UNIVERSITY OF CALIFORNIA SAN DIEGO

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

A thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Bioengineering

by

Kristin J Varner

Committee in charge:

Professor Benjamin Smarr, Chair
Professor Brian Aguado
Professor Kevin King

2024

The thesis of Kristin J Varner is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

Thesis Approval Page iii
Table of Contents iv
List of Figures vi
List of Tables vii
Acknowledgements viii
Abstract of the Thesis ix
Introduction 1
Results 4
Cohort and MET Data Foundational Analysis 4
Variability Metrics of MET Sums 7
Analysis of Cyclicity 8
Analysis of Weekly Rhythms 10
Analysis of Age 14
Generalized Additive Model of Those Features Found to Have Significant Impact on Variability of CDI of 24-Hour MET Sums Across Individuals: Sex, Age, and Weekly Rhythm 19
Discussion 21
Methods 24
Data source 24
Data Preprocessing 24
Subjects 24
Data Filling 25
Statistical Methods 27
Cohort and MET Data Foundational Analysis 27
Line plot and histogram of two individual's MET values 27
Daily MET sums 28
Whole population mean and standard deviation of 24-hour MET Sums 28
Mean and variability metrics of MET sums by sex and time state 28
Coefficient of variation (CV) 29
Proportional variability index (PV) 29
Consecutive disparity index (CDI) 29
Analysis of Age, Cyclicity, and Weekly Rhythm 30
Analysis of Cyclicity 30
Analysis of Weekly Rhythms 31
Analysis of Age 32
Generalized Additive Model of Those Features Found to Have Significant Impact on Variability of CDI of 24-Hour MET Sums Across Individuals: Sex, Age, and Weekly Rhythm 32
Declarations 34
Ethics Statement 34
Data Use Statement 34

Competing Interests 34
Appendix 36
Supplementary Table 1 36
Supplementary Table 2 37
Supplementary Figure 1 38
Supplementary Figure 2 39
Supplementary Figure 3 40
Supplementary Figure 4 41
Supplementary Figure 5 42
References 43

LIST OF FIGURES

Figure 1: Longitudinal plots of a three-week interval of metabolic equivalents (MET) data with a histogram. Plot of all individuals' mean and standard deviation of 24-hour daily MET sums. Violin plots of male and female individual means and standard deviations for MET sum metrics. 6

Figure 2: Violin plots of female and male distribution of coefficient of variation, proportional variability index, and consecutive disparity index for MET sum metrics. All female, all male, acyclic female, cyclic female, and all acyclic individuals (of either sex) distributions of consecutive disparity index (CDI). 9

Figure 3: Heatmap of relative activity for every individual, individuals with a weekend high effect, and individuals with a weekend low effect. Distribution of consecutive disparity index for the female and male whole population, weekend effect population, and patternless population. 13

Figure 4: Boxenplot of consecutive disparity indices for each sex-age category with Bonferroni corrected significance annotations for comparisons across sex within age bin. Female and male distribution of consecutive disparity index in each age bin. 17

Figure 5: Generalized additive model fitted factor functions for sex, weekend rhythm, and age with confidence intervals. Boxplot of consecutive disparity index of 24-hour MET sums for each unique group. Stacked histogram of the number of individuals in CDI bins labeled by group. 20

LIST OF TABLES

Table 1: Mean and Standard Deviation Statistics by Sex and Time State: Population median of male and female intra-individual mean and standard deviations for each time state with Mann-Whitney comparison across sex distributions for each time state and statistical metric. 7

Table 2: Variability Metric Statistics by Sex and Time State: Median of male and female variability metric distributions with Mann-Whitney comparison and Bonferroni corrected significance annotations for comparisons across sex for each time state and metric. 10

Table 3: Age Bin Statistics: Diagonal: 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. Above/right of diagonal: Cohen's d effect sizes of the comparisons that were significantly different. 16

Table 4 and Table 5: Female (Blue, Top) and Male (Red, Bottom) Age Bin Statistics: The median consecutive disparity index of each age bin, *p*-value of the post hoc Dunn's test comparing each age group, and Cohen's d effect sizes of the comparisons that were significantly different. 18

Table 6: Kruskal Wallis *p*-value (test statistic) for CDI distribution comparison between the whole female or male population and the female or male whole population without each age group. 18

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Benjamin Smarr and Lauryn Bruce who have both provided me with excellent mentorship and have created and fostered a uniquely supportive environment for students studying women's health.

In addition, I would like to acknowledge all co-authors for their contributions to this thesis. Thank you to Lauryn and Benjamin for your contributions to the conception and design of this study. Thank you to Ashley E. Mason, Anoushka Chowdhary, Leena Pandya, Benjamin, Stephan Dilchert, and Frederick M. Hecht for your contributions to participant recruitment and wearable and survey data collection. Thank you to Subhasis Dasgupta and Ilkay Altintas for your contributions to system development and data storage. Thank you to Lauryn and Severine Soltani for your contributions to data cleaning. Thank you to Lauryn, Severine, and Benjamin for your contributions to data analysis. Thank you to Lauryn, Benjamin, Ashley, Wendy Hartogensis, Frederick, and Stephan for your contributions to the manuscript preparation. Lastly, thank you to all co-authors for your contributions to the manuscript review and editing.

The content of this thesis including the abstract, introduction, results, discussion, methods, declarations, and appendix is currently being prepared for publication. Kristin 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. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

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

by

Kristin J Varner

Master of Science in Bioengineering

University of California San Diego, 2024

Professor Benjamin Smarr, Chair

Sex is an important consideration in biomedical research. Efforts to expand sex inclusion have had some success, but females are still underrepresented in both animal and human biomedical research despite increasing evidence in support of sex-inclusive study

design. Hesitancy to include female subjects is partially due to the hypothesis that ovarian rhythms increase female variability and weaken statistical power. We recently used continuous skin temperature data from wearable devices to test this hypothesis and found that the data did not support the hypothesis that females, cycling or not, reduce statistical power. However, ovarian rhythms are not the only timescale of change that might shape variability in human data. Additionally, whereas temperature is linked to endocrine and physiological systems, physical activity captured by wearables may be more related to behavioral patterns that exist independently of endogenous physiological rhythms, and so is worthy of investigation separate from temperature.

Here we used minute-level metabolic equivalent task (MET) data spanning 206 days each from 596 individuals to explore physical activity (PA), focusing on comparing the scale of sex differences in variability of PA to the scale of differences in variability arising at the timescales of days, weeks, menstrual cycles, and decades of life. We report that females have lower intra-individual variability than males as a whole and the presence of menstrual cycles did not increase variability. PA patterns reflective of behavioral patterns were found on weekly time scales and across decades of life. The exclusion of either sex was not supported by our analysis.

INTRODUCTION

Persistent sex bias in subject selection for biomedical research in humans and its detrimental impact on women's healthcare has been thoroughly described previously [1–4]. The harmful trend of excluding 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; this attention has led to marked improvements in cohort selection [5,6]. While progress has been achieved, many researchers still fail to include subjects of both sexes in experiments, and those who do often fail to perform sex-stratified analyses [5].

Historically, sex bias arose in part from concerns about statistical power: if female variance is similar to male variance but also includes menstrual cycle variability, then females should be expected to generate data with higher overall variability than males for comparable observations or tasks, leading to weakened statistical power for comparable tests [7]. Numerous recent investigations have rejected this hypothesis in mice [8–11]. To our knowledge, two investigations of this hypothesis have been performed in humans [12,13]. In the most recent study, we tested the hypothesis using continuous longitudinal data for the first time. We used a data set of minute-level skin temperature recordings from 600 individuals (300 males, 300 females) covering six months each and found that these data supported the inclusion of females with analysis specific to the presence or absence of relevant physiological differences such as the presence of menstrual cycles in females [13]. That is, temperature cycling linked to menstrual cycles [14] is present in a subset of individuals identifying as female, and while the variability is not obviously greater at multiple timescales in any of these groups, the means and the temporal structure of the variability do predictably differ by group. This recent work suggests two subsequent questions: 1) are the findings similar in other modalities than skin temperature; and 2) do other timescales of change besides the menstrual cycle contribute to structured variability in group-specific ways?

Previous studies have demonstrated that multiple timescales of change can interact to give rise to non-random structure in variability of human data [12,13,15]. This temporal structure arises specifically from interactions between physiological rhythms, such as menstrual and circadian rhythms, societal phenomena such as the 5-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 semi-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 will yield reduced power for detecting real effects by treating all sources of variability as equivalent. By contrast, when the non-random variability is accounted for, the residual variability is by definition lower, and the statistical power is improved for the same analysis. In this way, the historical concern about menstrual cycle variability was in part due to an underlying concern of structured but historically unaccountable variability from menstrual cycles.

The emergence of digital tools in daily life has led to a rapid change in the amount of longitudinal data that can be easily generated on an individual study subject [16]. It is now not only feasible but also relatively inexpensive to generate continuous, longitudinal data at high temporal resolution over large populations by using wearable sensor devices. The increasing scale of data that is feasibly generatable also allows for the accounting of previously unaccountable sources of variability. Our previous work took advantage of this by using minute-level skin temperature data, generated by Oura Ring device users in the real world, to test and reject the hypothesis that females can be expected to be statistically more variable than males [13]. We used temperature following previous work that indicated that skin temperature can be used to identify physiological changes, especially menstrual cycles [17]. We confirmed this and developed tools to allow us to identify cyclic females from their skin temperature patterns across months. We also found that cyclic and acyclic females showed substantially different patterns of change over time, so “menstrually cyclic” was a more informative label than “sex” when predicting the structure of variability in an

individual's skin temperature over time.

Here we sought to use the same cohort of subjects as this previous analysis to assess the effect of sex, ovarian rhythms, and non-ovarian rhythms on variability of physical activity, and identify groups or subgroups with significantly unique structures of variability. This analysis is a contrast to the previous in that activity is primarily a behavioral variable, and temperature is primarily a physiological variable. The Oura Ring generated activity data at the same 1-minute resolution as temperature, allowing us to reuse the cohort with cyclic and acyclic labels identified in the previous study and omit only four individuals due to insufficient activity data (596 individuals: 298 males and females). Oura Ring reports activity in the form of metabolic equivalent tasks (METs) [18]. METs are used to express the intensity of an activity as multiples of the MET achieved while at rest (1.0), where 2.0 METs is roughly twice as intense as 1.0 MET [19]. Here we quantify intra-individual variability patterns, and then assess the relative amplitude of change of these metrics by sex and timescale.

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 (Fig. 1A, Fig. 1B). We observed variation in MET values between awake and asleep states with increased MET during awake time periods, as expected (Fig. 1A and Fig. 1B, left). Finding that the distribution of MET values appeared highly dependent on asleep or awake state (Fig. 1A and Fig. 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, Fig. 1C, Fig. 1D). However, we observed an apparent increase in male mean of 24-hour MET sums at the upper extreme (Fig. 1C). Consistent with this observation, a comparison of individuals' mean of 24-hour MET (Fig. 1C, right) revealed that the 60 males with the largest average 24-hour MET sum had a significantly larger average than the top 60 females (Kruskal Wallis, H = 10.25, p = 1.37E-3, Cohen's d (d) = 0.56). We also observed differences between male and female 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, Fig. 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). 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, ** $p < 0.01$). Violin plots of male and female individual means (D) and standard deviations (E) for 24-hour MET sums, awake time state MET sums, and asleep time state MET sums. (Mann-Whitney, Bonferroni corrected significance annotations for 3 comparisons - n.s.: non-significant, ** $p < 3.33E\text{-}3$, **** $p < 3.33E\text{-}5$)

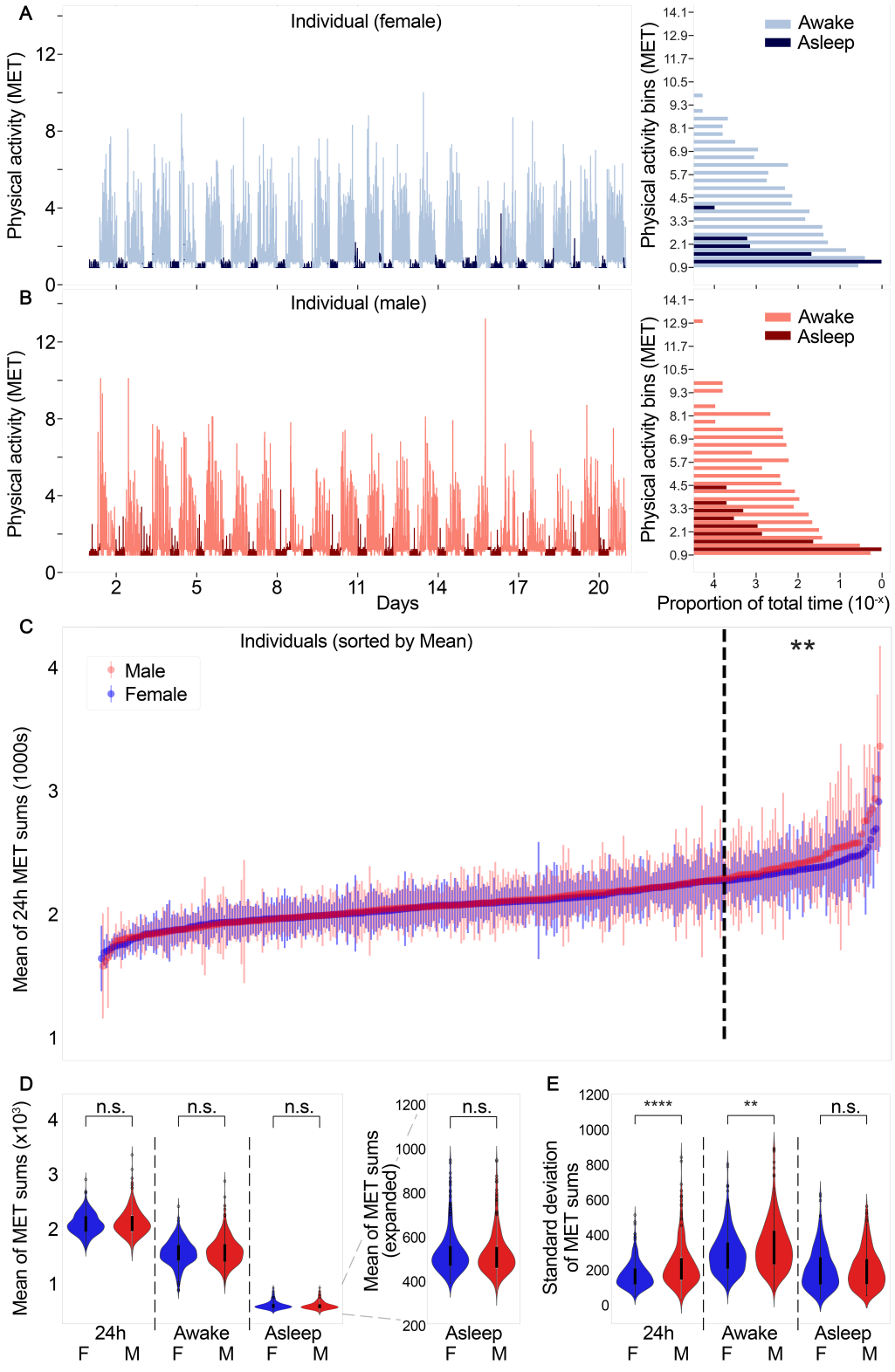

A
Individual (female)
Physical activity (MET)
Physical activity bins (MET)
Awake
Asleep
B
Individual (male)
Days
Proportion of total time (10-x)
C
Individuals (sorted by Mean)
Male
Female
**
Mean of 24h MET sums (1000s)
D
Mean of MET sums (x10³)
n.s.
Mean of MET sums (expanded)
24h
Awake
Asleep
F
M
E
Standard deviation of MET sums
****
**

Table 1. Mean and Standard Deviation Statistics by Sex and Time State: Population median of male and female intra-individual mean and standard deviations for each time state with Mann-Whitney comparison across sex distributions (p-value (U statistic)) for each time state and statistical metric. (Mann-Whitney, Bonferroni corrected significance annotations for three comparisons - ** $p < 3.33E\text{-}3$, **** $p < 3.33E\text{-}5$)

	Female Median			Male Median			Mann-Whitney Comparison		
	24-hour	Awake	Asleep	24-hour	Awake	Asleep	24-hour	Awake	Asleep
Mean	2108	1595	506	2113	1610	502	0.548 (4.31E4)	0.527 (4.31E4)	0.314 (4.65E4)
Standard Deviation	163	287	190	201	322	185	**** 5.37E-10 (3.14E4)	** 6.61E-4 (3.72E4)	0.854 (4.40E4)

Variability Metrics of MET Sums

In addition to standard deviation, we used three other metrics to analyze intra-individual 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 [13]. We included CV and PV here for the same validation and focused on CDI because it is the most appropriate metric of variability for this data due to its accounting of chronological order and its lack of dependence on the mean for its calculation. 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 (Fig. 2A-B, Table 2). 24-hour MET sum CDI was significantly larger for males than females (Fig. 2C, Table 2). In all three of these metrics, asleep MET sum variability was not significantly different across sexes (Fig. 1D, Table 1, Fig. 2A-C, Table 2).

The Cohen's d effect size between male and female CDI of 24-hour MET sums was 0.49. The difference between the median CDIs was 0.014, whereas the interquartile range (IQR) of 24-hour CDI was 0.035 for females and 0.046 for males. The difference between

medians was about three times smaller than the IQRs, showing that most of the variability in CDI was not explained by sex.

Analysis of Cyclicity

Cyclic females and all acyclic people (male or female) did not have significantly different mean 24-hour MET sums (Kruskal Wallis, H = 0.456, p = 0.499, data not shown) or significantly different CDI of 24-hour MET sums (Kruskal Wallis H = 1.03, p = 0.309, Fig. 2D). However, we found a significant difference between male, cyclic female, and acyclic female CDI of 24-hour MET sums (Kruskal Wallis, H = 32.4, p = 9.41E-8, Fig. 2D). A Dunn's test showed that females were less variable than males, regardless of cyclic status (males vs. cyclic females: p = 5.88E-3, d = 0.35. males vs. acyclic females: p = 2.18E-8, d = 0.53), and that cyclic females and acyclic females were not significantly different (p = 0.092). Furthermore, removing cyclic females from the female population did not significantly reduce the whole female CDI of 24-hour MET sums (Kruskal Wallis, H = 0.752, p = 0.386) (Fig. 2D).

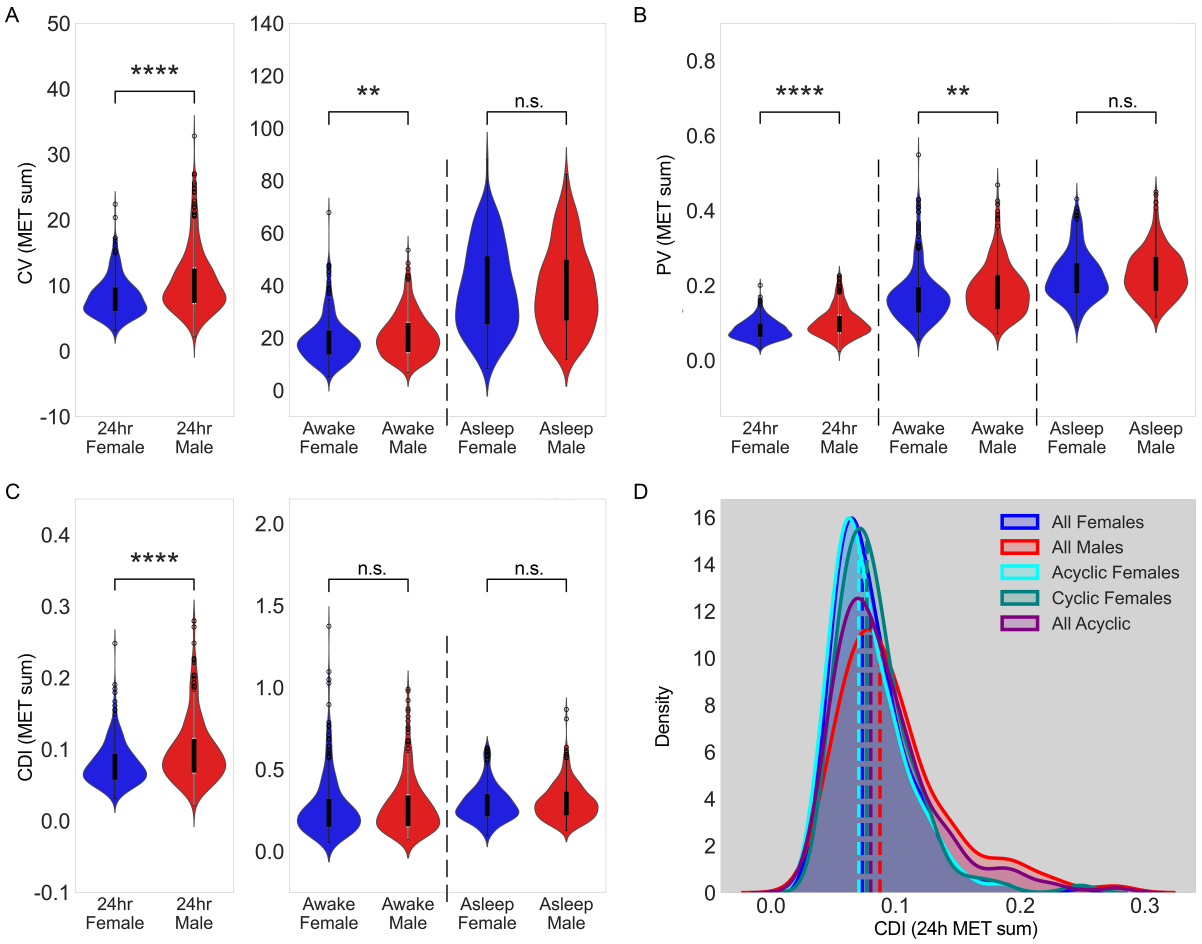


Figure 2. Violin plots of female (blue) and male (red) distribution of A) coefficient of variation (CV), B) proportional variability index (PV), and C) consecutive disparity index (CDI) for 24-hour MET sums, awake time state MET sums, and asleep time state MET sums. (Mann-Whitney, Bonferroni corrected significance annotations for 3 comparisons - * $p < 0.017$, ** $p < 3.33E\text{-}3$, *** $p < 3.33E\text{-}4$, **** $p < 3.33E\text{-}5$) D) All female (blue), all male (red), acyclic female (teal), cyclic female (blue-green), and all acyclic individuals of either sex (purple) distributions of CDI. Group median CDI: dashed vertical lines.

Table 2. Variability Metric Statistics by Sex and Time State: Median of male and female variability metric distributions with Mann-Whitney comparison across sex (p-value (U statistic)) for each time state and metric. (Mann-Whitney, Bonferroni corrected significance annotations for 3 comparisons - * $p < 0.017$, ** $p < 3.33E\text{-}3$, *** $p < 3.33E\text{-}4$, **** $p < 3.33E\text{-}5$)

	Female Median			Male Median			Mann-Whitney Comparison		
	24-hour	Awake	Asleep	24-hour	Awake	Asleep	24-hour	Awake	Asleep
Coefficient of Variation	7.72	18.0	36.2	9.07	19.7	37.3	**** 3.83E-11 (3.05E4)	** 2.22E-3 (3.80E4)	0.676 (4.35E4)
Proportional Variability Index	0.078	0.163	0.214	0.092	0.179	0.228	**** 7.46E-10 (3.15E4)	** 9.26E-4 (3.74E4)	0.078 (4.07E4)
Consecutive Disparity Index	0.073	0.229	0.274	0.087	0.240	0.284	**** 5.56E-8 (3.30E4)	0.343 (4.24E4)	0.328 (4.23E4)

Analysis of Weekly Rhythms

Agglomerative clustering of 4 months of data per individual across the whole cohort revealed clusters of individuals sharing prominent weekly structures (Fig. 3A). Two clusters of individuals with weekend patterns were identified: a ‘weekend high’ group (Fig. 3B) and a ‘weekend low’ group (Fig. 3C). The three clusters without weekend patterns are referred to as ‘patternless’ clusters.

Significant differences in mean of 24-hour MET sums existed between individuals in the weekend high cluster, weekend low cluster, and the patternless clusters (Kruskal Wallis: $H = 9.18$, $p = 0.010$, 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: $p = 6.97E\text{-}3$. weekend high vs. patternless: $p = 0.014$). Cohen’s d effect sizes between significant groups were 0.48 (weekend high vs. weekend low) and 0.25 (weekend high vs. patternless).

Next, we grouped the individuals with any weekend pattern (weekend high or weekend low) to examine variability. The group of individuals in the weekend cluster had significantly larger CDI of 24-hour MET sums than individuals in the patternless cluster (Kruskal Wallis, H = 10.1, p = 1.46E-3, data not shown). The CDI IQR of the weekend cluster and patternless cluster was 0.048 and 0.042, respectively. The difference in the medians (weekend cluster median: 0.086, patternless cluster median: 0.077) between the two populations was 0.0089 and the Cohen's d effect size was 0.28. The difference between the medians was about five times smaller than the IQRs of the groups thus weekly rhythm did not explain most of the variability in activity. Additionally, weekly rhythms explained less variability than sex where the difference between the median CDI of the male and female population was only three times smaller than the IQR.

We found significant effects of sex and cluster on 24-hour MET sum CDI (Kruskal Wallis, H = 34.6, p = 1.48E-7, Fig. 3D). Males have larger CDI of 24-hour MET sums than females in the same cluster (Dunn's test: patternless cluster: p = 7.16E-5, Cohen's d (d) = 0.43. weekend cluster: p = 3.18E-3, d = 0.51). Additionally, males in the weekend cluster had significantly larger 24-hour MET sum CDI than females from the patternless cluster (Dunn's test, p = 3.22E-8, d = 0.81); however, females in the weekend cluster did not have significantly larger 24-hour MET sum CDI than males in the patternless cluster (Dunn's test, p = 0.242). We found no significant effect between clusters within sex on 24-hour sum CDI: males in the weekend cluster did not differ from males in the patternless cluster (Dunn's test, p = 0.020, Bonferroni corrected significance level: * p < 8.33E-3), nor did females in the weekend cluster differ from females in the patternless cluster (Dunn's test, p = 0.062). Whole population male and female CDI distributions were not significantly affected by the removal of individuals with weekend patterns (Kruskal Wallis, all females vs. females without weekend clusters: H = 0.579, p = 0.447, all males vs. males without weekend clusters: H = 1.02, p = 0.312) (Fig. 3D).

Figure 3. A) Heatmap of relative activity for every individual across 4 consecutive months. Relative activity was defined as arctan(2*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).

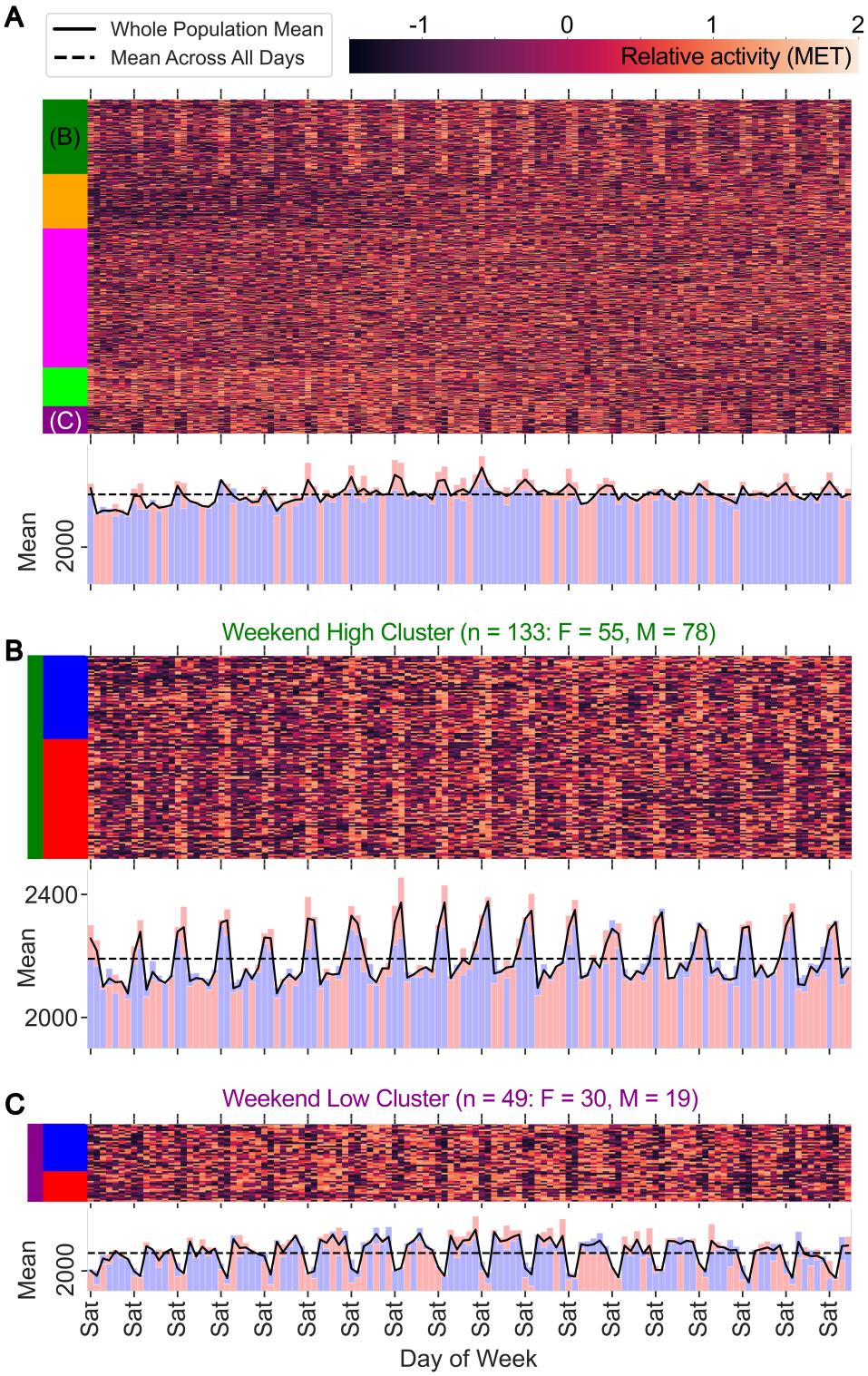
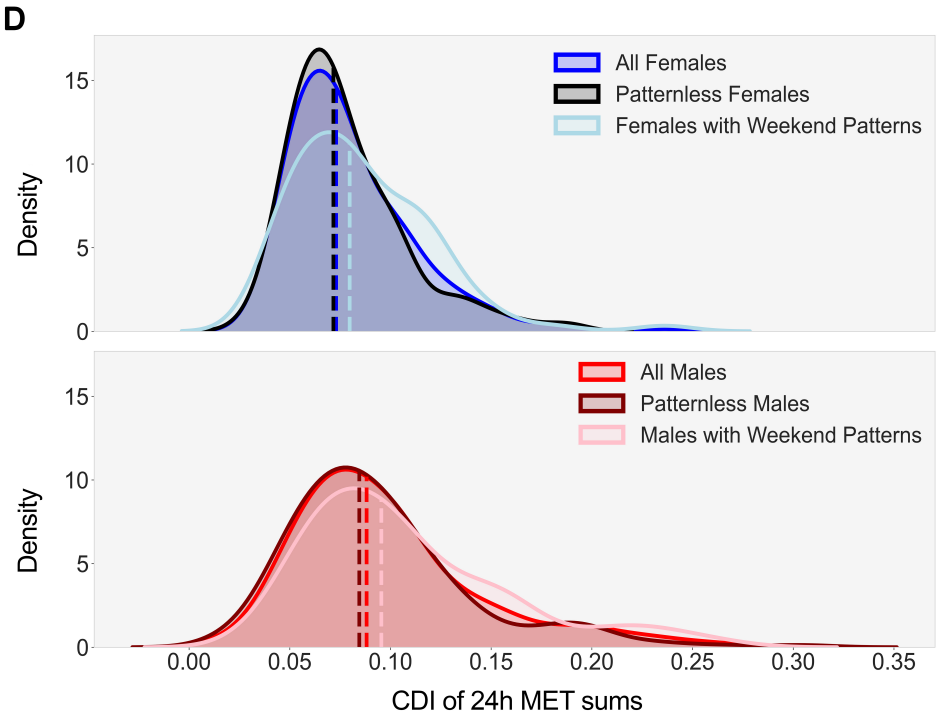

A
Whole Population Mean
Mean Across All Days
-1
0
1
2
Relative activity (MET)
(B)
(C)
Mean
2000
Weekend High Cluster (n = 133: F = 55, M = 78)
B
2400
Mean
2000
Weekend Low Cluster (n = 49: F = 30, M = 19)
C
Mean
2000
Sat
Day of Week

Figure 3 Continued. D) Distribution of consecutive disparity index calculated from four consecutive months for the female and male whole population, weekend pattern population, and patternless population. Vertical dashed lines represent the population median CDI.



Analysis of Age

We found significant differences in mean 24-hour MET sums across age groups (all individuals, Kruskal Wallis: H = 24.3, p = 1.90E-4, 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 = 1.43E-5, d = 0.68. 70-79 vs. 50-59: p =4.56E-4, d = 0.55), 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 = 2.68E-3, d = 0.39). 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 = 40.6, p = 1.15E-7, 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 interquartile ranges by age bin were: 20-29 = 0.049, 30-39 = 0.037, 40-49 = 0.047, 50-59 = 0.053, 60-69 = 0.033, 70-79 = 0.041. The average difference between medians in age groups that had significantly different CDI of 24-hour MET sum was 0.019. The average difference between medians was about 2.0 to 2.5 times smaller than the IQRs for each age group. Therefore, age did not explain most of the variability in CDI of 24-hour MET sums. However, age explained a similar amount of variability as sex, where the difference between medians was about three times smaller than the IQR.

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 when a Bonferroni correction for six comparisons was applied (Kruskal Wallis: 30-39 M vs. 30-39 F: $H = 8.62$, $p = 3.32\text{E-}3$, $d = 0.61$. 40-49 M vs. 40-49 F: $H = 8.64$, $p = 3.29\text{E-}3$, $d = 0.61$. Fig. 4A). We further note that while the remaining post hoc 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: 20-29 M vs. 20-29 F: $H = 0.96$, $p = 0.327$. 50-59 M vs. 50-59 F: $H = 0.78$, $p = 0.378$. 60-69 M vs. 60-69 F: $H = 6.58$, $p = 0.010$. 70-79 M vs. 70-79 F: $H = 6.38$, $p = 0.012$. Fig. 4A, Table 4, Table 5).

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 20-29 and 50-59 (Fig. 4B, Table 4). Cohen's d effect sizes for these differences were between 0.635 and 0.845 (Table 4). Males aged 70-79 were significantly less variable than males aged 30-39 with a 0.665 Cohen's d effect size (Fig. 4B, Table 5). No single age group significantly increased the CDI of the whole male or female population (Table 6).

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 and starred. Above/right of diagonal: lightly shaded cells show Cohen's d effect sizes of the comparisons that were significantly different. (Kruskal Wallis, Bonferroni corrected significance annotations for 15 comparisons * $p < 3.33E\text{-}3$, ** $p < 6.67E\text{-}4$, *** $p < 6.67E\text{-}5$, **** $p < 6.67E\text{-}6$)

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.082					0.562
30-39	0.375	0.087			0.442	0.647
40-49	0.682	0.195	0.081			0.506
50-59	0.643	0.674	0.382	0.089	0.417	0.622
60-69	4.75E-3	2.00E-04**	0.016	1.01E-03*	0.070	
70-79	3.46E-05***	4.71E-07****	1.90E-04**	4.12E-06****	0.184	0.065

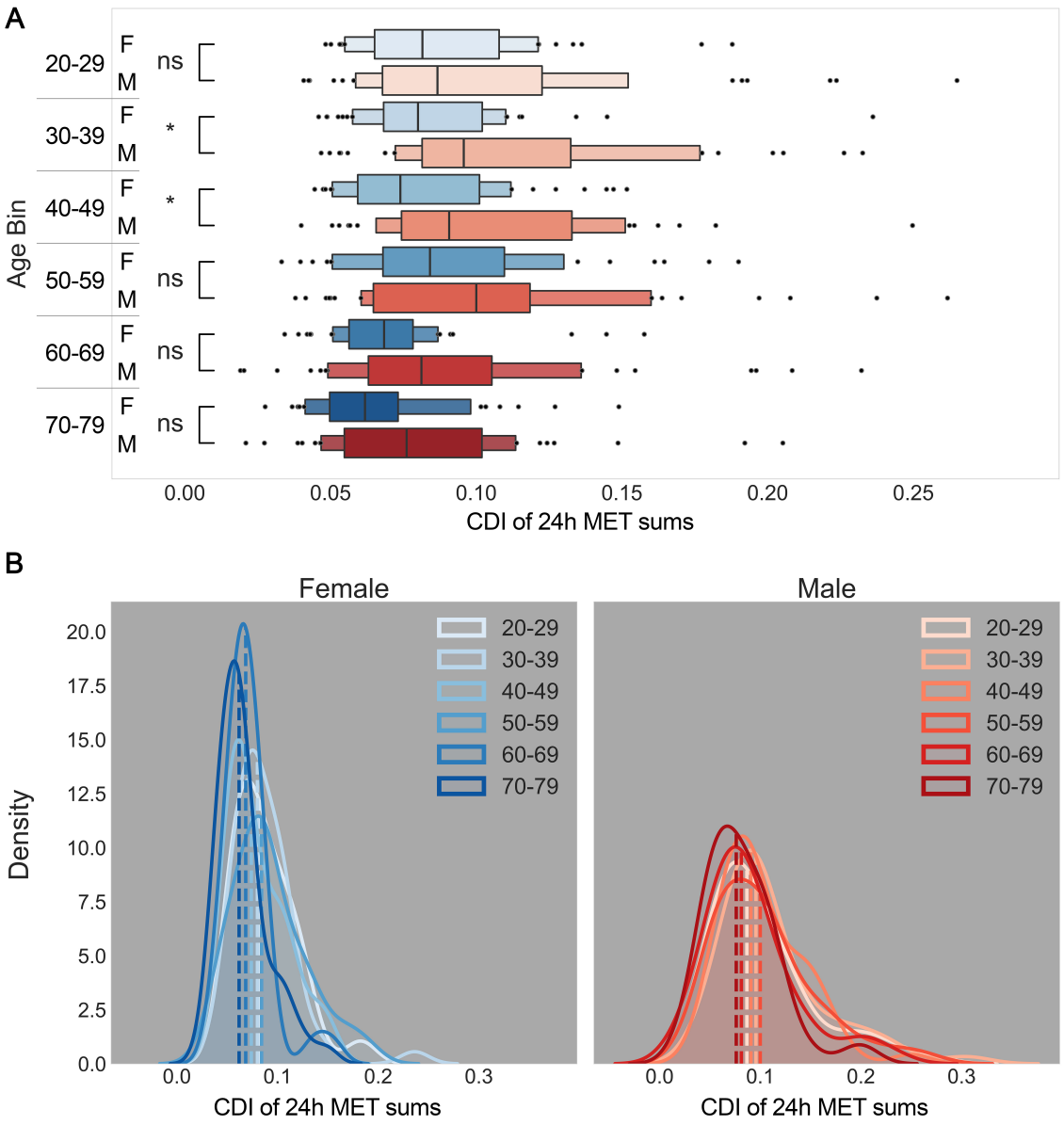


Figure 4. A) Boxenplot of consecutive disparity indices for each sex-age category (Kruskal Wallis, Bonferroni corrected significance annotations for 6 comparisons - * $p < 8.3E\text{-}3$, ns = not significant). B) Female and male distribution of consecutive disparity index (CDI) in each age bin. Dashed 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 and starred. Above/right of diagonal: lightly shaded cells show Cohen's d effect sizes of the comparisons that were significantly different. (Kruskal Wallis, Bonferroni corrected significance annotations for 15 comparisons * $p < 3.33E\text{-}3$, ** $p < 6.67E\text{-}4$, *** $p < 6.67E\text{-}5$, **** $p < 6.67E\text{-}6$)

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.082				0.635	0.803
30-39	0.882	0.080				0.734
40-49	0.167	0.217	0.074			
50-59	0.768	0.658	0.095	0.084	0.694	0.845
60-69	2.62E-3 *	4.23E-3	0.104	1.00E-3*	0.068	
70-79	2.48E-5 ***	4.72E-5 ***	4.50E-3	7.13E-6 ***	0.221	0.062

	20-29	30-39	40-49	50-59	60-69	70-79
20-29	0.087					
30-39	0.176	0.096				0.665
40-49	0.555	0.448	0.091			
50-59	0.724	0.315	0.809	0.100		
60-69	0.289	0.015	0.098	0.156	0.081	
70-79	0.049	8.45E-4*	0.010	0.020	0.361	0.076

Table 6. Kruskal Wallis *p*-value (test statistic) for CDI distribution comparison between the whole female or male population and the female or male whole population without each age group.

	Removed Age Group					
	20-29	30-39	40-49	50-59	60-69	70-79
Female	0.508 (0.44)	0.554 (0.35)	0.99 (0.0002)	0.434 (0.61)	0.453 (0.56)	0.194 (1.68)
Male	0.954 (0.003)	0.486 (0.49)	0.741 (0.11)	0.822 (0.051)	0.659 (0.19)	0.384 (0.76)

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

A generalized additive model (GAM) was used to summarize the contributions of sex, age, and weekly rhythm on the variability of CDI of 24-hour MET sums across individuals (Fig. 5). Unique combinations of the categories across the variables resulted in 24 groups (e.g., Female, 20-29, weekend pattern was one group) for which the model predicted a CDI value. Each of the variables had a significant effect on the model prediction (sex: p-value = 2.06E-7, weekly rhythm: p-value = 0.013, age: p-value = 1.26E-5). Model parameters indicated that sex and specific age groups had the greatest effect on CDI out of these categories: sex (Fig. 5A) had an overall effect of ±0.0091 (decreased for females and increased for males), weekend patterns (Fig. 5B) had an overall effect of ±0.0043 (decreased for patternless and increased for weekend patterns), and age group (Fig. 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 model was 11.3%, indicating a low proportion of null deviance explained by the model. This is consistent with our analyses that found the difference in median CDIs between groups to be smaller than the size of the interquartile ranges (IQRs) of the groups themselves (see sections on Sex, Weekly Rhythms, and Age, above; 2x, 3x, and 5x respectively). Together, both of these analyses indicated that even those timescales of change that were significant sources of variability were not also substantial sources of variability that would likely confound statistical comparisons across individuals from different groups. GAM analysis added that the intersection of sex at specific age groups (30-39, 50-59, 60-69, and 70-79) affected the GAM prediction the most, but further confirmed that no such group is in itself a substantial source of variability in the population. Model predictions did not align with unique values for each group and there was significant overlap between groups in CDI range (Fig. 5D-E).

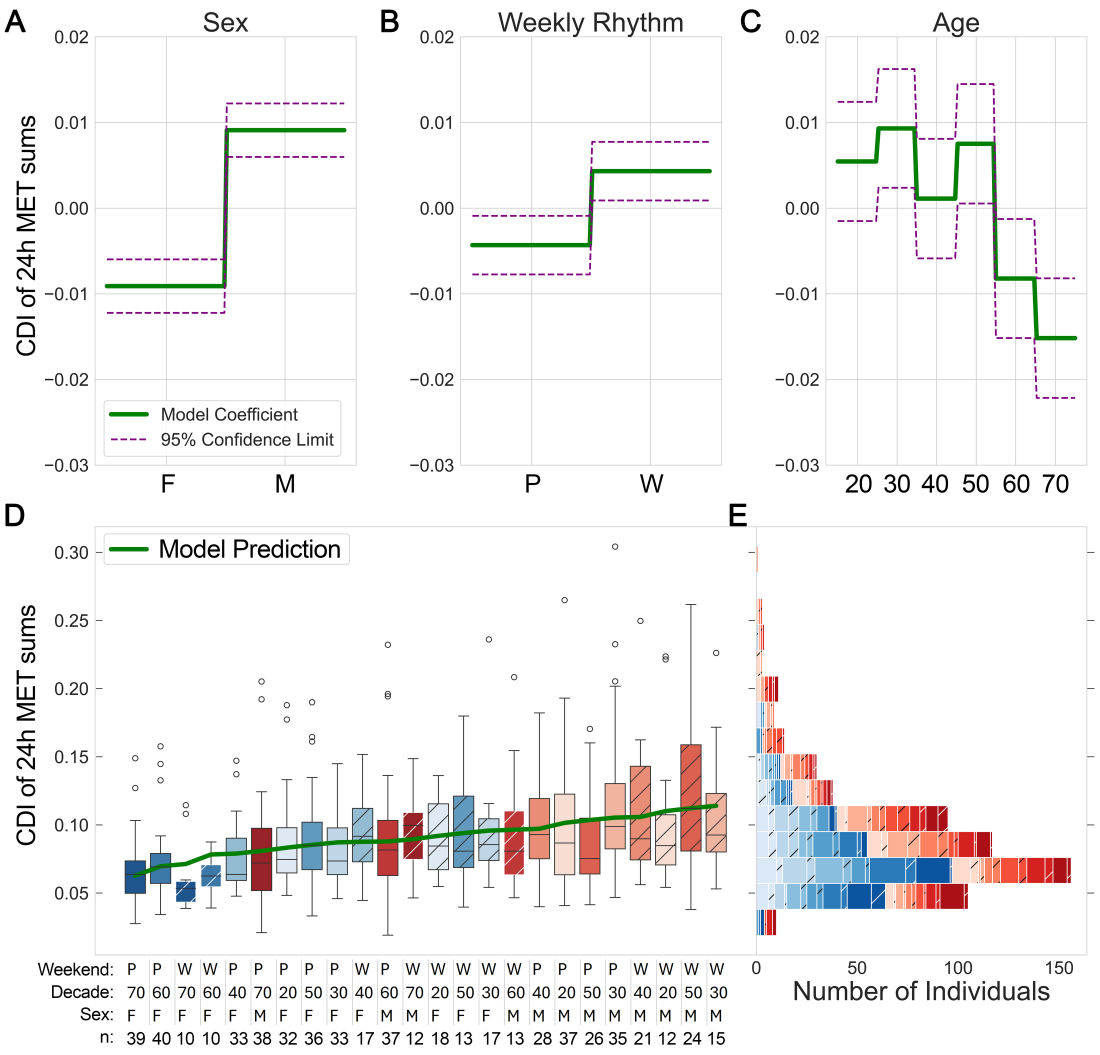


Figure 5. GAM fitted factor functions for sex (A), weekend rhythm (B), and age (C) with confidence intervals. D) Boxplot of consecutive disparity index of 24-hour MET sums for each unique group in order of the model prediction (green line) for that category. Age-Sex categories are colored as they were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

DISCUSSION

Here we found evidence to reject the hypothesis that it is statistically expedient to exclude female subjects from human behavioral biomedical research on the basis that menstrual cycles substantially increase the intra-individual variability of physical activity (PA). Rather, we found that females have significantly less intra-individual variability than males regardless of their cycling status. Sex alone appears to be a poor proxy for behavioral patterns that impact statistical comparisons, consistent with our findings for physiological patterns in temperature [13].

Furthermore, menstrual cyclicity is not the only timescale that appears to bear consideration when categorizing research subjects based on the structure of the variability over time; indeed, cyclicity had less overall significant impact than did any other variable analyzed. Weekly rhythm and age had significant effects on intra-individual variability, where menstrual cycling alone did not. Males tended to show higher variability across these two other timescales, but sex differences were most pronounced in midlife (30s and 40s). We also found that sex differences existed in the presence of weekly rhythms; interestingly, those with weekend effects were more likely to be male, though both sexes were represented in this group. The different effects seen in time of day, menstrual cycling, weekly patterns, and age on mean and intraindividual variability all suggest that sex alone is not an effective proxy for the presence of structured variability or of the intra-individual variability that may affect statistical analysis. Instead, more specific phenotypes defined by intersections of these categories might allow for a more nuanced accounting of explainable variability per individual in a given study (for instance, our analyses suggest that knowing that someone is a male in their 30s with weekend effects would be more useful in anticipating from whom they will be most statistically different than any of those individual categories alone).

Our multivariate analysis revealed that older females with weekend patterns appear to have the least intraindividual variability of all subject phenotypes (Fig. 5D), perhaps in-

dicating stronger behavioral routines in this group. Ironically, older females are historically even more understudied than females broadly [20,21], but would appear to have mitigated concerns about increased variability eroding statistical comparisons more than any other group, including the most included 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 variability that would preclude statistical comparison. Rather, this highlights that assumptions about who should be excluded 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). As longitudinal data become more prevalent in human studies, recognizing the need to cluster individuals into phenotypes with similar patterns of change across time should allow for improved precision in statistical comparisons without requiring the exclusion of any sex or specific phenotype. While the multivariate analysis suggests that sex and age affect intra-individual variability the most out of the four variables studied, none of these variables alone nor the intersection of these variables reliably predicted intra-individual variability, suggesting that no group is so different from the others as to warrant a need for statistical exclusion.

The key assertion is that in the context of PA, which is at present the most commonly used longitudinal measure for humans, we found no support for the hypothesis that females broadly are more variable than males. This study aligns with our previous findings about sex and menstrual impacts on variability in continuous temperature data [13]. As those analyses and these stemmed from the same cohort, it is possible that new cohorts would show different distributions, and so expanded studies like this 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 3 characteristics of age, sex, and weekly rhythm had a substantially reduced N, and so may well not be reliably representative of the broader population of older women. 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. As such, while MET does not provide insights into total absolute activity or types of activity, METs change as a function of intensity of activity and nevertheless provide a means of assessing different timescales of behavioral change across individuals' data, as we analyzed here. While METs have been found to have systematic inaccuracies in energy expenditure estimates due to their reliance on body weight for calculation [22], this does not affect the relative change we analyzed in intra-individual variability. 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 any application or field of research.

In conclusion, our findings support sex-based and age-based analysis in biomedical research, while rejecting the exclusion of females, males, weekend types, or any other specific intersectional phenotype from biomedical research based on assumptions of increased variability. This variability can increasingly be captured and analyzed on relevant timescales of change. We therefore suggest that broad categorization as proxies for variability, as in sex being a broad proxy for menstrual cyclicity, may be able to be improved in some cases. Empirical clustering of observed patterns of variability into physiologically derived phenotypes ("physiotypes") is more descriptive of variability than previously used broad categorizations. This is not to reject the use of sex as a variable, but rather, to say that within each sex there is substantial diversity and that new data sources make some of this diversity newly measurable. There are groups of individuals whose variability can be defined through the use of longitudinal data from an increasing number of sources on an increasing number of variables. In time, we hope this new abundance allows for more precise definitions of physiological diversity than broader categories like sex allow for.

METHODS

Data Source

Data originates from the TemPredict Study [18]. 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.

Data Preprocessing

High resolution (per minute and per five minutes) and nightly aggregated data were provided by Oura Ring. Data was stored in large parquet files on the San Diego supercomputer (SDSC) and accessed through the Nautilus Portal [23]. Nightly data, also referred to as sleep summary data, was 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 was recorded at 1-min intervals for 24-hours per day. These data were date-time indexed and normalized to '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. Triaxial accelerometers were used to estimate metabolic equivalents (MET) at 60s resolution during both sleep and wake periods [18]. The exact MET calculation is proprietary to Oura Ring and not known to us.

Subjects

Subjects were obtained by filtering methods described in "Variability of temperature measurements recorded by a wearable device by biological sex" [13]. Briefly, 62,653 subjects were determined to have suitable physiological and demographic data. Survey response to the question 'What is your biological sex? Male, Female, Other (please describe)." was used to determine participants' sex.

Filtering for individuals who have all data type files and temperature data present 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, and 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 [24]. All MET values below 0.9 were dropped and participants were evaluated for missingness over 206 days between April and October 2020. Four participants, two of each sex, with percent missingness above 29% were removed. 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 present in each age bin: 20-29, 30-39, 40-49, 50-59, 60-69, 70-79.

Data Filling

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 (An interval of 1440 describes a full missing day (60 minutes x 24 hours = 1440 minutes)). 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.

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 occur.

2. Linear interpolation was achieved with the interpolate method from the Python package pandas (*pandas.DataFrame.interpolate*, version 2.2.1, https://panas.pydata.org/))[25]. 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 on the daily sum.

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 simulated intervals of missingness were then filled using each of the four filling methods. After filling, the predicted values in the sleep state dataframe 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 difference between the predicted and actual values of the test indexes. 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 Appendix: 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 Appendix: Supplementary Figure 4). The best filling method was applied to each dataset before any analyses were performed.

The sum of absolute difference 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: $H = 1.97$ $p = 0.374$, Appendix: Supplementary Figure 3. Sleep State: $H = 0.256$, $p = 0.880$, Appendix: Supplementary Figure 5).

Statistical Methods

P-values were reported as $p = 0.0XX$ for values greater than or equal to 0.010 and using scientific notation as $p =$ XE-Y for Y values greater than or equal to three. A Bonferroni correction was applied to all analyses that compare more than two groups, such that the threshold for significance (0.05) was divided by the number of comparisons made.

The Cohen's d Effect size was used to describe the magnitude of the difference between two significantly different populations [26]. Cohen's d Effect size was calculated with the pingouin Python package (*pingouin.compute_effsize*, version 0.5.4, https:// pingouin-stats.org/) [27].

The interquartile range (IQR) and the difference between group medians were used to approximate the proportion of variability that a variable (sex, age, etc) explains. If the IQRs of two significantly different groups were smaller than the difference between those groups' medians, the variable that differentiates those groups would be considered to explain most of the variability between those groups.

Cohort and MET Data Foundational Analysis

Line plot and histogram of two individual's MET values

A random subset of 20 consecutive days of data from two randomly selected individuals of each sex was chosen for this analysis. MET values were examined at minute-level

resolution. Histograms show the percent 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. The percent time spent in each bin is shown in log scale.

Daily MET sums

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 because the source of variability of daily MET sums is different in each state. Variability in 24-hour sums is due to sleep movement, awake movement, and intentional exercise, where movement is considered PA that results in a MET value of 1.5 or less and intentional exercise is considered PA that results in a MET value greater than 1.5 [24]. Variability in awake daily sums is due to time spent awake, intentional exercise, and awake movement. Variability in asleep daily sums is due to sleep duration and sleep movement.

Whole population mean and standard deviation of 24-hour MET Sums

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 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 with the SciPy Python library (*scipy.stats*, version 1.11.2, https://scipy.org/) [28] between the top 60 males and the top 60 females.

Mean and variability metrics of MET sums by sex and time state

The mean and four variability metrics of daily MET sums were calculated for each person and time state (24-hour, awake, asleep) from all 206 days. Distributions for each sex were constructed from each member's mean or variability metric. Distributions were compared using the Mann-Whitney-Wilcoxon two-sided test (SciPy Python library, *scipy.stats*,

version 1.11.2, https://scipy.org/) [28] with Bonferroni correction for 3 comparisons per metric (alpha values and annotation calculated manually). Four 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.

Coefficient of variation (CV)

CV is a common metric for describing temporal variability [29]. Here it describes a participant's standard deviation across all 206 days relative to their mean across all 206 days;

$$CV = \frac{\sigma}{mean}$$

CV is limited by its sensitivity to rare events and its dependence on the mean [29].

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 [29–32];

$$PV = \frac{2\sum 1 - \frac{min(z_i, z_j)}{max(z_i, z_j)}}{n(n-1)}$$

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 [29].

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 [29]. CDI describes time series

variability through the average rate of change between consecutive values;

$$CDI = \frac{1}{n-1} \sum_{i=1}^{n-1} \left| ln \frac{p_{i+1}}{p_i} \right|$$

where n = length of time series and p_i = value in series at time i [29].

Analysis of Age, Cyclicity, and Weekly Rhythm

All analyses including age, cyclicity, and weekly rhythms as variables focus exclusively on the CDI of daily 24-hour MET sums. 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. The CDI variability metric was chosen due to its accounting for chronological order.

Analysis of Cyclicity

Every participant's cyclic status was determined through methods described in "Variability of temperature measurements recorded by a wearable device by biological sex" (Bruce 2023)[13]. Briefly, autocorrelation profiles were generated from nightly maximum temperature recordings. Cyclic individuals' temperature trend deviation autocorrelation signals show wave-like patterns, while acyclic individuals do not exhibit such patterns. Profiles were classified as cyclic or acyclic by hierarchical clustering of pairwise distances between signals (pairwise distances calculated with dynamic time warping). 105 participants in this cohort were classified as cyclic, and 193 were classified as acyclic.

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). 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). The effect of cyclic females on the variability of the whole female population was calculated via a Kruskal Wallis test (*scipy.stats*, version 1.11.2, https://scipy.org/) [28] between the whole female population and the acyclic female population (All females vs females without

cyclic females).

Analysis of Weekly Rhythms

The weekly rhythm of every participant was determined by agglomerative clustering of four consecutive months of z-scored 24-hour MET sum data. Agglomerative clustering was performed using the scikit learn Python package (*sklearn.cluster.AgglomerativeClustering*, version 1.1.3, https://scikit-learn.org/stable/) [33]. To determine the optimal number of clusters and confirm that patterns were not artifacts of data filling, we examined a hierarchically clustered heatmap of the unfilled data and observed two groups with different weekly patterns: relatively high (weekend high) and relatively low (weekend low) 24-hour MET sums on weekends. Four clusters allowed for the recovery of these two groups in agglomerative clustering of the filled data. To confirm the presence of the pattern observed on the heatmap, we calculated the average 24-hour MET sum for each day in the consecutive four months across all participants, across participants in the weekend high cluster, and across participants in the weekend low cluster. These averages were visualized as a line plot with the mean across all days in that group layered on top. Compared to the averages across all individuals, the weekend low and weekend high group's averages were visually depressed or elevated respectively on weekends.

Mean and CDI of 24-hour MET sums were calculated for each individual over the four consecutive months used to categorize their weekly rhythm and compared across weekly rhythm categories. The means were compared across weekend high, weekend low, and patternless (no weekly rhythm) groups while the CDI was only compared across weekend pattern (the aggregated group of individuals with either weekend high or weekend low rhythm) and patternless groups. CDI was only compared across the presence or absence of a weekly 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 were also compared across groups of individuals with unique combinations of sex and weekly rhythm (weekend pattern male, weekend pattern

female, patternless male, and patternless female). The effect of weekly rhythm on the variability of the whole male and female population was calculated via a Kruskal Wallis test (*scipy.stats*, version 1.11.2, https://scipy.org/) [28] between the whole population and the patternless population (all females vs. patternless females and all males vs. patternless males).

Analysis of Age

Mean and CDI of 24-hour MET sums were calculated for each individual over all 206 days and compared across age categories. CDI of 24-hour MET sums were also compared across sex groups in the same age category and across age categories within the same sex group. A boxenplot (Seaborn Python library, *seaborn.boxenplot*, version 0.12.2, https://seaborn.pydata.org/index.html) [34], or letter-value plot, was used to visually compare males and females within age groups. 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 via a Kruskal Wallis test (*scipy.stats*, version 1.11.2, https://scipy.org/) [28] between the whole population and the population without that age group (Populations compared when the 20-29 age group was removed: whole female population vs. female population aged 30-79).

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

Previous studies have utilized generalized additive models (GAMs) to predict health outcomes using sex and/or age as features [35,36]. In this study, a GAM was used to rank the effect of variables on CDI and detect groups with outlier variability. A generalized additive model was built in Python using the package pyGAM (*pygam.LinearGAM*, version

0.9.1, https://pygam.readthedocs.io/en/latest/) [37]. A factor term was fit to sex, age, and weekly rhythm categories (Sex: female, male. Age: 20-29, 30-39, 40-49, 50-59, 60-69, 70-79. Weekly Rhythm (WR): weekend patterns, patternless). Cyclic categories were left out of this analysis because acyclic and cyclic individuals were not found to have significantly different CDIs. A factor term was fit for each variable (sex, age, weekly rhythm) resulting in the following GAM structure:

$$g(E(CDI)) = \beta_0 + f_{Sex}(Sex) + f_{WR}(WR) + f_{Age}(Age)$$

where g is the link function and β_0 is the intercept of the model. 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.

DECLARATIONS

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 and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of 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.

Competing Interests

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

The content of this thesis including the abstract, introduction, results, discussion, methods, declarations, and appendix is currently being prepared for publication. Kristin 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. The thesis author was the primary investigator and author of this material.

APPENDIX

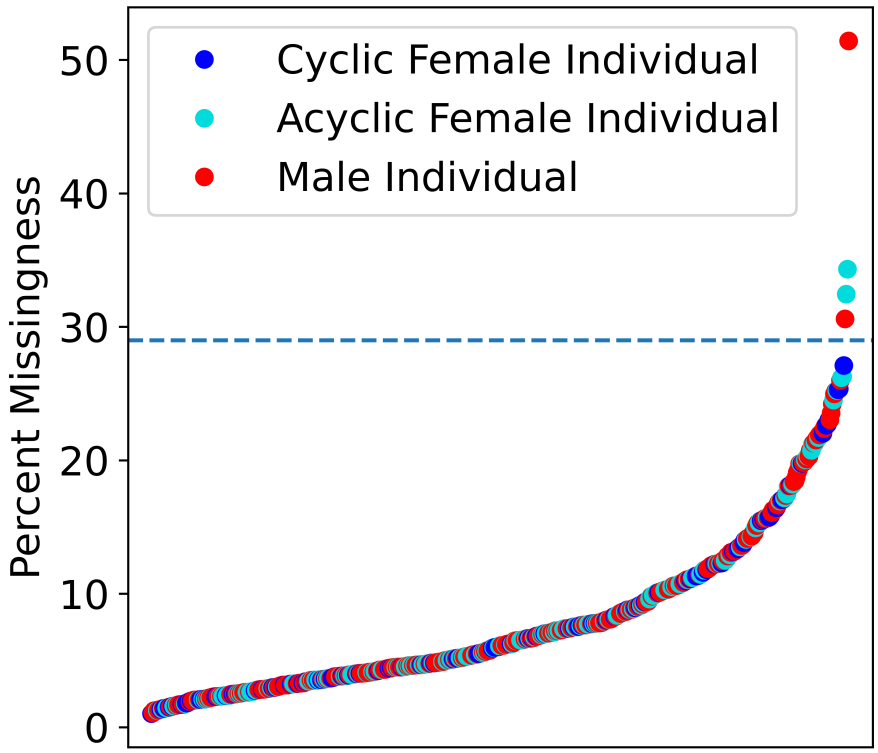
Tables recording population standard deviations of each sex for each MET sum metric and for each sex subgroup for 24-hour MET sums. Population standard deviations are presented here for their relevance to power analysis.

Supplementary Table 1. The standard deviation of the population distributions of mean 24-hour daily MET sum, awake daily MET sum, and asleep daily MET sum calculated from 206 days

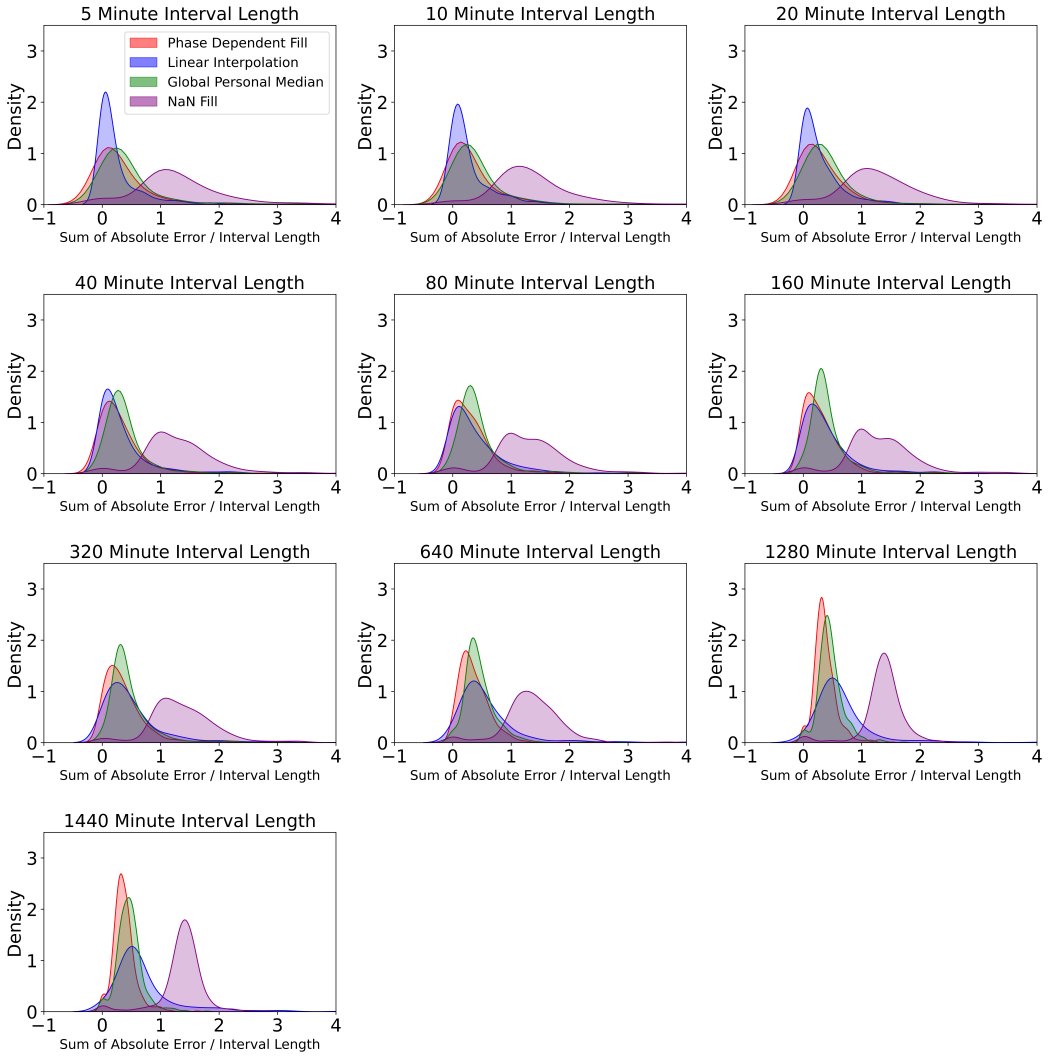
	Population Standard Deviation		
Daily Sum Metric	24-Hour	Awake	Asleep
Male	235	254	97.7
Female	194	230	99.6

Supplementary Table 2. Population standard deviations of all subgroups studied calculated from the population distributions of mean 24-hour daily MET sum. Weekend group individual means were calculated from 4 consecutive months of data and all other group's individual means were calculated from all 206 days in the dataset.

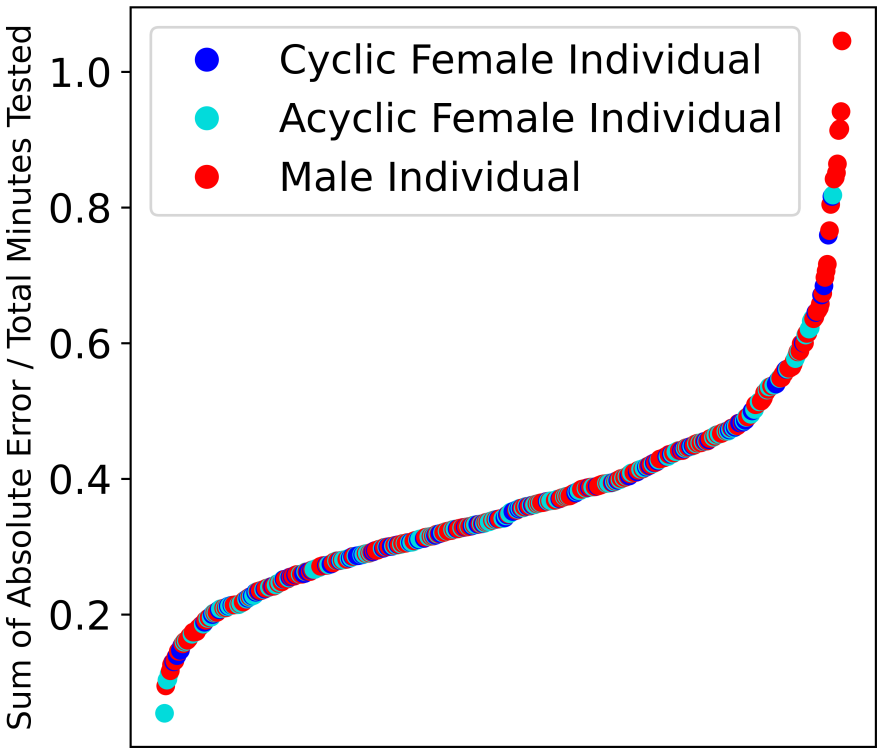
	Female Population Standard Deviation	Male Population Standard Deviation
Whole population	194	235
Without Weekend Patterns	212	240
With Weekend Patterns	200	235
Weekend High	201	248
Weekend Low	187	150
Cyclic	175	N/A
Acyclic	203	235
Age 20-29	175	278
Age 30-39	172	220
Age 40-49	170	221
Age 50-59	219	238
Age 60-69	226	247
Age 70-79	177	169



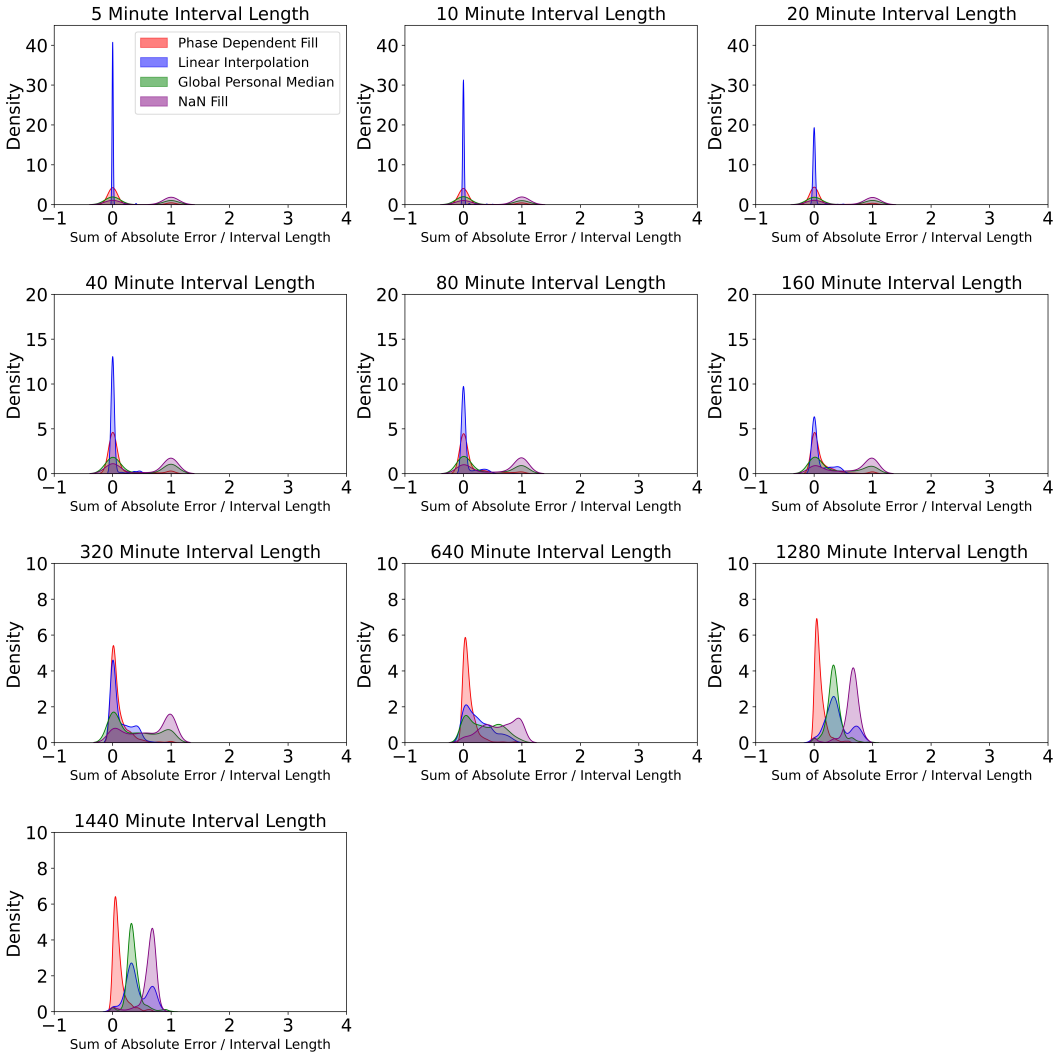
Supplementary Figure 1. Percent of values that were missing in each individual from the starting cohort sorted from least to greatest and labeled by their sex and cyclic status. Individuals above the horizontal dashed line at 29% were removed from the cohort for this analysis, leaving 596 total individuals with less than 29% total missingness in their data.



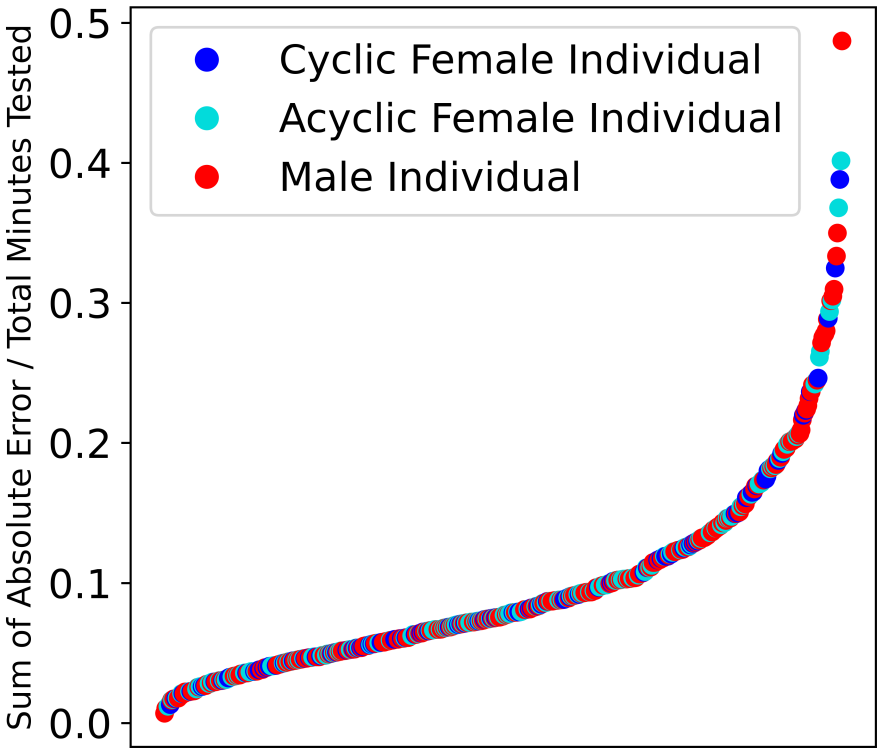
Supplementary Figure 2. Kernel density estimate of the error generated by each filling method on simulated missing MET data. Error is shown as sum of the absolute error divided by the interval length. Each distribution is composed of exactly 596 measurements, one of each interval length for every individual.



Supplementary Figure 3. Sum of the absolute error (divided by the total minutes tested - 3995) generated in each individual by linear interpolation in intervals of length 5 to 40 and the error generated by the phase-dependent filler in intervals length 80 to 1440 in simulated missing MET data. Unfilled data results in a sum of absolute error to total minutes tested proportion of at least 0.9.



Supplementary Figure 4. Kernel density estimate of the error generated by each filling method on simulated missing sleep state data. Error is shown as sum of the absolute error divided by the interval length. Each distribution is composed of exactly 596 measurements, one of each interval length for every individual.



Supplementary Figure 5. Sum of the absolute error (divided by the total minutes tested - 3995) generated in each individual by linear interpolation in intervals of length 5 to 320 and the error generated by the phase-dependent filler in intervals length 640 to 1440 in simulated missing sleep state data. Unfilled data results in a sum of absolute error to total minutes tested proportion of 0 to 1.

REFERENCES

[1] Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. Surgery 2014;156:508–16. https://doi.org/10.1016/j.surg.2014.07.001.

[2] 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. https://doi.org/10.1016/j.addr.2021.05.014.

[3] 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:e196700. https://doi.org/10.1001/jamanetworkopen.2019.6700.

[4] Hamberg K. Gender Bias in Medicine. Womens Health (Lond Engl) 2008;4:237–43. https://doi.org/10.2217/17455057.4.3.237.

[5] Zucker I, Prendergast BJ, Beery AK. Pervasive Neglect of Sex Differences in Biomedical Research. Cold Spring Harb Perspect Biol 2021:a039156. https://doi.org/10.1101/cshperspect.a039156.

[6] Biden J. Remarks of President Joe Biden – State of the Union Address As Prepared for Delivery. The White House 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/ (accessed March 18, 2024).

[7] Zucker I, Beery AK. Males still dominate animal studies. Nature 2010;465:690–690. https://doi.org/10.1038/465690a.

[8] 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;8:7. https://doi.org/10.1186/s13293-016-0125-3.

[9] 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;7:34. https://doi.org/10.1186/s13293-016-0087-5.

[10] Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. Neuroscience & Biobehavioral Reviews 2014;40:1–5. https://doi.org/10.1016/j.neubiorev.2014.01.001.

[11] 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;13:41. https://doi.org/10.1186/s13293-022-00451-1.

[12] 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. https://doi.org/10.1186/s13293-021-00375-2.

[13] Bruce LK, Kasl P, Soltani S, Viswanath VK, Hartogensis W, Dilchert S, Hecht FM, Chowdhary A, Anglo C, Pandya L, Dasgupta S, Altintas I, Gupta A, Mason AE, Smarr BL. Variability of temperature measurements recorded by a wearable device by biological sex. Biology of Sex Differences 2023;14:76. https://doi.org/10.1186/s13293-023-00558-z.

[14] Regidor P-A, Kaczmarczyk M, Schiweck E, Goeckenjan-Festag M, Alexander H. Identification and prediction of the fertile window with a new web-based medical device using a vaginal biosensor for measuring the circadian and circamensual core body temperature. Gynecological Endocrinology 2018;34:256–60. https://doi.org/10.1080/09513590.2017.1390737.

[15] Klein A, Viswanath VK, Smarr B, Wang EJ. Detecting Periodic Biases in Wearable-Based Illness Detection Models. ICLR 2023 Workshop on Time Series Representation Learning for Health, 2023.

[16] Huhn S, Axt M, Gunga HC, Maggioni MA, Munga S, Obor D, Sié A, Boudo V, Bunker A, Sauerborn R, Bärnighausen T, Barteit S. The Impact of Wearable Technologies in Health Research: Scoping Review. JMIR Mhealth Uhealth 2022;10:e34384. https://doi.org/10.2196/34384.

[17] Grant A, Smarr B. Feasibility of continuous distal body temperature for passive, early pregnancy detection. PLOS Digit Health 2022;1:e0000034. https://doi.org/10.1371/journal.pdig.0000034.

[18] Mason AE, Hecht FM, Davis SK, Natale JL, Hartogensis W, Damaso N, Claypool KT, Dilchert S, Dasgupta S, Purawat S, Viswanath VK, Klein A, Chowdhary A, Fisher SM, Anglo C, Puldon KY, Veasna D, Prather JG, Pandya LS, Fox LM, Busch M, Giordano C, Mercado BK, Song J, Jaimes R, Baum BS, Telfer BA, Philipson CW, Collins PP, Rao AA, Wang EJ, Bandi RH, Choe BJ, Epel ES, Epstein SK, Krasnoff JB, Lee MB, Lee SW, Lopez GM, Mehta A, Melville LD, Moon TS, Mujica-Parodi LR, Noel KM, Orosco MA, Rideout JM, Robishaw JD, Rodriguez RM, Shah KH, Siegal JH, Gupta A, Altintas I, Smarr BL. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study. Sci Rep 2022;12:3463. https://doi.org/10.1038/s41598-022-07314-0.

[19] Hills AP, Mokhtar N, Byrne NM. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. Front Nutr 2014;1:5. https://doi.org/10.3389/fnut.2014.00005.

[20] Bernard MA, Clayton JA, Lauer MS. Inclusion Across the Lifespan: NIH Policy for Clinical Research. JAMA 2018;320:1535–6. https://doi.org/10.1001/jama.2018.12368.

[21] Rochon PA, Mason R, Gurwitz JH. Increasing the visibility of older women in clinical research. The Lancet 2020;395:1530–2. https://doi.org/10.1016/S0140-6736(20)30849-7.

[22] Tompuri TT. Metabolic equivalents of task are confounded by adiposity, which disturbs objective measurement of physical activity. Frontiers in Physiology 2015;6. https://doi.org/10.3389/fphys.2015.00226.

[23] Purawat S, Dasgupta S, Song J, Davis S, Claypool KT, Chandra S, Mason A, Viswanath V, Klein A, Kasl P, Wen Y, Smarr B, Gupta A, Altintas I. TemPredict: A Big Data Analytical Platform for Scalable Exploration and Monitoring of Personalized Multimodal Data for COVID-19. 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA: IEEE; 2021, p. 4411–20. https://doi.org/10.1109/BigData52589.2021.9671441.

[24] Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett Jr. DR, Tudor-Locke C, Greer JL, Vezina J, Whitt-Glover MC, Leon AS. 2011 Compendium of Physical Activities: A Second Update of Codes and MET Values. American College of Sports Medicine 2012;2012:126–7. https://doi.org/10.1016/j.yspm.2011.08.057.

[25] The Pandas Development Team. pandas-dev/pandas: Pandas 2024. https://doi.org/10.5281/zenodo.10697587.

[26] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates; 1988.

[27] Vallat R. Pingouin: statistics in Python. JOSS 2018;3:1026. https://doi.org/10.21105/joss.01026.

[28] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors, Vijaykumar A, Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.

[29] 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:e02527. https://doi.org/10.1002/ecs2.2527.

[30] Heath JP. Quantifying temporal variability in population abundances. Oikos 2006; 115:573–81. https://doi.org/10.1111/j.2006.0030-1299.15067.x.

[31] Heath JP, Borowski P. Quantifying Proportional Variability. PLoS ONE 2013;8:e84074. https://doi.org/10.1371/journal.pone.0084074.

[32] McArdle BH, Gaston KJ. The Temporal Variability of Densities: Back to Basics. Oikos 1995;74:165. https://doi.org/10.2307/3545687.

[33] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

[34] Waskom M. seaborn: statistical data visualization. JOSS 2021;6:3021. https://doi.org/10.21105/joss.03021.

[35] Clements MS, Armstrong BK, Moolgavkar SH. Lung cancer rate predictions using generalized additive models. Biostatistics 2005;6:576–89. https://doi.org/10.1093/biostatistics/kxi028.

[36] 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;2019:343–52.

[37] Daniel S, Brummitt C, Abedi H. dswah/pyGAM: v0.8.0 2018. https://doi.org/10.5281/zenodo.1476122.