Journal for the Measurement of Physical Behaviour, 2026, 9, 1-14
https://doi.org/10.1123/jmpb.2025-0018

Human Kinetics
ORIGINAL RESEARCH

Addressing Missing Data in Accelerometer Studies: Evaluating the Performance of Imputation Methods for Longitudinal Data

Noemi Berliner, Maik Bieleke, Julia Schüler, and Fridtjof W. Nussbeck
Universtät Konstanz, Konstanz, Germany

Adequate handling of missing data on physical activity assessments is crucial in longitudinal accelerometer studies. This study aimed to evaluate the effectiveness of various imputation methods for handling missing data in an empirical application which utilizes wearable accelerometers. We employed a simulation approach to assess performance under different missing data scenarios including Missing Completely at Random, Missing at Random, and Missing Not at Random for a longer study period (6 weeks). Our findings revealed that mean imputation and hot-deck imputation applied with a fine degree of matching criteria (participant, day of the week, and time of day) outperformed discard-based methods under Missing Completely at Random and Missing at Random conditions as they produced the smallest bias and best precision. Notably, no imputation methods performed well under Missing Not at Random scenarios. We recommend conducting simulation studies tailored to specific study designs to compare imputation methods, implement strategies for improving data quality, gather information on nonwear periods, and ensure continuous monitoring and participant compliance thereby reducing bias in activity level estimates. If a simulation study is not feasible, we recommend to impute data relying on mean or hot-deck approaches with the finest possible degree of matching criteria.

Keywords: wearables, physical activity, ProPELL

Key Points

- Mean and hot-deck imputation with fine matching criteria outperformed discard-based methods for accelerometer data under Missing Completely at Random and Missing at Random conditions.
- Fine-grained matching on participant, weekday, and time improved imputation performance.
- None of the evaluated imputation methods performed well under Missing Not at Random scenarios.

Physical activity (PA) plays a vital role in both mental and physical health (Forte et al., 2023). To understand temporal processes and dynamic relations, it is necessary to assess PA over extended periods. Accelerometer devices are frequently used in medical and public health research, as they provide real-life, high-resolution data while avoiding recall biases inherent in self-reports (Karas et al., 2019; LeBlanc & Janssen, 2010; Vetter et al., 2023). Despite challenges such as participant compliance and data management (e.g., high upfront costs, device distribution, technical issues, high data complexity; Karas et al., 2019), accelerometers generally provide a minimally intrusive and relatively cost-efficient method, making them suitable for large-scale longitudinal multicenter and multinational studies (e.g., Kim et al., 2018; Menai et al., 2017).

Acceleration data is principally processed into metrics like step counts (Chen & Bassett, 2005; John & Freedson, 2012) and subsequently aggregated for fixed-length time windows, known as epochs. When compared with self-reports, accelerometers provide a more detailed assessment of PA but can suffer from missing data (Prince et al., 2008) and cannot be used to capture a person's PA retrospectively.

Bieleke https://orcid.org/0000-0003-2586-1416
Schüler https://orcid.org/0000-0002-7790-0491
Nussbeck https://orcid.org/0000-0003-4002-8916
Berliner (noemi.berliner@uni-konstanz.de) is corresponding author, https://orcid.org/0009-0003-1183-1651

Reasons for Missing Data in Accelerometer Studies

Missing data in accelerometer studies can arise from device malfunction (e.g., battery depletion or water damage), participant noncompliance (e.g., forgetting or choosing not to wear the device; Cho et al., 2021), but will also arise when the batteries of the devices have to be charged. That is, nonwear is typically not missing completely at random. Missingness often follows patterns influenced by participant behavior or context such as occurring more frequently at the start and end of the day (Xu et al., 2018). Sociodemographic and health factors also play a role: higher body mass index, race/ethnicity (with non-Hispanic Black and Hispanic individuals at greater risk than non-Hispanic Whites), lower education level, smoking, and using street drugs have been linked to missing/invalid data (Loprinzi et al., 2013). In children, being a boy, overweight status, health limitations or disabilities, disadvantaged family backgrounds, and low PA levels have been linked to less reliable data (Rich et al., 2013).

Study Objectives and Relevance

With this simulation study, we evaluate the performance of various imputation methods for missing data in "active steps," a sensor-derived measure of PA, within a longitudinal study context. This study offers insights into challenges in longitudinal accelerometer

research and recommendations on handling missing data. Specifically, we investigate how different imputation methods perform under missing data scenarios that might be expected in practice including Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) scenarios.

First, we describe the empirical application providing the accelerometer data; the Promoting Physical Exercise in Lab and Life (ProPELL) study (Bieleke et al., 2022). We outline the steps of data processing, aggregation, and creation of a full data set, as well as how we introduced missing data according to the different scenarios. Finally, we report the various imputation strategies and the evaluation of their performance with the goal to identify the most appropriate imputation method for "active steps." This should enable us to create a data set suitable for the analysis of further research questions within the mentioned research project but also inform other researchers with comparable data sets about the most appropriate strategy to deal with missing data.

Defining and Handling Missing Data

The starting point for data analysis but also data preparation is the definition of the epochs. That is, the degree of temporal resolution must be determined, resulting in a specified epoch (time interval) length (e.g., 1 hr, 1 day, or 1 week). The choice of temporal resolution depends on the research question—whether short-term processes of PA or stable activity levels are of interest—as well as on the availability of computational power. Then, nonwear time must be identified. Ideally, an additional indicator—such as heart rate—allows for a direct distinction between zero active steps caused by inactivity and those resulting from nonwear (e.g., based on the absence of heart rate data). If no additional indicator is available, algorithms can classify consecutive zero counts as either sedentary time or nonwear periods (Evenson & Terry, 2009).

After the identification of nonwear time, three scenarios must be addressed: (a) a time interval is fully observed, (b) no data are recorded at all, or (c) data are partially observed.

(a) For fully observed time intervals, the average number of active steps adequately represents the individual's activity level. (b) Intervals with no data are considered missing. (c) For partially observed time intervals, a decision must be made about whether there is sufficient data to estimate the average number of steps for that interval. This is typically done using a wear-time threshold, which varies depending on the time interval, sample, and study design. Migueles et al. (2017) offer a comprehensive review of accelerometer data processing protocols and recommend practical thresholds for valid days (≥10 hr during waking hours) and valid weeks (≥ 4 valid days). Threshold choice affects data quality and interpretation (Herrmann et al., 2014).

Imputation Methods for Missing Data

Several approaches have been applied to handle missing data in accelerometer research (Butera et al., 2019; Kapphahn et al., 2022; Lee, 2013; Lee & Gill, 2018; Tackney et al., 2023). These range from simpler discard-based methods, such as complete case analysis, to more advanced methods like multiple imputation.

Complete Case Analysis

Complete case analysis excludes participants with any missing data. While unbiased under MCAR, it introduces bias, when missing is MAR or MNAR (Schafer & Graham, 2002). Furthermore, complete case analysis reduces the sample size and, hence, statistical power; even more so as in many studies only a small subset of participants will provide complete data. In accelerometer studies, nonwear periods are commonly observed for all participants and even have to occur if charging of the device is necessary. As a result, complete case analysis would exclude the entire sample. However, in data sets with minimal amounts of missing data (e.g., 7% during waking hours in Borgundvaag et al., 2017) estimates may remain robust (Borghese et al., 2019).

Single Imputation

Single Imputation replaces missing values with reasonable values, resulting in a complete cases data set. Common approaches include last observation carried forward (LOCF; Samuels et al., 2011), mean replacement, regression-based imputation (Xu et al., 2018), and an individual-centered approach (Kang et al., 2009).

Multiple Imputation

Multiple Imputation (Rubin, 1987) consists of three steps: First, an imputation method is applied to generate a complete data set. This process is repeated multiple times to create several imputed data sets (e.g., 10 times). Second, analyses are performed on each data set. Third, the (10) results are combined. Multiple imputation allows not only to estimate the target parameter but also to consider the precision of the estimation.

Using Simulations to Evaluate Imputation Methods

The various imputation methods have strengths and limitations, posing the question of which method is best suited for a given data set. To address this question, simulation studies have become a valuable tool for evaluating the effectiveness of various imputation methods, offering insights into potential biases and the appropriateness of the imputed values. Simulation studies allow an evaluation of new methods and comparison of alternative methods especially when assumptions are violated (Morris et al., 2019).

Several simulation studies have evaluated the appropriateness of imputation methods in the context of accelerometer data, often focusing on short (typically 1 week) measurement periods and incorporating covariates into the imputation models. Table 1 provides a summary of studies including the sample characteristics, missing data mechanisms, applied imputation methods, and main findings.

These studies collectively emphasize the importance of aligning imputation strategy with the data structure, missingness pattern, and research question.

Our simulation study differs from previous studies in two aspects. First, we extended the observation period. We evaluate the performance of various imputation methods over an extended 6-week observation period. This allows capturing weekly patterns. Second, we did not use covariates for imputation. Our approach imputes missing data solely using time-related variables, without including other potentially relevant covariates. Thereby, the imputed data set remains broadly applicable to research questions across various domain, as it avoids introducing biases or dependencies specific to a particular set of covariates. It allows the same imputed data set to be used for analyses involving different variables without recalculating imputations tailored to specific research questions. To our mind, this is especially useful if multiple

Table 1 Overview of Studies Evaluating Imputation Methods for Accelerometer Data

Study (year)	Sample	Missing data	Imputation methods	Key findings
Lee and Gill (2018)	NHANES 2003–2004; $N = 218$ participants; 7 days	MAR; entire days are missing, 15.5% missing data	ZIPLN mixture model; GLM; imputation on minute-level	ZIPLN handles excess zeros and autocorrelation well. Parametric and semi-parametric imputation under ZIPLN performed robustly
Xu et al. (2018)	NHANES, overweight postmenopausal breast cancer survivors; $N = 328$; 1–15 days	Mimic missing data patterns observed in the population, 12.5%–25% missing data	Variance-weighted regression; subject-specific intercept/slope models; EM; K-nearest neighbor; imputation on day-level and minute-level	Not accounting for missingness destabilizes estimates; methods weighting by wear-time and modeling individual-level slopes improved precision
Maeda et al. (2019)	NHANES; age, M (SD) = 34.5 (23.1); BMI, M (SD) = 26.1 (7.1); $N = 50$, 100, 200, 500; 7 days	Mimic the minute-level missing data patterns observed in the population; 26.0%–28.6% missing data	Average of valid days as default method, within-minute average (Alhassan et al., 2022), day-level imputation (Lee, 2013) and minute-level imputation (Lee & Gill, 2018); imputation on day-level and minute-level	Within-minute and day-level methods most accurately estimated average counts per day when ≤3 valid days
Butera et al. (2019)	OPACH cohort; older women; $N = 2{,}550$; 7 days	MAR conditional on time-of-day, in-bed status, BMI, age; 25% missing data	Hot-deck multiple imputation (self- and nonself donor pools, stratified by time, in-bed status, BMI, age); imputation on epoch level; imputation on day-segment-level	Hot-deck MI reduced bias and improved confidence interval coverage compared to complete- or available-case methods
Kapphahn et al. (2022)	Simulated Stanford GOALS; children; $N = 100$; 7 days	MAR and MNAR scenarios at epoch/daily levels; 38%–53% missing data	Complete-case, single imputation, regression-based imputation with covariates: BMI, demographic variables, lag; imputation on minute-level	MI methods produced lower bias and improved associations between mean counts and BMI compared to discard-based methods and SI methods
Tackney et al. (2023)	PACE-UP trial; $N = 360$ in three treatment groups; age 45–75; 7 days at baseline and follow-up periods	Mimic missing data patterns observed in the population	Tobit-like parametric MI; nonparametric donor-based MI matching age, sex, BMI, imputation on 5-s level	Nonparametric MI delivered most accurate mean step counts and estimates of the effect of treatment

Note. NHANES = National Health and Nutrition Examination Survey; OPACH = Women's Health Initiative Objective Physical Activity and Cardiovascular Health Study; ZIPLN = Zero-Inflated Poisson Log-Normal; GLM = Generalized Linear Model; EM = Expectation-Maximization; BMI = Body Mass Index; MI = Multiple Imputation; MAR = Missing At Random; MNAR = Missing Not At Random; BMI = body mass index.

research groups use the same central variable (here "active steps") but in different analyses.

Furthermore, we examine the performance of these methods under different missingness mechanisms commonly encountered in practice, including MCAR, MAR, and MNAR.

Methods

ProPELL—Data

This simulation study is based on the ProPELL project (Bieleke et al., 2022), which was designed to explore the effects of physical exercise on both physiological and psychological parameters. The ProPELL study is in accordance with the Declaration of Helsinki and approved by the ethics committee of the University of Konstanz (reference number: 31/2022).

In total, $N = 75$ students ($M = 22.8$ years, $SD = 2.9$) participated in a 21-week randomized controlled trial study at the University of Konstanz, Germany, in 2022/2023. Out of 354 individuals who passed strict exclusion criteria, including body mass index over 30 and various psychological and physical conditions, the least physically active participants were selected for the study. Participants in the experimental group ($n = 43$) underwent an 8-week jump training, while participants in the control group ($n = 32$) received no training. This was followed by an 8-week observation of PA and exercise behavior in both groups. Besides physiological and psychological parameters measured at baseline (pretraining), after intervention (posttraining), and follow-up, continuous measurements of activity-related variables (e.g., step count, heart rate) were obtained throughout the study period.

To conduct our simulation study, we constructed a full data set using 6 weeks of ProPELL data from 29 individuals in the control group (three individuals of the control group dropped out). Since the aim of this simulation is to evaluate the appropriateness of imputation methods over a prolonged observation period, we chose a 6-week dataset. Six weeks allow us to identify weekly patterns in active steps and the data set reflects variability across day- and time-specific intervals. We used data from the control group to avoid the influence of any intervention effects and did not expect changes in typical behavioral patterns.

In the following, we outline the steps of the simulation study. (I) We generated a complete data set, free of missing values, which represents participants' activity levels over 6 weeks. (II) We introduced missing data patterns based on various scenarios. (III) We implemented different imputation methods, ranging from simple replacement approaches to more advanced techniques incorporating varying degrees of matching and grouping criteria. (IV) We evaluated the imputation methods by comparing the imputed data sets with the original complete data set. All R scripts used in the simulation are available at the Open Science Framework: https://osf.io/ncejd/?view_only = 54156af9a2434dbbbc59fff00229d3a2.

Active Steps Measurement

Throughout the study period, participants were required to wear an activity tracker that recorded activity data. Specifically, we used a Polar Vantage V2 sports watch, which computes an "active steps" variable. Active steps are accumulated when movement is detected, with steps counted for all activities (including nonstepping activities, such as cycling, and swimming). In 60-s time epochs, the steps are summed and registered (Polar Research Center, 2021).

Degree of Temporal Resolution

We set the temporal resolution for imputation and analysis to an hourly level. This allows us to capture variation of PA levels throughout the day, while remaining treatable for the 6-week data set. Accordingly, the complete data set comprises 29,232 data points (29 persons × 6 weeks × 7 days × 24 hr). In total, we found less than 1% of missing data (i.e., 205 hr without active steps being recorded for less than 20 min, see also below). Across participants, the missing data were distributed as follows: the mean number of missing hours per participant was 7.3, the median was 4, with a minimum of 1 and a maximum of 41 hr.

Creating a Complete Dataset

In the first step, we created a data set to ensure that it is both complete and representative of the participants' activity levels, by selecting the most informative weeks and replacing missing values where necessary.

1. Exclusion of dropped participants: Initially, the three participants who dropped out were excluded from the data set.
2. Selection of data-rich weeks: We identified the 6 weeks with the most complete data for each participant in the control group. Specifically, we calculated the total wear time per week for each participant. This was done by summing the daily wear times across each week, with the total expressed in hours. From the summarized wear times, we selected the 6 weeks with the highest total wear time for each participant, maintaining the order of the weeks (e.g., Weeks 1, 2, 3, 6, 8, and 9). By focusing on these weeks, we ensured that the analysis was based on the most comprehensive data available. In preliminary analyses, we examined whether participants exhibited systematic increases or decreases in PA (as measured by active steps) over the study period. This was done both graphically and by regressing daily active step counts on time. There were no trends in the data. The presence of such trends could bias the evaluation of imputation methods by favoring those that account for elapsed study time such as LOCF potentially leading to an overestimation of their performance. Ensuring temporal stability in activity levels allowed us to isolate the impact of missingness scenarios on the performance of different imputation methods.
3. Calculation of average hourly activity level: We differentiated between three cases: complete records, time intervals with less than 20 min of valid data and time intervals with more than 20 min of valid data within the complete hour. The 20-min threshold corresponds to a third of an hour, aligning with the less restrictive threshold of eight valid hours out of 24 in a day (Evenson & Terry, 2009). If the wear time within any given hour was less than 20 min, the entire hour was treated as missing data. For time intervals meeting the minimum wear time criterion, the average hourly activity level for each participant was calculated through the following steps: First, the total number of active steps within each 1-hr interval was summed. Subsequently, to account for variations in how long the device was worn during each hour, the total active steps were divided by the wear time within that hour. The resulting value (active steps per minute of wear time) was then multiplied to scale it to a standardized hourly rate, for example, 200 steps recorded from 4:00 to 4:40 p.m. results in an hourly rate of 300 steps for the 4:00 p.m. time interval.

4. Addressing missing data points: Next, we addressed the remaining missing data in the standardized active steps variable, comprising the 205 missing data entries out of 29,232 total data points (less than 1%). To complete the data set, missing values were replaced using data from subsequent weeks that were not initially selected among the six data-rich weeks, representing active steps that could have occurred. The replacement was conducted by matching the participant's ID, the day of the week, and the hour of the day.

By selecting the most data-rich weeks and filling in missing values with appropriate substitutes, we created a complete data set, that presumably reflects the participants' actual activity patterns best.

Inducing Missing Data

We introduced missing data based on seven distinct scenarios to investigate how different missing data mechanisms affect the performance of various imputation techniques. Testing imputation methods under different missing data patterns is essential, as the performance of these methods can vary depending on the underlying missingness mechanism. By simulating a range of realistic missing data scenarios, we are able to systematically assess the appropriateness and robustness of imputation techniques across conditions. We set the amount of missing data equal to the amount of missing data in the experimental group (6.44%).

1. MCAR-1: Missing data was introduced completely at random (MCAR) in 1-hr blocks. This approach ensured that missingness was random across all hours, without any specific pattern or relation to time or activity level.
2. MCAR-3: Similar to the MCAR-1 method, missing data was introduced randomly, but in 3-hr blocks. By using 3-hr blocks, this method created longer gaps in data that are still completely at random. However, this may more closely simulate real-world scenarios where devices are not worn for extended periods.
3. MAR: Missing data was introduced in hour-long blocks, with the probability of missingness being related to the time of day. The number of missing data points was adjusted to match the corresponding interval in the 6-week experimental group data set. That is more missing at daytime (e.g., 8.29% at 1 p.m.) than during night (e.g., 5.17% at 3 a.m.).
4. MPAT (Missing Pattern): Missing data was introduced to mimic realistic scenarios by copying the missing pattern of the experimental group during their first 6 weeks. Each participant in the full data set was randomly assigned to a participant from the experimental group, and hours were set to missing, if for the corresponding experimental group participant data was missing according to the abovementioned criteria. In this way, we ensured that previously inserted data were not removed again because of the individual's own missingness pattern. This case could potentially favor imputation techniques considering day and time of day (see below).
5. MNAR: Missing data was related to the number of active steps itself, with lower activity intervals being more likely to be missing. We calculated a score for each hourly time interval by subtracting the number of active steps from the maximum number of steps and then dividing the result by the maximum steps. To this score, we added a normally distributed error term, $\varepsilon \sim N(0, 0.1)$. The hourly time intervals were then ranked based on these scores, and the 6.44% with the highest scores were designated as missing, again yielding the same percentage of missingness.
6. MNAR-MCAR: We combined mechanisms to introduce missing data. 50% of the missing data was based on MNAR (see "5. MNAR") and the remaining 50% was randomly selected (MCAR, see "1. MCAR-1") from the remaining data points.
7. MTRAIN (Missing during Training): To simulate regular training sessions where fitness trackers were not worn (e.g., during team sports), we identified the 10% most active hours for each participant. If a participant had the same high-activity hour on the same day across four or more weeks, these hours were considered scheduled training and were set to missing, resulting in a MNAR mechanism. Note that the overall percentage of missingness (0.63%) differs from the missing rate of all previous six data sets.

Impute Data

Four commonly used imputation methods were employed to handle missing data: replacement with zero (RZ), LOCF, hot-deck imputation, and mean imputation (MN). Hot-deck imputation and MN were each applied using four different levels of precision.

The RZ method represents a straightforward approach to address missing data by substituting all missing values with zeros. This method assumes that an individual is not physically active when not wearing the accelerometer device. Replacing missing data with zero can introduce bias by underestimating PA and can distort the underlying data distribution. Furthermore, the substitution of missing values with zero reduces variability in the data by artificially adding a constant value. In certain cases, replacing missing values with zero may be a logical choice. For example, in a study examining changes in PA levels of patients after surgery, it may be reasonable to replace missing data points during periods of bed rest with zeros.

LOCF is a method used to handle missing data in longitudinal studies or time-series data. In LOCF, the last observed value for a particular variable is carried forward to replace all subsequent missing values for that same variable. LOCF assumes that there is no change in the variable after the last observation, which may not be true. It does not account for trends, patterns, or fluctuations in the data over time. The method can underestimate the variability within a data set by artificially creating a flat trend where actual changes may have occurred. LOCF might be appropriate when the data points are close in time, and there is a reasonable expectation that values remain relatively stable over time.

Hot-deck imputation (HD) is a method used to handle missing data in data sets by replacing missing values with observed values from similar records within the same data set. The donor's value is typically chosen at random among the records that meet the matching criteria. We used four different degrees of precision and based the donor pool on the following matching criteria:

a. Donor values could be any observed values from any participant in the full data set (Hot deck; HD).
b. Donor values were selected from other time points within the same individual's data (Hot deck, id; HD-ID).
c. Donor values were restricted to matching both participant ID and hour of day (Hot deck, id, hour; HD-ID-H).
d. Donor values were restricted to matching participant ID, hour of day, and day of the week (Hot deck, id, hour, day; HD-ID-H-D).

Since the replacement values come from the same data set, hot-deck imputation helps maintain the distribution, variability, and structure of the data, unlike mean or median imputation, which can

distort the distribution. The random selection of the value out of the donor pool allows for variation when imputation is replicated multiple times.

MN is a simple method used to handle missing data by replacing missing values in a data set with the mean of the observed values for that variable. It reduces the variance of the variable by pulling all missing values toward the mean, which can underestimate the true variability in the data. MN methods assume that the underlying distribution of the data is normal. If the data is skewed or follows a different distribution, the imputed values may not be accurate.

We used MN with an added normally distributed error. This is an enhancement over simple MN that aims to preserve the variability of the original data. This method replaces missing values with the mean of the observed data for a variable but adds a random error term drawn from a normal distribution to mimic the natural variability in the data. The variance of the normal distribution corresponds to the variance of the observed data. The added error introduces variability, ensuring that the imputation results differ when performed multiple times. We used four different conditions with increasing precision to group the available data before calculating the means and variances. The grouping criteria corresponded to the matching criteria for the hot-deck imputation:

a. The mean and *SD* were calculated across all participants and all time points. The same global mean with additional error terms (independently drawn for every participant) was used to impute missing data (MN).
b. The mean and *SD* were calculated separately for each participant using all of their observed data. All missing values for a participant were imputed using their personal distribution, meaning the same participant-specific mean was used but with independently drawn error terms (Mean imputation, id; MN-ID).
c. The mean and *SD* were computed using the participant's available data within the same 1-hr time interval. All missing values for a participant were imputed with a random draw from their personal time-specific distribution (Mean imputation, id, hour; MN-ID-H).
d. The mean and *SD* were computed using the participant's available data within the same 1-hr time interval and the same day of the week. All missing values for a participant were imputed with a random draw from their personal time-specific and day-specific distribution (Mean imputation, id, hour, day; MN-ID-H-D).

Multiple imputation was employed to evaluate the precision and variability of the imputation methods by generating 10 imputed data sets for each applicable method. Specifically, this procedure was applied to all variants of the hot-deck imputation and MN methods, where missing values were replaced based on random draws (hot-deck variants and MN variants). Each imputed data set was created independently, using the same missing data patterns but different random draws, allowing us to capture both the within- and between-imputation variability. Importantly, we did not proceed to the pooling phase typically used in multiple imputation, as our aim was not to estimate a specific model parameter but to evaluate the general performance and stability of imputation methods. By refraining from fitting a specific analysis model, we avoided tailoring the imputation process to a particular analytical outcome, thereby ensuring that the resulting imputed data sets remain broadly applicable to a wide range of research questions.

Evaluation of Imputation Performance

A satisfactory imputation result approximates the true (original) data without introducing systematic bias or increasing or decreasing variability. Specifically, performance is indicated by (a) minimal average deviation between imputed and true values, and (b) low variance in these deviations.

In order to assess the performance of each imputation method, we calculated the differences between the imputed and original values once for single imputations and separately for each of the multiple imputations. The distributions of these differences were visualized using boxplots, allowing for comparison across different imputation methods and missing data mechanisms. Mean signed differences (MSD; Catellier et al., 2005) were calculated by averaging the differences between the original and imputed values:

$$\mathrm{MSD} = \frac{\sum_{j=1}^{\mathrm{Nmvp}} (\text{original value}_j - \text{replacement value}_j)}{\mathrm{Nmvp}},$$

with Nmvp representing the total number of missing value points. A close-to-zero MSD represents a smaller bias of the imputed values.

Root mean square differences (RMSD; Catellier et al., 2005) were calculated for each combination of missing data mechanism and imputation approach with a smaller RMSD indicating a smaller variance of deviances.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{j=1}^{\mathrm{Nmvp}} (\text{original value}_j - \text{replacement value}_j)^2}{\mathrm{Nmvp}}},$$

with Nmvp representing the total number of missing value points.

We anticipated varying patterns of activity levels depending on the time of day. During sleep, activity levels are typically low and consistent, whereas daytime activity levels are more variable due to different types of activities. Given that these differences in activity patterns could influence the effectiveness of various imputation methods, we decided to conduct separate analyses for different time periods: the entire day, daytime (9 a.m. to 10 p.m.), and nighttime (1 a.m. to 8 a.m.). We left out those hours of a day with a less clear activity pattern (see Figure 1).

Multiple imputation was used to assess the precision and variability of the imputation methods under different missing data scenarios. We calculated the differences between the imputed and original values for each of the 560 imputed datasets (8 imputation methods × 7 missing data scenarios × 10 iterations).

The evaluation of multiple imputation was performed visually. Figure 2 (in the "Results" section) displays the distribution of differences for each iteration, highlighting the variability of underestimation and overestimation within a single iteration, as well as the variability across different iterations. For a specific iteration, the shape of the distribution indicates the mean deviation and the variation of underestimation or overestimations of active steps. Different iterations of the same imputation technique are depicted using different colors. Ideally, distributions for each iteration should be centered at zero with minimal spread, and the differences between iterations should be small, indicating consistent imputation performance.

Results

We used data from 29 participants in the control group of the ProPELL study (16 females, 13 males). Participants had a mean age of 22.76 years (*SD* = 2.34; range: 20–31) and were all university

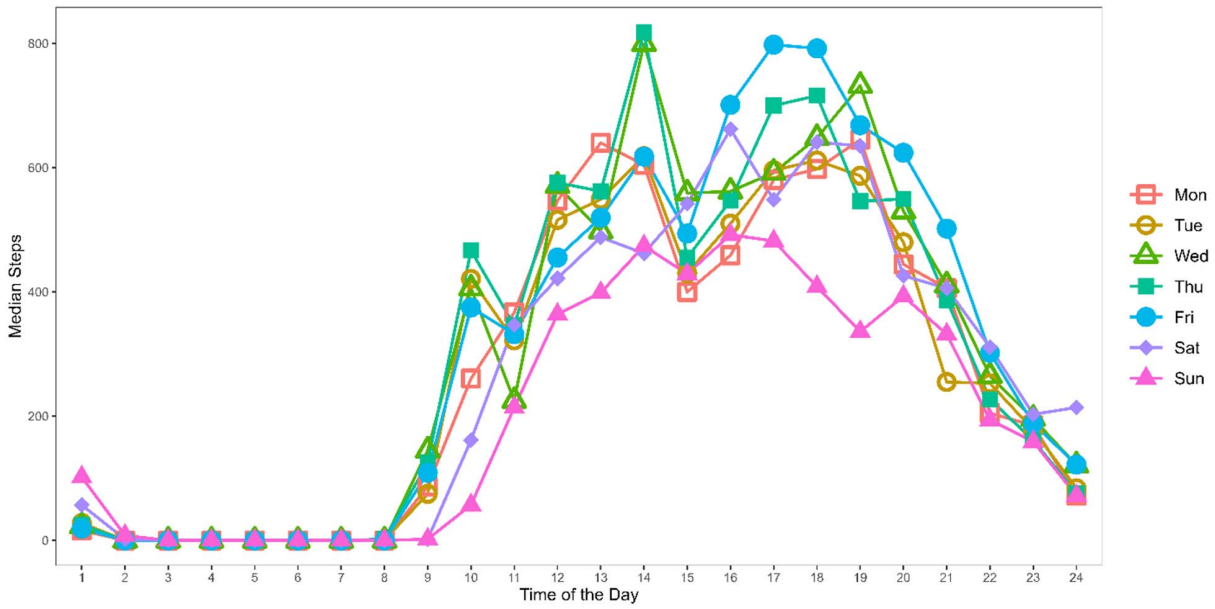

Figure 1 — Median steps by time of day and day of the week.

students. All had a high level of education: Six held a university degree (Bachelor), and 23 had a higher education entrance qualification (e.g., Abitur or Fachhochschulreife). The mean body mass index was 23.38 ($SD = 2.81$).

Physical Activity

In the completed data set and across individuals, active steps fell in the range from 0 to 9,153 steps per time interval (1 hr) with a median of 140 steps per time interval. Interindividual differences in the activity level are reflected in individual medians ranging from 32 to 464 steps per time interval.

Physical activity fluctuated throughout the day, with the most active period occurring between 10 a.m. and 9 p.m. Figure 1 illustrates the median steps by time of day and day of the week. On weekends, activity levels increased later in the day compared to weekdays, with Sunday afternoons being notably least active.

Missings

For the simulation study, we determined the missingness rate based on the first 6 weeks of the experimental group. Out of a total of 29,232 hr, 1,884 hr were coded as missing according to the 20-min criterion, resulting in 6.44% of the 1-hr intervals being classified as missing.[1] The rate of missingness was not uniform throughout the day. Figure 3 presents the distribution of missingness over a day, aggregated across participants and days. The highest missing rate occurred at 1 p.m., with over 8% of the data points missing within this 1-hr interval.

Notably, the amount and duration of nonwear time varied. All but one participant fell within the range from 6 to 126 hr of nonwear time over the 6 week period with a median of 26 hr missing. One participant had a higher nonwear time of 791 hr.

Differences in Imputation Performance

To provide a clear comparison of the imputation approaches, we start with single imputation and subsequently present the results using multiple imputation.

To compare the performance of the different imputation approaches in different missing data scenarios, we plotted boxplots of the underestimation/overestimation for each combination of missing data mechanism and imputation approach (Figure 4) based on the data of the entire day. Each panel represents one missingness scenario, imputation approaches are shown on the x-axis, and differences between original and imputed value are depicted on the y-axis. With the exception of the MTRAIN scenario, all medians were close to the zero, indicating that no central tendency of underestimation/overestimation is present. For MTRAIN, all medians were positive, indicating that all imputation methods underestimated the level of PA. The interquartile range was small in general. Outliers were found for each combination of missingness scenario and imputation approach. As expected, the RZ approach showed only positive values for the differences of original and imputed values.

Results for MSD and root mean squared deviation (RMSD) are presented in Figure 5. The appropriateness of the imputation approach varies depending on the missing data mechanism:

The RZ approach notably underestimated active steps, a finding that aligns with its expected performance characteristics. This method typically exhibits high MSD and RMSD values, indicating that it is not suitable in most scenarios. However, in the context of MNAR data, the RZ approach performed relatively well. This is likely because the likelihood of missing data during low activity periods increases, making the replacement of missing values with zero a pragmatic choice.

Conversely, the LOCF method demonstrated strong performance under MCAR and MAR conditions. When the missing data was based on the pattern of the experimental group, LOCF resulted in poorer imputation performance compared to the HD and MN imputation methods.

Hot-deck methods (HD) exhibit mixed performance, with methods using finer matching criteria (HD-ID-H and HD-ID-H-D) often outperforming those with coarser matching criteria. For MCAR-1, all hot-deck methods performed well. Under MPAT and MNAR, HD-ID-H and HD-ID-H-D demonstrated better performance than HD or HD-ID.

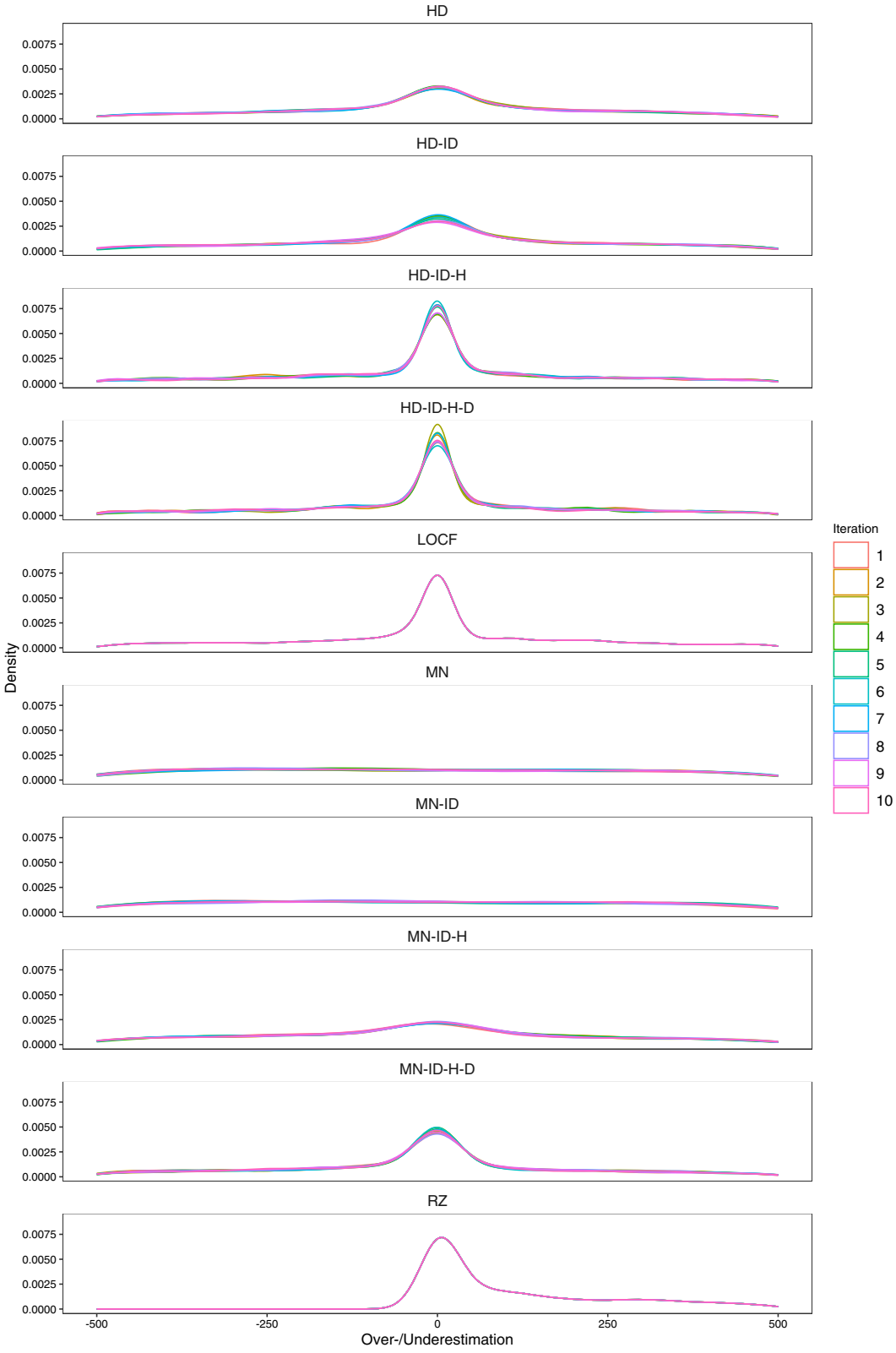

Figure 2 — Distribution of overestimation and underestimation across iterations using multiple imputation. *Note*. This figure shows the distribution of overestimation and underestimation across 10 iterations using multiple imputation for HD imputation and MN methods. Each line represents the distribution of differences between imputed and true values for one iteration. Ideally, distributions for each iteration are centered near zero with narrow spread reflecting minimal bias and small variability within an iteration. Across-iteration variability is represented by the differences between lines, which reflect fluctuations in imputation results across iterations. When lines overlap closely, it indicates consistent imputation performance. Deterministic imputation methods (LOCF and RZ) are displayed as single imputation results. The missing data mechanism was MAR, and the data spans the full 24-hr period. 10 different imputation approaches: RZ, LOCF, HD imputation with different donor pools (no criteria [HD], same participant [HD-ID], same participant and same 1-hr time slot [HD-ID-H], same participant, same 1-hr time slot, and same day of the week [HD-ID-H-D]), and MN with different grouping variables (no criteria [MN], same participant [MN-ID], same participant and same one-hr time slot [MN-ID-H], same participant, same 1-hr time slot, and same day of the week [MN-ID-H-D]). MN = mean imputation; LOCF = last observation carried forward; RZ = replacement with zero; MAR = Missing at Random; HD = hot deck. See online article for color version of the figure.

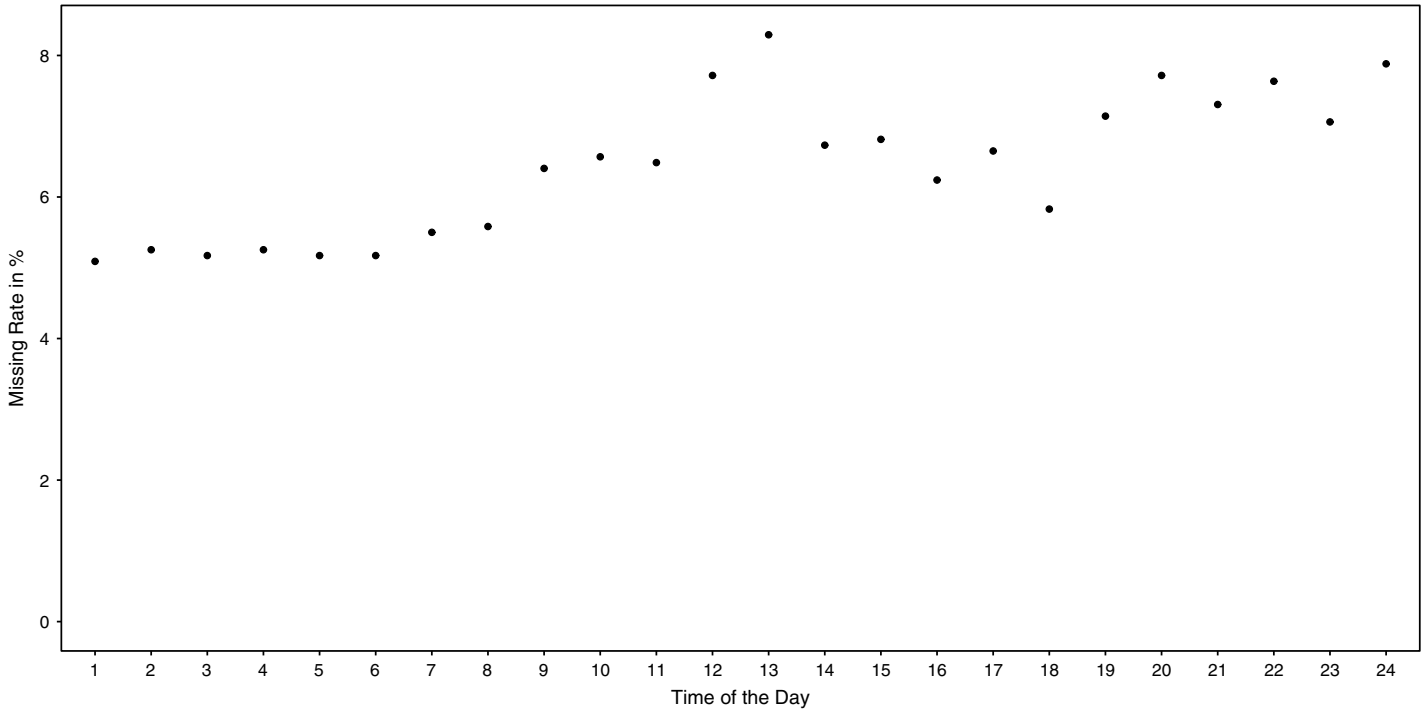

Figure 3 — Hourly distribution of missing data across participants and days. *Note*. This figure shows the percentage of missing data points for each hourly interval, aggregated across participants and days. The missing rate peaks at 1:00 p.m., with more than 8% of data points missing during this interval. See online article for color version of the figure.

MN exhibited variable results. For both MCAR and MAR scenarios, MN-ID-H consistently emerged as the best performer for both 24-hr and daytime data. In nighttime scenarios, MN-ID-H-D also proved to be a robust option. In contrast, MN displayed significant bias under the MNAR mechanism across all cases.

Overall, bias was found to be lower for MCAR compared with other missing data mechanisms, which is consistent with theoretical expectations. For the MAR mechanisms, MN-ID-H performed adequately across both variations. In both MNAR conditions, where the likelihood of missing data increases during periods of either high or low activity, no imputation method produced acceptable estimations. Notably, imputation accuracy generally decreased for daytime data, with higher MSD and RMSD values, compared with nighttime data. This trend is likely attributable to increased variability in the data observed during daytime hours.

Finally, no clear patterns emerged regarding RMSD across imputation methods. However, it was noted that the RMSD was lower for MNAR 24-hr data. This finding is explainable, as periods of low activity were intentionally set to missing. Thus, the deleted values were primarily near zero, which restricted variability in the original values that were marked as missing.

Next, we present the results of the multiple imputations. We explored the precision of the imputation methods by running the process multiple times for the methods where either a random error was included in the imputation process or a random draw of a donor pool. For RZ and LOCF, we again present the distributions of the single imputations for comparison. We illustrate the distributions of overestimation and underestimation for the 24-hr data under the MAR condition in Figure 2. Each panel corresponds to one missing data scenario; each line shows the distribution of differences between imputed and true values for a single iteration. Ideally, distributions should be centered around zero (indicating no systematic overestimation or underestimation) and low variance (indicating high precision). The distributions of the overestimation/underestimation did not vary substantially between iterations. Notably, the distribution for MN was flatter compared with both hot deck and LOCF implying more variation and, hence, imprecision. The distribution of RZ was skewed, as differences could only be positive values by definition. Results for other missing data scenarios were comparable. Supplementary illustrations can be accessed via the Open Science Framework: https://osf.io/ncejd/?view_only = 54156af9a2434dbbbc59fff00229d3a2. Only for MTRAIN, variability between iterations was increased, but distributions still overlapped to a large extent. This suggests that while variability due to imputation uncertainty is reflected by multiple imputations, the overall performance of the imputation methods remained consistent across multiple iterations.

Discussion

In this simulation study, we examined the appropriateness of various imputation methods for handling missing data in a longitudinal accelerometer empirical data set. Our findings underscore the critical importance of selecting suitable imputation techniques to accurately estimate PA levels. The results indicate that while certain methods, such as MN, and hot-deck imputation applied with a fine degree of matching criteria (participant and time of the day), performed well under specific missing data mechanisms (e.g., MCAR and MAR), others, like the RZ approach, were often inadequate. We recommend using multiple imputation with hot-deck imputation applied with a fine degree of matching criteria for the ProPELL project or projects with comparable properties. This recommendation is based on the performance patterns observed in our evaluation of imputation methods across various missing data

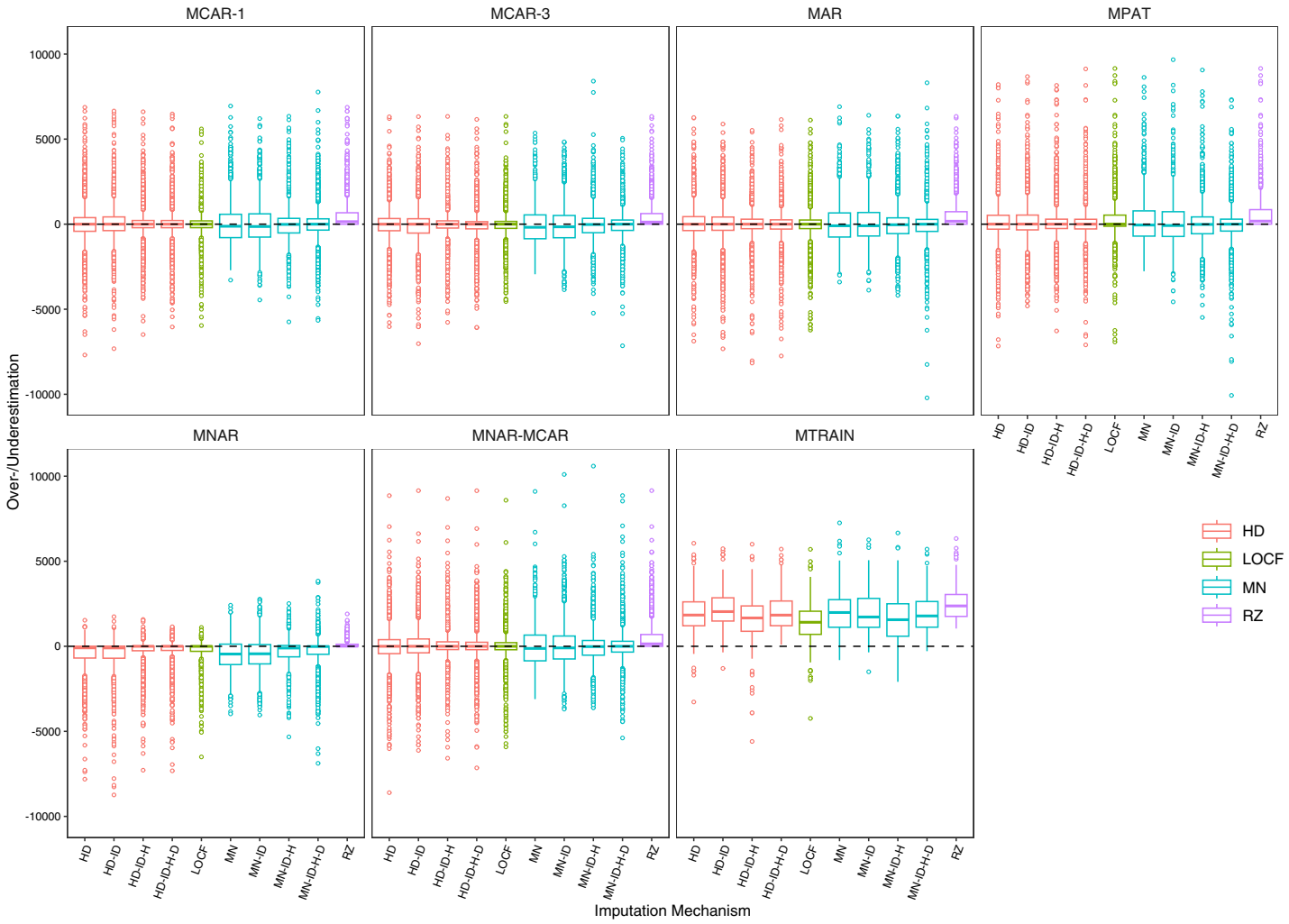

Figure 4 — Comparison of imputation approaches across different missing data scenarios. *Note.* This figure presents boxplots illustrating the distribution of differences between imputed and true values for each combination of missing data mechanism and imputation approach based on data of the entire day. A median close to zero indicates minimal systematic bias in the imputation, while a narrow interquartile range and few(er) outliers reflect lower variability in the differences between imputed and true values. Seven different missing data mechanism: MCAR in 1 hr-periods (MCAR-1); MCAR in 3 hr-periods (MCAR-3); MAR, MPAT, MNAR, combination of MNAR and MCAR (MNAR-MCAR); and MTRAIN. 10 different imputation approaches: RZ, LOCF, HD imputation with different donor pools (no criteria [HD], same participant [HD-ID], same participant and same 1-hr time slot [HD-ID-H], same participant, same 1-hr time slot, and same day of the week [HD-ID-H-D]), and MN with different grouping variables (no criteria [MN], same participant [MN-ID], same participant and same 1-hr time slot [MN-ID-H], same participant, same 1-hr time slot, and same day of the week [MN-ID-H-D]). MN = mean imputation; MAR = Missing at Random; MCAR = Missing Completely at Random; MPAT = Missing based on realistic Pattern; MNAR = Missing Not at Random; MTRAIN = Missing during Training; LOCF = last observation carried forward; RZ = replacement with zero; HD = hot deck. See online article for color version of the figure.

mechanisms. In Figure 4, boxplots demonstrate that hot-deck imputation with specific matching criteria (HD-ID-H and HD-ID-H-D) produces medians close to zero and narrow interquartile ranges, indicating both low bias and high precision in the imputed values. Figure 2 further confirms the robustness of the recommended method. For hot-deck imputation with detailed matching, the distributions are consistently centered around zero with minimal variation across iterations. This approach provides an unbiased estimation of active steps, and the variation due to imputation is reflected in the multiple data sets. For MNAR, none of the imputation approaches provided satisfactory results. This underscores the critical importance of investigating the underlying reasons for missing data, such as nonwear in accelerometer studies. Furthermore, the variability in imputation performance across different scenarios emphasizes the potential for bias if missing data is not addressed appropriately.

We conducted this simulation study with the goal of identifying the most appropriate imputation method for handling missing data of "active steps" in an extended observation period thereby enabling the creation of a data set that accurately reflects participants' PA levels. By including only time-related variables for imputation and excluding other potentially relevant covariates, the imputed data set remains broadly applicable to research questions across various domains. Especially in large-scale studies involving interdisciplinary collaboration across multiple research groups, this approach enables a common data basis for all research groups, streamlines workflows, and ensures seamless link between results. For example, in the ProPELL study, a straightforward imputation

A. 24-hours dataset

	MCAR-1		MCAR-3		MAR		MPAT		MNAR		MNAR-MCAR		MTRAIN	
	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD
HD	0	1262	-44	1177	28	1224	115	1370	-453	1011	1	1244	2059	2440
HD-ID	9	1238	-48	1170	22	1215	85	1385	-462	1039	10	1229	2062	2445
HD-ID-H	-2	1114	-50	1050	-15	1108	35	1284	-305	831	19	1086	1695	2184
HD-ID-H-D	5	1093	-55	1024	-42	1129	42	1246	-297	819	19	1080	2036	2334
LOCF	0	951	-7	954	-17	1082	252	1252	-281	763	4	1034	1356	1899
MN	11	1245	-56	1180	29	1226	108	1367	-456	1022	10	1237	2031	2435
MN-ID	6	1234	-50	1166	14	1209	87	1395	-455	1013	9	1225	2053	2423
MN-ID-H	-3	1120	-53	1048	-8	1119	43	1256	-307	852	18	1087	1695	2218
MN-ID-H-D	6	1164	-55	1062	-33	1195	4	1274	-302	872	26	1137	1977	2274
RZ	542	1054	482	928	557	1025	642	1238	103	218	542	1039	2563	2776

B. Daytime dataset

	MCAR-1		MCAR-3		MAR		MPAT		MNAR		MNAR-MCAR		MTRAIN	
	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD
HD	28	1501	-63	1278	40	1332	115	1563	-633	1222	36	1351	1715	2227
HD-ID	72	1285	-76	1299	16	1303	107	1508	-637	1240	-63	1401	1818	2224
HD-ID-H	29	1389	-76	1295	9	1278	71	1509	-608	1169	15	1350	2057	2358
HD-ID-H-D	70	1512	-76	1360	16	1408	231	1552	-718	1302	39	1465	1763	2210
LOCF	45	1263	25	1192	-4	1303	-202	1492	-463	967	-8	1255	1332	1909
MN	-31	1389	-63	1396	-16	1352	91	1552	-673	1272	187	1427	1706	2232
MN-ID	29	1424	-37	1366	-42	1363	113	1565	-626	1246	39	1414	1679	2186
MN-ID-H	-28	1499	-67	1313	-16	1451	2	1501	-578	1277	67	1481	1847	2180
MN-ID-H-D	113	1457	-103	1396	-16	1406	113	1561	-574	1170	90	1440	1789	2317
RZ	897	1404	793	1224	858	1297	999	1564	205	323	892	1374	2580	2807

C. Night dataset

	MCAR-1		MCAR-3		MAR		MPAT		MNAR		MNAR-MCAR	
	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD	MSD	RMSD
HD	10	597	-11	512	-24	438	33	541	-75	453	17	645
HD-ID	19	539	-30	612	15	341	5	539	-44	298	23	633
HD-ID-H	16	503	-18	475	-13	442	18	496	-74	425	3	646
HD-ID-H-D	-3	627	-17	529	-21	487	-17	632	-73	422	39	600
LOCF	-10	396	-11	346	-24	524	27	296	-69	457	23	533
MN	-1	636	-14	592	-24	501	-10	546	-66	427	31	611
MN-ID	30	534	-26	580	-2	441	-26	527	-69	432	22	651
MN-ID-H	8	562	-3	587	-24	526	-4	561	-89	526	9	586
MN-ID-H-D	8	553	19	568	28	511	-11	603	-80	444	62	616
RZ	96	419	73	410	72	299	77	444	20	71	115	472

Figure 5 — Results: Comparison of imputation methods by MSD and RMSD. *Note.* MSD < 0: The imputed values result in an average overestimation of the activity level. MSD > 0: The imputed values result in an average underestimation of the activity level. A MSD close to zero indicates smaller bias in the imputed values, while a smaller RMSD indicates lower variability of deviations. Table colors range from green (more desirable values) to red (less favorable values). MSD = mean signed differences; RMSD = root mean square differences. See online article for color version of the figure.

approach that does not account for covariates is suitable for addressing diverse research questions across fields such as sports science and psychology, where research questions and related variables may diverge but rely on shared core data.

Limitations

Generally, our primary focus was on comparing the performance of different imputation approaches. We did not systematically vary other potentially influential factors, such as the proportion of missing data, the length of the observation period, the granularity of the time intervals, or the threshold criteria used for defining wear time. Additionally, we did not incorporate trends in PA or intervention effects, which may be relevant in longitudinal or experimental designs. Future research should explore how variations in these factors influence imputation performance. Furthermore, we used a set of (multiple) imputation techniques; yet, there are additional techniques like machine learning-based methods (e.g., MissForest (Stekhoven & Bühlmann, 2012) which we did not consider. Future research should investigate the applicability and performance of additional imputation techniques for imputing PA data.

Furthermore, we fixed the missing data rate for each mechanism at 6.44%, based on the ProPELL data. However, in practice, researchers often encounter a wide range of missing data rates depending on participants' compliance, as well as the chosen temporal resolution and wear-time threshold criteria (Lee & Gill, 2018).

A common design for PA studies involves a 1-week period of measurement (Kapphahn et al., 2022). In studies measuring intervention effectiveness, baseline and follow-up periods also typically span 1 week (Tackney et al., 2023). Not all of the presented imputation approaches are feasible for short study periods, as the available data for donor pools may be insufficient particularly for approaches with stringent matching criteria. Future research is needed to determine the minimum number of weeks required for imputation methods to produce satisfactory results.

As we used minute-long recordings of accelerometer data from the Polar watch, and our imputation and analysis were based on hour-long time intervals, we cannot tell, if other temporal resolutions might be more appropriate depending on the population or research focus. For instance, Trost et al. (2005) suggested using shorter epoch lengths to measure activities of children. The research question can also influence the required resolution; for example, when examining long-term activity changes, a lower resolution might suffice, whereas a higher resolution is necessary to assess short-term activities (e.g., exercise napping) and their effects. Additionally, the threshold for sufficient wear-time must align with the time interval length and study period.

Researchers should cautiously transfer results from other simulation studies, as results often depend on the specific scenario involved. It is crucial to compare one's own study in terms of research question, study design, analysis model, and sample characteristics. In particular, researchers should carefully compare their sample to that of the simulation study with regard to missing data rate, sample size, and activity patterns, as different populations (e.g., older adults or children) may yield different results concerning choice of imputation approach. Our study was based on data from highly educated university students with relatively low activity levels which may limit generalizability to other groups. Many studies comparing imputation methods are based on data from the National Health and Nutrition Examination Survey (Herrmann et al., 2014; Lee & Gill, 2018; Liu et al., 2016). However, results from these studies stem from a specific sample (in terms of gender, age, and activity levels) and may not be generalizable to all accelerometer studies. We recommend searching for simulation studies that closely match the available data. If no suitable study is found, researchers should conduct their own simulation studies, or, at least perform sensitivity analyses by comparing the results of different imputation approaches.

Finally, and related to the previous point, we used "active steps" as the central outcome in our study. It is important to consider that PA studies may also rely on other accelerometer-derived metrics, such as activity counts and time spent in different intensity categories based on cut points (e.g., light, moderate, vigorous PA). Previous research indicates that both the choice of imputation technique and the selection of cut points can introduce considerable variability and potential bias in estimates and classifications of PA levels, which underscores the need for careful methodological considerations when extending findings to alternative metrics (Herrmann et al., 2013; Lee & Gill, 2018; Lyden et al., 2017; Troiano et al., 2008). Therefore, we refrain from any generalization to these cases but recommend to impute the original metric (e.g., active steps) first and categorize the data in a second step if this is possible.

Final Remarks

As health research focuses more on the long-term effects of interventions, such as establishing healthy habits, it becomes essential to investigate long-term processes. With the increasing availability of intensive data from wearable devices, measuring individual physical health parameters over a long time period and with high temporal resolution becomes possible. Effectively handling missing data from wearable devices is crucial not only for measuring PA through step counts but also for other health metrics like heart rate and blood pressure. Proper data handling in these areas will strengthen the reliability of health research outcomes and provide accurate insights into the time-dependent effects of interventions.

In addition to selecting an appropriate imputation approach, we recommend implementing strategies to improve data quality. As the results highlight, when data is MNAR, the estimation of activity levels is compromised. To gather information on nonwear periods, ambulatory assessments should also be conducted. Participants should be asked when they did not wear the device and what activities they engaged in during that time.

In the MTRAIN scenario, bias was particularly elevated. To mitigate this participants should be asked to create a timetable outlining a typical week which would help identify scheduled activities, including regular exercise sessions even during periods when participants do not wear their device. This approach can provide valuable context for interpreting missing data and help reduce bias in activity estimates.

To further enhance data quality continuous monitoring of data transfer should be implemented to allow for prompt detection of noncompliance. If necessary, participants can then receive timely reminders to wear the watch during the study helping to minimize data gaps. Additionally, offering a bonus payout for high compliance could serve as an effective incentive to ensure participants adhere to the study protocols ultimately reducing instances of nonwear time and improving the overall quality of the data collected.

We conclude that MN and hot-deck imputation applied with a fine degree of matching criteria (participant, day of the week, and time of day) outperformed discard-based methods under MCAR

and MAR conditions given the particular data situation of our study. We also learned that much more information about potential causes of missing data is needed (e.g., scheduled training sessions), but also context in general (e.g., depicted by a weekly timetable), allowing to choose the best imputation method. Furthermore, monitoring compliance of participants with prompt reminders to wear/turn on the recording device seems a good option to reduce the amount of missing data. Implementing these strategies together with carefully chosen imputation methods enhances the quality of longitudinal PA measurements from wearable devices and supports robust health research conclusions.

Notes

1 When the threshold is set to 30 min, 6.61% of the time intervals are missing. A more stringent threshold of 40 min results in 7.37% missing time intervals, while an even stricter threshold of 60 min leads to 9.60% missing time intervals.

Acknowledgment

The authors gratefully acknowledge funding from the Committee on Research (AFF; Ausschuss für Forschungsfragen) at the University of Konstanz for the research initiative "ProPELL—Promoting Physical Exercise in Lab and Life."

References

Alhassan, S., Sirard, J.R.,Bieleke, M., Fischer, U., Gruber, M., Kanning, M., Keim, D., Mier, D., Pruessner, J., & Schüler, J. (2022). Promoting physical exercise in lab and life. Identifier DRKS00029727. https://drks.de/search/en/trial/DRKS00029727

Bieleke, M., Fischer, U., Gruber, M., Kanning, M., Keim, D., Mier, D., Pruessner, J., & Schüler, J. (2022). Promoting physical exercise in lab and life. Identifier DRKS00029727. https://drks.de/search/en/trial/DRKS00029727

Borghese, M.M., Borgundvaag, E., McIsaac, M.A., & Janssen, I. (2019). Imputing accelerometer nonwear time in children influences estimates of sedentary time and its associations with cardiometabolic risk. *International Journal of Behavioral Nutrition and Physical Activity, 16*(1), Article 770. https://doi.org/10.1186/s12966-019-0770-0

Borgundvaag, E., McIsaac, M., Borghese, M.M., & Janssen, I. (2017). Imputing accelerometer nonwear time when assessing moderate to vigorous physical activity in children. *Journal of Physical Activity & Health, 14*(11), 852–860. https://doi.org/10.1123/jpah.2016-0706

Butera, N.M., Li, S., Evenson, K.R., Di, C., Buchner, D.M., LaMonte, M.J., LaCroix, A.Z., & Herring, A.H. (2019). Hot deck multiple imputation for handling missing accelerometer data. *Statistics in Biosciences, 11*(2), Article 225. https://doi.org/10.1007/s12561-018-9225-4

Catellier, D., Hannan, P., Murray, D., Addy, C., Conway, T., Yang, S., & Rice, J. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise, 37*(11), S555–S562. https://doi.org/10.1249/01.mss.0000185651.59486.4e

Chen, K.Y., & Bassett, D.R. (2005). The technology of accelerometry-based activity monitors: Current and future. *Medicine & Science in Sports & Exercise, 37*(11), S490–S500. https://doi.org/10.1249/01.mss.0000185571.49104.82

Cho, S., Ensari, I., Weng, C., Kahn, M.G., & Natarajan, K. (2021). Factors affecting the quality of person-generated wearable device data and associated challenges: Rapid systematic review. *JMIR mHealth and uHealth, 9*(3), Article 20738. https://doi.org/10.2196/20738

Evenson, K.R., & Terry, J.W., Jr. (2009). Assessment of differing definitions of accelerometer nonwear time. *Research Quarterly for Exercise and Sport, 80*(2), 355–362. https://doi.org/10.1080/02701367.2009.10599570

Forte, P., Teixeira, J.E., Portella, D.L., & Monteiro, D. (2023). Editorial: Towards a psychophysiological approach in physical activity, exercise, and sports. *Frontiers in Psychology, 14*, Article 670. https://doi.org/10.3389/fpsyg.2023.1191670

Herrmann, S.D., Barreira, T.V., Kang, M., & Ainsworth, B.E. (2013). How many hours are enough? Accelerometer wear time may provide bias in daily activity estimates. *Journal of Physical Activity and Health, 10*(5), 742–749. https://doi.org/10.1123/jpah.10.5.742

Herrmann, S.D., Barreira, T.V., Kang, M., & Ainsworth, B.E. (2014). Impact of accelerometer wear time on physical activity data: A NHANES semisimulation data approach. *British Journal of Sports Medicine, 48*(3), 278–282. https://doi.org/10.1136/bjsports-2012-091410

John, D., & Freedson, P. (2012). ActiGraph and actical physical activity monitors: A peek under the hood. *Medicine & Science in Sports & Exercise, 44*(1S), S86–S89. https://doi.org/10.1249/MSS.0b013e3182399f5e

Kang, M., Rowe, D.A., Barreira, T.V., Robinson, T.S., & Mahar, M.T. (2009). Individual information-centered approach for handling physical activity missing data. *Research Quarterly for Exercise and Sport, 10*, Article 546. https://doi.org/10.1080/02701367.2009.10599546

Kapphahn, K.I., Banda, J.A., Haydel, K.F., Robinson, T.N., & Desai, M. (2022). Simulation-based evaluation of methods for handling nonwear time in accelerometer studies of physical activity. *Journal for the Measurement of Physical Behaviour, 25*, Article 30. https://doi.org/10.1123/jmpb.2021-0030

Karas, M., Bai, J., Str□czkiewicz, M., Harezlak, J., Glynn, N.W., Harris, T., Zipunnikov, V., Crainiceanu, C., & Urbanek, J.K. (2019). Accelerometry data in health research: Challenges and opportunities: Review and examples. *Statistics in Biosciences, 11*(2), 210–237. https://doi.org/10.1007/s12561-018-9227-2

Kim, K., Nikzad, N., Quer, G., Wineinger, N.E., Vegreville, M., Normand, A., Schmidt, N., Topol, E.J., & Steinhubl, S. (2018). Real world home blood pressure variability in over 56,000 individuals with nearly 17 million measurements. *American Journal of Hypertension, 31*(5), 566–573. https://doi.org/10.1093/ajh/hpx221

LeBlanc, A.G.W., & Janssen, I. (2010). Difference between self-reported and accelerometer-measured moderate-to-vigorous physical activity in youth. *Pediatric Exercise Science, 22*(4), 523–534. https://doi.org/10.1123/pes.22.4.523

Lee, J.A., & Gill, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Statistical Methods in Medical Research, 27*(2), 490–506. https://doi.org/10.1177/0962280216633248

Lee, P.H. (2013). Data imputation for accelerometer-measured physical activity: The combined approach. *American Journal of Clinical Nutrition, 97*(5), 965–971. https://doi.org/10.3945/ajcn.112.052738

Liu, B., Yu, M., Graubard, B.I., Troiano, R.P., & Schenker, N. (2016). Multiple imputation of completely missing repeated measures data within person from a complex sample: Application to accelerometer data in the National Health and Nutrition Examination Survey. *Statistics in Medicine, 35*(28), 5170–5188. https://doi.org/10.1002/sim.7049

Loprinzi, P.D., Cardinal, B.J., Crespo, C.J., Brodowicz, G.R., Andersen, R.E., & Smit, E. (2013). Differences in demographic, behavioral, and biological variables between those with valid and invalid accelerometry data: Implications for generalizability. *Journal of Physical Activity and Health, 10*(1), 79–84. https://doi.org/10.1123/jpah.10.1.79

Lyden, K., Keadle, S.K., Staudenmayer, J., & Freedson, P.S. (2017). The activPALTM accurately classifies activity intensity categories in healthy adults. *Medicine and science in sports and exercise, 49*(5), 1022–1028. https://doi.org/10.1249/MSS.0000000000001177

Maeda, H., Cho, C.C., Cho, Y., & Strath, S.J. (2019). Comparing methods for using invalid days in accelerometer data to improve physical activity measurement. *2*(1), 4–12. https://doi.org/10.1123/jmpb.2018-0015

Menai, M., Brouard, B., Vegreville, M., Chieh, A., Schmidt, N., Oppert, J.-M., Lelong, H., & Loprinzi, P.D. (2017). Cross-sectional and longitudinal associations of objectively-measured physical activity on blood pressure: Evaluation in 37 countries. *Health Promotion Perspectives, 7*(4), 190–196. https://doi.org/10.15171/hpp.2017.34

Migueles, J.H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J.R., & Ortega, F.B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine, 47*(9), 1821–1845. https://doi.org/10.1007/s40279-017-0716-0

Morris, T.P., White, I.R., & Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Polar Research Center. (2021). *Polar activity tracking: Background, feedback and validity*. =https://www.polar.com/img/static/whitepapers/pdf/polar-qPxM5uAMwVf5dVf40dIzlbYb10FXeNGZogoactivity-tracking-white-paper.pdf?srsltid=AfmBOopnc-htvFl9mTAnmqPxM5uAMwVf5dVf40dIzlbYb10FXeNGZogo

Prince, S.A., Adamo, K.B., Hamel, M., Hardt, J., Connor Gorber, S., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity, 5*(1), Article 56. https://doi.org/10.1186/1479-5868-5-56

Rich, C., Cortina-Borja, M., Dezateux, C., Geraci, M., Sera, F., Calderwood, L., Joshi, H., & Griffiths, L.J. (2013). Predictors of non-response in a UK-wide cohort study of children's accelerometer-determined physical activity using postal methods. *BMJ Open, 3*(3), Article 2290. https://doi.org/10.1136/bmjopen-2012-002290

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696

Samuels, T.Y., Raedeke, T.D., Mahar, M.T., Karvinen, K.H., & DuBose, K.D. (2011). A randomized controlled trial of continuous activity, short bouts, and a 10,000 step guideline in inactive adults. *Preventive Medicine, 52*(2), 120–125. https://doi.org/10.1016/j.ypmed.2010.12.001

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Stekhoven, D.J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tackney, M.S., Williamson, E., Cook, D.G., Limb, E., Harris, T., & Carpenter, J. (2023). Multiple imputation approaches for epoch-level accelerometer data in trials. *Statistical Methods in Medical Research, 32*(10), 1936–1960. https://doi.org/10.1177/09622802231188518

Troiano, R.P., Berrigan, D., Dodd, K.W., Mâsse, L.C., Tilert, T., & Mcdowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine & Science in Sports & Exercise, 40*(1), 181–188. https://doi.org/10.1249/mss.0b013e31815a51b3

Trost, S.G., McIver, K.L., & Pate, R.R. (2005). Conducting accelerometer-based activity assessments in field-based research. *Medicine & Science in Sports & Exercise, 10*, Article 98. https://doi.org/10.1249/01.mss.0000185657.86065.98

Vetter, V.M., Özince, D.D., Kiselev, J., Düzel, S., & Demuth, I. (2023). Self-reported and accelerometer-based assessment of physical activity in older adults: Results from the Berlin Aging Study II. *Scientific Reports, 13*(1), 10047. https://doi.org/10.1038/s41598-023-36924-5

Xu, S.Y., Nelson, S., Kerr, J., Godbole, S., Patterson, R., Merchant, G., Abramson, I., Staudenmayer, J., & Natarajan, L. (2018). Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity. *Statistical Methods in Medical Research, 27*(4), 1168–1186. https://doi.org/10.1177/0962280216657119