Sex Differences in Variability of Physical Activity Measurements Across Multiple Timescales Recorded by a Wearable Device

Kristin J Varner, Lauryn Keeler Bruce, Severine Soltani, Wendy Hartogensis, Stephan Dilchert, Frederick M Hecht, Anoushka Chowdhary, Leena Pandya, Subhasis Dasgupta, Ilkay Altintas, Amarnath Gupta, Ashley E Mason, Benjamin L Smarr

Submitted to: Journal of Medical Internet Research
on: September 07, 2024

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Sex Differences in Variability of Physical Activity Measurements Across Multiple Timescales Recorded by a Wearable Device

Kristin J Varner[1] MS; Lauryn Keeler Bruce[2, 3] MS, PhD; Severine Soltani[3] BS; Wendy Hartogensis[4] MPH, PhD; Stephan Dilchert[5] PhD; Frederick M Hecht[4] MD; Anoushka Chowdhary[4] BS; Leena Pandya[4] PhD; Subhasis Dasgupta[6] PhD; Ilkay Altintas[7, 6] PhD; Amarnath Gupta[7, 6] PhD; Ashley E Mason[4*] PhD; Benjamin L Smarr[7, 1*] PhD

[1] Shu Chien-Gene Lay Department of Bioengineering University of California San Diego La Jolla US
[2] UC San Diego Health Department of Biomedical Informatics University of California San Diego La Jolla US
[3] Department of Bioinformatics and Systems Biology University of California San Diego La Jolla US
[4] Osher Center for Integrative Health University of California San Francisco San Francisco US
[5]Department of Management Zicklin School of Business Baruch College, The City University of New York New York US
[6] San Diego Supercomputer Center University of California San Diego La Jolla US
[7] Hal?c?o?lu Data Science Institute University of California San Diego La Jolla US
[*]these authors contributed equally

Corresponding Author:
Benjamin L Smarr PhD

Hal?c?o?lu Data Science Institute
University of California San Diego
3234 Matthews Ln
La Jolla
US

Abstract

Background: Biological sex is an important consideration in biomedical research, yet females are still underrepresented in both human and animal biomedical research. Hesitancy to include female subjects is partially due to the hypothesis that biological rhythms driven by menstrual cycles, and occurring on the timescale of roughly 28 days, increase biological variability and weaken statistical power.

Objective: We aimed to determine if variability of physical activity (PA) is affected by biological sex, and if so, whether having menstrual cycles (as indicated by temperature rhythms) contributes to increased female PA variability. We then sought to compare the effect of sex and menstrual cycles on PA variability to the effect of PA rhythms on the timescales of days and weeks and to the effect of non-rhythmic temporal structure in PA on the timescale of decades of life (age).

Methods: We used minute-level metabolic equivalent task (MET) data collected using a wearable device across a 206-day study period for each of 596 individuals as an index of physical activity (PA) to assess the magnitudes of variability in PA accounted for by biological sex and temporal structure on different timescales. We represented intraindividual variability in PA with consecutive disparity index (CDI).

Results: Females (regardless of whether they had menstrual cycles) demonstrated lower intraindividual variability in PA than males (Kruskal-Wallis, H=29.51, P<.001). Furthermore, people with menstrual cycles did not have greater intraindividual variability than people without menstrual cycles (Kruskal-Wallis, H=0.54, P=.46). PA rhythms differed at the weekly timescale: individuals with increased or decreased PA on weekends had larger intraindividual variability (Kruskal-Wallis, H=10.13, P=.001). Additionally, intraindividual variability differed by decade of life, with older age groups tending to have less variability in PA (Kruskal-Wallis, H=40.55, P=1x10-7, Bonferroni corrected significance threshold for 15 comparisons: P=3x10-3). A generalized additive model (GAM) predicting CDI of 24-hour MET sums (variability of PA) showed that sex, age, and weekly rhythm accounted for only 11% of PA variability.

Conclusions: The exclusion of people from biomedical research based on their biological sex or the presence of menstrual cycles is not supported by our analysis. Menstrual cycles did not significantly affect female PA variability. Temporal structures in PA on other timescales had significant effects on both female and male PA. Our findings highlight the potential for emerging longitudinal data sources to allow for phenotyping of individuals by their temporal structure on relevant timescales. This may improve precision in statistical and machine learning models as an alternative to excluding any groups.

(JMIR Preprints 07/09/2024:66231)
DOI: https://doi.org/10.2196/preprints.66231

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Kristin J Varner[1], MS; Lauryn Keeler Bruce[2,3], MS; Severine Soltani[3], BS; Wendy Hartogensis[4], MPH, PhD; Stephan Dilchert[5], PhD; Frederick M Hecht[4], MD; Anoushka Chowdhary[4], BS; Leena Pandya[4], PhD; Subhasis Dasgupta[6], PhD; Ilkay Altintas[6,7], PhD; Amarnath Gupta[6,7], PhD; Ashley E Mason[4]*, PhD; Benjamin L Smarr[1,7]*, PhD

1. Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, San Diego, California, United States of America
2. UC San Diego Health Department of Biomedical Informatics, University of California San Diego, San Diego, California, United States of America
3. Bioinformatics and Systems Biology, University of California San Diego, San Diego, California, United States of America
4. Osher Center for Integrative Health, University of California San Francisco, San Francisco, CA, USA
5. Department of Management, Zicklin School of Business, Baruch College, The City University of New York, New York, NY, USA
6. San Diego Supercomputer Center, University of California San Diego, San Diego, California, United States of America
7. Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, California, United States of America

*These authors contributed equally: Ashley E. Mason and Benjamin L Smarr.

Corresponding author: Benjamin L Smarr
Address: 9500 Gilman Drive, La Jolla, CA, 92093, US
Phone: (858) 822-4536
Email: bsmarr@ucsd.edu

Sex Differences in Variability of Physical Activity Measurements Across Multiple Timescales Recorded by a Wearable Device

Abstract

Background: A significantly lower proportion of females participate in sufficient daily activity when compared to males despite the known health benefits of exercise. Investment in female sports and exercise medicine research may help close this gap, yet females are underrepresented in this research. Hesitancy to include female subjects is partially due to assumptions that biological rhythms driven by menstrual cycles, and occurring on the timescale of roughly 28 days, increase intraindividual biological variability and weaken statistical power. An analysis in continuous skin temperature data measured using a commercial wearable device found that temperature cycles indicative of menstrual cycles did not substantially increase variability in female skin temperature. Here we explore physical activity (PA) data as a variable more related to behavior, whereas temperature is more reflective of physiological changes.

Objective: We aimed to determine if intraindividual variability of PA is affected by biological sex, and if so, whether having menstrual cycles (as indicated by temperature rhythms) contributes to increased female intraindividual PA variability. We then sought to compare the effect of sex and menstrual cycles on PA variability to the effect of PA rhythms on the timescales of days and weeks, and to the effect of non-rhythmic temporal structure in PA on the timescale of decades of life (age).

Methods: We used minute-level metabolic equivalent task (MET) data collected using a wearable device across a 206-day study period for each of 596 individuals as an index of PA to assess the magnitudes of variability in PA accounted for by biological sex and temporal structure on different timescales. We represented intraindividual variability in PA with consecutive disparity index (CDI).

Results: Females (regardless of whether they had menstrual cycles) demonstrated lower intraindividual variability in PA than males (Kruskal-Wallis, H=29.51, *P*<.001). Furthermore, people with menstrual cycles did not have greater intraindividual variability than people without menstrual cycles (Kruskal-Wallis, H=0.54, *P*=.46). PA rhythms differed at the weekly timescale: individuals with increased or decreased PA on weekends had larger intraindividual variability (Kruskal-Wallis, H=10.13, *P*=.001). Additionally, intraindividual variability differed by decade of life, with older age groups tending to have less variability in PA (Kruskal-Wallis, H=40.55, *P*=1x10^{-7}, Bonferroni corrected significance threshold for 15 comparisons: *P*=3x10^{-3}). A generalized additive model (GAM) predicting CDI of 24-hour MET sums (intraindividual variability of PA) showed that sex, age, and weekly rhythm accounted for only 11% of the population variability in intraindividual PA variability.

Conclusions: The exclusion of people from PA research based on their biological sex, age, the presence of menstrual cycles, or the presence of weekly rhythms in PA is not supported by our analysis.

Keywords: Wearables; Activity; Sex as a Biological Variable; Time Series Variance; Timescales of Change; Metabolic Equivalents; Metabolic Equivalent Task; Sex Differences

Introduction

Regular physical activity (PA) compared to inactivity has been associated with a lower risk of all-cause mortality in both males and females [1]. Yet, a meta-analysis reported that PA decreased in several nations between 1995-2017 [2]. While this decrease has occurred equally in males and females, females are less likely to participate in sufficient exercise than males [3–5]. An evaluation of insufficient activity (participating in at least 150 minutes of moderate-intensity or 75 minutes of vigorous-intensity PA per week) from 1.9 million participants found that 27.5% of participants did not participate in sufficient activity where women had significantly higher rates of inactivity than men (31.7% vs. 23.4%)[3]. As females have been shown to derive greater risk reduction than males for an equivalent increase in exercise [1], it is important to identify the causes of the sex/gender gap in PA. While the reasons for the gap are not well understood [5], it has been attributed to many factors including children's exposure to rigid gender norms, women's concerns about stereotypes, lack of leisure time, and importantly, lack of investment in women's and girl's sports [4]. These knowledge gaps pervade sports and exercise science research. An analysis of three major sports and exercise medicine journals over three years (2011-2013) found that just 39% of participants in 1382 original research articles were female [6]. A subsequent analysis analyzing 5,621 studies from six sport and exercise journals (inclusive of the three journals in the previous study) examined the 7 years following the previous study (2014-2020) and reported a lower proportion of total female participants (34%) and a significantly higher number of studies including only male (~1630, 31%) versus only females (~315, 6%) participants [7]. Exclusion of female participants from sports and exercise medicine studies is partially attributed to the assumption that ovarian hormones (or menstrual cycles) increase intraindividual PA variability in females thereby increasing the difficulty in interpreting results (due to increased intraindividual variability contributing to greater interindividual variability) or complicating methodology to account for changes in ovarian hormones [8–11]. This assumption also implies that results generated by male subjects are generalizable to females: if male and female baseline physiology is the same but females have more intraindividual variability, they increase population-level (interindividual) variability, decrease statistical power, and their inclusion provides no benefit. However, the hypothesis that male results generalize to females (or that they have the same baseline physiology) has repeatedly been shown to be false[1,12–14]. This in itself should motivate the inclusion of females, but as female participation in sports and exercise research is still low relative to male participation [6,7], it is important to assess the extent to which menstrual cycles and other biological and social rhythms interfere with researchers' ability to analyze PA. Building on previous work exploring physiological variability from distal skin temperature measured by a commercial wearable [15], here we explore the intraindividual variability in physical activity (PA) between sexes using longitudinal PA measurements from 298 males and 298 females who were using Oura Rings during 2020.

Numerous animal studies have rejected the hypothesis that females are more variable in both physiology and behavior [16–19], but far fewer investigations of this hypothesis have been performed in humans [15,20]. This is in part due to historical difficulty in generating longitudinal data sets that were also big enough to be representative of both sexes broadly. The emergence of digital tools such as wearable devices (wearables) in daily life has led to a rapid change in the amount of longitudinal data that can be easily generated on individual study subjects. Data from wearables provides unique opportunities to explore physiological and behavioral variability between sexes both across populations and within individual time series data [21].

In our previous work, we used continuous longitudinal distal skin temperature data generated by Oura Ring users *in situ*, to test the hypothesis that females are statistically more physiologically variable than males [15]. Temperature was chosen as prior work indicates that skin temperature can be used to identify physiological changes, such as a 28-day oscillating skin temperature pattern generated by menstrual cycles [22,23]. Using a data set of minute-level skin temperature data from 600 individuals (300 males, 300 females) over six months, we developed a tool capable of determining cyclic status, where female data which showed a roughly 28-day pattern in nightly maximum temperature were labeled as cyclic, and those without were labeled as acyclic. We also found that cyclic and acyclic individuals, whether female or male, showed substantially different patterns of change over time, such that cyclic status was a more informative label than sex when predicting the structure of variability in an individual's skin temperature over time.Our analyses led us to reject the hypothesis that females, cyclic or acyclic, should be excluded due to concerns over statistical power, even though our findings also supported the use of sex as a biological variable (SABV) in analyses. That is, body temperature changes linked to menstrual cycles [24] were present in a subset of individuals who self-reported as biological females, and while the variability was not substantially greater at multiple timescales in any of these groups, the means and the temporal structure of temperature predictably differed by biological sex and cyclic status. Here we seek to recapitulate these analyses on the same population but assessing PA, as this measure is less closely tied to hormonal changes physiologically, and instead more reflective of behavioral changes.

Previous studies have demonstrated that multiple timescales of change can interact to give rise to non-random structure in intraindividual variability of human time series data [15,20,25]. This temporal structure arises specifically from interactions between physiological rhythms, such as menstrual and circadian rhythms, societal phenomena such as the 7-day work week, and non-rhythmic temporal scales such as aging. To the extent that variability is non-random, it is by definition at least partially predictable. If not accounted for in experimental design, then non-random (unaccounted) variability will be combined with random (unaccountable) variability to the effect that statistical tests—by treating all sources of variability as equivalent—will yield reduced power for detecting real effects. By contrast, when non-random variability is accounted for, residual variability is by definition lower, and statistical power is improved for the same analysis. Even while sources and structures of male variability are not well characterized[13], the contribution these other timescales of change impart on variability, is not often considered; without a direct comparison, we cannot know impactful these other timescales of change are to PA analyses as compared to the effects of menstrual cycles.

Here we used the same cohort of subjects as in our previous analysis of temperature [15] to assess the effect of sex, cyclic status, and temporal structures in PA on other timescales of change on intraindividual PA variability. Specifically, we seek to determine if the presence of roughly 28-day cyclic temperature patterns we previously identified correlates with increased intraindividual variability in PA measurements, and to quantify the extent that these roughly 28-day cycles affect statistical analysis of PA. Additionally, we seek to ascertain if temporal structure occurring on other timescales besides menstrual cycles (*e.g.*, weeks and decades) contribute to intraindividual PA variability. Oura Ring reports activity in the form of metabolic equivalent tasks (METs) [26] where METs express the intensity of an activity as multiples of the MET recorded at rest [27]. Using these measurements, we quantified individual daily PA and intraindividual variability in PA and found that biological sex, cyclic status, and weekly and decadal temporal structures in PA do not explain most of the intraindividual variability in PA.

Methods

Data source

Data originated from the TemPredict Study [26]. Physiological data were collected using the wearable device Oura Ring (Oura Health Oy, Oulu, Finland), and self-reported demographic information such as sex and age were collected via survey.

Subjects

Subjects were identified by filtering methods described in "Variability of temperature measurements recorded by a wearable device by biological sex" [15]. Briefly, 62,653 subjects were determined to have suitable physiological and demographic data. Responses to the survey question 'What is your biological sex? Male, Female, Other (please describe)." was used to determine participants' sex.

Filtering for subjects with data files for all data types and for whom temperature data were available for all months between January and November 2020 narrowed the subject number to 7,915. Next, subjects who had less than 70% average daily completeness in temperature were eliminated. We chose to filter out subjects with less than 70% average daily completeness to increase the likelihood that both sleep and wake states were captured in the data (sleep usually covers ~33% of a day). A cohort of 600 individuals was chosen from the final list such that 50 individuals of each sex were present in six 10-year age bins spanning 20 to 79 years old.

Additional filtering of the subjects was performed for this analysis. The lower limit of real MET recordings is 0.9, which occurs when a person is asleep [28]. All MET values below 0.9 were dropped (due to non-wear time artifacts) and participants were evaluated for missingness over 206 days between April and October 2020. Four participants, two of each sex, with a percent missingness of MET data above 29% were removed (see Supplementary Figure 1). The final data consisted of 206 consecutive days for 596 individuals: 298 females and 298 males. Six age bins were represented equally with 49-50 individuals of each sex in each age bin: 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79.

Data preprocessing

High resolution (per minute and per five minutes) and nightly aggregated data were generated by the Oura Ring. Data was stored in large parquet files on the San Diego supercomputer (SDSC) and accessed through the Nautilus Portal [29]. We expected MET to vary by sleep state (whether an individual is awake or asleep), therefore we labeled minute-level data with asleep/awake labels. Nightly data, also referred to as sleep summary data, were stored as a single parquet file for each participant. These data contained sleep-related information such as sleep time start and sleep time end. The longest sleep duration for each day was used to label measurements as asleep. All other times were labeled as awake.

High-resolution distal body temperature and metabolic equivalent task (MET) data were recorded at 1-minute intervals for 24 hours per day. These data were date-time indexed and normalized to subjects' local time. Duplicate time points were removed and the remaining time points were annotated as awake or asleep.

MET was calculated by Oura Ring before data were transferred to us for analysis. Tri-axial accelerometers were used to estimate metabolic equivalents (MET) at 60s resolution during both sleep and wake periods [26]. The exact MET calculation is proprietary to Oura Ring and not known to us; however, Oura Ring (Gen 2) activity measurements displayed high correlation when previously validated against multiple accelerometers [30].

Data filling

Missing sleep state data and MET data were filled for all 596 participants. Sleep state data described the sleep state (awake or asleep) at every minute for every participant. MET data contained the MET value at every minute for every participant.

To limit the artifacts resulting from filling, we assessed the accuracy of four filling methods on several intervals of missingness. An interval of missingness describes the number of consecutive minutes for which there are missing values (i.e., an interval of 1440 describes a full missing day). The intervals tested were 5, 10, 20, 40, 80, 160, 320, 640, 1280, and 1440 minutes. The filling methods tested were: 1) a phase-dependent filler, 2) linear interpolation, 3) global personal mean filling, and 4) zero filling (or NaN-fill).

1. The phase-dependent filler constructs a 'median week' from the median value of each minute on each day of the week across half of the dataset (103 days) for each subject (2 median weeks per subject). If no median value exists for a minute in the constructed median week, a value was forward-filled from the median value of the preceding minute. The minutes without data in the 103-day period from which the week of median values was constructed were filled based on the minute and day of the week in which they occurred.
2. Linear interpolation was achieved with the interpolate method from the Python package pandas (*pandas.DataFrame.interpolate*, version 2.2.1)[31]. A two-way limit direction was used such that missing data from the first minute in the data could be filled.
3. The global personal median method finds the median value for each person across the

entire dataset and fills the missing values with this median value.

4. The zero-filling method fills all missing values with zero. This method was included because the sum of MET values was used to summarize daily activity. Zero fill equates to the effect of not filling these values because NaNs are treated as zeros during daily summation (*i.e.* not a number or NaN-fill).

To test the accuracy of the filling methods on each interval, a test data frame was constructed. For each participant, simulated missing data were constructed by inserting intervals of missingness starting at randomly chosen minutes. Each participant had 3,995 extra missing data points composed of 5, 10, 20, 40, 80, 160, 320, 640, 1280, and 1440 length intervals of missingness. The intervals were allowed to overlap and occur on the same day. The simulated intervals of missingness were then filled using each of the four filling methods. After filling, the predicted values in the sleep state data frame were rounded to zero or one to reflect a prediction of awake or asleep.

The performance of each method for each person on each interval size was evaluated by the sum of the absolute differences between the predicted and actual values of the test indexes. Because some participants did not have enough data, some simulated missing data had indeterminate error (the ‘actual’ value was missing): 0.25% of the simulated missing data in the MET filling test had indeterminate error and 0.49% of the simulated missing data in the sleep state filling test had indeterminate error. The best method for each interval size was determined by the smallest sum of absolute differences across all individuals. In the MET dataset, the best method for intervals of missingness of size less than or equal to 40 minutes was linear interpolation and for intervals with size greater than 40 minutes the best method was phase-dependent filler (Error data shown in Supplementary Figure 2). In the sleep state dataset, the best method for intervals of missingness less than or equal to 320 minutes was linear interpolation and for intervals of missingness greater than 320 minutes the best method was phase-dependent filler (Error data shown in Supplementary Figure 3). The best filling method for each interval of missingness was applied to each dataset before any analyses were performed.

The sum of absolute differences across all test intervals (filling error) was not significantly different between males, cyclic females, and acyclic females in the sleep state and MET data tests (Kruskal-Wallis: MET: H=1.97 *P*=.37, Supplementary Figure 4. Sleep State: H=0.26, *P*=.88, Supplementary Figure 5).

Filled data were used for every analysis described below, except where explicitly described to not have been used (see Analysis by Weekend Rhythm in Physical Activity).

Statistical Methods

Kruskal Wallis H tests, Bonferroni correction, and post-hoc Dunn’s tests

Population differences were determined using a Kruskal-Wallis H test between population distributions of the relevant metric (mean, standard deviation, etc). Python was used to carry out all Kruskal-Wallis tests (SciPy Python library, *scipy.stats.kruskal*, version 1.11.2)[32]. In the case that three or more populations were compared, a Bonferroni correction was manually applied to all analyses that compared more than two groups, such that the threshold for significance (*P*=.05) was divided by the number of comparisons made. If the significance threshold was met and groups were compared with a single Kruskal-Wallis test, a post-hoc Dunn’s test was performed using Python (scikit-posthocs Python package, *scikit_posthocs.posthoc_dunn*, version 0.9.0)[33] to identify the pair-wise population comparisons that met significance. Although the shape of distributions for male subjects tended to be wider than distributions for females, median values were used to determine the population with the larger metric. The results from these tests and/or the distributions compared with these tests were shown in most of the figures and tables (Figure 1C-E, Figure 2 A-D, Figure 3D, Figure 4A-B, Supplemental Figures 2-5, Tables 1-6) Population standard deviations of the subpopulations described were calculated for their relevance to power analysis (Supplementary Table 1-2).

Modified Cohen's d

As the distributions in these analyses were non-normal, a modified Cohen's *d* effect size (d_m) was used to describe the magnitude of the difference between two significantly different populations (shown as P1 and P2 here)[34]:

$$d_m = (|\text{median}(P1) - \text{median}(P2)|) / (\text{mean}(\text{IQR}(P1), \text{IQR}(P2))) \quad \mathbf{(1)}$$

where IQR(P#) represents the interquartile range of the population (IQR- the difference between the 75th and 25th percentile values). This modification to the Cohen's *d* effect size compares medians instead of means and IQR instead of standard deviation to accommodate calculations appropriate for skewed distributions.

The modified Cohen's *d* effect size approximates the proportion of population variability accounted for by a characteristic (sex, age, etc). For example, if d_m =1, the difference in the medians is equal to the mean of the two population IQRs which means that there is little overlap of values and the characteristic accounts for a significant proportion of the variability between those populations. Modified Cohen's *d* was calculated between subpopulations that were significantly different by either a Kruskal-Wallis or post-hoc Dunn's test (Figure 1C, Figure 2C-D, Figure 3D, Figure 4B, Tables 2-5).

Effect of subpopulations

To determine if a subpopulation contributes a significant amount of variability to a whole population, we first identified two groups of subjects: the whole population and the whole population excluding the subpopulation of interest. The second group is itself a subset of the whole population, which makes statistical comparisons problematic: the whole population contains every value in the subset. To avoid making comparisons between identical values, we calculated the interquartile ranges of the 24-hour MET sums for each day for each group. This generated two lists of 206 IQRs representing each group's variability across the 206 days in this study. The two lists were compared with a Kruskal-Wallis test to evaluate whether a whole population changed when a subpopulation was excluded. If the whole population had significantly larger IQRs than the whole population with the subpopulation of interest excluded, then the subpopulation was considered to have imparted a significant amount of variability on the whole population. This test was performed on distributions shown in Figure 2D, Figure 3D, and results from this test are shown in Table 6. If a subpopulation did impart a significant amount of variability on the whole population, we used Lehr's rule to calculate the difference in sample size required to detect the same effect (with 80% power and significance level 0.05) when the group was included or excluded[35]:

$$n = 16(s^2)/(\mu_1 - \mu_2)^2 \quad \mathbf{(2)}$$

where n is the sample size required, s^2 is the variance of the population tested, and (μ_1 - μ_2) is the difference in means between each population. We used the median IQR across all 206 days as a proxy for s and tested multiple values for (μ_1 - μ_2): 40 (approximately the difference in 24-hour MET sums resulting from a 20-minute walk), 100 (approximately the difference in 24-hour MET sums resulting from 20-minutes of moderate intensity activity) and 180 (approximately the difference in 24-hour MET sums resulting from 20-minutes of high intensity of activity). We chose these values to represent a difference that may be significant to health.

Kernel density estimate plots

Kernel density estimate plots were used to ensure distributions were visually comparable despite differences in group size and to enable comparisons of idealized distributions. Plotting was performed in Python using the seaborn library (*seaborn.kdeplot*, version 0.12.2)[36]) with the default kernel (Gaussian) and bandwidth smoothing method (Scott's Rule). The bandwidth scaling parameter (bw_adjust) was adjusted per distribution to create visually smoother plots and estimation ranges were limited to real values. Kernel density estimate plots are displayed in Figure 2D, Figure 3D, and Figure 4B.

Cohort and MET Data Foundational Analysis

To visually inspect the effect of time of day on activity, a random subset of 20 consecutive days of data from two randomly selected individuals of each sex was chosen to represent a MET value time series and distribution (Figure 1A and Figure 1B). Finding that MET values were highly dependent on awake or asleep state as expected, MET values were summed for each day (206 total) over 24 hours, awake time states, and asleep time states to summarize the total daily physical activity (PA) for each person in each state. These states were considered separately in subsequent analyses because the source of variability of daily MET sums is different in each state. We considered five drivers of variability including awake movement, intentional exercise, sleep movement, time spent asleep, and time spent awake. The first three drivers of variability are associated with a state (awake/asleep) and a MET range. Sleep movement occurs while asleep and at a MET above 0.9 (sleep results in a MET value of 0.9[28]), awake movement occurs while awake and at a MET between 1.0 and 1.5 (resting while awake results in a MET value of 1.0 and intentional exercise results in a MET value greater than 1.5 [28]), and intentional exercise occurs while awake and at a MET above 1.5. Time spent awake and time spent asleep refer to the number of minutes per day that a person spends awake and asleep. In contrast to 24-hour MET sums, where the number of values being summed is always 1440 (24 hours x 60 minutes), awake and asleep daily MET sums vary by the number of values being summed per day due to varying amounts of time spent in those states each day. The possible sources of variability in 24-hour sums are sleep movement, awake movement, and intentional exercise. The possible sources of variability in awake daily sums are time spent awake, intentional exercise, and awake movement. The possible sources of variability in asleep daily sums are sleep duration and movement while asleep.

A PA summary of all participants across all 206 days was constructed from the mean and standard deviation of the 206 daily 24-hour MET sums. Individuals in each sex population were sorted by the mean of 24-hour MET sums and represented as a point and line representing +/- one intraindividual standard deviation, such that individuals at the same rank in each population could be compared. Noticing a divergence between the populations in the individuals with the largest means, we performed a Kruskal-Wallis test between the top 60 males and the top 60 females (Figure 1C).

Whole population distributions of male and female mean and standard deviation across all 206 days for 24-hour, awake, and asleep MET sums were compared using a Kruskal-Wallis test with a Bonferroni correction for three comparisons (three MET sum metrics each for mean and standard deviation) (Figure 1D and Figure 1E).

Variability Metrics of MET Sums

In addition to standard deviation, we used three other metrics to analyze intraindividual variability: coefficient of variation (CV), proportional variability index (PV), and consecutive disparity index (CDI). In prior work, we used CV and PV as controls to validate the statistical findings from the CDI analyses [15]. We included CV and PV here for the same validation and focused on CDI because it is the most appropriate metric of intraindividual variability for these data because it accounts for chronological order and is not dependent on the mean for its calculation. Further analyses used only CDI as a variability metric. Whole population distributions of male and female CV, PV, and CDI across all 206 days for 24-hour, awake, and asleep MET sums were compared using a Kruskal-Wallis test with a Bonferroni correction for three comparisons (three MET sum metrics each for CV, PV, and CDI).

Coefficient of variation (CV)

CV is a common metric for describing temporal variability [37]. Here it describes a participant's standard deviation(σ) across all 206 days relative to their mean across all 206 days;

$$CV = \sigma / \text{mean} \quad \textbf{(3)}$$

CV is limited by its sensitivity to rare events and its dependence on the mean [37] (Figure 2A, Table 2).

Proportional variability index (PV)

The proportional variability index (PV) was developed to solve some of the limitations of CV. PV quantifies variability by calculating the average percent difference between all combinations of measurements [37–40];

$$PV = 2(\Sigma(1-(\min(z_i, z_j)/\max(z_i, z_j))) / n(n-1) \quad \textbf{(4)}$$

where n = total number values, z = a list of values on which pairwise comparisons are calculated, i and j = indices of any two different values. PV improves upon CV because it is not mean-dependent and it is less sensitive to rare events [41] (Figure 2B, Table 2).

Consecutive disparity index (CDI)

The consecutive disparity index (CDI) was developed to improve PV by accounting for the chronological order of measurements in a time series [41]. CDI describes time series variability through the average rate of change between consecutive values;

$$CDI = (1/(n-1))\ \Sigma^{n-1}_{i=1} |\ln(p_{i+1}/p_i)| \quad \textbf{(5)}$$

where n = length of time series and p_i = value in series at time i [41] (Figure 2C-D, Figure 3D, Figure 4A-B, Figure 5A-E, Table 2-5).

Analysis of Physical Activity by Cyclic Status

Every participant's cyclic status (a label of cyclic describes the presence of a roughly 28-day temperature rhythm generated by menstrual cycles) was determined through methods described in “Variability of temperature measurements recorded by a wearable device by biological sex” [15]. Briefly, autocorrelation profiles were generated from nightly maximum temperature recordings (not shown). Only cyclic individuals' temperature trend deviation autocorrelation signals show wave-like structure. Profiles were classified as cyclic or acyclic by hierarchical clustering of pairwise distances between signals (pairwise distances calculated with dynamic time warping)(not shown). Hierarchical clustering classified 193 females in this cohort as acyclic, 105 females as cyclic, and all but one male as acyclic. The temperature trend deviation autocorrelation signal for the cyclic classified male individual did not show a wave-like structure and the male was manually reclassified as acyclic. 102 out of the 105 females classified as cyclic were between the ages of 20 and 49 and 3 were between 50 and 59. 48 of the 193 females that were classified as acyclic were under the age of 50, 27 were under the age of 40, and 46 between 50 and 59.

Analysis of physical activity by cyclic status focused on the CDI variability metric and daily 24-hour MET sum metric. 24-hour MET sums were chosen for analysis to focus on the overall variability that is due to PA, in contrast to asleep or awake sums that vary with time spent in the state, as described in the ‘Cohort and MET Data Foundational Analysis’. The CDI variability metric was chosen due to its accounting for chronological order, as described in the Variability Metrics of MET Sums section.

The autocorrelation and clustering techniques used to classify subjects as cyclic or acyclic were also used to determine if cyclic people had unique structures in daily 24-hour MET sums such as a 28-day structure.

Mean and CDI of 24-hour MET sums were calculated for each individual over all 206 days present in the data and compared across cyclic status (cyclic females vs all acyclic individuals of either sex, Kruskal-Wallis test). CDI of 24-hour MET sums were also compared across groups of individuals with unique combinations of sex and cyclic status (acyclic male, cyclic female, acyclic female. Kruskal-Wallis test with Bonferroni correction for three comparisons and post-hoc Dunn's test (Figure 2D). Cyclic and acyclic females of the same age were compared to control for the uneven age distributions between the two groups (cyclic females aged 20-59 vs acyclic females aged 20-59 and cyclic females aged 20-49 vs acyclic females aged 20-49, Kruskal-Wallis test). The effect of cyclic females on the variability of the whole female population was calculated using IQR distributions as described in the ‘Statistical Methods’ section.

Analysis by Weekend Rhythm in Physical Activity

Analysis by weekend rhythm in PA focused on the CDI variability metric and daily 24-hour MET sum metric. 24-hour MET sums were chosen for analysis to focus on the overall variability that is due to PA, in contrast to asleep or awake sums that vary with time spent in the state, as described in the ‘Cohort and MET Data Foundational Analysis’ methods section. The CDI variability metric was chosen due to its accounting for chronological order, as described in the ‘Variability Metrics of MET Sums’ section.

To determine if PA rhythms existed on a weekly timescale, we examined a hierarchically clustered heatmap (seaborn Python library, *seaborn.clustermap*, version 0.12.2)[36]) of unfilled and intraindividual z-scored 24-hour MET sum data (not shown). Hierarchical clustering of unfilled (non-imputed) data ensured that clustered structures were not artifacts of data filling (*e.g.*, the median week imputation in the phase dependent filling method may introduce weekly rhythms), and z-scoring highlighted groups with similar patterns of change regardless of their baseline PA. Hierarchical clustering was performed on four consecutive months of data. The same four months were chosen for every individual to avoid days with larger proportions of missing data at the beginning and end of the study period. We observed two groups with different weekly PA rhythms on the heatmap: one group with high 24-hour MET sums on weekends relative to themselves and one group with low 24-hour MET sums on weekends relative to themselves. These rhythms were defined as weekend rhythms, where the group with relatively high 24-hour MET sums on weekends was further identified as the weekend high PA rhythm group, and the second group was identified as the weekend low PA rhythm group.

Convinced that weekend rhythms were not artifacts of data filling, we performed agglomerative clustering on filled MET data (filling methods described in ‘Data filling’) to identify individuals with weekend high and weekend low PA rhythms. Agglomerative clustering was performed on four consecutive months (the same months used in the hierarchical clustering) of the filled and intraindividual z-scored 24-hour MET sum data using the *scikit-learn* Python package (*sklearn.cluster.AgglomerativeClustering*, version 1.1.3)[42]. Clustering into five groups (Figure 3A) allowed for the recovery of both the weekend high PA rhythm group (Figure 3B) and the weekend low PA rhythm group (Figure 3C), herein referred to as the weekend high cluster and the weekend low cluster.

To confirm the presence of the weekend rhythms observed on the heatmap (Figure 3A-C top), we calculated the average 24-hour MET sum for each day in the consecutive four months across all participants (Figure 3A bottom), across only participants in the weekend high cluster (Figure 3B bottom), and across only participants in the weekend low cluster (Figure 3C bottom). These averages were visualized as a line plot with the mean across all days in that group layered on top (Figure 3A-C bottom).

To assess the differences between people with different weekend rhythms and without weekend rhythms (patternless), mean and CDI of 24-hour MET sums were calculated for each individual over the four consecutive months used to cluster the individuals by PA rhythm. The means were compared across weekend high, weekend low, and patternless clusters (Kruskal-Wallis test, Bonferroni correction for three comparisons, and post-hoc Dunn’s test) while the CDI was only compared across weekend rhythm (the aggregated group of individuals with either weekend high or weekend low PA rhythm) and patternless clusters (Kruskal-Wallis test between two groups). CDI was only compared across the presence or absence of a weekend rhythm because the direction of change in 24-hour MET sums on the weekend does not affect the CDI.

CDI of 24-hour MET sums was also compared across groups of individuals with unique combinations of sex and PA rhythm (weekend pattern male, weekend pattern female, patternless male, and patternless female; Kruskal-Wallis test, Bonferroni correction for six comparisons, and post-hoc Dunn’s test, Figure 3D). The effect of weekend rhythms on the variability of the whole male and female population was calculated using IQR distributions as described in the ‘Statistical Methods’ section.

Analysis of Age

Analysis of age focused on the CDI variability metric and daily 24-hour MET sum metric. 24-hour MET sums were chosen for analysis to focus on the overall variability that is due to PA, in contrast to asleep or awake sums that vary with time spent in the state, as described in the 'Cohort and MET Data Foundational Analysis' methods section. The CDI variability metric was chosen due to its accounting for chronological order, as described in the 'Variability Metrics of MET Sums' section.

Mean and CDI of 24-hour MET sums were calculated for each individual over all 206 days and compared across age categories (Kruskal-Wallis test, Bonferroni correction for 15 comparisons, and post-hoc Dunn's test, Table 3). CDI of 24-hour MET sums were also compared across sex groups in the same age category (Kruskal-Wallis test, Bonferroni correction for six comparisons(six age groups), and post-hoc Dunn's test, Figure 4A) and across age categories within the same sex group (Kruskal-Wallis test, Bonferroni correction for 15 comparisons, and post-hoc Dunn's test, Figure 4B and Table 4-5). A boxenplot (seaborn Python library, *seaborn.boxenplot*, version 0.12.2)[36], or letter-value plot, was used to visually compare males and females within age groups (Figure 4A). A boxenplot is similar to a boxplot, but represents the whiskers as a variable number of quantiles. If quantiles are sufficiently unique, meaning that they do not include values from other quantiles, they are represented as a box. This leaves 5-8 outliers on each side.

The effect of each age group on the variability of the whole male or female population was calculated using IQR distributions as described in the 'Statistical Methods' section (Table 6).

Generalized Additive Model (GAM) of Those Features Found to Have Significant Impact on CDI of 24-Hour MET Sums Across Individuals: Sex, Age, and Weekend Rhythm

Previous studies have utilized generalized additive models (GAMs) to predict health outcomes using sex and/or age as features [43,44]. In this study, a GAM was used to rank the effect of variables on CDI of 24-hour MET sums and detect groups with outlier intra indivdiual variability (Figure 5A-E). A generalized additive model was built in Python using the package pyGAM (*pygam.LinearGAM*, version 0.9.1)[45].

Three initial models were tested: a model with an identity link and a factor term for all variables analyzed in this paper (sex, age, weekend rhythm, and cyclic status), all variables and all two-way interactions (sex-age, age-weekend rhythm, etc.), and all variables with all two-way and all three-way interactions (sex-age-cyclic status, etc.). Model performance was assessed using the likelihood ratio pseudo-R-squared metric which represents the proportional reduction in the deviance and was shown as a percent for this analysis. The final model does not include cyclic status as its effects were not significant (see Results), thus factor terms were fit to sex, age, and weekend rhythm categories (Sex: female, male. Age: 20-29, 30-39, 40-49, 50-59, 60-69, 70-79. Weekend Rhythm (WR): weekend rhythm, patternless, Figure 5A-C). This resulted in the following GAM structure:

$$G(E(CDI)) = \beta_0 + f_{sex}(sex) + f_{WR}(WR) + f_{age}(age) \quad \textbf{(6)}$$

where g is an identity link function and β_0 is the intercept of the model. Individual feature importance was determined by the magnitude of the coefficients in each level of the factor terms and by the change in null deviance when each feature was left out.

Results

Cohort and MET Data Foundational Analysis

As an initial comparison of MET between sexes, we visually assessed minute-level MET value time series and distributions for two representative individuals (Figure 1A, Figure 1B). We observed variation in MET values between awake and asleep states with increased MET during awake time periods, as expected (Figure 1A and Figure 1B, left). Finding that the distribution of MET values appeared highly dependent on asleep or awake state (Figure 1A and Figure 1B right), further comparisons used daily aggregated MET values separated into sums over either 24 hours, only awake time periods, or only asleep time periods. Female and male distributions of mean 24-hour, awake, and asleep daily MET sums over the 206 days overall were not significantly different (Table 1, Figure 1C, Figure 1D). However, we observed an apparent increase in the male mean of 24-hour MET sums at the upper extreme (Figure 1C). Consistent with this observation, a comparison of individuals' mean of 24-hour MET (Figure 1C, right) revealed that the 60 males with the largest average 24-hour MET sum had a significantly higher average than the top 60 females (Kruskal-Wallis, H=10.25, *P*=.001, Modified Cohen's d (d_m) = 0.34). We also observed differences between male and female intraindividual variability: the standard deviation for individual males was significantly larger than that of individual females for both the standard deviation of awake and 24-hour MET sums (Table 1, Figure 1E).

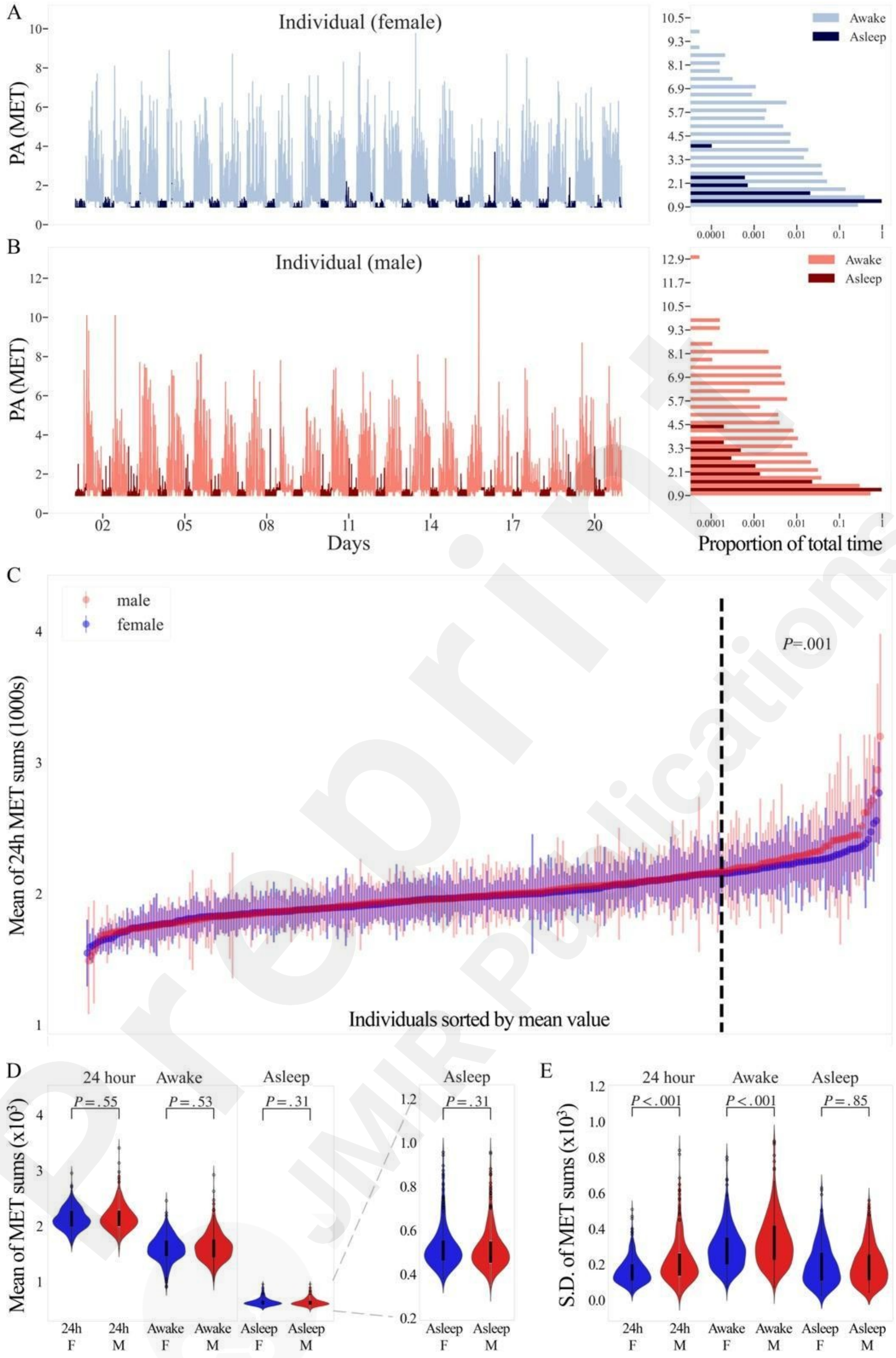


Figure 1. Longitudinal plot of a representative three-week interval of minute-level MET data (left) from (A) one female (blue) and (B) one male (red) with the histogram of the MET values for each separated by awake (light) and asleep (dark) values (right). MET values were examined at minute-level resolution. Histograms show the percent time (percent time is shown on a log scale and referenced in the figure as 'Proportion of total time') spent in 37 bins of MET values while awake or asleep. MET values range from 0.9 to 16.1 and each bin is 0.4 METs in size. C) Plot of all individuals' (n=596) mean (dot) and standard deviation (vertical line) of 24-hour daily MET sums, sorted by mean. The dashed line separates the 60 individuals in each sex with the largest means from the rest of the population. The top 60 were subsequently compared across sex (Kruskal-Wallis test). D) Violin plots of male and female individual means and E) standard deviations for 24-hour MET sums, awake time state MET sums, and asleep

time state MET sums (Kruskal-Wallis test, Bonferroni corrected significance threshold for three comparisons: *P*=0.02).

Table 1. Mean and Standard Deviation Statistics by Time State: Kruskal-Wallis test across sex for mean and standard deviation of each time state. (Bonferroni corrected significance threshold for three comparisons: *P*=0.02).

Statistic		Kruskal-Wallis Test Across Sex		
		P-value	H statistic	Sex with Larger Median
	MET Sum			
Mean				
	24-hour	.55	0.36	Male
	Awake	.53	0.40	Male
	Asleep	.31	1.01	Female
Standard Deviation				
	24-hour	<.001	38.54	Male
	Awake	<.001	11.60	Male
	Asleep	.85	0.03	Female

Variability Metrics of MET Sums

Four intraindividual variability metrics were measured: standard deviation, coefficient of variation, proportional variability index, and consecutive disparity index. The most appropriate metric of variability for our analyses was the consecutive disparity index because of its accounting for chronological order and non-dependence on the mean for calculation. Other metrics were included as controls to validate the statistical findings from consecutive disparity index analyses. Further analyses used only CDI as a variability metric.

CV and PV of male individuals were significantly larger than female individuals for awake and 24-hour MET sums (Figure 2A-B, Table 2). 24-hour MET sum CDI was significantly larger for males than females (Figure 2C, Table 2, Modified Cohen's *d* (d_m) = 0.35). In all three of these metrics, asleep MET sum intraindividual variability was not significantly different across sexes (Figure 1D, Table 1, Figure 2A-C, Table 2).

Analysis of Physical Activity by Cyclic Status

Neither 28-day (or near 28-day) temporal structures nor any unique temporal structure in daily 24-hour MET sums were identified in cyclic people. Cyclic females and all acyclic subjects (male or female) did not have significantly different mean 24-hour MET sums (Kruskal-Wallis, H=0.46, *P*=.50, data not shown) or significantly different CDI of 24-hour MET sums (Kruskal-Wallis H=1.03, *P*=.31, Figure 2D). However, we found a significant difference between male, cyclic female, and acyclic female CDI of 24-hour MET sums (Kruskal-Wallis, H=32.36, *P*<.001, Figure 2D). A Dunn's test showed that females were less variable within individual than males, regardless of cyclic status (males vs. cyclic females: *P*=.006, Modified

Cohen’s *d* (d_m) = 0.27. males vs. acyclic females: *P*<.001, d_m = 0.41), and that cyclic females and acyclic females were not significantly different (*P*=.09). Cyclic females and acyclic females of the same age were also compared to confirm that the uneven age distribution between the two groups did not contribute to there being no statistical difference between the groups (Kruskal-Wallis, cyclic females age 20-59 (n=105) vs acyclic females age 20-59 (n=94): H=2.30, *P*=.13. Cyclic females aged 20-49 (n=102) vs acyclic females aged 20-49 (n=48): H=0.53, *P*=.47). We then compared the population variability of the whole female population and the female population excluding cyclic females, as described in the methods in the ‘Effect of subpopulations’ subsection of the ‘Statistical Methods’ section. Removing cyclic females from the female population did not significantly reduce the whole female population variability 24-hour MET sums (Kruskal-Wallis, H=0.12, *P*=.73).

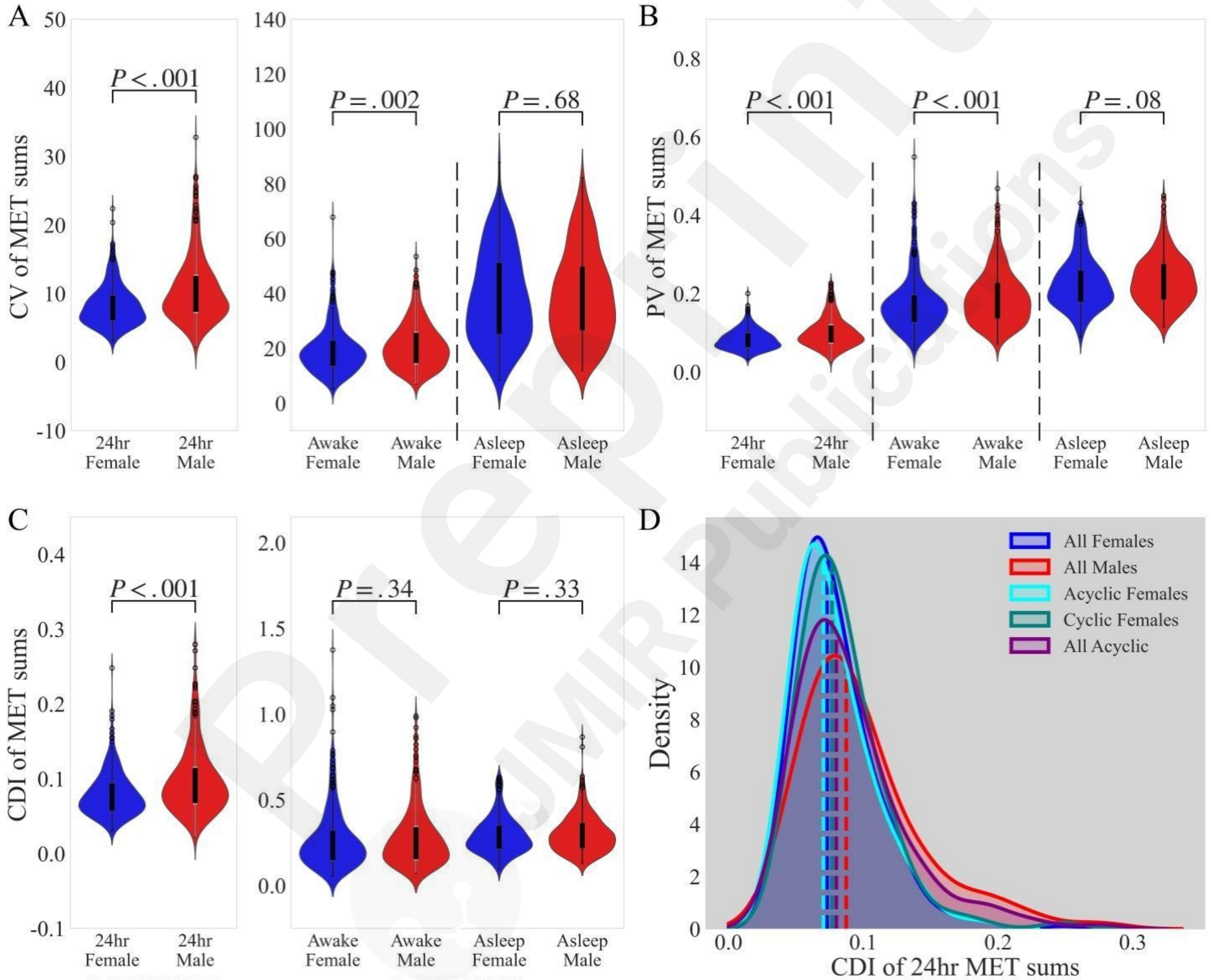


Figure 2. Violin plots of female (blue) and male (red) distribution of A) coefficient of variation (CV), B) proportional variability index (PV), and C) consecutive disparity index (CDI) for 24-hour MET sums, awake time state MET sums, and asleep time state MET sums. (Kruskal-Wallis, Bonferroni corrected significance threshold for three comparisons: *P*=0.02), and D) kernel density estimate plots of all female (blue), all male (red), acyclic female (teal), cyclic female (blue-green), and all acyclic individuals of either sex (purple). Group median CDI: dashed vertical lines.

Table 2. Variability Metrics by Time State: Kruskal-Wallis test across sex for coefficient of variation (CV), proportional variability index (PV), and consecutive disparity index (CDI) of

each time state. (Bonferroni corrected significance threshold for three comparisons: *P*=0.02).

Statistic		Kruskal-Wallis Test Across Sex		
		P-value	H statistic	Sex with Larger Median
	MET Sum			
CV				
	24-hour	<.001	43.70	Male
	Awake	.002	9.36	Male
	Asleep	.68	0.17	Male
PV				
	24-hour	<.001	37.90	Male
	Awake	<.001	10.97	Male
	Asleep	.08	3.12	Male
CDI				
	24-hour	<.001	29.51	Male
	Awake	.34	0.90	Male
	Asleep	.33	0.96	Male

Analysis by Weekend Rhythm in Physical Activity

Agglomerative clustering of four months of data per individual across the whole cohort revealed clusters of individuals sharing prominent PA rhythms on a weekly timescale (Figure 3A). Two clusters of individuals with weekend rhythms were identified: a ‘weekend high’ cluster (Labeled dark green in Figure 3A and Figure 3B) and a ‘weekend low’ cluster (Labeled purple in Figure 3A and Figure 3C). The three clusters without weekend rhythms are referred to as ‘patternless’ clusters (Labeled with orange, pink, and light green in Figure 3A).

Significant differences in the means of 24-hour MET sums existed between individuals in the weekend high cluster, weekend low cluster, and the patternless clusters (Kruskal-Wallis: H=9.18, *P*=.01, Bonferroni corrected significance threshold: *P*=.02, data not shown). The weekend high cluster had significantly larger mean 24-hour MET sums than the weekend low cluster and the patternless clusters (Dunn’s test: weekend high vs. weekend low: *P*=.007. weekend high vs. patternless: *P*=.01). Modified Cohen’s *d* effect sizes (d_m) between significantly different groups were 0.41 (weekend high vs. weekend low) and 0.22 (weekend high vs. patternless).

Next, we grouped the individuals with any weekend rhythm (weekend high or weekend low) to examine intraindividual variability. The cluster of individuals with either weekend rhythm had significantly larger CDI of 24-hour MET sums than individuals in the patternless clusters (Kruskal-Wallis, H=10.13, *P*=.001, d_m = 0.20, data not shown). The modified Cohen’s *d* (d_m) between male and female CDI of 24-hour MET sums was 0.35, suggesting that sex explained more intraindividual variability than PA rhythms on the weekly timescale.

We found significant effects of sex and weekend rhythm on 24-hour MET sum CDI (Kruskal-Wallis, Bonferroni corrected significance threshold: *P*=.008, H=34.60, *P*<.001, Figure 3D). Males have larger CDI of 24-hour MET sums than females in the same cluster (Dunn’s test: patternless cluster: *P*<.001, d_m = 0.32. weekend rhythm cluster: *P*=.003, d_m = 0.51). Additionally, males in the weekend rhythm cluster had significantly larger 24-hour MET sum CDI than females from the patternless clusters (Dunn’s test, *P*<.001, d_m = 0.49); however, females in the weekend rhythm cluster did not have significantly larger 24-hour MET sum CDI than males in the patternless clusters (Dunn's test, *P*=.24). We found no significant effect between clusters within sex on 24-hour sum CDI: males in the weekend rhythm cluster did not differ from males in the patternless clusters (Dunn's test, *P*=.02), nor did females in the weekend rhythm cluster differ from females in the patternless clusters (Dunn's test, *P*=.06).

We compared the variability of the whole male and female populations to the populations excluding individuals with weekend rhythms using the strategy described in the methods section in the 'Effect of subpopulations' subsection of the 'Statistical Methods' section. Excluding individuals with weekend rhythms did not reduce the population variability of 24-hour MET sums of either the whole male or female population (Kruskal-Wallis, Bonferroni corrected significance threshold: *P*=.025, all females vs. females without weekend rhythm clusters: H=2.62, *P*=.11. all males vs. males without weekend rhythm clusters: H=4.46, *P*=.03).

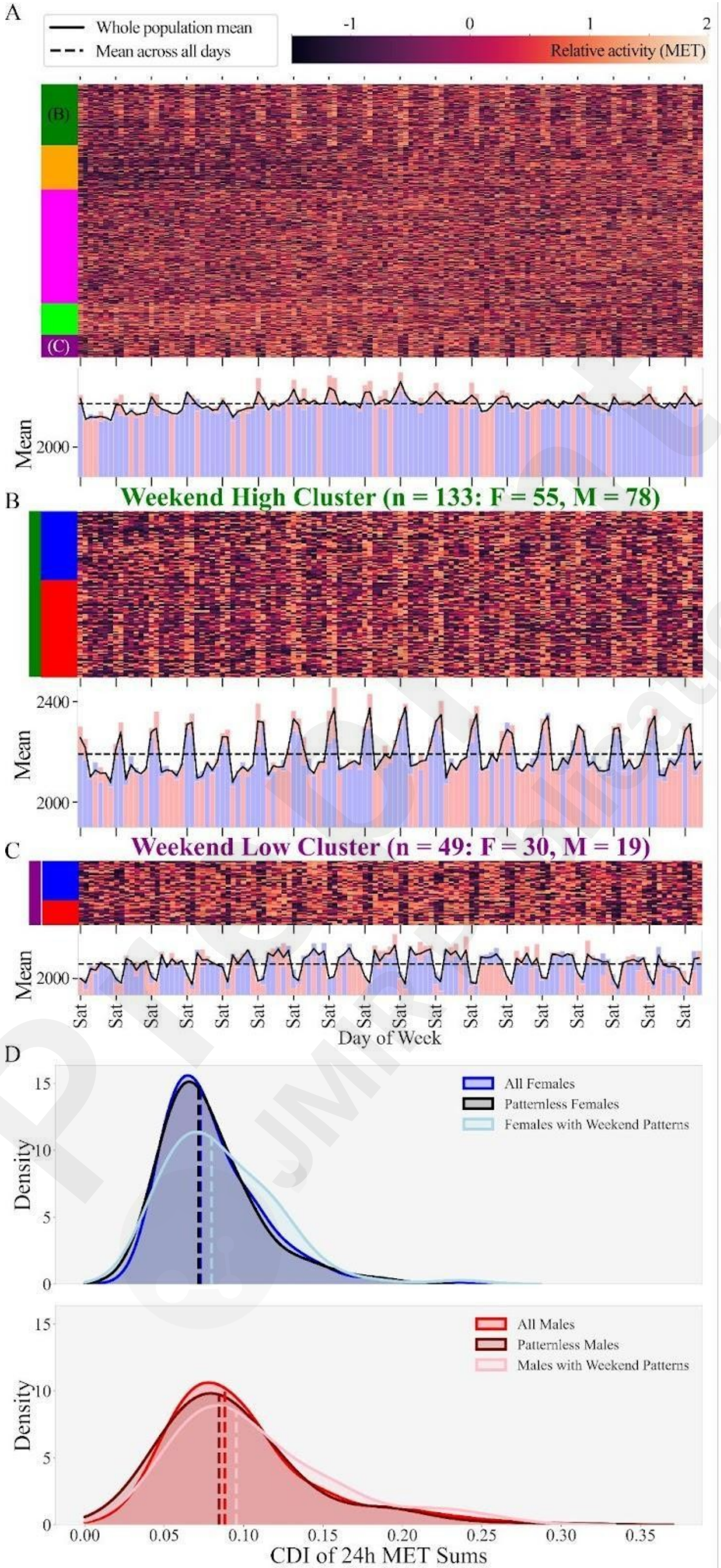


Figure 3. A) Heatmap of relative activity for every individual across four consecutive months. Relative activity was defined as arctan(2*intraindividual z-score(daily 24-hour MET sum)). Relative activity values above 2 and below -1.5 are colored with the lightest and darkest values

respectively. Individuals are sorted by agglomerative cluster number and clusters are demarcated by the colors in the bar to the left of the heatmap. The line and layered barplot below each heatmap show the daily mean 24-hour MET sum across all individuals in the connected heatmap (solid black line), the mean 24-hour MET sum across all days in the four-month period (dashed black line), and the daily 24-hour MET sum mean of the males (red) and females (blue) where the sex with the lower mean for each day was layered on top. Magnification of the dark green cluster: weekend high heatmap (B); and dark purple cluster: weekend low heatmap (C). Heatmap rows, representing one individual each, are all of equal size so that the height of the heatmap is representative of the number of people in the cluster. Individuals are labeled and sorted by sex (blue box on the left of the heatmap for female, red for male). D) Kernel density estimate plot of consecutive disparity index calculated from four consecutive months for the female and male whole population, weekend cluster population, and other clusters. Vertical dashed lines represent the population median CDI.

Analysis of Age

We found significant differences in mean 24-hour MET sums across age groups (Kruskal-Wallis: H=24.30, *P*=2x10^{-4}, Bonferroni corrected significance threshold for 15 comparisons: *P*=3x10^{-3}, data not shown). Individuals aged 70-79 had significantly smaller mean 24-hour daily MET sums than individuals aged 30-39 and 50-59 (Dunn's test: 70-79 vs. 30-39: *P*=1x10^{-5}, d_m = 0.54. 70-79 vs. 50-59: *P*=5x10^{-4}, d_m = 0.39), and individuals aged 60-69 had significantly smaller mean 24-hour daily MET sums than individuals aged 30-39 (Dunn's test: 60-69 vs. 30-39: *P*=3x10^{-3}, d_m = 0.28). Other comparisons of mean 24-hour MET sums between age groups were not statistically significant (data not shown).

Differences in CDI of 24-hour MET sums existed across age groups (Kruskal-Wallis, H=40.55, *P*=1x10^{-7}, Bonferroni corrected significance threshold for 15 comparisons: *P*=3x10^{-3}, Table 3). Individuals aged 70-79 had significantly smaller CDI of 24-hour MET sums than individuals aged 20-29, 30-39, 40-49, and 50-59 (Table 3). Individuals aged 60-69 had significantly smaller CDI of 24-hour MET sums than individuals aged 30-39 and 50-59 (Table 3). The modified Cohen's *d* (d_m) between the groups that were significantly different ranged from 0.36 to 0.56, suggesting that age explained more intraindividual variability than sex (d_m = 0.35) and weekly rhythm (d_m = 0.20).

Having found a significant effect of sex and also of age bin, we carried out pairwise comparisons of sex within each age bin and found that males in the 30-39 group and the 40-49 group had significantly higher 24-hour MET sum CDI than females in the same age groups (Kruskal-Wallis, Bonferroni corrected significance threshold for six comparisons: *P*=.008: 30-39 M vs. 30-39 F: H=8.62, *P*=.003, d_m = 0.37. 40-49 M vs. 40-49 F: H=8.64, *P*=.003, d_m = 0.33. Figure 4A). We further note that while the remaining comparisons were not significant, the trend in every age group was toward the same direction of difference, with males having higher median CDI at all ages (Kruskal-Wallis, Bonferroni corrected significance threshold for six comparisons: *P*=.008: 20-29 M vs. 20-29 F: H=0.96, *P*=.33. 50-59 M vs. 50-59 F: H=0.78, *P*=.38. 60-69 M vs. 60-69 F: H=6.58, *P*=.01. 70-79 M vs. 70-79 F: H=6.38, *P*=.01. Figure 4A).

Females aged 70-79 were significantly less variable than females aged 20-29, 30-39, and 50-59; females aged 60-69 were significantly less variable than females aged 50-59 (Figure 4B, Table 4). Modified Cohen's *d* effect sizes for these differences were between 0.50 and 0.69 (Table 4). Males aged 70-79 were significantly less variable than males aged 30-39 with a 0.40 modified Cohen's *d* effect size (Figure 4B, Table 5).

We compared the variability of the whole male and female populations excluding each single age group using the strategy described in the methods section in the 'Effects of subpopulations’ subsection of the ‘Statistical Methods’ section. The IQR distributions composed of the daily IQRs of population 24-hour MET sums were not significantly different between: a) the whole population and b) the population without any single age group, except in one comparison (Table 6). The whole female population and the female population without individuals aged 60-69 have significantly different IQRs of 24-hour MET sums, such that the female population variability was increased by the presence of females aged 60-69 (Table 6, d_m=0.18). Using Lehr’s rule, we calculated the effect of the increased population variability caused by females aged 60-69 on the approximate required sample size to detect a statistically significant difference. We found that to detect a difference of 40 (approximately the difference in 24-hour MET sums resulting from a 20-minute walk), the exclusion of females aged 60-69 results in a sample size reduction from 1088 to 1047 (3.8% reduction). For a difference of 100 (approximately the difference in 24-hour MET sums resulting from 20-minutes of moderate intensity of activity), the exclusion results in a sample size reduction from 174 to 167 (4.0% reduction), and for a difference of 180 (approximately the difference in 24-hour MET sums resulting from 20-minutes of high intensity of activity), the exclusion results in a sample size reduction from 54 to 52 (3.7% reduction).

Table 3. Age Bin Statistics. Diagonal (dark-shaded cells): The median consecutive disparity index (CDI) of each age bin. Below/left of diagonal: *P*-value of the post-hoc Dunn’s test comparing each age group, significant comparisons are lightly shaded. Above/right of diagonal: lightly shaded cells show the modified Cohen’s *d* effect sizes of the comparisons that were significantly different. (Kruskal-Wallis, Bonferroni corrected significance threshold for 15 comparisons: P=3x10^{-3})

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.082					0.38
30-39	.38	0.087			0.47	0.56
40-49	.68	.20	0.081			0.36
50-59	.64	.67	.38	0.089	0.43	0.51
60-69	5x10^{-3}	2x10^{-4}	.02	1x10^{-3}	0.07	
70-79	4x10^{-5}	5x10^{-7}	2x10^{-4}	4x10^{-6}	.18	0.065

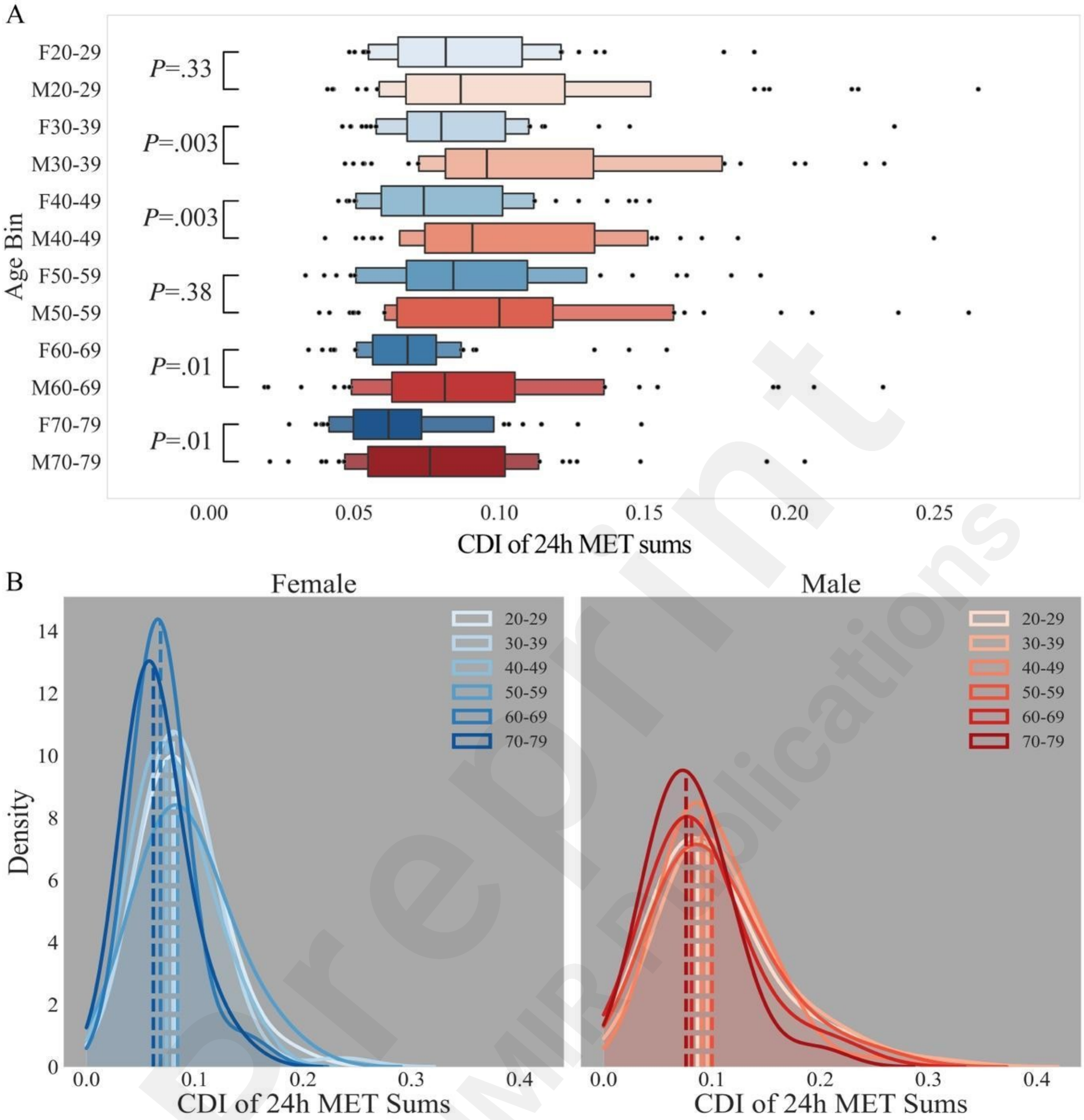


Figure 4. A) Boxenplot of consecutive disparity indices for each sex-age category (Kruskal-Wallis, Bonferroni corrected significance threshold for six comparisons: *P*=.008). B) Female and male kernel density estimate plots of consecutive disparity index (CDI) in each age bin. Dashed lines represent the median CDI of the sex-age population.

Table 4 and Table 5. Female (Blue, Top) and Male (Red, Bottom) Age Bin Statistics. Diagonal (dark-shaded cells): The median consecutive disparity index of each age bin. Below/left of diagonal: p-value of the post-hoc Dunn's test comparing each age group, significant comparisons are lightly shaded. Above/right of diagonal: lightly shaded cells show the modified Cohen's *d* effect sizes of the comparisons that were significantly different. (Kruskal-Wallis, Bonferroni corrected significance threshold for 30 comparisons: *P*=1.7x10^{-3})

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.082					0.60

30-39	.88	0.080				0.64
40-49	.17	.22	0.074			
50-59	.77	.66	.10	0.084	0.50	0.69
60-69	$3x10^{-3}$	$4x10^{-3}$	.10	$1x10^{-3}$	0.068	
70-79	$3x10^{-5}$	$5x10^{-5}$	$5x10^{-3}$	$7x10^{-6}$	.22	0.062

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.087					
30-39	.18	0.096				0.40
40-49	.56	.45	0.091			
50-59	.72	.32	.81	0.100		
60-69	.29	.02	.10	.16	0.081	
70-79	.05	$9x10^{-4}$	.01	.02	.36	0.076

Table 6. Kruskal-Wallis Test of daily IQRs (n=206) between the whole female or male population and the female or male whole population with one age group removed. (Bonferroni corrected significance threshold for 12 comparisons: *P*=.004)

Sex		Kruskal-Wallis Test	
		P-value	H statistic
	Removed Age Group		
Male			
	20-29	.57	0.32
	30-39	.57	0.33
	40-49	.06	3.57
	50-59	.75	0.10
	60-69	.03	4.75
	70-79	.007	7.40
Female			
	20-29	.68	0.17
	30-39	.20	1.65
	40-49	.03	4.81
	50-59	.09	2.89
	60-69	<.001	11.11
	70-79	.007	7.17

Generalized Additive Model (GAM) of Those Features Found to Have Significant Impact on CDI of 24-Hour MET Sums Across Individuals: Sex, Age, and Weekend Rhythm

A GAM was used to summarize the contributions of sex, age, cyclic status, and weekend rhythm on CDI of 24-hour MET sums across individuals. Three initial models were tested to find the best model for explaining population variability in CDI while retaining interpretability: 1) a model with an identity link and a factor term for all variables analyzed in this paper (sex, age, weekend rhythm, and cyclic status); 2) all variables and all two-way interactions (*e.g.* sex-age, age-weekend rhythm, etc.); 3) all variables with all two-way and all three-way interactions (*e.g.* sex-age-cyclic status, etc.). The first model explained 11.5% of the null deviance, but the cyclic status term was not significantly different from zero (*P*=.17). The last two models explained 1.6% and 2.8% more of the null deviance than the first model, where again cyclic status was not significant (second model *P*=.63, third model *P*=.84). These analyses support our finding that acyclic and cyclic individuals did not have significantly different CDI. Given the marginal increase in null deviance explained for the substantial increase in model complexity (7 and 11 additional relational features, respectively) and the increased difficulty of interpreting the models with multiple interaction terms (4 terms in the first vs 7 & 11 in the second and third, respectively), the first model was chosen for further interrogation. To construct the final model, the cyclic status variable was removed from the first model since the term was not significantly different from zero, leaving the final variables as sex, age, and weekend rhythm.

Unique combinations of the categories (physiological phenotypes) across the final variables resulted in 24 phenotype groups (*e.g.*, Female, 20-29, weekend rhythm) for which the model predicted a CDI value. Each of the variables had a significant effect on the model prediction (sex: *P*<.001, weekend rhythm: *P*<.01, age: *P*<.001). The null deviance explained by the final model decreased by 4.9% by the exclusion of sex as a feature, by 4.7% by the exclusion of age as a feature, and by 0.92% by the exclusion of weekend rhythm as a feature, indicating that sex and age are the most important features in this model for predicting CDI. Coefficient magnitudes indicated that sex and specific age bins had the greatest effect on CDI out of these categories: sex (Figure 5A) had an overall effect of ±0.0091 (decreased for females and increased for males), weekend rhythm (Figure 5B) had an overall effect of ±0.0043 (decreased for patternless and increased for weekend rhythms), and age bin (Figure 5C) had an overall effect of 0.0093 to -0.015 (20-29: 0.0055, 30-39: 0.0093, 40-49: 0.0011, 50-59: 0.0075, 60-69: -0.0082, 70-79: -0.015). However, the overall deviance explained by the final model was 11.3%, indicating a low proportion of null deviance explained by the model. This is consistent with our modified Cohen's *d* analyses that found the difference in median CDI between categories to be smaller than the size of the interquartile ranges (IQRs) of the categories themselves (see sections on Sex, Weekly Rhythms, and Age, above; modified Cohen's *d* (d_m) = 0.35, 0.20, and 0.36-0.56 respectively). Together, both of these analyses indicated that even timescales of change that were significant sources of variability in CDI were not substantial sources of variability that would likely weaken statistical power. GAM analysis added that the intersection of sex at specific age bins (30-39, 50-59, 60-69, and 70-79) affected the GAM prediction the most, but further confirmed that no such category is in itself a substantial source of variability in the population. Model predictions did not align with unique values for each phenotype group and there was significant overlap between groups in CDI range (Figure 5D-E).

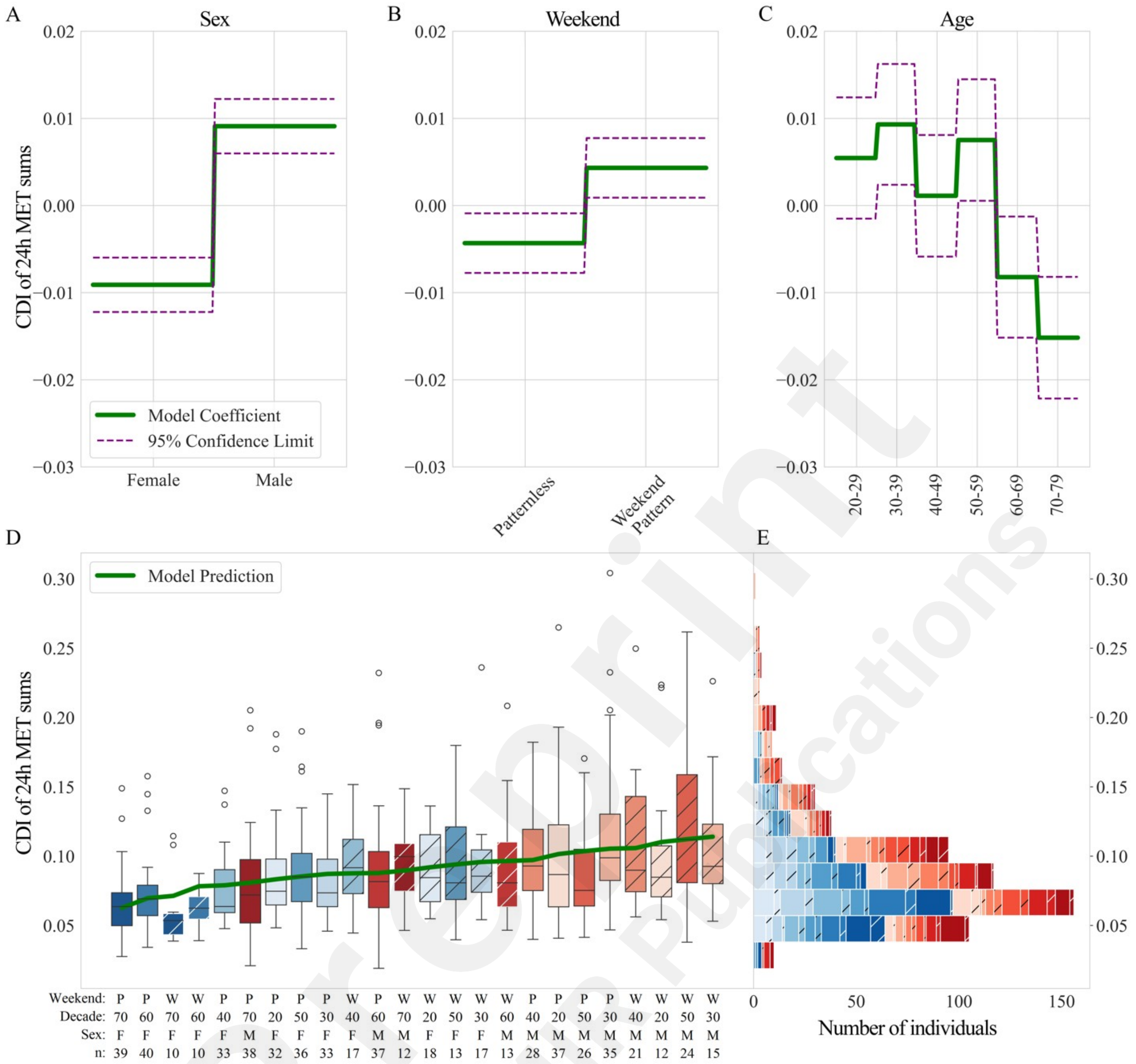


Figure 5. GAM fitted factor functions for sex (A), weekend rhythm (B), and age (C) with confidence intervals. D) Boxplot of consecutive disparity index of 24-hour MET sums for each unique phenotype group in order of the model prediction (green line) for that phenotype group. Age-sex categories are colored as they were in Figure 4A and hatching indicates the presence of a weekend rhythm in individuals in the labeled group (P = patternless, W = weekend rhythm). E) Stacked histogram (ordered for visual clarity) of the number of individuals in CDI bins labeled by phenotype group, highlighting the overlap of each group in most bins.

Discussion

Principal Results

In this work, we found evidence to reject the hypothesis that it is necessary to exclude women as research subjects when assessing PA-related behaviors. Sex and cyclic status were found to represent different populations, and neither sex nor menstrual cycles substantially increased the intraindividual variability of PA. Rather, we found that females have significantly less intraindividual variability than males, regardless of their cyclic status. This study also demonstrates the exclusion of either sex is unwarranted, as the overall difference in intraindividual PA variability was small. However, this work did reinforce the utility of sex as a biological variable (SABV), as we found differences by sex in the contributions of different timescales (weekends and age) to the patterns of change in PA over time.

Males and females showed no significant differences between mean 24-hour MET sums, but the 60 most active males were significantly more active than the 60 most active females. The standard deviation, coefficient of variation (CV), proportional variability index (PV), and consecutive disparity index (CDI) of 24-hour MET sums were all significantly different by sex. Because CDI captures local changes instead of only global structure, we deemed CDI the best indicator of continuous intraindividual variability for time series data. Cyclic status had no effect on CDI of 24-hour MET sums and no temporal structures on the timescales of menstrual cycles were found in cyclic people (i.e. the roughly 28-day rhythms in these individuals' temperature data[15] were not reflected in their PA).

We did find that some people in the dataset had temporal structure on the timescales of weeks. The people with weekend rhythms were found to have higher intraindividual variability (CDI of 24-hour MET sums) than people without weekend rhythms (patternless), regardless of sex. However, within sex, people with weekend rhythms did not have significantly different intraindividual variability than those without weekend rhythms, nor did their inclusion increase the population variability of the whole male or female populations. Males were more intraindividually variable than females regardless of weekend rhythm. Without SABV analysis, we may have concluded that CDI was significantly different between individuals with and without weekend patterns when the actual cause of this deviation appears to be due to the fact that male PA is more variable within individuals than female PA.

We also found that sex differences existed in the presence of weekend rhythms. Interestingly, those with weekend effects were more likely to be male, though both sexes were represented in this category (85 females and 97 males had weekend rhythms). This may be because weekends play a large role in modulating behavior. For example, work schedules may inhibit PA during weekdays, leading some individuals to make up their PA debt on weekends. Others may have active work schedules, and seek to rest and recuperate on weekends. One study found that individuals who were more active on weekdays than on weekends had lower education and were more likely to work manual occupations than those who were consistently inactive [46]. A higher group membership of males (55 females and 78 males) in the weekend high group may also support the finding that females have higher rates of inactivity [3] if increased activity on the weekend is due to participation in exercise.

Age did not have a consistent effect on intraindividual variability. When the data was sex-disaggregated, females aged 70-79 and 60-69 were less variable than a few of the younger age bins, but only one age bin was different between males: males 70-79 were less variable than males aged 30-39. This decrease in intraindividual variability in the oldest age groups we analyzed is likely caused by increased sedentary behavior with increased age [47]. Additionally, males in the 40-49 and 30-39 age bins were more intraindividually variable than females in the same age groups. This, again, is in contrast to the results when all individuals of both sexes were considered in statistical tests. If the results were not sex-disaggregated, we may have concluded that male intraindividual variability across age bins looks similar to female intraindividual variability when it evidently does not. The lack of difference across age bins in males appears to be caused by increased population variability of CDI of 24-hour MET sums within each age bin when compared to females. We note that females aged 60-69 were the only group to significantly increase the population variability of the whole female population. We used this group to test the hypothesis that excluding minority groups that significantly increased whole population variability would meaningfully improve statistical power for the included groups. We found a change in sample size of less than 5% for computed comparisons. We argue that the benefits from reducing, for example, a 200 person study to a 192 person study is less than the value of including a whole other group so that findings apply broadly to more people.

The effects of weekend rhythms and age, along with the lack of effects due to cyclic status, on intraindividual variability all suggest that sex alone is not an effective proxy for the presence of temporal structure or the intraindividual variability that may affect statistical analysis. In our final analysis, we employed a multivariate (GAM) model that determined that while sex, weekend rhythm, and age have significant effects on intraindividual PA variability, only 11.3% of the population variability in CDI of 24-hour MET sums can be explained by these phenotypes. The analysis showed that age and sex had similar effects on intraindividual PA variability and that weekend rhythm had a much smaller effect comparatively. Cyclic status did not have a significant effect (consistent even in the more complex models), and in fact had less effect than any other timescale studied. The analysis also highlights the potential usefulness of intersectional phenotypes in showing that they provide more information about an individual than single phenotypes. Indeed, digital twinning is emerging as a computational approach for providing precision insights into health by grouping “similar” individuals (similar based on many potential features of their data) and then identifying signs or treatments specific to that group, as opposed to being limited to more classical demographics like sex or ethnicity alone [48,49]. As these approaches mature, timescales of change like menstrual cycle, weekend patterns, and circadian rhythms might prove to be useful features by which to define similarity. Even when the intraindividual variability is roughly equal across such groups (we found only 11.3% of intraindividual variability can be accounted for by the various timescales in this work) the behaviors or needs of groups with different dynamics may still differ due to differing physiology.

Older females with weekend rhythms appear to have the least intraindividual variability of all subject phenotypes (Figure 5D), perhaps indicating stronger behavioral routines in this phenotype group. Ironically, older females, who are historically even more understudied than females broadly [50,51], would appear to have mitigated concerns about increased intraindividual variability eroding statistical comparisons more than any other group, including the most historically overrepresented population of midlife males. This is not an argument that men should be excluded—no group should be excluded from research, and no groups in our models exhibited an overwhelming amount of intraindividual variability that would reduce power in statistical comparison. Rather, this highlights that assumptions about who should be excluded in the interest of minimizing population variability and maximizing statistical power may have made statistical inference harder rather than easier (and may still be doing so when numerical examinations of these assumptions are absent in any given field of study). While the multivariate analysis suggests that sex and age most affect intraindividual variability among the four variables studied, none of these variables alone, nor their intersection, reliably predicted intraindividual variability. This suggests that no group is so different from the others as to warrant statistical exclusion.

The key assertion is that in the context of PA, which is at present the most commonly available longitudinal physiological measure for humans, we found no support for the hypothesis that females broadly are more variable than males.

Limitations

This study aligns with our previous findings about sex and menstrual cycle impacts on variability in continuous temperature data [15]. As those analyses and the analyses presented here stemmed from the same cohort, it is possible that new cohorts would show different distributions, and so additional studies would help identify the stability and context for variability in different phenotypes and populations. For example, we do not suggest that all older females are less variable than all young males - indeed the least variable phenotype across the three characteristics of age, sex, and weekend rhythm had a substantially reduced N (Figure 5D), and so may well not be reliably representative of the broader population of older females. Instead, we suggest that our longitudinal analyses find this to be the case in this modality (PA), in this data set.

Additionally, it is worth noting that MET is not the same as steps, but is instead an adjusted measure of activity, conditioned by the weight of the individual. While MET does not provide insights into total absolute activity or types of activity, METs change as a function of intensity of activity and thus provide a means of assessing different timescales of behavioral change across individuals' data, as we analyzed here. Although METs have been found to have systematic inaccuracies in energy expenditure estimates due to their reliance on body weight for calculation [52], this does not affect the relative change we analyzed in intraindividual variability. Furthermore, while the exact formula for the calculation of METs is proprietary to Oura Ring and not known to us; Oura Ring (Gen 2) activity measurements displayed high correlation when previously validated against multiple accelerometers [30]. As always, we encourage further study using different metrics to more fully describe the variability landscape from as many angles as might be relevant to other applications or fields of research.

Comparison with Prior Work

This work joins a growing body of analyses that support the inclusion of both sexes in biomedical research [13,15–20,53–57]. Persistent sex bias in subject selection for biomedical research in humans and its detrimental impact on women's healthcare has been thoroughly described previously [53–56]. The harmful exclusion of women and females as subjects has received increased attention in the past decade - including specific mention as a problem in the 2024 Presidential State of the Union Address [58]. Public attention to this issue along with the U.S. [59] and international [60,61] policy changes affecting the inclusion of females has led to marked improvements in cohort equity [13,62], however, many researchers still fail to include subjects of both sexes in experiments, and those who do, often fail to perform SABV analyses [13,16,59]. Researchers' resistance to including females in both animal and human studies in biomedical research stems from the same concerns seen in sports and exercise medicine: including females will increase intraindividual measurement variability due to hormone fluctuations, and thus reduce statistical power [57]. Our results support inclusion of female subjects, consistent with many other studies that found female subjects do not reduce the statistical power of experiments due to substantial variability (*e.g.* [16–20]). Both this work and our previous work in temperature found that sex does affect variability, but that cyclic status alone does not account for the difference between males and females [15]. Neither segregation by sex nor segregation by cyclic status in and of themselves seems to be a useful control for overall variability in these modalities [15]. As a result, our work suggests that exclusion for the sake of preserving statistical power is neither necessary nor justified.

While this study is related to sex bias in biomedical research at large, the findings presented here are most applicable and comparable to behavioral research (here considered a subset of biomedical research) and epidemiological research in PA because the variability metric used (consecutive disparity index (CDI) of daily MET sums) approximates the amount of total exercise and movement in a day without consideration for types of activity or physiological processes.

In regards to epidemiological research in PA, our findings did not reflect the general consensus that females are less active than males [3–5]. However, as discussed above, METs have been found to have systematic inaccuracies in energy expenditure estimates [52], and may therefore inaccurately measure the amount of PA. Another potential cause for this discrepancy is that people who use wearables are more likely to be active than those who don't [63,64].

The effects of menstrual cycles on exercise performance have been studied previously, and the results are largely conflicting and inconclusive [9,10]. While this work does address PA variability in people with roughly 28-day temperature cycles, it is unique to these studies in metric: these studies assess exercise performance metrics such as strength and endurance, and our analyses examine the intraindividual variability of a daily summary of behavior or physical activity. This study also does not examine specific stages of the menstrual cycle or exercise performance metrics, however, the fact that 28-day temporal patterns in 24-hour MET sums do not exist at least suggests that if changes in exercise performance caused by cycling exist, they do not significantly affect behavior or total amount of physical activity.

Instead of finding temporal structures on menstrual timescales, we found temporal structures on weekly timescales confirming the findings from other recent accelerometry studies that found weekly rhythms in PA [46,65]. While this study did not use raw accelerometer data, it expands on previous studies in cohort age diversity [46] and the length of the study period [46,65]. However, these previous studies have focused on total amounts of activity rather than the presence of rhythms and are not directly comparable to this work. Weekend rhythms are not the main thrust of our work, but these findings may be of interest to those studying activity patterns

Conclusions

In conclusion, our findings support sex-based and age-based analysis in biomedical research involving PA, while rejecting the exclusion of females, males, weekend rhythm types, or any other specific intersectional phenotype from biomedical research based on assumptions of increased intraindividual variability of PA interfering with statistical power.

Acknowledgments

This effort was funded under MTEC solicitation MTEC-20-12-Diagnostics-023 and the USAMRDC under the Department of Defense (#MTEC-20-12-COVID19-D.-023). The #StartSmall foundation (#7029991), and Oura Health Oy (#134650) also provided funding for this work. The views expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy of the Department of Defense, or the U.S. Government. Lauryn R. Keeler Bruce, MS, was funded by the National Library of Medicine T15LM011271.

The authors thank the San Diego Supercomputer Center's Sherlock team, especially Sandeep Chaudry, for supporting secure data management.

Ethics Statement

The University of California San Francisco (UCSF) Institutional Review Board (IRB, IRB# 20-30408) and the U.S. DOD Human Research Protections Office (HRPO, HRPO# E01877.1a) approved of all study activities, and all research was performed in accordance with relevant guidelines and regulations and the Declaration of Helsinki. All participants provided informed electronic consent. We did not compensate participants for participation.

Data Use Statement

Oura’s data use policy does not permit us to make wearable device data (collected via the Oura Ring) available to third parties. We can make self-report data available; please contact Ashley E. Mason and Benjamin L. Smarr to obtain an application to obtain these data.

Conflict of Interest

A.E.M. has received remuneration for consulting work from Ouraring Inc. but declares no non-financial competing interests. B.L.S. has received remuneration for consulting work from, and has a financial interest in, Ouraring Inc. but declares no other non-financial competing interests.

A. E. M., PhD, and B. L. S., PhD, are listed as co-inventors on patent applications as follows: 17/357,922, filed June 24, 2021, entitled "ILLNESS DETECTION BASED ON TEMPERATURE DATA," status is pending; PCT/US21/39260, filed June 25, 2021, entitled "ILLNESS DETECTION BASED ON TEMPERATURE DATA," status is expired; and 17/357,930, filed June 24, 2021, entitled "HEALTH MONITORING PLATFORM FOR ILLNESS DETECTION," status is pending. These were all filed as of July 2021 by Oura Health Oy on behalf of UCSD. All applications cover the use of wearable device data to detect illness onset.

References

1. Ji H, Gulati M, Huang TY, et al. Sex Differences in Association of Physical Activity with All-Cause and Cardiovascular Mortality. *J Am Coll Cardiol*. 2024;83(8):783-793. doi:10.1016/j.jacc.2023.12.019
2. Conger SA, Toth LP, Cretsinger C, et al. Time Trends in Physical Activity Using Wearable Devices: A Systematic Review and Meta-analysis of Studies from 1995 to 2017. *Medicine & Science in Sports & Exercise*. 2022;54(2):288-298. doi:10.1249/MSS.0000000000002794
3. Guthold R, Stevens GA, Riley LM, Bull FC. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet Global Health*. 2018;6(10):e1077-e1086. doi:10.1016/S2214-109X(18)30357-7
4. Time to tackle the physical activity gender gap. *The Lancet Public Health*. 2019;4(8):e360. doi:10.1016/S2468-2667(19)30135-5
5. Guthold R, Willumsen J, Bull FC. What is driving gender inequalities in physical activity among adolescents? *J Sport Health Sci*. 2022;11(4):424-426. doi:10.1016/j.jshs.2022.02.003
6. Costello JT, Bieuzen F, Bleakley CM. Where are all the female participants in Sports and Exercise Medicine research? Accessed December 17, 2024. https://onlinelibrary.wiley.com/doi/10.1080/17461391.2014.911354
7. Cowley ES, Olenick AA, McNulty KL, Ross EZ. “Invisible Sportswomen”: The Sex Data Gap in Sport and Exercise Science Research. Published online September 21, 2021. doi:10.1123/wspaj.2021-0028
8. Elliott-Sale KJ, Minahan CL, de Jonge XAKJ, et al. Methodological Considerations for Studies in Sport and Exercise Science with Women as Participants: A Working Guide for Standards of Practice for Research on Women. *Sports Med*. 2021;51(5):843-861. doi:10.1007/s40279-021-01435-8
9. McNulty KL, Elliott-Sale KJ, Dolan E, et al. The Effects of Menstrual Cycle Phase on Exercise

Performance in Eumenorrheic Women: A Systematic Review and Meta-Analysis. *Sports Med*. 2020;50(10):1813-1827. doi:10.1007/s40279-020-01319-3

10. Colenso-Semple LM, D’Souza AC, Elliott-Sale KJ, Phillips SM. Current evidence shows no influence of women’s menstrual cycle phase on acute strength performance or adaptations to resistance exercise training. *Front Sports Act Living*. 2023;5:1054542. doi:10.3389/fspor.2023.1054542
11. Smith ES, McKay AKA, Ackerman KE, et al. Methodology Review: A Protocol to Audit the Representation of Female Athletes in Sports Science and Sports Medicine Research. *Int J Sport Nutr Exerc Metab*. 2022;32(2):114-127. doi:10.1123/ijsnem.2021-0257
12. Szadvári I, Ostatníková D, Babková Durdiaková J. Sex differences matter: Males and females are equal but not the same. *Physiology & Behavior*. 2023;259:114038. doi:10.1016/j.physbeh.2022.114038
13. Zucker I, Prendergast BJ, Beery AK. Pervasive Neglect of Sex Differences in Biomedical Research. *Cold Spring Harb Perspect Biol*. 2022;14(4):a039156. doi:10.1101/cshperspect.a039156
14. Zucker I, Prendergast BJ. Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biol Sex Differ*. 2020;11(1):32. doi:10.1186/s13293-020-00308-5
15. Bruce LK, Kasl P, Soltani S, et al. Variability of temperature measurements recorded by a wearable device by biological sex. *Biology of Sex Differences*. 2023;14(1):76. doi:10.1186/s13293-023-00558-z
16. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neuroscience & Biobehavioral Reviews*. 2014;40:1-5. doi:10.1016/j.neubiorev.2014.01.001
17. Smarr BL, Grant AD, Zucker I, Prendergast BJ, Kriegsfeld LJ. Sex differences in variability across timescales in BALB/c mice. *Biology of Sex Differences*. 2017;8(1):7. doi:10.1186/s13293-016-0125-3
18. Becker JB, Prendergast BJ, Liang JW. Female rats are not more variable than male rats: a meta-analysis of neuroscience studies. *Biology of Sex Differences*. 2016;7(1):34. doi:10.1186/s13293-016-0087-5
19. Smarr B, Kriegsfeld LJ. Female mice exhibit less overall variance, with a higher proportion of structured variance, than males at multiple timescales of continuous body temperature and locomotive activity records. *Biology of Sex Differences*. 2022;13(1):41. doi:10.1186/s13293-022-00451-1
20. Smarr BL, Ishami AL, Schirmer AE. Lower variability in female students than male students at multiple timescales supports the use of sex as a biological variable in human studies. *Biol Sex Differ*. 2021;12:32. doi:10.1186/s13293-021-00375-2
21. Huhn S, Axt M, Gunga HC, et al. The Impact of Wearable Technologies in Health Research: Scoping Review. *JMIR Mhealth Uhealth*. 2022;10(1):e34384. doi:10.2196/34384
22. Maijala A, Kinnunen H, Koskimäki H, Jämsä T, Kangas M. Nocturnal finger skin temperature in menstrual cycle tracking: ambulatory pilot study using a wearable Oura ring. *BMC Women’s Health*. 2019;19(1):150. doi:10.1186/s12905-019-0844-9
23. Grant A, Smarr B. Feasibility of continuous distal body temperature for passive, early pregnancy detection. Yoon D, ed. *PLOS Digit Health*. 2022;1(5):e0000034. doi:10.1371/journal.pdig.0000034
24. Baker FC, Siboza F, Fuller A. Temperature regulation in women: Effects of the menstrual cycle. *Temperature*. 2020;7(3):226-262. doi:10.1080/23328940.2020.1735927
25. Klein A, Viswanath VK, Smarr B, Wang EJ. Detecting Periodic Biases in Wearable-Based Illness Detection Models. In: *ICLR 2023 Workshop on Time Series Representation Learning for Health*. ; 2023. https://openreview.net/forum?id=W0pLyiSuSSa

26. Mason AE, Hecht FM, Davis SK, et al. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study. *Sci Rep*. 2022;12(1):3463. doi:10.1038/s41598-022-07314-0
27. Hills AP, Mokhtar N, Byrne NM. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Front Nutr*. 2014;1:5. doi:10.3389/fnut.2014.00005
28. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: A Second Update of Codes and MET Values. *American College of Sports Medicine*. 2012;2012:126-127. doi:10.1016/j.yspm.2011.08.057
29. Purawat S, Dasgupta S, Song J, et al. TemPredict: A Big Data Analytical Platform for Scalable Exploration and Monitoring of Personalized Multimodal Data for COVID-19. In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE; 2021:4411-4420. doi:10.1109/BigData52589.2021.9671441
30. Kristoffersson A, Lindén M. A Systematic Review of Wearable Sensors for Monitoring Physical Activity. *Sensors (Basel)*. 2022;22(2):573. doi:10.3390/s22020573
31. The Pandas Development Team. pandas-dev/pandas: Pandas. Published online February 23, 2024. doi:10.5281/zenodo.10697587
32. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2
33. Terpilowski MA. scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software*. 2019;4(36):1169. doi:10.21105/joss.01169
34. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum Associates; 1988.
35. Lehr R. Sixteen S-squared over D-squared: A relation for crude sample size estimates. *Statistics in Medicine*. 1992;11(8):1099-1102. doi:10.1002/sim.4780110811
36. Waskom M. seaborn: statistical data visualization. *JOSS*. 2021;6(60):3021. doi:10.21105/joss.03021
37. Fernández-Martínez M, Vicca S, Janssens IA, Carnicer J, Martín-Vide J, Peñuelas J. The consecutive disparity index, D: a measure of temporal variability in ecological studies. *Ecosphere*. 2018;9(12):e02527. doi:10.1002/ecs2.2527
38. Heath JP. Quantifying temporal variability in population abundances. *Oikos*. 2006;115(3):573-581. doi:10.1111/j.2006.0030-1299.15067.x
39. Heath JP, Borowski P. Quantifying Proportional Variability. Zhang SD, ed. *PLoS ONE*. 2013;8(12):e84074. doi:10.1371/journal.pone.0084074
40. McArdle BH, Gaston KJ. The Temporal Variability of Densities: Back to Basics. *Oikos*. 1995;74(1):165. doi:10.2307/3545687
41. Fernández-Martínez M, Vicca S, Janssens IA, Carnicer J, Martín-Vide J, Peñuelas J. The consecutive disparity index, *D* : a measure of temporal variability in ecological studies. *Ecosphere*. 2018;9(12):e02527. doi:10.1002/ecs2.2527
42. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825-2830.
43. Clements MS, Armstrong BK, Moolgavkar SH. Lung cancer rate predictions using generalized additive models. *Biostatistics*. 2005;6(4):576-589. doi:10.1093/biostatistics/kxi028
44. Cui Z, Fritz BA, King CR, Avidan MS, Chen Y. A Factored Generalized Additive Model for Clinical Decision Support in the Operating Room. *AMIA Annu Symp Proc*. 2020;2019:343-352.
45. Servén D, Brummitt C, Abedi H. dswah/pyGAM: v0.8.0. Published online October 31, 2018. doi:10.5281/ZENODO.1476122
46. Kany S, Al-Alusi MA, Rämö JT, et al. Associations of “Weekend Warrior” Physical Activity

With Incident Disease and Cardiometabolic Health. *Circulation*. 2024;150(16):1236-1247. doi:10.1161/CIRCULATIONAHA.124.068669

47. Pollard TM, Wagnild JM. Gender differences in walking (for leisure, transport and in total) across adult life: a systematic review. *BMC Public Health*. 2017;17(1):341. doi:10.1186/s12889-017-4253-4
48. Shen M di, Chen S bing, Ding X dong. The effectiveness of digital twins in promoting precision health across the entire population: a systematic review. npj Digit Med. 2024;7(1):1-10. doi:10.1038/s41746-024-01146-0
49. Smarr BL. AI for precision medicine must keep non-random complexity in mind to support equity in outcomes. 2024 IEEE 20th International Conference on e-Science (e-Science). Published online September 16, 2024:1-7. doi:10.1109/e-Science62913.2024.10678664
50. Bernard MA, Clayton JA, Lauer MS. Inclusion Across the Lifespan: NIH Policy for Clinical Research. *JAMA*. 2018;320(15):1535-1536. doi:10.1001/jama.2018.12368
51. Rochon PA, Mason R, Gurwitz JH. Increasing the visibility of older women in clinical research. *The Lancet*. 2020;395(10236):1530-1532. doi:10.1016/S0140-6736(20)30849-7
52. Tompuri TT. Metabolic equivalents of task are confounded by adiposity, which disturbs objective measurement of physical activity. *Frontiers in Physiology*. 2015;6. doi:10.3389/fphys.2015.00226
53. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. *Surgery*. 2014;156(3):508-516. doi:10.1016/j.surg.2014.07.001
54. Madla CM, Gavins FKH, Merchant HA, Orlu M, Murdan S, Basit AW. Let's talk about sex: Differences in drug therapy in males and females. *Advanced Drug Delivery Reviews*. 2021;175:113804. doi:10.1016/j.addr.2021.05.014
55. Feldman S, Ammar W, Lo K, Trepman E, van Zuylen M, Etzioni O. Quantifying Sex Bias in Clinical Studies at Scale With Automated Data Extraction. *JAMA Netw Open*. 2019;2(7):e196700. doi:10.1001/jamanetworkopen.2019.6700
56. Hamberg K. Gender Bias in Medicine. *Womens Health (Lond Engl)*. 2008;4(3):237-243. doi:10.2217/17455057.4.3.237
57. Zucker I, Beery AK. Males still dominate animal studies. *Nature*. 2010;465(7299):690-690. doi:10.1038/465690a
58. The White House. Remarks of President Joe Biden -- State of the Union Address As Prepared for Delivery. The White House. March 8, 2024. Accessed March 18, 2024. https://www.whitehouse.gov/briefing-room/speeches-remarks/2024/03/07/remarks-of-president-joe-biden-state-of-the-union-address-as-prepared-for-delivery-2/
59. Woitowich NC, Woodruff TK. Implementation of the NIH Sex-Inclusion Policy: Attitudes and Opinions of Study Section Members. *Journal of Women's Health*. 2019;28(1):9-16. doi:10.1089/jwh.2018.7396
60. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev*. 2016;1(1):2. doi:10.1186/s41073-016-0007-6
61. Heidari S, Fernandez DGE, Coates A, et al. WHO's adoption of SAGER guidelines and GATHER: setting standards for better science with sex and gender in mind. *The Lancet*. 2024;403(10423):226-228. doi:10.1016/S0140-6736(23)02807-6
62. Mazure CM, Jones DP. Twenty years and still counting: including women as participants and studying sex and gender in biomedical research. *BMC Women's Health*. 2015;15(1):94. doi:10.1186/s12905-015-0251-9
63. Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-Based Wearable Activity Trackers Increase Physical Activity Participation: Systematic Review and Meta-Analysis.

JMIR mHealth and uHealth. 2019;7(4):e11819. doi:10.2196/11819

64. Kyytsönen M, Vehko T, Anttila H, Ikonen J. Factors associated with use of wearable technology to support activity, well-being, or a healthy lifestyle in the adult population and among older adults. *PLOS Digit Health*. 2023;2(5):e0000245. doi:10.1371/journal.pdig.0000245
65. Suorsa K, Leskinen T, Rovio S, et al. Weekday and weekend physical activity patterns and their correlates among young adults. *Scandinavian Journal of Medicine & Science in Sports*. 2023;33(12):2573-2584. doi:10.1111/sms.14475

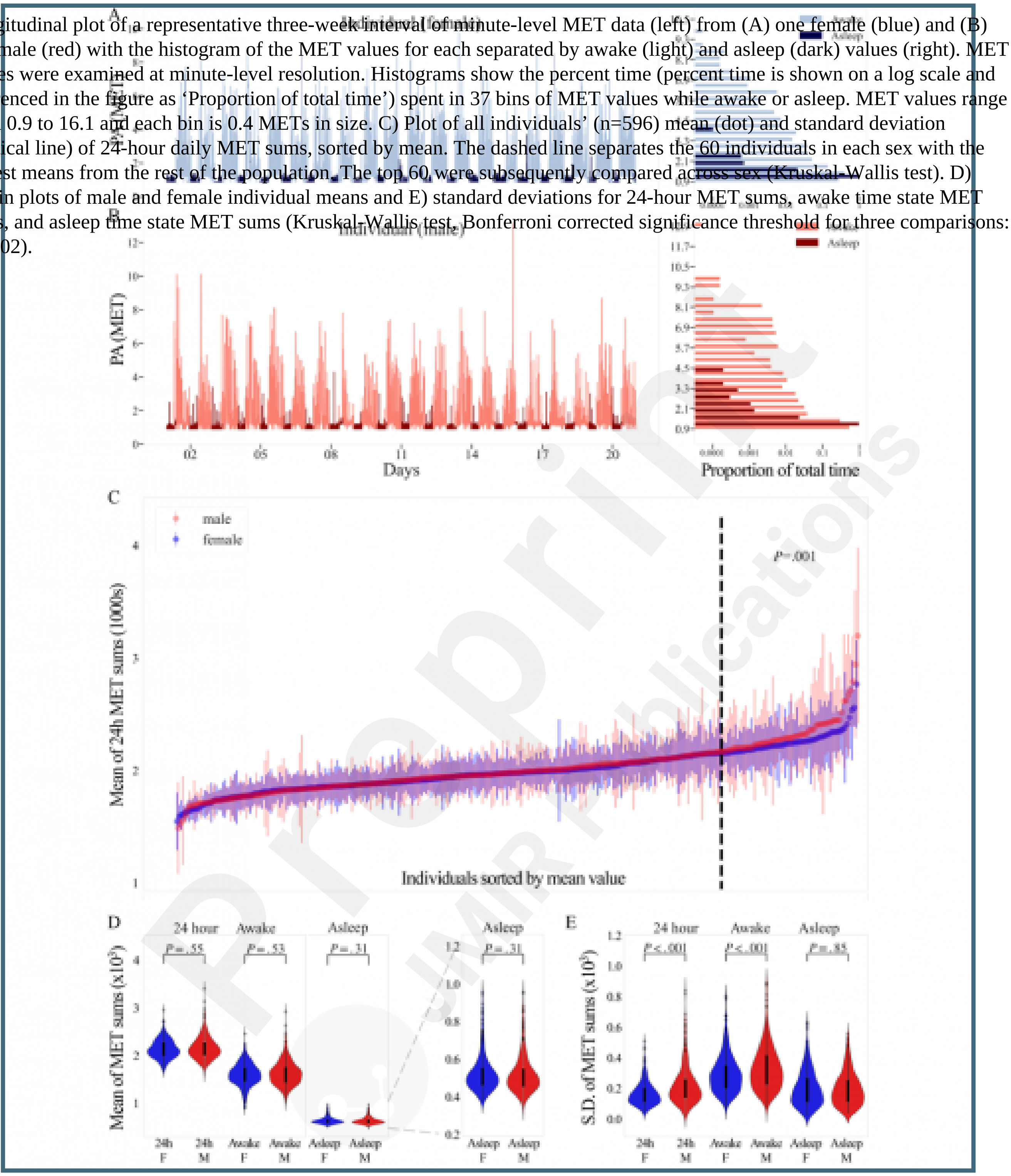
Abbreviations

MET: Metabolic equivalent task, or metabolic equivalents
PA: Physical activity
CV: Coefficient of variation
PV: Proportional variability index
CDI: Consecutive disparity index
IQR: Interquartile range
GAM: Generalized additive model
WR: Weekend Rhythm

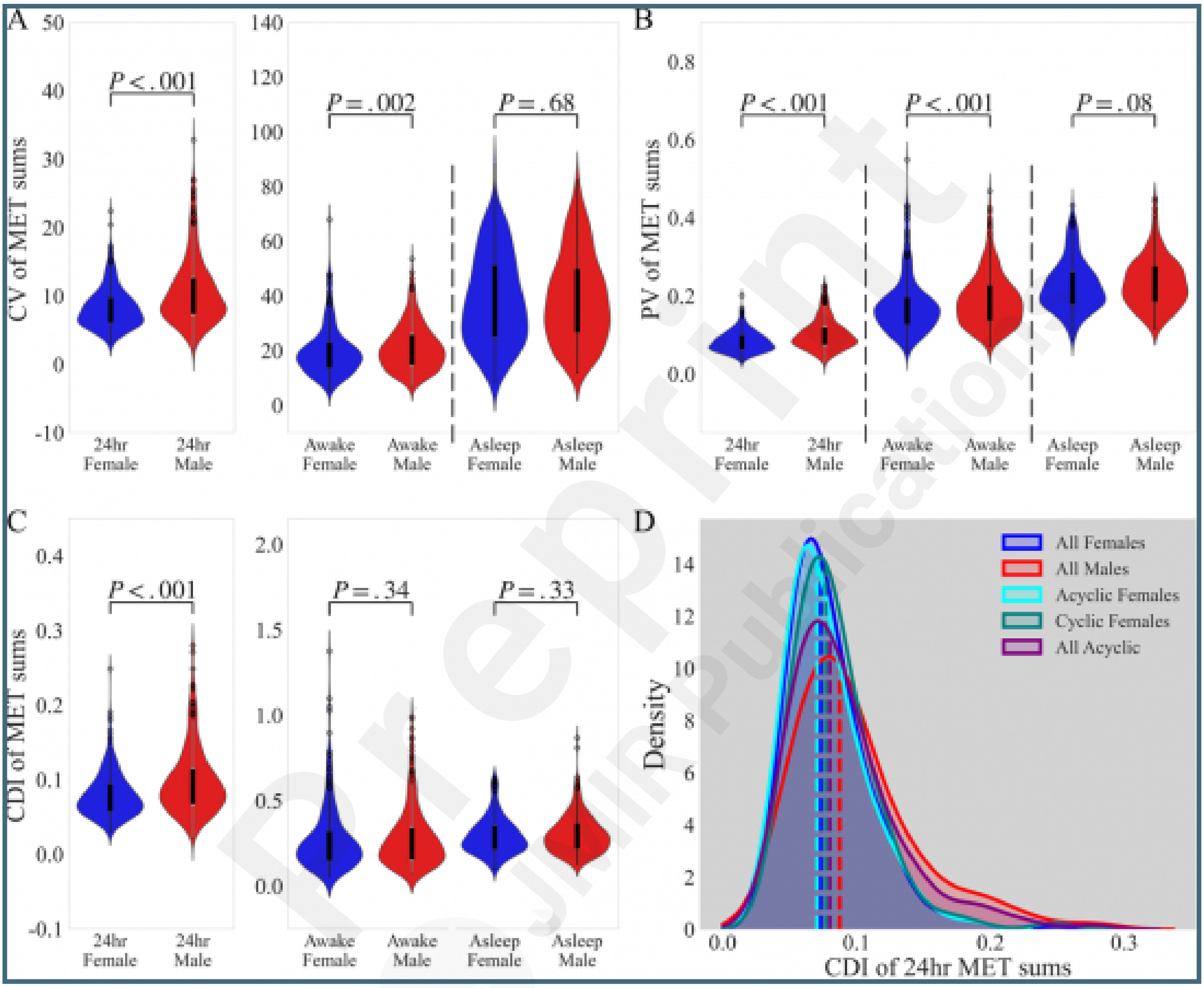
Supplementary Files

Figures

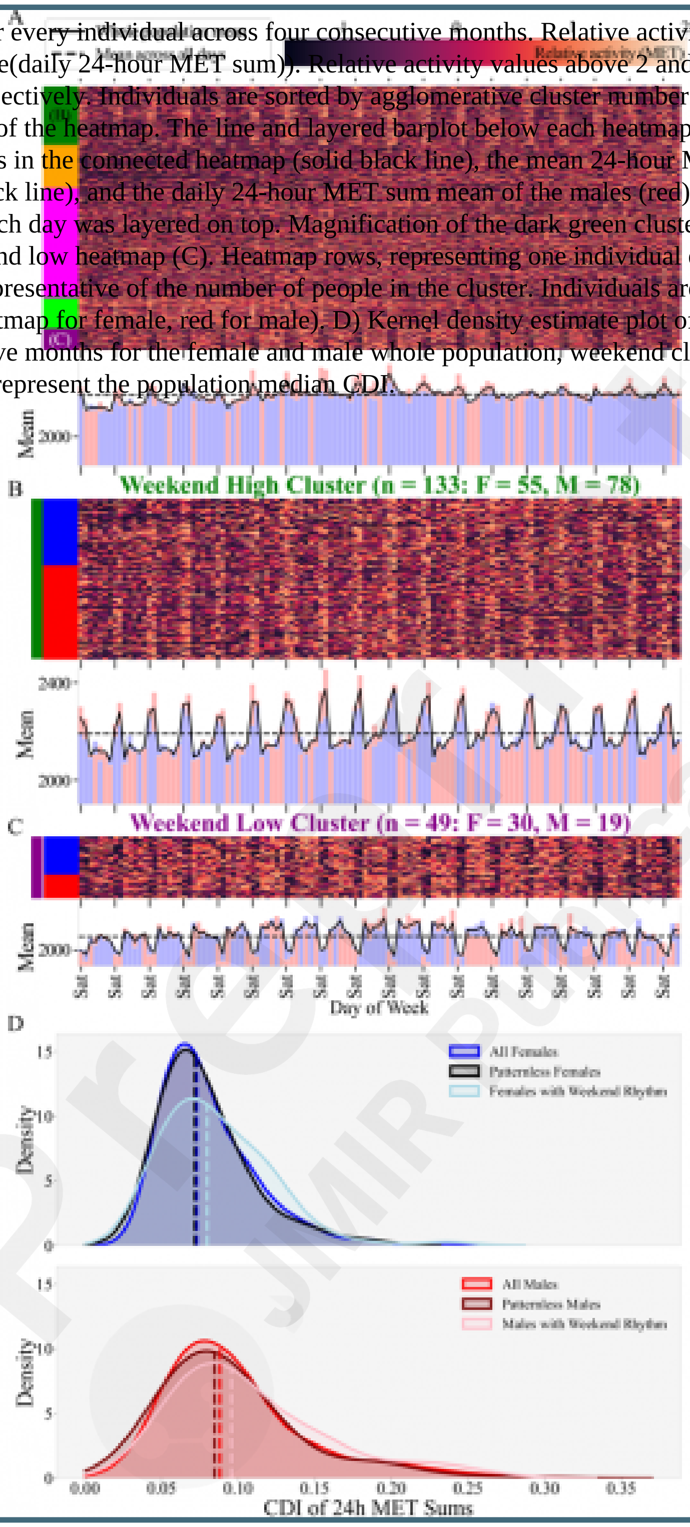
Longitudinal plot of a representative three-week interval of minute-level MET data (left) from (A) one female (blue) and (B) one male (red) with the histogram of the MET values for each separated by awake (light) and asleep (dark) values (right). MET values were examined at minute-level resolution. Histograms show the percent time (percent time is shown on a log scale and referenced in the figure as ‘Proportion of total time’) spent in 37 bins of MET values while awake or asleep. MET values range from 0.9 to 16.1 and each bin is 0.4 METs in size. C) Plot of all individuals’ (n=596) mean (dot) and standard deviation (vertical line) of 24-hour daily MET sums, sorted by mean. The dashed line separates the 60 individuals in each sex with the largest means from the rest of the population. The top 60 were subsequently compared across sex (Kruskal-Wallis test). D) Violin plots of male and female individual means and E) standard deviations for 24-hour MET sums, awake time state MET sums, and asleep time state MET sums (Kruskal-Wallis test, Bonferroni corrected significance threshold for three comparisons: P=0.02).



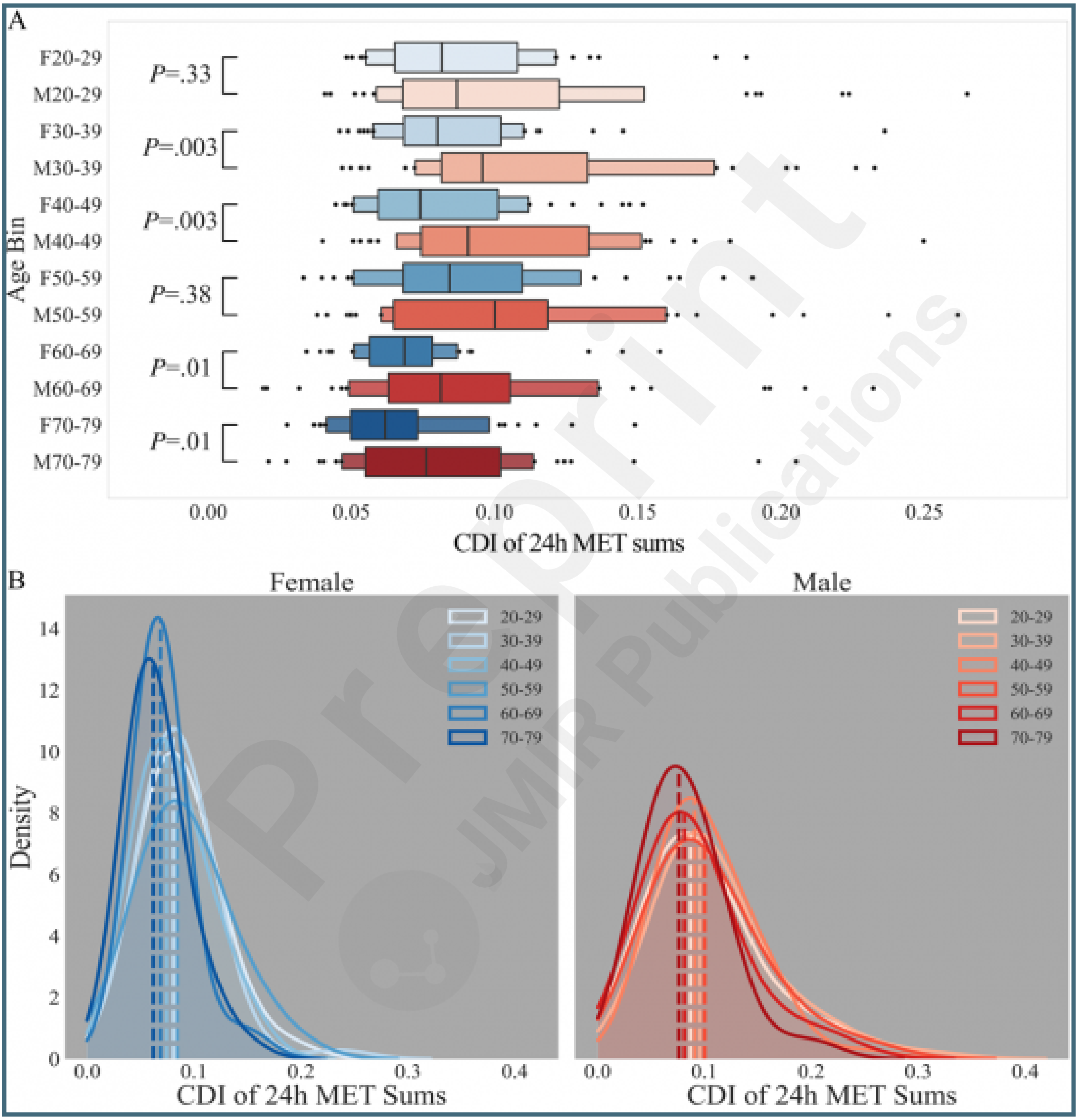
Violin plots of female (blue) and male (red) distribution of A) coefficient of variation (CV), B) proportional variability index (PV), and C) consecutive disparity index (CDI) for 24-hour MET sums, awake time state MET sums, and asleep time state MET sums. (Kruskal-Wallis, Bonferroni corrected significance threshold for three comparisons: P=0.02), and D) kernel density estimate plots of all female (blue), all male (red), acyclic female (teal), cyclic female (blue-green), and all acyclic individuals of either sex (purple). Group median CDI: dashed vertical lines.



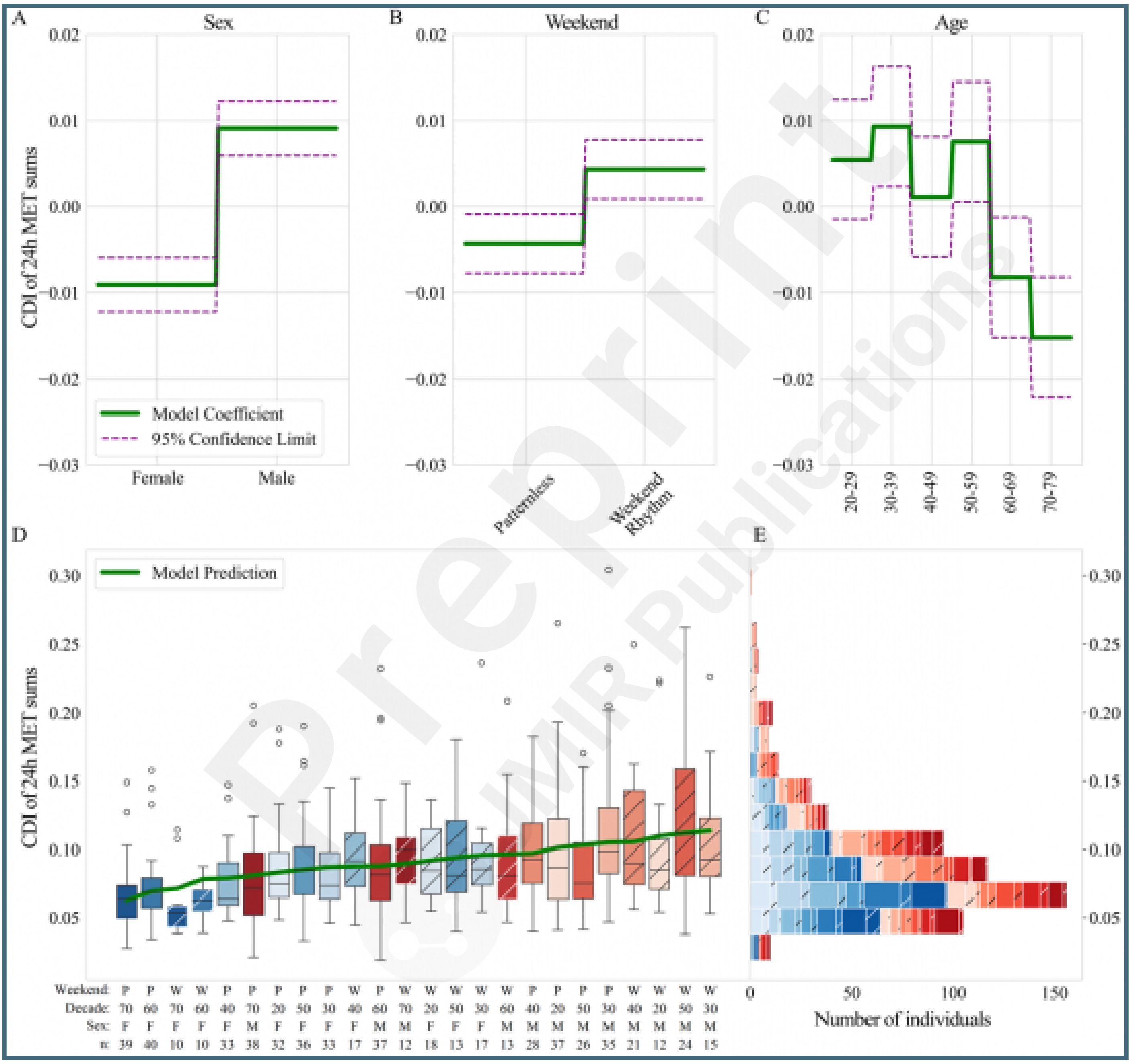
Heatmap of relative activity for every individual across four consecutive months. Relative activity was defined as arctan(2*intraindividual z-score(daily 24-hour MET sum)). Relative activity values above 2 and below -1.5 are colored with the lightest and darkest values respectively. Individuals are sorted by agglomerative cluster number and clusters are demarcated by the colors in the bar to the left of the heatmap. The line and layered barplot below each heatmap show the daily mean 24-hour MET sum across all individuals in the connected heatmap (solid black line), the mean 24-hour MET sum across all days in the four-month period (dashed black line), and the daily 24-hour MET sum mean of the males (red) and females (blue) where the sex with the lower mean for each day was layered on top. Magnification of the dark green cluster: weekend high heatmap (B); and dark purple cluster: weekend low heatmap (C). Heatmap rows, representing one individual each, are all of equal size so that the height of the heatmap is representative of the number of people in the cluster. Individuals are labeled and sorted by sex (blue box on the left of the heatmap for female, red for male). D) Kernel density estimate plot of consecutive disparity index calculated from four consecutive months for the female and male whole population, weekend cluster population, and other clusters. Vertical dashed lines represent the population median CDI.



A) Boxenplot of consecutive disparity indices for each sex-age category (Kruskal-Wallis, Bonferroni corrected significance threshold for six comparisons: P=.008). B) Female and male kernel density estimate plots of consecutive disparity index (CDI) in each age bin. Dashed lines represent the median CDI of the sex-age population.



GAM fitted factor functions for sex (A), weekend rhythm (B), and age (C) with confidence intervals. D) Boxplot of consecutive disparity index of 24-hour MET sums for each unique phenotype group in order of the model prediction (green line) for that phenotype group. Age-sex categories are colored as they were in Figure 4A and hatching indicates the presence of a weekend rhythm in individuals in the labeled group (P = patternless, W = weekend rhythm). E) Stacked histogram (ordered for visual clarity) of the number of individuals in CDI bins labeled by phenotype group, highlighting the overlap of each group in most bins.



Multimedia Appendixes

Tables recording population standard deviations of each sex for each MET sum metric and for each sex subgroup for 24-hour MET sums, and figures related to data filling (Supplementary Figures 1-5). Population standard deviations are presented here for their relevance to power analysis.
URL: http://asset.jmir.pub/assets/ef00e0a0490c476097090b063dbd228e.doc