at multiple physical activity (where physical activity means any ADL) intensity levels (previously defined): sedentary, light, moderate, vigorous. *Time spent in activity*, measured in minutes, is the total time spent at a particular activity intensity in a day. *Energy expenditure* is the total energy used in a particular activity; it is measured in $MET * min$. *Relative amount of activity at intensity* is the percentage of a person's total activity that was done at a particular intensity (thus, sedentary, light, moderate, and vigorous levels sum to 100%). We take this metric to see the distribution of activity intensity. *Number of activity periods* measures the number of periods with distinct activity intensity (*e.g.*, starting with sedentary activity, then switching to vigorous activity, then engaging in light activity; constitutes three periods), and *Duration of activity period* measures the median duration of the aforementioned activity periods. These measurements are all calculated once per day, thus making for a total of seven of these measurements per subject.

6.2.2 Nighttime ADLs (Asleep): For sleep patterns, *Bedtime* and *Rise time* respectively measure the time a person went to bed to sleep for the night and got out of bed in the morning. We take both daily and median

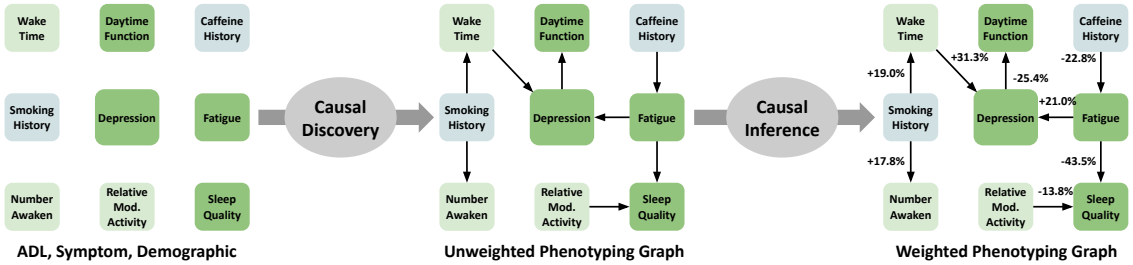

Fig. 4. Variables are input to causal discovery, which outputs an unweighted phenotyping graph, where directed edges show relationships between variables. The unweighted graph is input to causal inference, which outputs a weighted graph, with ATEs as edge weights. [Positive, negative] ATE means an increase in the variable at the tail is associated with an [increase, decrease] in the variable at arrowhead.

measurements to get perspective on the effect of variations in ADLs on different days, as well as general waking and sleeping patterns. Similarly, *Wake duration* and *Sleep duration* measure how long a person was awake and asleep, respectively; these are taken daily. However, there is a distinction between sleeping and simply lying in bed, so we also measure *Time spent in bed* (not necessarily sleeping) on a daily basis. *Number of awakenings* is the total number of times a person awoke from sleep over the course of the one-week study. *Sleep efficiency* is the ratio of the time spent sleeping to the time spent in bed, calculated daily. We measure *Total light exposure* (in *lux*) daily via the smartwatch light sensor, to find out if light influences activity levels or sleep patterns, as suggested in previous studies [86, 87].

6.3 Demographics

6.3.1 Basic: We record *Age*, *Gender*, and *Race*. A summary of these variables is listed in Table 1. The reason that Caucasians and Females make up a majority of the participants is that MS most commonly occurs among Caucasians and Females [53].

6.3.2 Medical: We also take features commonly collected in healthcare — *BMI*, *Year MS started* (year effects of MS first happened), *Year of MS diagnosis*, *Smoking history* (yes or no), *Alcohol history* (scale of 0 (none) - 2 (frequent drinker)), *Caffeine history* (consumes caffeine or doesn't consume caffeine), *Exercise history* (active or not active), levels of *General health* (scale of 1 (most unhealthy) - 5 (most healthy)).

6.3.3 Social: Additionally, social factors are associated with different outcomes of disease management and symptom progression [56, 57, 88, 89]. As such, we record *Marital status* (married or unmarried), *Health insurance* (in Bronze, Silver, Gold, or Platinum tier plan, as defined by U.S. Government [90]), *Annual income levels* (scale of 1 (bottom 20% of earners) - 5 (top 20% of earners)), and *Employment status* (scale of 1 (unemployed) - 5 (working 40+ hrs/week)).

7 GRAPH-BASED STATISTICAL ANALYSIS

Now that we have the necessary data, we can proceed to identify the relevant digital behavioral phenotypes of MS symptoms in daily life. To satisfy this goal, we design a two-step graph-based statistical analysis framework.

7.1 Causal Discovery

7.1.1 Problem Definition: The first step is to discern the relationships among the digital behavioral phenotypes (ADLs) and MS symptoms in our dataset. For this task, we leverage causal discovery algorithms, whose ability to

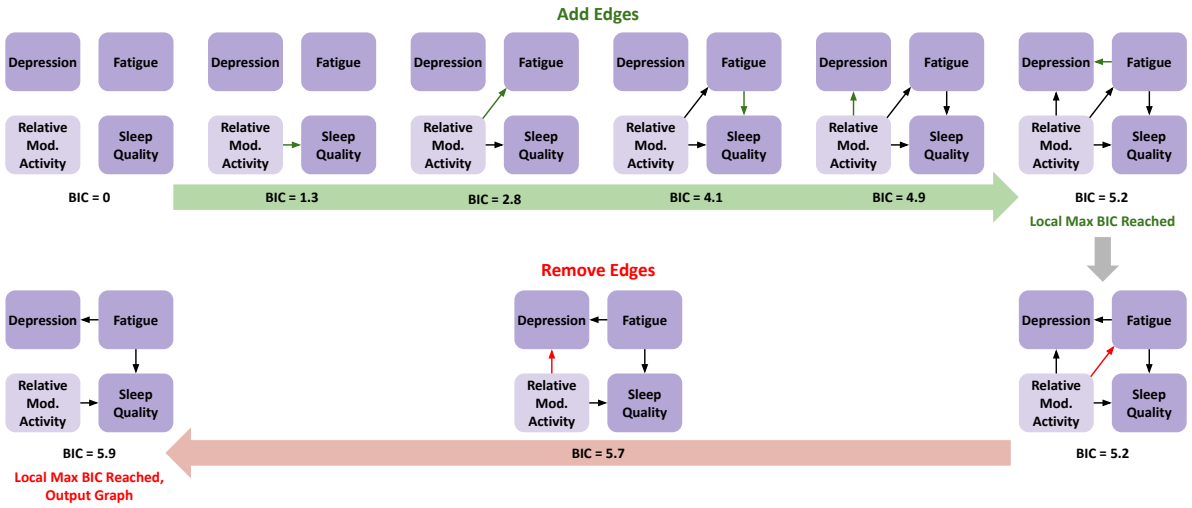

Fig. 5. Toy example for Fast Greedy Equivalence Search (causal discovery), illustrating the two-step iterative maximization process with respect to BIC.

discover relationships (outputted as a graph) from observational data has been verified both in previous healthcare applications [91, 92] and through mathematical analysis [93, 94]. Their advantage is that they account for the covariance/confounding relationships among variables, as shown by the outputted graph's edges.

7.1.2 Fast Greedy Equivalence Search: We use the fGES (fast greedy equivalence search) causal discovery algorithm, which has been successful in previous health applications and is mathematically optimal. It takes a dataset as input and outputs the dataset's underlying most probable causal relationships as a causal structure graph (where directed edge $A \rightarrow B$ means A "causes" changes in B) [1] [91, 93–96].

Bayesian Information Criterion: In our system, fGES aims to find the causal structure graph with the highest Bayesian Information Criterion (BIC) score [97], a widely-used metric that approximates the likelihood that our collected dataset is true given a particular causal structure graph of the relationships among ADLs, MS symptoms, and demographics [95]. BIC is defined: [93, 95, 97–99]:

$$BIC = 2 * ln(P(data|\hat{\theta}, \mathcal{G})) - k * ln(n), \tag{2}$$

with the variables defined below [93, 95]:

- $data$ is the dataset of 57 features from Table 2. It has a total of (# features) $\times$ (# subjects) $\times$ $\frac{(\#\text{ observations})}{\text{subject}}$ = $57 \times 30 \times 7 = 11970$ entries.

[1] **Causality:** As noted in [51], distinguishing cause from effect is an open question in such analyses. Although causal discovery (and causal inference) *suggest* highly likely causal relationships, the directionality (*e.g.*, $A \rightarrow B$) of these relationships is not definitive. To determine if A truly causes B would take a controlled clinical experiment, which we do not conduct, due to our previously stated concerns regarding the genuineness of the ADLs. Nonetheless, we still find these algorithms to be highly useful in our digital behavioral phenotyping relationship analysis, due to their built-in control for covariates/confounding variables. We simply do not make claims of causality from the output; rather, we claim correlation.

- $\mathcal{G}$ is a partially built causal Bayesian network (aka possible causal structure graph), with one node for each variable (including ADLs, MS symptoms, and demographics; making for a total of 57 nodes) from $data$ and edges representing associations between variables.
- n is the number of observations (*i.e.*, (# subjects) $\times$ (# days) $= 30 \times 7 = 210$) in $data$.
- P, the (marginal) likelihood function, is the probability that $data$ is valid given the causal structure graph $\mathcal{G}$ and the corresponding parameters $\hat{\theta}$ (*i.e.*, maximum likelihood estimators) to $\mathcal{G}$ which maximize P. More specifically, P is the integral with respect to $\mathcal{G}$'s parameters (*i.e.*, $d\theta$) of a joint conditional probability distribution among all 57 variables in $data$, making P the integrated product ($\prod$) of the conditional probabilities that each of the 57 variables would have the $data$ values that were actually observed (*i.e.*, "mini"-likelihood functions) given a set of parent nodes (*i.e.*, phenotypes) as defined by the graph edges on $\mathcal{G}$. In this case, the conditional probability (*i.e.*, "mini"-likelihood) function for each (random) variable is assumed to be Gaussian (*i.e.*, normal). (For more information, please refer to Equations (1, 3, 4) in [93]).
- $\hat{\theta}$ consists of the set of maximum likelihood estimators, *i.e.*, the parameters which specify each of the variables' conditional Gaussian distributions such that P is maximized. Since there is a conditional probability distribution for each variable, and the conditional probability distribution is known to be Gaussian; for one variable, the maximum likelihood estimator would consist of the mean, the variance, and the covariances with its parents (*i.e.*, phenotypes) on $\mathcal{G}$.
- Thus, k, the size of $\hat{\theta}$, would be: (# means) + (# variances) + (# covariances) = (# variables) + (# variables) + (# edges on $\mathcal{G}$) $= 57 + 57 +$ (# edges on $\mathcal{G}$) $= 104 +$ (# phenotyping relationships on $\mathcal{G}$).

The first term in BIC is meant to maximize the likelihood that the data is true given the graph of digital behavioral phenotyping relationships (*i.e.*, good fit, accurate digital behavioral phenotypes). The second term penalizes complex models in favor of simpler models (*i.e.*, fewer edges) [97].

Iterative Maximization Process: To maximize the BIC score in our application, fGES starts with an empty graph that contains all the nodes representing the ADLs, MS symptoms, and demographics; but no edges representing relationships. It then finds the causal structure graph with maximum BIC in two steps. First, it iteratively adds edges to the graph by considering every possible directed edge (*i.e.*, digital behavioral phenotyping relationship), and using a greedy policy (with respect to the BIC score) to choose the next edge to add. At some iteration, the BIC score reaches a local maximum, so adding edges can no longer increase it. Now, entering the second step, the algorithm iteratively deletes edges (using a similar greedy policy) to further maximize the BIC score and reach another local maximum [93–95]. The output of fGES is the resultant graph, with directed edges representing relationships among ADLs, MS symptoms, and demographics.

Note on Dimensionality: We also note that finding the causal structure graph could be a computationally expensive problem with a large search space (the total number of possible edges is $|E| = O(|V|^2)$, where V represents the set of all nodes, making the total number of possible graphs $O(2^{|E|}) = O(2^{|V|^2})$. To combat this, we place a reasonable limit that the degree of the graph cannot exceed 5 (*i.e.*, each node has at most 5 edges). Furthermore, the greedy policy w.r.t. edges in the iterative maximization process is actually already designed to reduce the search space of possible causal structure graphs from an exponentially large to a polynomially large space [93].

7.1.3 Implementation: We use the Tetrad library's Java implementation of fGES, run on a standard desktop CPU [60]. As for hyperparameters, we input a prior knowledge graph that indicates trivial forbidden causal relationships (*e.g.*, daily-life activities like vigorous exercise cannot cause demographic characteristics like age).

7.1.4 Causal Discovery in MSLife: For *input* to fGES, we provide the processed ADL, symptom, and demographic data. As *output*, fGES gives the optimal causal structure graph showing the digital behavioral phenotyping relationships among MS symptoms, ADLs, and demographics (where edge $A \rightarrow B$ means that A is related to B).

7.2 Causal Inference

7.2.1 Problem Definition: Knowing the existence of digital behavioral phenotyping relationships of MS symptoms is merely the first step — the second, equally important step is to know their strengths. When monitoring MS symptoms via their digital behavioral phenotypes (*i.e.*, ADLs), we should focus our monitoring on those phenotypes that are known to have the strongest relationships, as those are the best indicators of symptom development. Thus, to quantify the strength of the relationships among ADLs, MS symptoms, and confounding demographic factors; we leverage causal inference algorithms, whose correctness for this task has been mathematically and empirically verified [32].

7.2.2 Propensity Score Matching: We use a standard causal inference algorithm called Propensity Score Matching, which calculates the change in an outcome variable Y that would result from changing a treatment (*i.e.*, independent) variable W, all other factors held constant. This change in the outcome variable is known as the average treatment effect (ATE), which is commonly used to measure the strength of relationships [32, 61].

Challenges in Calculating ATE: When calculating ATE, one of the most important considerations is to find a means by which to control for the effects of the confounding variables (a.k.a. covariates). Without controlling, confounding variables could bring spurious effects into ATE.

7.2.3 Definition of Propensity Score: Thus, to account for the influence of the confounding variables, propensity score matching uses the propensity score to quantify the closeness of the values of the confounding variables. Propensity score, $e(x)$, is defined as the conditional probability that a subject received a particular treatment given that a certain set of values for the confounding variables was observed. Mathematically:

$$e(x) = P(W = 1|X = x), \tag{3}$$

with the variables defined as follows [32, 61, 100]:

- W represents a particular treatment (*e.g.*, moderate physical activity), with $W = 1$ indicating that the subject received it, and $W = 0$ indicating that the subject did not receive it, as treatments in causal inference are typically assumed to be binary. To match this standard assumption, for each "treatment" variable, we adopt its mean value across all patients as a threshold, assigning values above that threshold as $W = 1$ (high levels) and values below as $W = 0$ (low levels).
- $X = x$ indicates that the confounding variables (*e.g.*, age, smoking history) had the values in set x.
- P is the conditional probability function, calculated via a logistic regression on the possible values of W as the range (*i.e.*, $[0, 1]$) and the possible values of X as the domain.

Need for Propensity Score: Without the propensity score, controlling for the effects of the covariates/confounders can become a complicated high-dimensional problem, since we would have to match similar values for every single one of the covariates. The propensity score reduces this high-dimensional problem to a problem in a scalar space, while still accounting for all the covariates; since it is simply a multivariable-input scalar-output logistic function with range (0, 1).

7.2.4 Estimation of Outcome under Treatment: Now, in order to determine the effect which treatment W has on the outcome variable Y, we wish to estimate the value $\hat{Y}_i(1)$ which the outcome variable Y in subject i would have taken on if subject i had received treatment W, and then compare it to the value of the outcome variable $\hat{Y}_i(0)$ had subject i not received treatment W.

Without loss of generality, we assume that subject i did receive treatment W, so $W_i = 1$ (this argument can be similarly applied for the no-treatment case $W_i = 0$). Thus, we know that $\hat{Y}_i(1) = Y_i$. The question is now, how do we estimate $\hat{Y}_i(0)$?

To achieve this, we create a set of "neighbors" which are close to subject i in all respects (*i.e.*, covariate values), except that $W_{neighbor} = 0$. Using this set of neighbors as a synthetic control, we can estimate the value the outcome variable would have taken on, had subject i not received treatment W. (The exact number of neighbors varies from implementation to implementation; in our system, we set it to be 1.)

To create the aforementioned set of "neighbors", we match subjects from opposite treatment groups with the closest propensity scores. When propensity scores are close, covariate values are generally close as well. Thus, the propensity score is used as a scalar distance metric for covariate values, such that we can match subjects by the closeness of covariate values without having to do high-dimensional calculations [100].

So, for each subject i in the dataset for which $W_i = 1$, we calculate a set N_i which contains the subject(s) j that were the closest to subject i in propensity score and with $W_j = 0$. Now, for each subject i, given this set N_i of "neighbor" subjects j for which $W_j = 0$, we can estimate the value $\hat{Y}_i(w)$ which the outcome variable Y in subject i would have taken on if it had not received treatment W as follows [32]:

$$\hat{Y}_i(0) = \frac{1}{N(i).length} \sum_{j \in N_i} Y_j; \hat{Y}_i(1) = Y_i. \tag{4}$$

Similarly (WLOG), for the case where $W_i = 0$:

$$\hat{Y}_i(1) = \frac{1}{N(i).length} \sum_{j \in N_i} Y_j; \hat{Y}_i(0) = Y_i. \tag{5}$$

7.2.5 Calculation of Average Treatment Effect: Now, we can present the formula for average treatment effect, where S is the set of all the subjects:

$$ATE = \frac{1}{S.length} \sum_{i \in S} (\hat{Y}_i(1) - \hat{Y}_i(0)). \tag{6}$$

7.2.6 Implementation: We use the Python implementation of propensity score matching from the DoWhy library, run on a standard desktop CPU [61]. To match the assumption in causal inference that treatment (independent) variables are binary, we adopt a threshold-based binarization algorithm on the selected treatment variables. For each variable, its mean value across all subjects is used as the binary threshold. Observations below that threshold are assigned a value of 0 to indicate low levels of that variable, while observations above that threshold are assigned a value of 1 to indicate high levels. As for hyperparameters, we set the number of neighbors $N(i).length$ to be 1.

7.2.7 Causal Inference in MSLife: As *input* to Propensity Score Matching, we provide our processed dataset and the causal structure graph. We run propensity score matching on each phenotyping pair $W \rightarrow Y$ in the causal structure graph, with W as the treatment variable (*e.g.*, relative levels of moderate physical activity), and Y as the outcome variable (*e.g.*, sleep quality); other variables which are covariates of W are treated as the confounding variables X. As *output*, we receive the ATEs for each phenotyping pair in the graph, allowing us to quantify the strength of the relationships among ADLs, MS symptoms, and demographic confounders.

7.3 Weighted Graph of Digital Behavioral Phenotyping Relationships

As the final output of our graph-based statistical analysis, we produce a weighted graph of digital behavioral phenotyping relationships. This is created by assigning the ATEs (from causal inference) as the weights of the

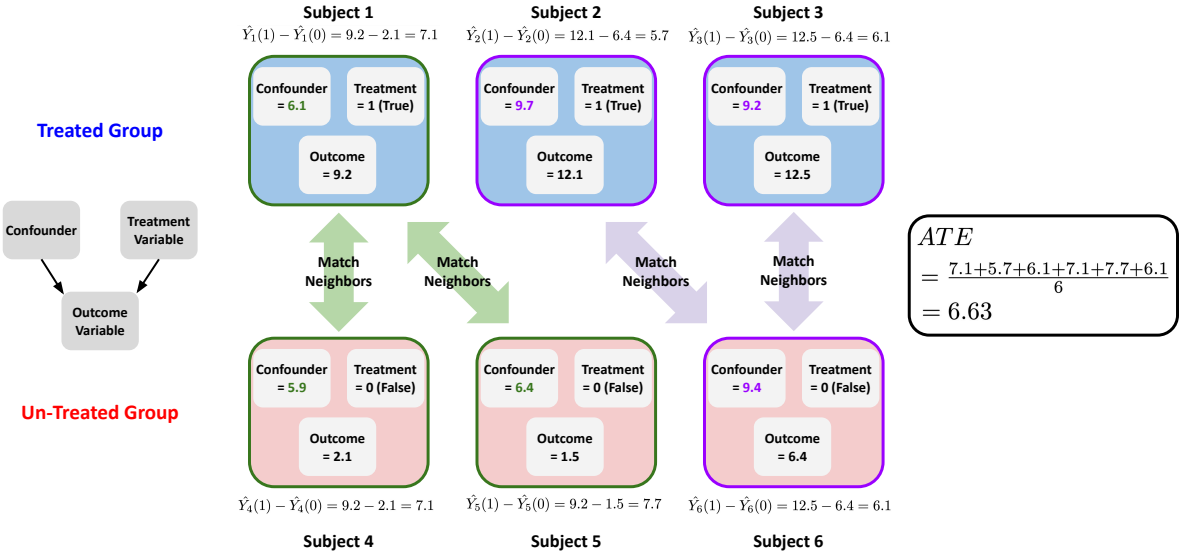

Fig. 6. Generic toy example for Propensity Score Matching (causal inference). It matches subjects from opposite treatment groups which have similar covariate values, as determined by propensity score. It then compares the values of the outcome variable in these matched "neighbors" to estimate the treatment effect.

edges in the original graph (from causal discovery). [2] [3] This shows both the existence and strength of the digital behavioral phenotyping relationships among MS symptoms, ADLs, and demographics.

We highlight that our graph-based approach accounts and controls for the simultaneous interactions among confounding, treatment, and outcome variables. Thus, the outputted relationships between pairs of adjacently linked variables (*e.g.*, $A \longrightarrow B$) in the graph are free from the influence of additional confounding factors.

8 FINDINGS: PHENOTYPES OF MS SYMPTOMS

We present the results of our graph-based statistical analysis. After applying the above algorithms on our collected data, we obtain a weighted graph describing the digital behavioral phenotypes of MS symptoms; with 57 nodes and 102 directed edges, having corresponding average treatment effects (ATE) as edge weights. Because our graph-based approach controls for confounders, the relationships presented in the following subsections represent the pure relationships between the treatment and outcome variables only.

Medically significant relationships, with their ATEs, are listed in Table 3. (We normalized the values for ATE to represent the percent change in the value of the outcome variable, to enable uniform comparison across

[2] **ATE as Edge Weight:** We note that the covariances computed as part of the maximum likelihood estimators ($\hat{\theta}$) of the causal discovery step may seem like they could be used as edge weights. Yet, ATE and the covariances are actually computed with two different (albeit adjacent) aims. As detailed in this section, ATE is a widely used metric of relationship strength that controls for the effects of variables that are *already* known to be confounders. In contrast, the covariances in $\hat{\theta}$ are used to find out *what* the confounding variables are in the first place.

[3] **Bayesian vs. Frequentist Approaches:** We are aware that at first glance, it may seem that integrating the fGES algorithm for causal discovery (based on the Bayesian Information Criterion) with the Propensity Score Matching algorithm for causal inference (based on a more frequentist interpretation) to create our weighted graph mixes the perpetually at-odds Bayesian and frequentist interpretations of statistics. Actually, BIC, while having Bayesian in its name, is not a strictly Bayesian approach. As noted by Schwarz (its formulator), it can actually also be applied outside a Bayesian paradigm, since Schwarz specifically designed it to be independent of the prior [97]. This is what enables the accepted practice of combining causal discovery and causal inference algorithms in one integrated framework [61].

Table 3. Weighted edges from the graph of phenotyping relationships, presented as a table for readability. We use fGES to discover the existence of edges (which indicate phenotyping relationships). We use propensity score matching to calculate ATEs.

Treatment Variable →	**Outcome Variable**	**ATE (%)**
Caffeine history	Sleep quality	-80.1
Sleep Quality	Depression	-69.7
Year MS diagnosed	Fatigue	-58.8
Bed time	Energy expenditure in vigourous activity	-54.4
Fatigue	Sleep quality	-43.5
Disability	Time spent in moderate activity	-42.7
General health	Fatigue	-33.6
Time spent in bed	Going to bed time	-32.8
Type of circadian rhythm	Restorative sleep	-29.8
Depression	Level of daytime function	-25.4
Relative amount of moderate activity	Time spent sleeping	-24.9
Caffeine history	Fatigue	-22.8
Time spent in bed	Sleep efficiency	-21.8
Time spent in activity	Number of awakenings	-21.1
Age	Restorative sleep	-14.5
Relative amount of moderate activity	Sleep quality	-13.8
Annual income	Rise (wake) time	-13.1
Time spent in bed	Energy expenditure in light activity	-13.1
Year MS start	Rise (wake) time	-12.4
Going to bed time	Number of awakenings	-6.3
Rise (wake) time	Mood	-6.1
Employment status	Sleep quality	+8.1
Caffeine history	Mood	+11.7
Relative amount of sedentary activity	Time spent sleeping	+13.4
General health	Mood	+15.6
Going to bed time	Sleep efficiency	+15.6
Female Gender	Going to bed time	+17.4
Smoking history	Number of awakenings	+17.8
Smoking history	Rise (wake) time	+19.0
Fatigue	Depression	+21.0
Race (Non-Caucasian)	Fatigue	+26.1
Rise (wake) time	Depression	+31.3
Unemployment	Depression	+38.0
Restorative sleep	Mood	+38.1
Duration of awakeness	Time spent in light physical activity	+46.7

relationships.) We especially discuss our findings related to three of the most important MS symptoms, which are Depression, Fatigue, and Sleep Quality [5, 6, 6, 12, 74].

8.0.1 Phenotype Hypotheses: If our approach is valid, our results should be mostly congruent with prior medical knowledge (with the expectation that we may discover some previously unknown relationships). Thus, we briefly list a few obvious relationships which we expected the graph to contain: a positive association between fatigue and depression [101, 102], a negative association between caffeine consumption and sleep quality [103], a negative

association between caffeine consumption and fatigue [103]. (As detailed in the following subsections, these all hold true in our analysis.)

8.1 Phenotypes of Depression

8.1.1 ADLs: We find it notable that later rise times (waking up times) are associated with higher levels of depression. Indeed, previous studies show that oversleeping can often occur in tandem with depression [104, 105].

8.1.2 Demographics: The data showed that unemployment is associated with increased depression. This is in accordance with previously established links between unemployment and depression [106].

8.1.3 Symptoms: Higher fatigue is connected with an increase in depression, which makes sense, as lack of energy is often linked with people feeling depressed [101, 102]. It is also notable that depression is linked with decreased levels of daytime function. This is reasonable, as when people are depressed, they have less motivation to perform day-to-day activities [101]. Finally, we find that better sleep quality is associated with lower depression. This makes sense, considering the well-documented links between depression and sleep quality [107].

8.2 Phenotypes of Fatigue

8.2.1 ADLs: While we did not find any direct edges on the phenotyping graph between fatigue and ADLs, we note that by virtue of fatigue's other edges, it is still related to ADLs. For instance, later rise times are associated with higher fatigue, as later rise times are associated with increased depression, which is associated with higher fatigue.

Note: The more graph edges we consider (*e.g.*, 2, 3, 4 edges; as opposed to just 1 edge distance), the more such "indirect" relationships we can find, for fatigue or any other symptom. *e.g.*, As seen in Table 9, when considering a distance of 3 edges, there are 9 ADL phenotypes of fatigue, 13 for depression, and 17 for sleep quality. Thus, the absence of a direct edge from an ADL to a symptom in the graph does *not* imply that they are unrelated. For simplicity and brevity's sake, most of the relationships detailed in this section will be 1-edge relationships.

8.2.2 Demographics: We find that patients who are diagnosed with MS later reported lower levels of fatigue. This is reasonable, as previous work has shown a similar association between time living with MS and fatigue [108]. We find that high levels of general health are related to a reduction in fatigue levels. We also find that non-Caucasians experienced higher levels of fatigue than Caucasians. This may possibly be due to genetic variations and warrants further investigation. Additionally, we find that a history of caffeine consumption is associated with lower levels of fatigue, in accordance with the knowledge that caffeine is a stimulant [103].

8.2.3 Symptoms: As noted in the above section, there is a positive relationship between depression and fatigue (when one increases, the other increases). Furthermore, high fatigue is associated with a decrease in sleep quality. This could be because when people feel too fatigued, normal amounts of sleep are not sufficient to restore their full energy.

8.3 Phenotypes of Sleep

8.3.1 ADLs: We find that higher relative levels of moderate physical activity (as compared to other intensities) are linked to less time spent sleeping. A possible explanation for that is that moderate activity would not drain an MS patient's energy as much as vigorous activity, so they don't feel as tired and thus sleep less. Additionally, higher relative levels of moderate physical activity are related to a decrease in sleep quality. Again, this may be because moderate activity does not provide the high-intensity exercise needed to improve sleep quality [28, 109, 110]. Moreover, we note that spending more time in the bed is linked with decreased sleep efficiency. This is perhaps because human circadian rhythms result in natural tendencies to sleep and wake for certain amounts of time

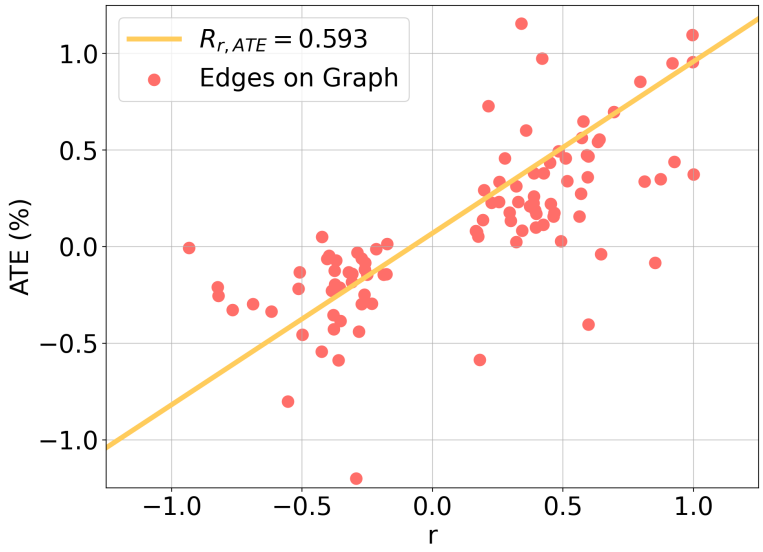

Fig. 7. Pearson correlation coefficient vs. average treatment effect, for variables that have an edge between them on the graph of phenotyping relationships. The values follow a strong positive linear relationship ($R_{r,ATE} = 0.593$), indicating that our causal inference submodule correctly infers magnitudes of phenotyping relationships.

[111], so after too much time spent in bed, the sleep is no longer deep, thus decreasing efficiency. We also note that MS patients who went to bed later experienced higher levels of sleep efficiency. This is perhaps due to the fact that going to bed later would mean that they are more tired when they go to bed and thus would sleep more deeply.

8.3.2 Demographics: We find that a history of high caffeine consumption is related to a reduction in the quality of sleep. This is not surprising, as caffeine is a stimulant, and thus impairs the ability to sleep [103]. Furthermore, we find that MS patients who had a history of smoking had more awakenings. We believe that this is because nicotine, a chemical found in cigarettes, is a stimulant [112, 113]. We also find older age is linked with lower restorative sleep levels. This may be somehow related to the fact that older people sleep less [114].

8.3.3 Symptoms: As previously noted, high fatigue is connected with a decrease in sleep quality.

9 STATISTICAL SIGNIFICANCE ANALYSIS

9.1 Experimental Design

9.1.1 Motivation: Typically, healthcare findings are considered significant if they pass a p-value test [115–117] and/or show a strong Pearson correlation [38, 118]. Thus, to validate our identified phenotypes, we compare the results of our graph-based analysis to these accepted standard statistical models, with the expectation that both analysis methods will generally identify the same relationships to be significant.

9.1.2 Statistical Models Used: We calculate the Pearson correlation coefficients (r), an accepted standard measure of the correlation between two variables [118], for each pair of variables linked by a direct edge on the graph of

phenotyping relationships. They contain three dimensions of information that can validate the correctness of our phenotyping relationships:

(1) $|r|$ is a standard measure of the strength (*i.e.*, magnitude) of relationships, with $|r| \geq 0.4$ indicating that a relationship is reasonably strong [118].
(2) $sign(r)$ is a standard measure of the direction of relationships [118].
(3) The corresponding p-value for r is a standard measure of the statistical significance of relationships, with p-value ≤ 0.05 being statistically significant (*i.e.*, beyond random chance) [115].

9.2 Validation of Causal Discovery

For the phenotyping relationships from the graph (*i.e.*, causal discovery results) to be significant, they should satisfy the precondition of having a strong, statistically significant correlation. Thus, we consider $|r|$ and p-value, where we should have $|r| \geq 0.4$ and p-value ≤ 0.05 [115]. Indeed, the 102 edges in the graph of phenotypes have average $|r| = 0.461$, with average p-value $= 9.1 * 10^{-4}$. Therefore, our phenotyping relationships generally meet the precondition of strong, statistically significant correlation; validating the correctness of our causal discovery submodule.

9.3 Validation of Causal Inference

We also must confirm that our calculations of the average treatment effect (*i.e.*, results of causal inference) give the correct sign and magnitude of the phenotyping relationships between variables. Thus, we consider $sign(r)$ and $|r|$ to quantify sign and magnitude. $sign(r)$ matches $sign(ATE)$ in 96 out of 102 (94%) edges in the graph, indicating that the causal inference algorithms gave the correct direction of the ATEs. Taking both $sign(r)$ and $|r|$ into account ($r = sign(r) * |r|$), we find that r has a strong positive linear correlation with ATE, with correlation coefficient $R_{r,ATE} = 0.593$; meaning that causal inference gave the correct magnitude of ATE as well. Thus, the correctness of our causal inference submodule in inferring strength of phenotyping relationships is also validated.

10 VALUE OF PHENOTYPES FOR IDENTIFYING MS SYMPTOMS

10.1 Experimental Design

10.1.1 Motivation: There exists a great need among medical professionals to monitor MS disease progression [119–121], which essentially consists of new symptoms developing and old ones worsening [122, 123]. (One major reason for monitoring symptoms is to optimize treatment based on symptoms experienced [2, 9, 10].) Digital phenotyping can provide a convenient, instantaneous way to quantify these important MS symptoms as they develop over time; as compared to burdensome clinic visits. Thus, in this section, we explore if our digital behavioral phenotypes can help track MS symptoms.

10.1.2 Phenotype-Based Feature Selection: To prove this, we leverage the graph of phenotypes to select features. Particularly, we defined a symptom's closely related phenotypes to be any features within 3 directed edges of the symptom on the graph. (For example, if E is the symptom and $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$, then B, C, D would be considered closely related phenotypes to E, but A would not.) We then adapt these features to train classification models to identify if MS patients have certain symptoms (*i.e.*, Depression, Fatigue, and Sleep Quality), in retrospective analysis. If our phenotyping relationships are correct, classifiers trained on our closely related phenotypes (*i.e.*, features) should outperform those trained on all features (*i.e.*, the variables named in Table 2).

10.1.3 Symptom Prediction: We train four traditional supervised machine learning models: support vector machine (SVM), Logistic Regression, k-nearest neighbors (KNN), and Random Forest (RF), to show that our approach of phenotype-based feature selection is classifier-agnostic.

Baselines: To establish a baseline, a dummy classifier makes randomized predictions based on the distribution of the training data. Furthermore, we compare our phenotype-based feature selection method against the standard ANOVA F-Value feature selection. (The F-Value is defined as the ratio of inter-group variance to intra-group variance in the outcome variable, with groups being defined by splitting the dataset based on values of a treatment variable. $F > 1$ indicates that splitting based on the chosen treatment variable produces changes in the outcome variable. Thus, ANOVA F-Value feature selection chooses treatment variables that yield the largest F [124–128].)

Symptoms: We choose Depression, Fatigue, and Sleep Quality which are the most famous symptoms in MS as the identified symptoms. We consider their overall measurements, which were assessed via clinical questionnaires (CES-D, FSS, PSQI [80, 81, 83]) at the beginning of the study, as listed in Table 2. We consider overall measurements instead of daily measurements since we want to predict whether a patient will develop this symptom persistently, as opposed to having it for just one day due to random variation in day-to-day events. (For example, a patient who had poor sleep quality for just one night, due to the fire alarm going off at midnight; would not be considered to have the *persistent* symptom of poor sleep quality.)

Furthermore, we binarize Depression, Fatigue, and Sleep Quality scores. This follows a common convention in MS research that the presence of these symptoms is treated as binary, where the score from clinical rating scales is converted to a binary output (*e.g.*, a patient is diagnosed as either having or not having depression) [129–136]. Specifically, we use the following thresholds previously determined in clinical research: for Depression, a CES-D score below 16/60 meant a person did not have depression, score at or above meant a person did [129, 132]; for Fatigue, an FSS score below 36/63 meant a person did not have fatigue, score at or above meant a person did [130, 135]; for Sleep Quality, a PSQI score less than 16/21 meant a person had poor sleep quality, score at or above meant good sleep quality [131, 136].

Train-Test Splits: Moreover, we adopt 6-fold cross-validation to train our model with standard performance metrics accuracy, precision, and recall. In each fold, we adopt features from 25 subjects for training and test on the remaining 5 subjects. We highlight that there is no overlap of subjects in the training and testing set.

10.1.4 Performance Metrics: We use performance metrics that are common for classification tasks in mobile health; where TP, TN, FP, FN stand respectively for the number of true positives, true negatives, false positives, and false negatives predicted [137, 138]:

- $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, which measures the fraction of samples correctly labeled as symptomatic or asymptomatic.
- $recall = \frac{TP}{TP+FN}$, which quantifies our classifiers' ability to identify all symptomatic individuals without misclassifying any symptomatic individuals as asymptomatic.
- $precision = \frac{TP}{TP+FP}$, which quantifies our classifiers' robustness against falsely identifying (misclassifying) asymptomatic individuals as symptomatic.

.

10.2 Experimental Results

10.2.1 Impact of Phenotype-Based Feature Selection: The results of this experiment are shown in Figure 8 and Table 4. Figure 8 shows the comparison of accuracy, precision, and recall when using *only* closely related phenotypes versus using all features or features selected via ANOVA f-values. We highlight that accuracy, precision, and recall for the group using closely related phenotypes are consistently the highest. Furthermore, Table 4 shows us the results using every single combination of MS symptom and ML classifier. Though performance is different among models, features selected via closeness of phenotyping relationships always achieve the best performance, thus showing that we found relevant phenotypes.

As for why our phenotype-based feature selection achieves the best performance, we believe it is due to the fact that we automatically account and control for the simultaneous interactions among confounding, treatment,

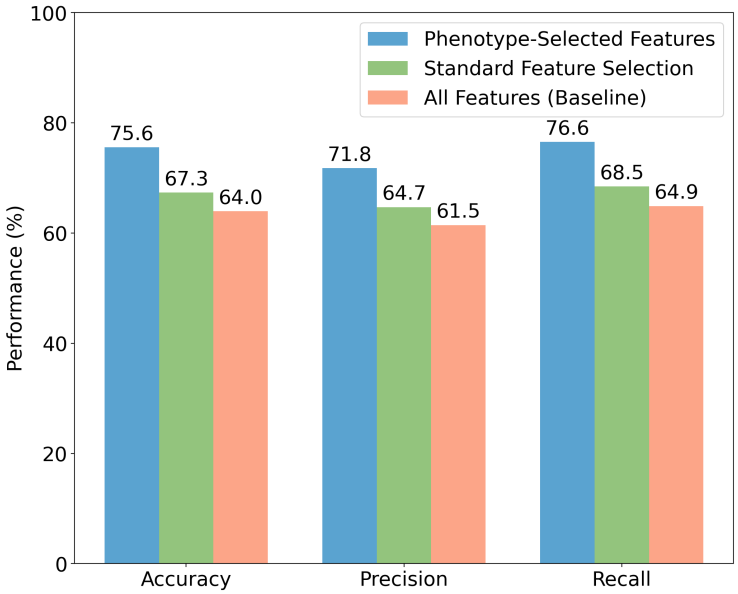

Fig. 8. Average performance across all symptoms (Depression, Sleep Quality, Fatigue) and classifiers (SVM, KNN, Log. Reg., RF) trained with phenotype-based, f-value (standard), and no feature selection. Phenotype-based feature selection results in the best classifier performance.

Table 4. Accuracy using phenotype-based, f-value, and no feature selection to train ML classifiers to predict MS symptoms. Numbers in parentheses **(#)** indicate the number of features used as predictors. Features that measure the same symptom as the predicted symptom are removed from input (*e.g.*, no "Diary Sleep Quality" when predicting "Overall Sleep Quality").

Accuracy (%)						
Symptom	**Feature Sel. (# feat.)**	**SVM**	**Logit**	**KNN**	**RF**	**Dummy**
Depression	**None (56)**	79.0	82.4	80.5	82.4	60.5
	Phenotype (35)	87.1	86.7	84.8	**90.0**	60.5
	F-Value (35)	80.0	84.8	84.8	78.6	60.5
Fatigue	**None (54)**	54.3	61.4	42.9	33.8	48.6
	Phenotype (21)	**72.9**	66.7	66.2	65.2	48.6
	F-Value (21)	65.2	62.4	47.6	31.9	48.6
Sleep Quality	**None (52)**	65.2	69.5	61.0	55.2	52.4
	Phenotype (33)	70.0	**74.3**	69.0	73.8	52.4
	F-Value (33)	66.7	71.4	66.2	68.6	52.4

and outcome variables via our graph-based analysis of phenotypes. Thus, the outputted relationships between pairs of adjacently-linked variables in the graph represent the true relationships between treatment and outcome variables, free from the influence of additional confounding factors. In contrast, ANOVA f-value feature selection compares inter- and intra-group variances of the outcome variable with different treatment variable splits but does not involve an explicit graph that includes all the outcome and treatment variables *and their simultaneous*

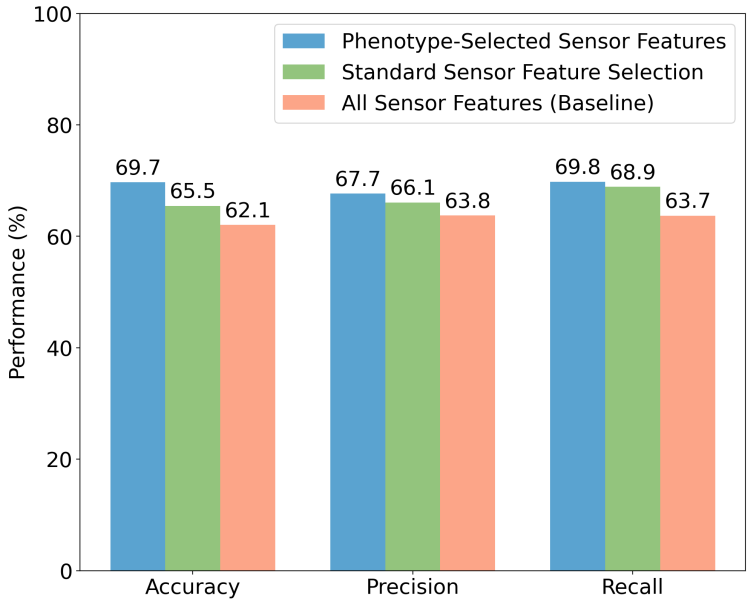

Fig. 9. Average performance across all symptoms and machine learning classifiers when trained *only on sensor data*. Performance is comparable to using all data, and phenotype-based feature selection still improves classifier performance.

Table 5. Accuracy with various feature selection methods (phenotype-based, f-value, none) to train various ML classifiers in predicting MS symptoms from *sensor data only*. This time, we exclude demographics and other symptoms from the input.

		Accuracy (%)				
Symptom	**Feature Sel. (# feat.)**	**SVM**	**Logit**	**KNN**	**RF**	**Dummy**
Depression	**None (29)**	71.0	79.5	74.3	66.7	60.5
	Phenotype (13)	**85.7**	82.4	81.4	72.9	60.5
	F-Value (13)	80.5	82.9	81.4	73.3	60.5
Fatigue	**None (29)**	59.0	50.0	40.0	43.3	48.6
	Phenotype (9)	56.2	54.8	52.9	**66.7**	48.6
	F-Value (9)	58.1	55.2	34.8	50.0	48.6
Sleep Quality	**None (29)**	70.5	73.3	62.4	57.6	52.4
	Phenotype (17)	**75.2**	73.3	68.6	66.7	52.4
	F-Value (17)	65.7	71.9	58.6	73.3	52.4

interactions with possible confounders. Lacking this explicit graph, we believe that the ANOVA f-value may not control for the influence of confounders as well as our phenotype graph-based method, thus accounting for the difference in performance. Furthermore, we believe that the ability of our phenotype graph-based method to account for intermediate linking variables in "indirect" relationships (*e.g.*, $A \rightarrow B \rightarrow C$), as opposed to ANOVA f-value not accounting for intermediate linking variables, may also play a role in the performance difference, as it makes our feature selection more nuanced.

10.2.2 Impact of Sensor Data: The previous experiment selects data from both the smartwatch and the clinical survey. However, we are motivated to explore the performance using data from smartwatches only (*i.e.*, only features from activities of daily life). If it works, we can passively monitor the variations of symptoms and remind caregivers of the needs of intervention without interrupting the daily lives of patients. (This is also the reason why we conducted our study using data collected in the wild, as opposed to from a controlled clinical setting.) For this purpose, we repeat the above experiment using only the smartwatch data. As shown in Figure 9 and Table 5, the main result that using phenotype-based feature selection yields the best results for symptom prediction still holds true.

Smaller Performance Improvement: However, we note that the improvement in performances with phenotype-based feature selection is smaller using only the sensor data. We believe this is because the graph of phenotyping relationships was generated with all features under consideration; thus, some important phenotyping relationships may have been missing in this sub-experiment due to using only sensor data. In regards to whether the smaller increase in performance could have been achieved due to chance, it may be plausible for phenotype-based vs. f-value feature selection on sensor data, as those results are very close (+4.2% accuracy, +1.6% precision, +0.9% recall). However, when comparing phenotype-based feature selection to no feature selection on the sensor data, the difference in performance is large enough (+7.6% accuracy, +3.9% precision, +6.1% recall) that we believe it was not due to chance. Overall, this is acceptable, as it simply indicates that *both* phenotype-based and f-value-based feature selection are valid, as compared to no feature selection at all on sensor data.

Comparison of Sensor vs. All Features: Overall, the sensor-only classifiers have slightly lower performance than the classifiers which had all categories of features (*i.e.*, sensor, symptom, and demographic data) available. This is perhaps because there are fewer features to base the predictions on in the sensor-only classifiers. Even so, the accuracy, precision, and recall when using *sensor data only* are still all within 6% of the results which use all categories of data (Fig. 8, 9). The fact that the classifier performances in the two experiments are comparable is particularly impressive, considering the fact that we have half the features available when using sensor data only. This indicates that sensor data alone still holds great phenotyping value for MS symptoms. Furthermore, we highlight that the *best* performing model across all experiments for predicting a patient's sleep quality is actually the SVM trained *only on sensor data selected via phenotyping relationships*, with an accuracy of 75.2%. This is because the sensor data contained much information related to sleep, wake, and activity patterns, all of which would be very relevant to sleep quality. In summary, the digital behavioral phenotype alone is sufficient as a first line of defense for tracking MS symptoms (which could be subsequently reported to physicians, who then decide appropriate actions).

10.2.3 Impact of Classifier Implementation: As shown in the previous experiments, phenotype-based feature selection is a classifier-agnostic approach towards tracking MS symptoms. Even so, Tables 4 and 5 show us that some classifiers clearly do perform better than others. In particular, we see that SVM with phenotype-selected features has the best prediction accuracy in 3 out of 6 cases: depression (sensor-only), fatigue (all features), and sleep quality (sensor-only). Some possible reasons for the high performance of SVM are that they generally work well with high-dimensional data, are robust against overfitting, and are applicable to many different kinds of data distributions. We also note that random forest (RF) is the highest performer in two cases: predicting fatigue from phenotype-selected sensor data (*i.e.*, digital behavioral phenotype), and predicting depression from phenotype-selected data. This could be because RF's built-in ensemble learning allowed this model to overcome the reduced number of input features by combining predictions of many decision trees.

10.2.4 User-Side Considerations: Towards the goals of convenience and scalability in MS symptom tracking, we highlight that under our approach, first-time users only need to use the wearables for one day in order to track their symptoms. We showed this in the above evaluations, via two experimental design choices: 1) We made sure that there is no overlap between the MS patients in the training and testing datasets. 2) We conducted all the

symptom identification using only one day of data at a time. (Thus, in each fold, we identified symptoms for each of 35 samples = 5 subjects $\times$ 7 $\frac{\text{samples}}{\text{subject}}$, while using the 175 samples from other subjects for training). In other words, our approach works out-of-the-box for first-time users to track MS symptoms within only 24 hours.

11 RELATED WORK

11.1 Digital Phenotyping for Multiple Sclerosis

Digital phenotyping for MS is also an emerging topic [11, 139, 140]. Midaglia et al., while not finding any specific digital behavioral phenotypes, showed the feasibility, as it relates to user adherence and satisfaction, of using mobile and wearable technologies to passively monitor and actively collect data from MS patients [139]. Chitnis et al. used a digital phenotyping approach to monitor general neurological disability in MS patients. To obtain the digital phenotyping data, they monitored patients in clinical settings with nine simultaneously worn wearable sensors, and in the wild with three simultaneously worn wearable sensors [11]. The biggest difference between our work and theirs is a matter of scalability: theirs required up to nine wearable sensors, while ours only requires a singular wearable sensor. Furthermore, in contrast to our work, they did not focus on finding behavioral phenotypes for a wide spectrum of specific debilitating symptoms (*e.g.*, depression) [12]. Tong et al used data from smartwatches, smart sleep trackers, and smart scales to predict the fatigue and health-related quality of life of MS patients [140]. Our work goes further by using digital phenotyping as a means to track not just fatigue, but a wide spectrum of significant symptoms like depression, sleep quality, and mood [5, 6, 34]. Furthermore, our approach emphasizes scalability, due to only requiring one smartwatch, as opposed to three smart devices.

11.2 Data Analysis in Digital Phenotyping

11.2.1 Non-MS:. In digital phenotyping for non-multiple sclerosis applications, there are many methods of data analysis.

Pearson Correlation-Based: StudentLife modeled smartphone sensor data's digital phenotyping relationships with mental health and educational outcomes among college students using Pearson correlation coefficients (r) [38]. SmartGPA also used Pearson's r to model the digital phenotyping relationships between mobile sensor data and undergraduate GPA, and lasso regularized linear regression [141] to predict GPAs from sensor data [142]. Regarding prediction models, SmartGPA's was designed for a regression task, while ours (for reasons described in 10.1) was designed for a classification task — which is not directly comparable. However, we can compare modeling of digital phenotyping relationships: While StudentLife's and SmartGPA's proposed method of Pearson's r can find correlations among pairs of variables, it is not designed to account for potential confounders/covariates [118]. In contrast, our graph-based approach automatically identifies and controls for effects of covariates (*e.g.*, time spent in bed, energy expenditure in light activity) [31, 32, 94].

Accounting for Few Covariates: R Wang et al.'s study did account for intra-subject longitudinal covariance by using the Generalized Linear Mixed Model (GLMM) [143] to model digital phenotyping relationships between subjects' longitudinal depression score (PHQ-4, collected weekly) and passive sensing (phone, wearable) data. Yet, they still used regular Pearson correlation to model the digital phenotyping relationships between passive sensing data and "general" depression scores (PHQ-8, collected once at the beginning and once at end of the study). They used lasso regularized linear regression to predict depression score [144]. Again, their regression models are not directly comparable to our classification models, but we can compare the analyses of digital phenotyping relationships: While their use of GLMM addressed intra-subject longitudinal covariance, it did not address inter-variable covariance among sensor data (*e.g.*, conversation duration, stationary time), and neither did their Pearson correlation in the non-longitudinal analysis. W Wang et al.'s study also used GLMM [143] to control for differences in mobile sensor model when analyzing correlations between personality traits and mobile sensing data, and gradient-boosted regression trees (GBRT) [145] to predict personality traits [58]. Again, the GLMM

did not account for the inter-variable covariance among different sensor data (*e.g.*, physical activity, location), and GBRT (a regression model) is not directly comparable to our classification models from 10.1. CrossCheck used passive sensing data to detect mental health changes in people with schizophrenia [52]. To model digital phenotyping relationships between sensor data and mental health variables, they used Generalized Estimating Equations (GEE), a method related to GLMM which accounts for longitudinal covariance within a subject [146]; to predict mental health variables, they also used GBRT. Yet, their GEE still did not account for the inter-variable covariance among sensor data, in contrast to our graph-based method.

Accounting for Covariates: The approaches described in MyTraces [51] and Tsapeli and Musolesi's analysis of the StudentLife dataset [147] do attempt to account for the covariates. Similar to our work, they used causal inference algorithms to quantify the strength of relationships among mobile sensor data and emotional states. To find the covariates which the causal inference should account for, they conducted a pre-analysis based on the Kendall rank correlation [148], naming variables that had high correlations with both the treatment and outcome variables of interest as covariates. [51, 147]. Neither of these works built models to predict emotional states. We build upon their approaches, but instead of using Kendall rank correlations to identify covariates, we use state-of-the-art causal discovery algorithms [93–95] to construct complete graphical models of the covariance relationships among all the variables. Furthermore, we also design classification models to predict/identify which subjects have MS symptoms.

11.2.2 MS-Specific: As it specifically relates to the digital phenotyping of MS, the data analysis methods have similar shortcomings. Chitnis et al. used Spearman correlation coefficients [149] to analyze associations between wearable sensor data and MS disability scores [11]. Spearman correlation coefficients are similar to Pearson correlation coefficients in that by default, they do not account for all possible confounders [118, 149]. Tong et al. focused on regression to predict fatigue (FSS) and health state (EQ-5D) from multimodal sensor data, using tree-based Adaboost Regressors [150]. While they achieved a strong prediction model, their closest attempt at an explicit analysis of digital phenotyping relationships was identifying those features which were selected over half the time by Adaboost in their prediction experiments [140]. In other words, their study did not focus on rigorous analysis of digital behavioral phenotyping relationships which explicitly controlled for covariates.

In summary, the improvement of our data analysis methodology over previous works lies in the dual combination of 1) accounting and controlling for the covariates/confounders via graph-based analysis of digital behavioral phenotyping relationships, and 2) doing so to address the important problem of digitally behaviorally phenotyping multiple sclerosis.

12 DISCUSSION

12.1 Methodological Contribution

In terms of methodology, the immediately obvious contribution of our work is the application of various state-of-the-art technologies towards addressing the complex problem of digital behavioral phenotyping of MS patients.

12.1.1 A Versatile, Generalizable Approach: Yet, our methodological contribution goes far beyond studying just MS (which is a significant problem in and of itself); it can be applied to *any* chronic disease which has multiple complex symptoms (*e.g.*, Parkinson's, Alzheimer's, Huntington's) [151]. Thus, in *MSLife*, we introduce a versatile, generalizable approach for many future studies in digital behavioral phenotyping.

Wearables for Data Collection: Through wearable sensors (GENEActiv smartwatch), we are able to continuously, passively collect complex real-life data regarding unscripted ADLs in the wild from MS patients. In contrast to traditional laboratory-based methods, the novelty of our method is that we are able to capture *genuine* ADLs, which is an important step to fully exploring the behavioral phenotyping of MS symptoms. Again, we emphasize

that this wearable-based methodology could plausibly be used to digitally behaviorally phenotype any chronic multisymptomatic disease (*e.g.*, Parkinson's, Alzheimer's).

Analysis Methods: We introduce powerful statistical analysis methods to the ubiquitous computing community. Particularly, our work is one of the first in the ubiquitous computing community to build a graphical network (via causal discovery and inference algorithms) to model the relationships between sensor data collected in the wild and clinical health data. The advantage of this graph-based framework is that it accounts and controls for the simultaneous interactions among confounding, treatment, and outcome variables; while traditional methods like Pearson correlation do not [118], as we discuss more in-depth in Section 11.2. Additionally, our graph-based framework can account for indirect relationships and identify the specific linking variables along those paths (*e.g.*, later rise times are related to higher levels of depression, which are linked with higher fatigue). In our particular work, we modeled relationships pertaining to MS patients; but we once again highlight that our graph-based framework is extensible to any study which wishes to model the underlying relationships in complex real-life data.

12.1.2 Methodological Insights from Findings: Our approach allowed us to discover various digital behavioral phenotypes that exist for MS symptoms. Many of these (*e.g.*, inverse relationship between the relative amount of moderate physical activity and sleep quality) may not have been as easily discoverable were it not for our wearable sensing + graph-based analysis methodology. Even those findings which may have been previously known to some degree (*e.g.*, depression and fatigue have a positive relationship) are valuable, as they actually confirm the validity of our methodology.

12.2 MS-Specific Implications

Based on our study and subsequent evaluation of the digital behavioral phenotyping results, *MSLife* has great implications regarding smartwatch-based approaches in healthcare for multiple sclerosis patients.

12.2.1 Monitor MS Progression: There exists a great need among medical professionals to monitor MS disease progression [119–121], which essentially consists of new symptoms developing and old ones worsening [122, 123]. As shown in Section 10.2, where we leveraged the digital behavioral phenotyping relationships to identify the presence of MS symptoms; our work provides a way to quantify these important MS symptoms as they develop over time, based on daily-life data passively collected by smartwatches. This smartwatch-based approach to monitoring disease progression is very convenient as compared to regular clinical visits, which are known to have high costs and patient burden. Furthermore, it instantaneously captures disease progression/symptom variation 24/7, which is another advantage over clinical visits. We envision using smartwatches as the first line of defense to continuously monitor patients for MS symptoms; and if possible symptoms are detected by the smartwatch-based systems (as demonstrated in Section 10.2's symptom identification analysis), they are reported to physicians who make judgments on further action.

Note on Study Design: We emphasize that being able to *conveniently* monitor MS progression is a major reason why we specifically conducted our study using sensor data collected in the wild, as opposed to collected in a controlled clinical setting. Since our purpose for using wearables is to eventually monitor MS progression with minimal patient burden, it should be done without interrupting their daily lives; going to a clinic to use these wearables for ADL data collection would defeat this purpose.

12.2.2 Precision Medicine: Treatment/intervention given to MS patients varies based on the symptoms they experience. For example, MS patients who experience depression may be prescribed antidepressant medications, while patients who experience fatigue may be prescribed medications like Symmetrel or undergo magnetic therapy [2, 9]. Thus, using smartwatches as a tool to track specific MS symptoms (as demonstrated in Section

10.2) can help medical professionals give appropriate interventions. These interventions will also be timely, as digital behavioral phenotyping allows clinicians to monitor patients 24/7 with a minimal patient burden. Overall, this is known as a precision medicine approach, which is a topic of recent interest [152–155].

12.3 Limitations

12.3.1 MS Diagnosis: While digital behavioral phenotyping can be used for diagnosing MS, studies that aim to do so must include a healthy cohort alongside the MS cohort. In contrast, the goal of our study was to track *symptoms* of MS (*i.e.*, monitoring disease progression among those who *already have* MS). Thus, we designed our study to have only an MS cohort.

12.3.2 MS Risk Prediction: Similarly, digital behavioral phenotyping can be used to predict the risk of developing MS, but this again requires both a healthy and MS cohort (see previous).

12.3.3 Specificity to MS (Control Group): The lack of a non-MS cohort may raise the possible concern that the results (*e.g.*, better sleep quality lowers depression) may not be specific to MS patients. Regardless of whether or not they are MS-specific, the important part is that these relationships are still nonetheless valid for MS patients; and more importantly, can still be used to help MS patients manage their symptoms through intervention and behavioral therapy.

12.3.4 Validity of Causality: We designed a graph-based statistical analysis framework based on verified causal discovery and inference algorithms. We emphasize that while these algorithms have been optimized to *suggest* highly likely causal relationships; further work should be done to be sure that these relationships are truly causal, and not merely correlational. The gold standard for showing causality is still to conduct a controlled, randomized experiment; causal discovery and inference algorithms should only be used as the first step in those situations where it is not currently feasible to do so. Thus, we refrain from making claims of causality; we keep our claims at correlation, which is still valuable knowledge.

12.3.5 Extreme Conditions: We did not investigate the impact of extreme conditions on the digital behavioral phenotyping outcome and MS symptom tracking. Examples of extreme conditions include: accidentally dropping the smartwatch (on the ground or into water), wearing the smartwatch while on airplane travel.

12.3.6 Feature Coverage: We have considered a wide variety of features (57 total) — spanning ADLs, MS symptoms, and demographics — in our study. Yet, as in any scientific study (especially an in the wild one), there is always the possibility of features (*e.g.*, geographic location) that were not considered but actually do impact the symptom outcome of MS patients.

12.3.7 Cohort Size and Study Duration: We understand that it is a rule of thumb that a larger group of participants with a longer duration would be helpful in the experimental study, and reduces the concern results could have been due to chance. Even so, we still believe our results to be valuable for a pilot study since 30 people is a comparable cohort size (MyTraces [51] had 28 users, StudentLife [38] had 48 users, CrossCheck [52] had 21 subjects), and our 168-hour study duration is a comparable duration (Wang et al.'s study lasted for 14 days [58], MyTraces lasted for 20 days [51], a plurality of subjects (43%) in SugarMate had 6-10 days of data collected [59]) to that of similar studies involving mobile health technologies. Furthermore, our framework is designed to be scalable to larger cohort sizes and longer durations (*e.g.*, 200 people, 28 days), even if our pilot study was on 30 patients for 168 hours.

12.4 Future Works

12.4.1 MS Diagnosis, MS Risk Prediction, Specificity to MS:. To address the first three limitations, we hope to conduct a future study involving both MS and non-MS cohorts (*i.e.*, control group). This will allow us to explore digital behavioral phenotyping for MS diagnosis and risk prediction. It will also allow us to explore how this study's phenotyping relationships are different for MS patients as compared to a general population.

12.4.2 Validity of Causality: Regarding the fourth limitation, an experiment in a controlled clinical environment is the gold standard for proving causality. But as previously stated, it may be difficult to capture genuine ADLs in such an environment. We leave this challenge to be solved by future works which wish to rigorously prove the causality.

12.4.3 Extreme Conditions: To address the fifth limitation, future studies could ask patients to report if and when such extreme conditions occur in their daily lives.

12.4.4 Feature Coverage: Addressing the sixth limitation, we encourage future studies to continue to consider more factors in their analysis.

12.4.5 Cohort Size and Study Duration: Finally, in regards to the seventh limitation, future studies should increase the sample size to be on the order of magnitude of 10^2 people and duration to be on the order of magnitude of 10^1 weeks.

13 CONCLUSION

In this paper, we presented *MSLife*, one of the first end-to-end approaches to explore digital behavioral phenotyping of MS symptoms in the wild. We deployed *MSLife* with a cohort of 30 MS patients across a one-week in the wild IRB-approved clinical pilot study. We utilized unobtrusive commodity smartwatch sensors to passively, continuously monitor potential digital behavioral phenotypes (*i.e.*, ADLs) in daily life. We then designed a graph-based analysis framework to discover digital behavioral phenotyping relationships among MS symptoms, ADLs, and demographic factors.

Regarding results, we discover 102 statistically significant phenotyping relationships (*e.g.*, later rise times are related to increased depression, history of caffeine consumption is linked with lower fatigue levels, higher relative levels of moderate physical activity are associated with decreased sleep quality). Furthermore, our retrospective analysis showed that these digital behavioral phenotypes were highly effective in tracking MS symptoms, outperforming baseline machine learning methods in classifying whether or not a patient has a particular MS symptom (*e.g.*, depression, fatigue, poor sleep quality), with respect to metrics: accuracy (75.6% vs 64.0%), precision (71.8% vs 61.5%), and recall (76.6% vs 64.9%).

Highlighting our methodological contributions and implications: (1) Methodologically, we contribute a novel, versatile, generalizable approach which can be applied to many future studies in digital behavioral phenotyping of chronic diseases. Our approach is one of the first to introduce powerful graph-based statistical analysis methods which account and control for covariates to the ubiquitous computing community. (2) Regarding MS-specific implications, *MSLife* paves the way for smartwatch-based approaches to monitoring MS progression and facilitating precision medicine.

ACKNOWLEDGMENTS

This work was in part supported by the U.S. National Science Foundation under Grant CNS-2050910.

REFERENCES

[1] C. A. Dendrou, L. Fugger, and M. A. Friese, "Immunopathology of multiple sclerosis," *Nature Reviews Immunology*, vol. 15, no. 9, pp. 545–558, 2015.

[2] "Multiple sclerosis," https://www.nccih.nih.gov/health/multiple-sclerosis, accessed: 11/08/2020.

[3] "How many people live with ms?" https://www.nationalmssociety.org/What-is-MS/How-Many-People.

[4] G. Adelman, S. G. Rane, and K. F. Villa, "The cost burden of multiple sclerosis in the united states: a systematic review of the literature," *Journal of medical economics*, vol. 16, no. 5, pp. 639–647, 2013.

[5] V. Janardhan and R. Bakshi, "Quality of life in patients with multiple sclerosis: the impact of fatigue and depression," *Journal of the neurological sciences*, vol. 205, no. 1, pp. 51–58, 2002.

[6] W. E. Fleming and C. P. Pollak, "Sleep disorders in multiple sclerosis," in *Seminars in neurology*, vol. 25, no. 01. Copyright© 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New …, 2005, pp. 64–68.

[7] J. H. Noseworthy, "Progress in determining the causes and treatment of multiple sclerosis," *Nature*, vol. 399, no. 6738, pp. A40–A47, 1999.

[8] D. M. Wingerchuk, C. F. Lucchinetti, and J. H. Noseworthy, "Multiple sclerosis: current pathophysiological concepts," *Laboratory investigation*, vol. 81, no. 3, pp. 263–281, 2001.

[9] "Multiple sclerosis information page," https://www.ninds.nih.gov/Disorders/All-Disorders/Multiple-Sclerosis-Information-Page, accessed: 11/08/2020.

[10] R. B. Schiffer and N. M. Wineman, "Antidepressant pharmacotherapy of depression associated with multiple sclerosis." *The American journal of psychiatry*, 1990.

[11] T. Chitnis, B. I. Glanz, C. Gonzalez, B. C. Healy, T. J. Saraceno, N. Sattarnezhad, C. Diaz-Cruz, M. Polgar-Turcsanyi, S. Tummala, R. Bakshi *et al.*, "Quantifying neurologic disease using biosensor measurements in-clinic and in free-living settings in multiple sclerosis," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–8, 2019.

[12] R. Siegert and D. Abernethy, "Depression in multiple sclerosis: a review," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 4, pp. 469–475, 2005.

[13] H. Hasselmann, J. Bellmann-Strobl, R. Ricken, T. Oberwahrenbrock, M. Rose, C. Otte, M. Adli, F. Paul, A. U. Brandt, C. Finke *et al.*, "Characterizing the phenotype of multiple sclerosis–associated depression in comparison with idiopathic major depression," *Multiple Sclerosis Journal*, vol. 22, no. 11, pp. 1476–1484, 2016.

[14] V. M. Leavitt, R. Brandstadter, M. Fabian, I. Katz Sand, S. Klineova, S. Krieger, C. Lewis, F. Lublin, A. Miller, G. Pelle *et al.*, "Dissociable cognitive patterns related to depression and anxiety in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 26, no. 10, pp. 1247–1255, 2020.

[15] E. De Meo, E. Portaccio, A. Giorgio, L. Ruano, B. Goretti, C. Niccolai, F. Patti, C. G. Chisari, P. Gallo, P. Grossi *et al.*, "Identifying the distinct cognitive phenotypes in multiple sclerosis," *JAMA neurology*, vol. 78, no. 4, pp. 414–425, 2021.

[16] K. Radhakrishnan, M. T. Kim, M. Burgermaster, R. A. Brown, B. Xie, M. S. Bray, and C. A. Fournier, "The potential of digital phenotyping to advance the contributions of mobile health to self-management science," *Nursing outlook*, vol. 68, no. 5, pp. 548–559, 2020.

[17] D. Kos, E. Kerckhofs, G. Nagels, M. D'hooghe, and S. Ilsbroukx, "Origin of fatigue in multiple sclerosis: review of the literature," *Neurorehabilitation and neural repair*, vol. 22, no. 1, pp. 91–100, 2008.

[18] H. Kaynak, A. Altintaş, D. Kaynak, Ö. Uyanik, S. Saip, J. Ağaoğlu, G. Önder, and A. Siva, "Fatigue and sleep disturbance in multiple sclerosis," *European Journal of Neurology*, vol. 13, no. 12, pp. 1333–1339, 2006.

[19] "Phenotype," https://www.genome.gov/genetics-glossary/Phenotype, accessed 05/09/21.

[20] T. Olsson, L. F. Barcellos, and L. Alfredsson, "Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis," *Nature Reviews Neurology*, vol. 13, no. 1, p. 25, 2017.

[21] M. W. Nortvedt, T. Riise, and J. Maeland, "Multiple sclerosis and lifestyle factors: the hordaland health study," *Neurological Sciences*, vol. 26, no. 5, pp. 334–339, 2005.

[22] D. Jakimovski, Y. Guan, M. Ramanathan, B. Weinstock-Guttman, and R. Zivadinov, "Lifestyle-based modifiable risk factors in multiple sclerosis: review of experimental and clinical findings," *Neurodegenerative Disease Management*, vol. 9, no. 3, pp. 149–172, 2019, pMID: 31116081. [Online]. Available: https://doi.org/10.2217/nmt-2018-0046

[23] L. J. White and R. H. Dressendorfer, "Exercise and multiple sclerosis," *Sports medicine*, vol. 34, no. 15, pp. 1077–1100, 2004.

[24] J. J. Veldhuijzen van Zanten, L. A. Pilutti, J. L. Duda, and R. W. Motl, "Sedentary behaviour in people with multiple sclerosis: Is it time to stand up against ms?" *Multiple Sclerosis Journal*, vol. 22, no. 10, pp. 1250–1256, 2016.

[25] S. D. Brass, P. Duquette, J. Proulx-Therrien, and S. Auerbach, "Sleep disorders in patients with multiple sclerosis," *Sleep medicine reviews*, vol. 14, no. 2, pp. 121–129, 2010.

[26] A. Andreasen, E. Stenager, and U. Dalgas, "The effect of exercise therapy on fatigue in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 17, no. 9, pp. 1041–1054, 2011.

[27] M. B. Rietberg, D. Brooks, B. M. Uitdehaag, and G. Kwakkel, "Exercise therapy for multiple sclerosis," *Cochrane database of systematic reviews*, no. 1, 2005.

[28] P.-Y. Yang, K.-H. Ho, H.-C. Chen, and M.-Y. Chien, "Exercise training improves sleep quality in middle-aged and older adults with sleep problems: a systematic review," *Journal of physiotherapy*, vol. 58, no. 3, pp. 157–163, 2012.
[29] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research," *JMIR mental health*, vol. 3, no. 2, p. e16, 2016.
[30] O. Lab, "Research areas," https://www.hsph.harvard.edu/onnela-lab/research/.
[31] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3, no. 1. Springer, 2016, p. 3.
[32] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," 2020.
[33] H. Lassmann, W. Brück, and C. F. Lucchinetti, "The immunopathology of multiple sclerosis: an overview," *Brain pathology*, vol. 17, no. 2, pp. 210–218, 2007.
[34] R. T. Joffe, G. P. Lippert, T. A. Gray, G. Sawa, and Z. Horvath, "Mood disorder and multiple sclerosis," *Archives of Neurology*, vol. 44, no. 4, pp. 376–378, 1987.
[35] J. Rooksby, A. Morrison, and D. Murray-Rust, "Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.
[36] V. W.-S. Tseng, N. Valliappan, V. Ramachandran, T. Choudhury, and V. Navalpakkam, "Digital biomarker of mental fatigue," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–5, 2021.
[37] L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, "Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.
[38] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "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*, 2014, pp. 3–14.
[39] "Compare products," https://www.activinsights.com/products/geneactiv/compare-products/.
[40] "Geneactiv original," https://www.activinsights.com/products/geneactiv/, Accessed 07/08/21.
[41] https://www.activinsights.com/wp-content/uploads/2015/11/GENEActiv-Brochure-2015.pdf, https://www.activinsights.com/wp-content/uploads/2015/11/GENEActiv-Brochure-2015.pdf, Accessed 07/20/21.
[42] D. Eslinger, A. V. Rowlands, T. L. Hurst, M. Catt, P. Murray, and R. G. Eston, "Validation of the genea accelerometer," 2011.
[43] "How geneactiv accelerometer research watches work," https://www.activinsights.com/technology/geneactiv/how-it-works/, accessed 07/20/21.
[44] "Geneactiv," https://www.activinsights.com/products/geneactiv/, Accessed 07/20/21.
[45] https://help.fitbit.com/articles/en_US/Help_article/1136.htm, title=How accurate are Fitbit devices?, note=Accessed 07/08/21.
[46] "Watch - apple," https://www.apple.com/watch/, accessed 07/08/21.
[47] O. Kantarci, A. Siva, M. Eraksoy, R. Karabudak, N. Sütlaş, J. Ağaoğlu, F. Turan, M. Özmenoğlu, E. Toğrul, M. Demirkiran *et al.*, "Survival and predictors of disability in turkish ms patients," *Neurology*, vol. 51, no. 3, pp. 765–772, 1998.
[48] J. C. Bot, F. Barkhof, C. H. Polman, G. L. à. Nijeholt, V. de Groot, E. Bergers, H. J. Ader, and J. A. Castelijns, "Spinal cord abnormalities in recently diagnosed ms patients," *Neurology*, vol. 62, no. 2, pp. 226–233, 2004. [Online]. Available: https://n.neurology.org/content/62/2/226
[49] U. Dalgas, E. Stenager, J. Jakobsen, T. Petersen, H. Hansen, C. Knudsen, K. Overgaard, and T. Ingemann-Hansen, "Fatigue, mood and quality of life improve in ms patients after progressive resistance training," *Multiple Sclerosis Journal*, vol. 16, no. 4, pp. 480–490, 2010.
[50] C. Christodoulou, L. Krupp, Z. Liang, W. Huang, P. Melville, C. Roque, W. Scherl, T. Morgan, W. MacAllister, L. Li *et al.*, "Cognitive performance and mr markers of cerebral injury in cognitively impaired ms patients," *Neurology*, vol. 60, no. 11, pp. 1793–1798, 2003.
[51] A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi, "Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, 2017.
[52] R. Wang, M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer *et al.*, "Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 886–897.
[53] "Who gets ms? (epidemiology)," https://www.nationalmssociety.org/What-is-MS/Who-Gets-MS, accessed 05/07/21.
[54] M. A. Hernán, S. S. Jick, G. Logroscino, M. J. Olek, A. Ascherio, and H. Jick, "Cigarette smoking and the progression of multiple sclerosis," *Brain*, vol. 128, no. 6, pp. 1461–1465, 2005.
[55] J. M. Greer and P. A. McCombe, "Role of gender in multiple sclerosis: clinical effects and potential molecular mechanisms," *Journal of neuroimmunology*, vol. 234, no. 1-2, pp. 7–18, 2011.
[56] L. J. Julian, L. Vella, T. Vollmer, O. Hadjimichael, and D. C. Mohr, "Employment in multiple sclerosis," *Journal of neurology*, vol. 255, no. 9, pp. 1354–1360, 2008.
[57] M. Catanzaro and C. Weinert, "Economic status of families living with multiple sclerosis." *International journal of rehabilitation research. Internationale Zeitschrift fur Rehabilitationsforschung. Revue internationale de recherches de readaptation*, vol. 15, no. 3, pp. 209–218, 1992.

[58] W. Wang, G. M. Harari, R. Wang, S. R. Müller, S. Mirjafari, K. Masaba, and A. T. Campbell, "Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018. [Online]. Available: https://doi.org/10.1145/3264951
[59] W. Gu, Y. Zhou, Z. Zhou, X. Liu, H. Zou, P. Zhang, C. J. Spanos, and L. Zhang, "Sugarmate: Non-intrusive blood glucose monitoring with smartphones," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017. [Online]. Available: https://doi.org/10.1145/3130919
[60] "Tetrad manual," http://cmu-phil.github.io/tetrad/manual/, 2019.
[61] "DoWhy: A Python package for causal inference," https://github.com/microsoft/dowhy, 2019.
[62] M. Littner, C. A. Kushida, W. M. Anderson, D. Bailey, R. B. Berry, D. G. Davila, M. Hirshkowitz, S. Kapen, M. Kramer, D. Loube, M. Wise, and S. F. Johnson, "Practice Parameters for the Role of Actigraphy in the Study of Sleep and Circadian Rhythms: An Update for 2002," *Sleep*, vol. 26, no. 3, pp. 337–341, 05 2003. [Online]. Available: https://doi.org/10.1093/sleep/26.3.337
[63] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," *Sleep*, vol. 26, no. 3, pp. 342–392, 2003.
[64] J. L. Martin and A. D. Hakim, "Wrist actigraphy," *Chest*, vol. 139, no. 6, pp. 1514–1527, 2011.
[65] M. E. Rosenberger, M. P. Buman, W. L. Haskell, M. V. McConnell, and L. L. Carstensen, "24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices," *Medicine and science in sports and exercise*, vol. 48, no. 3, p. 457, 2016.
[66] T. G. Pavey, S. R. Gomersall, B. K. Clark, and W. J. Brown, "The validity of the geneactiv wrist-worn accelerometer for measuring adult sedentary time in free living," *Journal of science and medicine in sport*, vol. 19, no. 5, pp. 395–399, 2016.
[67] M. Hildebrand, B. H. Hansen, V. T. van Hees, and U. Ekelund, "Evaluation of raw acceleration sedentary thresholds in children and adults," *Scandinavian journal of medicine & science in sports*, vol. 27, no. 12, pp. 1814–1823, 2017.
[68] F. Fraysse, D. Post, R. Eston, D. Kasai, A. V. Rowlands, and G. Parfitt, "Physical activity intensity cut-points for wrist-worn geneactiv in older adults," *Frontiers in Sports and Active Living*, vol. 2, 2020.
[69] "Physical activity guidelines for americans," https://health.gov/sites/default/files/2019-09/Physical_Activity_Guidelines_2nd_edition.pdf.
[70] "General physical activities defined by level of intensity," https://www.cdc.gov/nccdphp/dnpa/physical/pdf/PA_Intensity_table_2_1.pdf.
[71] A. Sadeh, "The role and validity of actigraphy in sleep medicine: an update," *Sleep medicine reviews*, vol. 15, no. 4, pp. 259–267, 2011.
[72] A. Sadeh, M. Sharkey, and M. A. Carskadon, "Activity-based sleep-wake identification: an empirical test of methodological issues," *Sleep*, vol. 17, no. 3, pp. 201–207, 1994.
[73] A. Bamer, K. Johnson, D. Amtmann, and G. Kraft, "Prevalence of sleep problems in individuals with multiple sclerosis," *Multiple Sclerosis Journal*, vol. 14, no. 8, pp. 1127–1130, 2008.
[74] L. B. Krupp, L. A. Alvarez, N. G. LaRocca, and L. C. Scheinberg, "Fatigue in multiple sclerosis," *Archives of neurology*, vol. 45, no. 4, pp. 435–437, 1988.
[75] R. H. Benedict, E. Wahlig, R. Bakshi, I. Fishman, F. Munschauer, R. Zivadinov, and B. Weinstock-Guttman, "Predicting quality of life in multiple sclerosis: accounting for physical disability, fatigue, cognition, mood disorder, personality, and behavior change," *Journal of the neurological sciences*, vol. 231, no. 1-2, pp. 29–34, 2005.
[76] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.
[77] J. H. Petajan and A. T. White, "Recommendations for physical activity in patients with multiple sclerosis," *Sports medicine*, vol. 27, no. 3, pp. 179–191, 1999.
[78] R. W. Motl, E. McAuley, and E. M. Snook, "Physical activity and multiple sclerosis: a meta-analysis," *Multiple Sclerosis Journal*, vol. 11, no. 4, pp. 459–463, 2005.
[79] A. Ng and J. KENT-BRAUN, "Quantitation of lower physical activity in persons with multiple sclerosis," *Medicine & Science in Sports & Exercise*, vol. 29, no. 4, pp. 517–523, 1997.
[80] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.
[81] L. B. Krupp, N. G. LaRocca, J. Muir-Nash, and A. D. Steinberg, "The fatigue severity scale: application to patients with multiple sclerosis and systemic lupus erythematosus," *Archives of neurology*, vol. 46, no. 10, pp. 1121–1123, 1989.
[82] "What is depression?" https://www.psychiatry.org/patients-families/depression/what-is-depression, accessed 07/24/21, https://www.psychiatry.org/patients-families/depression/what-is-depression.
[83] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research," *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.
[84] E. R. Chasens, S. J. Ratcliffe, and T. E. Weaver, "Development of the fosq-10: a short version of the functional outcomes of sleep questionnaire," *Sleep*, vol. 32, no. 7, pp. 915–919, 2009.
[85] https://www.serenitymedicalservices.com/wp-content/uploads/2020/01/CEREVES_FOSQ_10_ENG.pdf, Accessed 07/16/21.

[86] M. Westberg, M. Feychting, F. Jonsson, G. Nise, and P. Gustavsson, "Occupational exposure to uv light and mortality from multiple sclerosis," *American journal of industrial medicine*, vol. 52, no. 5, pp. 353–357, 2009.
[87] B. K. Mehta, "New hypotheses on sunlight and the geographic variability of multiple sclerosis prevalence," *Journal of the neurological sciences*, vol. 292, no. 1-2, pp. 5–10, 2010.
[88] N. M. Wineman, "Adaptation to multiple sclerosis: the role of social support, functional disability, and perceived uncertainty." *Nursing research*, 1990.
[89] M. Krokavcova, J. P. van Dijk, I. Nagyova, J. Rosenberger, M. Gavelova, B. Middel, Z. Gdovinova, and J. W. Groothoff, "Social support as a predictor of perceived health status in patients with multiple sclerosis," *Patient Education and Counseling*, vol. 73, no. 1, pp. 159–165, 2008.
[90] HealthCare.gov, "The 'metal' categories: Bronze, silver, gold platinum," https://www.healthcare.gov/choose-a-plan/plans-categories/, accessed 08/04/21, https://www.healthcare.gov/choose-a-plan/plans-categories/.
[91] X. Shen, S. Ma, P. Vemuri, and G. Simon, "challenges and opportunities with causal discovery algorithms: Application to alzheimer's pathophysiology," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.
[92] W. Chen, Y. Hu, X. Zhang, L. Wu, K. Liu, J. He, Z. Tang, X. Song, L. R. Waitman, and M. Liu, "Causal risk factor discovery for severe acute kidney injury using electronic health records," *BMC medical informatics and decision making*, vol. 18, no. 1, p. 13, 2018.
[93] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.
[94] C. Meek, "Graphical models: Selecting causal and statistical models," Ph.D. dissertation, PhD thesis, Carnegie Mellon University, 1997.
[95] C. for Causal Discovery, "Fast greedy equivalence search (fges) algorithm for continuous variables," https://www.ccd.pitt.edu//wiki/index.php?title=Fast_Greedy_Equivalence_Search_(FGES)_Algorithm_for_Continuous_Variables.
[96] R. Tu, K. Zhang, B. Bertilson, H. Kjellstrom, and C. Zhang, "Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 793–12 804.
[97] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 03 1978. [Online]. Available: https://doi.org/10.1214/aos/1176344136
[98] A. E. Raftery, "Bayesian model selection in social research," *Sociological methodology*, pp. 111–163, 1995.
[99] E. Wit, E. v. d. Heuvel, and J.-W. Romeijn, "'all models are wrong...': an introduction to model uncertainty," *Statistica Neerlandica*, vol. 66, no. 3, pp. 217–236, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2012.00530.x
[100] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[101] R. Swindle, K. Kroenke, and L. Braun, "Energy and improved workplace productivity in depression," *Investing in health: The social and economic benefits of health care innovation*, 2001.
[102] A. Gardner and R. G. Boles, "Mitochondrial energy depletion in depression with somatization," *Psychotherapy and psychosomatics*, vol. 77, no. 2, pp. 127–129, 2008.
[103] T. Roehrs and T. Roth, "Caffeine: sleep and daytime sleepiness," *Sleep medicine reviews*, vol. 12, no. 2, pp. 153–162, 2008.
[104] E. Horwath, J. Johnson, M. M. Weissman, and C. D. Hornig, "The validity of major depression with atypical features based on a community study," *Journal of affective disorders*, vol. 26, no. 2, pp. 117–125, 1992.
[105] F. M. Quitkin, "Depression with atypical features: diagnostic validity, prevalence, and treatment," *Primary care companion to the Journal of clinical psychiatry*, vol. 4, no. 3, p. 94, 2002.
[106] D. Dooley, R. Catalano, and G. Wilson, "Depression and unemployment: panel findings from the epidemiologic catchment area study," *American journal of community psychology*, vol. 22, no. 6, pp. 745–765, 1994.
[107] N. Tsuno, A. Besset, and K. Ritchie, "Sleep and depression." *The Journal of clinical psychiatry*, 2005.
[108] A. Lerdal, E. Celius, and T. Moum, "Fatigue and its association with sociodemographic variables among multiple sclerosis patients," *Multiple Sclerosis Journal*, vol. 9, no. 5, pp. 509–514, 2003.
[109] K. Gebel, D. Ding, T. Chey, E. Stamatakis, W. J. Brown, and A. E. Bauman, "Effect of moderate to vigorous physical activity on all-cause mortality in middle-aged and older australians," *JAMA internal medicine*, vol. 175, no. 6, pp. 970–977, 2015.
[110] T. G. Pavey, G. Peeters, A. E. Bauman, and W. J. Brown, "Does vigorous physical activity provide additional benefits beyond those of moderate?" *Medicine and science in sports and exercise*, vol. 45, no. 10, pp. 1948–1955, 2013.
[111] M. H. Vitaterna, J. S. Takahashi, and F. W. Turek, "Overview of circadian rhythms," *Alcohol Research & Health*, vol. 25, no. 2, p. 85, 2001.
[112] N. L. Benowitz, P. Jacob, R. T. Jones, and J. Rosenberg, "Interindividual variability in the metabolism and cardiovascular effects of nicotine in man." *Journal of Pharmacology and Experimental Therapeutics*, vol. 221, no. 2, pp. 368–372, 1982.
[113] N. L. Benowitz, H. Porchet, L. Sheiner, and P. Jacob III, "Nicotine absorption and cardiovascular effects with smokeless tobacco use: comparison with cigarettes and nicotine gum," *Clinical Pharmacology & Therapeutics*, vol. 44, no. 1, pp. 23–28, 1988.
[114] M. Hirshkowitz, K. Whiton, S. M. Albert, C. Alessi, O. Bruni, L. DonCarlos, N. Hazen, J. Herman, E. S. Katz, L. Kheirandish-Gozal *et al.*, "National sleep foundation's sleep time duration recommendations: methodology and results summary," *Sleep health*, vol. 1, no. 1, pp. 40–43, 2015.

[115] S. D. Barbara Illowsky, *Introductory Statistics*. OpenStax, 2013, https://openstax.org/books/introductory-statistics/pages/12-4-testing-the-significance-of-the-correlation-coefficient.
[116] J. Haines, M. Ter-Minassian, A. Bazyk, J. Gusella, D. Kim, H. Terwedow, M. A. PericakVance, J. Rimmler, C. Haynes, A. Roses *et al.*, "A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex," *Nature genetics*, vol. 13, no. 4, pp. 469–471, 1996.
[117] M. Debouverie, S. Pittion-Vouyovitch, S. Louis, F. Guillemin, and L. Group, "Natural history of multiple sclerosis in a population-based cohort," *European Journal of Neurology*, vol. 15, no. 9, pp. 916–921, 2008.
[118] S. D. Barbara Illowsky, *Introductory Statistics*. OpenStax, 2013, https://openstax.org/books/introductory-statistics/pages/12-3-the-regression-equation.
[119] N. Losseff, S. Webb, J. O'riordan, R. Page, L. Wang, G. Barker, P. S. Tofts, W. I. McDonald, D. H. Miller, and A. J. Thompson, "Spinal cord atrophy and disability in multiple sclerosis: a new reproducible and sensitive mri method with potential to monitor disease progression," *Brain*, vol. 119, no. 3, pp. 701–708, 1996.
[120] D. L. Arnold, G. T. Riess, P. M. Matthews, G. S. Francis, D. L. Collins, C. Wolfson, and J. P. Antel, "Use of proton magnetic resonance spectroscopy for monitoring disease progression in multiple sclerosis," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 36, no. 1, pp. 76–82, 1994.
[121] N. Sola-Valls, Y. Blanco, M. Sepúlveda, E. Martinez-Hernandez, and A. Saiz, "Telemedicine for monitoring ms activity and progression," *Current treatment options in neurology*, vol. 17, no. 11, pp. 1–13, 2015.
[122] C. E. Schwartz, B. R. Quaranto, B. C. Healy, R. H. Benedict, and T. L. Vollmer, "Cognitive reserve and symptom experience in multiple sclerosis: a buffer to disability progression over time?" *Archives of physical medicine and rehabilitation*, vol. 94, no. 10, pp. 1971–1981, 2013.
[123] I. Kister, T. E. Bacon, E. Chamot, A. R. Salter, G. R. Cutter, J. T. Kalina, and J. Herbert, "Natural history of multiple sclerosis symptoms," *International journal of MS care*, vol. 15, no. 3, pp. 146–156, 2013.
[124] B. J. Feir-Walsh and L. E. Toothaker, "An empirical comparison of the anova f-test, normal scores test and kruskal-wallis test under violation of assumptions," *Educational and Psychological Measurement*, vol. 34, no. 4, pp. 789–799, 1974.
[125] "Anova for regression," http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm, http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm, Accessed 07/20/21.
[126] "2.6 - the analysis of variance (anova) table and the f-test," https://online.stat.psu.edu/stat501/lesson/2/2.6, https://online.stat.psu.edu/stat501/lesson/2/2.6, Accessed 07/20/21.
[127] "Analysis of variance (anova)," https://www.investopedia.com/terms/a/anova.asp, https://www.investopedia.com/terms/a/anova.asp, Accessed 07/20/21.
[128] scikit-learn developers, "1.13. feature selection." [Online]. Available: https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection
[129] R. A. Lorenz, J. Koedbangkham, S. Auerbach, N. S. Alanazi, H. Lach, P. Newland, K. Pandey, and F. P. Thomas, "0949 The Relationships between Circadian Rhythm, Sleep Quality, Fatigue, and Depressive Symptoms Among Adults with Multiple Sclerosis (MS)," *Sleep*, vol. 42, no. Supplement_1, pp. A381–A382, 04 2019. [Online]. Available: https://doi.org/10.1093/sleep/zsz067.947
[130] A. J. Dittner, S. C. Wessely, and R. G. Brown, "The assessment of fatigue: a practical guide for clinicians and researchers," *Journal of psychosomatic research*, vol. 56, no. 2, pp. 157–170, 2004.
[131] J. Backhaus, K. Junghanns, A. Broocks, D. Riemann, and F. Hohagen, "Test–retest reliability and validity of the pittsburgh sleep quality index in primary insomnia," *Journal of psychosomatic research*, vol. 53, no. 3, pp. 737–740, 2002.
[132] S. B. Patten, J. M. Burton, K. M. Fiest, S. Wiebe, A. G. Bulloch, M. Koch, K. S. Dobson, L. M. Metz, C. J. Maxwell, and N. Jetté, "Validity of four screening scales for major depression in ms," *Multiple Sclerosis Journal*, vol. 21, no. 8, pp. 1064–1071, 2015.
[133] L. B. Strober and P. A. Arnett, "Depression in multiple sclerosis: The utility of common self-report instruments and development of a disease-specific measure," *Journal of clinical and experimental neuropsychology*, vol. 37, no. 7, pp. 722–732, 2015.
[134] J. R. Berger, J. Pocoski, R. Preblick, and S. Boklage, "Fatigue heralding multiple sclerosis," *Multiple Sclerosis Journal*, vol. 19, no. 11, pp. 1526–1532, 2013.
[135] P. O. Valko, C. L. Bassetti, K. E. Bloch, U. Held, and C. R. Baumann, "Validation of the fatigue severity scale in a swiss cohort," *Sleep*, vol. 31, no. 11, pp. 1601–1607, 2008.
[136] H. M. Bøe Lunde, T. F. Aae, W. Indrevåg, J. Aarseth, B. Bjorvatn, K.-M. Myhr, and L. Bø, "Poor sleep in patients with multiple sclerosis," *PloS one*, vol. 7, no. 11, p. e49996, 2012.
[137] H. Zhang, G. Guo, C. Song, C. Xu, K. Cheung, J. Alexis, H. Li, D. Li, K. Wang, and W. Xu, "Pdlens: smartphone knows drug effectiveness among parkinson's via daily-life activity fusion," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.
[138] H. Zhang, C. Song, A. Wang, C. Xu, D. Li, and W. Xu, "Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–16.

[139] L. Midaglia, P. Mulero, X. Montalban, J. Graves, S. L. Hauser, L. Julian, M. Baker, J. Schadrack, C. Gossens, A. Scotland, F. Lipsmeier, J. van Beek, C. Bernasconi, S. Belachew, and M. Lindemann, "Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study," *J Med Internet Res*, vol. 21, no. 8, p. e14863, Aug 2019. [Online]. Available: http://www.jmir.org/2019/8/e14863/

[140] C. Tong, M. Craner, M. Vegreville, and N. D. Lane, "Tracking fatigue and health state in multiple sclerosis patients using connnected wellness devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, Sep. 2019. [Online]. Available: https://doi.org/10.1145/3351264

[141] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[142] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell, "Smartgpa: how smartphones can assess and predict academic performance of college students," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 295–306.

[143] C. E. Mcculloch and J. M. Neuhaus, *Generalized Linear Mixed Models Based in part on the article "Generalized linear mixed models" by Charles E. McCulloch, which appeared in the Encyclopedia of Environmetrics.* American Cancer Society, 2013. [Online]. Available: https://www.onlinelibrary.wiley.com/doi/abs/10.1002/9780470057339.vag009.pub2

[144] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[145] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[146] J. W. Hardin and J. M. Hilbe, *Generalized estimating equations.* chapman and hall/CRC, 2002.

[147] F. Tsapeli and M. Musolesi, "Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach," *EPJ Data Science*, vol. 4, no. 1, p. 24, 2015.

[148] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[149] J. H. Zar, "Spearman rank correlation," *Encyclopedia of biostatistics*, vol. 7, 2005.

[150] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[151] "Neurodegenerative diseases," https://www.niehs.nih.gov/research/supported/health/neurodegenerative/index.cfm, accessed 07/26/21, https://www.niehs.nih.gov/research/supported/health/neurodegenerative/index.cfm.

[152] "The promise of precision medicine," https://www.nih.gov/about-nih/what-we-do/nih-turning-discovery-into-health/promise-precision-medicine, accessed: 05/04/2021.

[153] G. Bose and M. S. Freedman, "Precision medicine in the multiple sclerosis clinic: Selecting the right patient for the right treatment," *Multiple Sclerosis Journal*, vol. 26, no. 5, pp. 540–547, 2020.

[154] M. R. Hansen and D. T. Okuda, "Precision medicine for multiple sclerosis promotes preventative medicine," *Annals of the New York Academy of Sciences*, vol. 1420, no. 1, pp. 62–71, 2018.

[155] T. Chitnis and A. Prat, "A roadmap to precision medicine for multiple sclerosis," *Multiple Sclerosis Journal*, vol. 26, no. 5, pp. 522–532, 2020.