***<u>Objective</u>*: Compare different techniques for handling accelerometer missing data via a simulation study

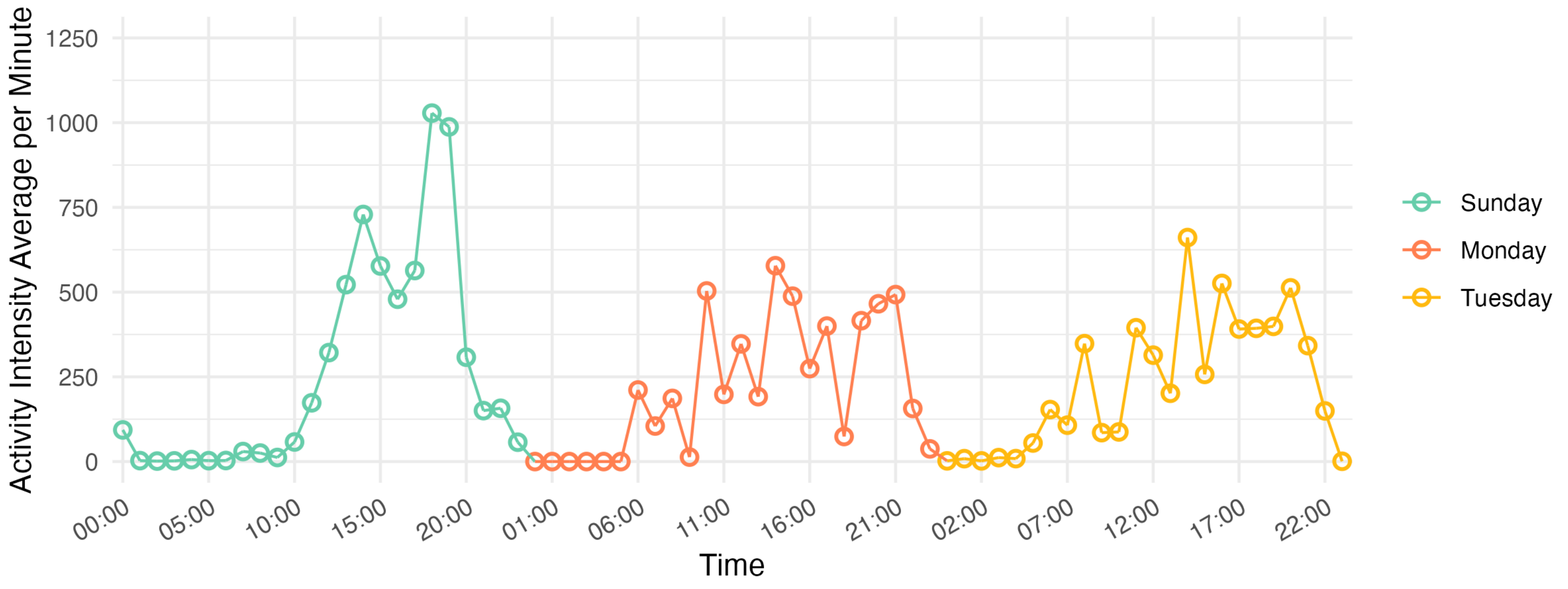
<u>NHANES Accelerometer Data:</u>

- Website link: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2005

- Variables of the dataset:
  - ✓ SEQN: respondent sequence number.
  - ✓ PAXDAY: day of the week. 1 indicates Sunday, 2 for Monday and so forth.
  - ✓ PAXN: records sequential observation number in minutes. The range starts with minute 1 on day1 (PAXN = 1) and end with the last minute of day 7 (PAXN = 10080).
  - ✓ PAXHOUR: hour of the observation.
  - ✓ PAXMINUT: minute within the hour of observation.
  - ✓ PAXINTEN: The intensity value recorded by the physical activity monitor.
  - ✓ Covariates: age, BMI, gender

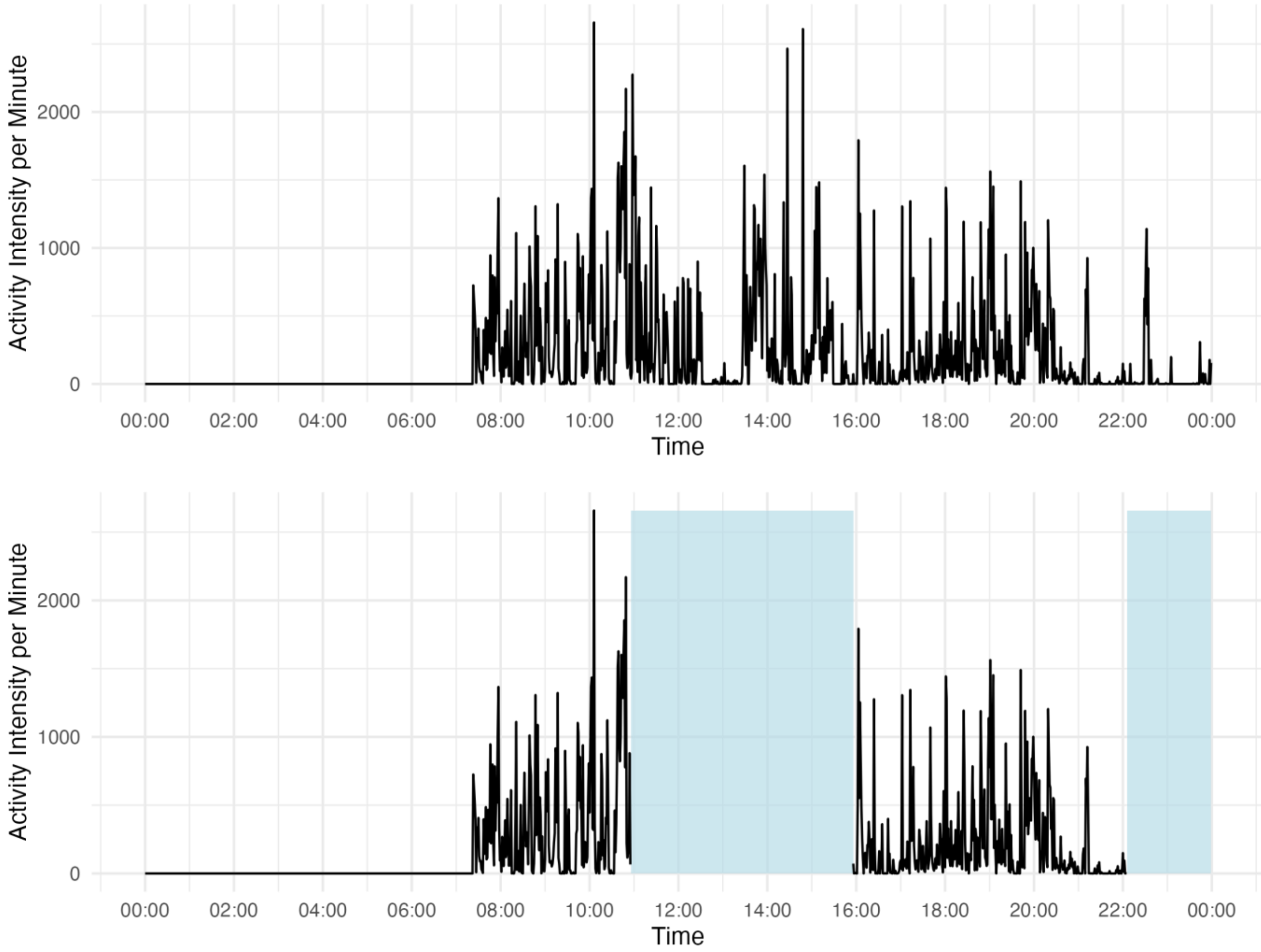
Preprocessing the Data:

- Subjects with less than 12 hours of device on any day across a week was excluded.
- Subjects with extremely high daily activity intensity (e.g., over 10,000 per minute) was excluded.
- After preprocessing the dataset, 105 participants was used for the subsequent study phase.
- Use one subject to illustrate the activity change for 24 hours and first three days



Missing Completely at Random Mechanism

- Define four groups with predetermined missing data levels.
- Randomly allocate subjects to these groups.
- Determine each subject's missing data block size from $N(360, 120^2)$.
- Calculate the total number of rows to sample for each subject as the product of wearable rows and the assigned missingness percentage.
- Use a while loop to iteratively select missing data blocks until the cumulative sampled rows reach the target.
- Ensure the loop correctly assigns non-overlapping missing data blocks.
- Replace data within the selected blocks with "NA".



Numerical Study

- The performance of two imputation methods were evaluated across 200 Monte Carlo replicates.

First approach: single imputation using a linear mixed model

- Specify an appropriate imputation model: linear mixed regression model, considering a random intercept and a random slope for total wear time to account for individual differences.

$$Y_{ij} = \beta_{0i} + \beta_1 A_i + \beta_2 B_i + \beta_3 G_i + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_{6i} T_{ij} + \varepsilon_{ij}$$

$$\beta_{0i} = \beta_0 + \tau_{0i} \quad \beta_{6i} = \beta_6 + \tau_{6i}$$

Y_{ij} is the daily sum of activity for the *j*th day of the *i*th subject

A_i is the age for the *i*th subject

B_i is the BMI for the *i*th subject

G_i is the gender for the *i*th subject

X_{4ij}, X_{5ij} represent the Saturday and Sunday indicator

T_{ij} is the total wear time for *i*th subject at *j*th day

Second approach: multiple imputation using predictive mean matching(PMM)

- Apply the PMM integrated with a linear mixed model, the idea is to find the closest candidates among the observed values where each of the missing value is replaced by this process.

1) Fit the linear imputation model using the subjects with complete data, estimate model parameters

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\sigma}_\varepsilon, \hat{\tau}_0$ and their variance-covariance matrix.

2) Randomly sample from the posterior predictive distribution of $\hat{\sigma}_\varepsilon^2$and produce a new set of coefficients σ_ε^{2*}.

3) Sample new fixed effect coefficients from the following multivariate normal (MVN) distribution,

$$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^* \sim MVN((\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6), var(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)\frac{\sigma_\varepsilon^{2*}}{\hat{\sigma}_\varepsilon^2})$$

4) Compute predicted values for observed responses:

$$\hat{y}_{ij}^{obs} = \beta_{0i}^* + \beta_1^* A_i + \beta_2^* B_i + \beta_3^* G_i + \beta_4^* X_{4ij} + \beta_5^* X_{5ij} + \beta_{6i}^* T_{ij}$$

5) Compute predicted values for missing responses:

$$\hat{y}_{ij}^{mis} = \beta_{0i}^* + \beta_1^* A_i + \beta_2^* B_i + \beta_3^* G_i + \beta_4^* X_{4ij} + \beta_5^* X_{5ij} + \beta_{6i}^* T_{ij}$$

6) For each missing y, identify the closest predicted values from the observed cases.

7) Randomly select one of these closest values and impute the missing y with the observed value.

Comparision Results

- PMM exhibits slightly more bias than single regression imputation for 20% and 30% missing data but performs similarly at 40% and 50% missing data.
- Additionally, PMM has a lower and more stable standard error across different levels of missing data compared to single regression imputation.