Simulation-Based Evaluation of Methods for Handling Nonwear Time in Accelerometer Studies of Physical Activity

Kristopher I. Kapphahn[1], **Jorge A. Banda**[2], **K. Farish Haydel**[3], **Thomas N. Robinson**[3], **Manisha Desai**[1]

[1]Quantitative Sciences Unit, Stanford University, Stanford, CA, USA

[2]Department of Public Health, Purdue University, West Lafayette, IN, USA

[3]Stanford Solutions Science Lab, Division of General Pediatrics, Department of Pediatrics and Stanford Prevention Research Center, Stanford University, Palo Alto, CA, USA

Abstract

Accelerometer data are widely used in research to provide objective measurements of physical activity. Frequently, participants may remove accelerometers during their observation period resulting in missing data referred to as nonwear periods. Common approaches for handling nonwear periods include discarding data (days with insufficient hours or individuals with insufficient valid days) from analyses and single imputation (SI) methods.

Purpose: This study evaluates the performance of various discard-, SI-, and multiple imputation (MI)-based approaches on the ability to accurately and precisely characterize the relationship between a summarized measure of accelerometer counts (mean counts per minute) and an outcome (body mass index).

Methods: Realistic accelerometer data were simulated under various scenarios that induced nonwear. Data were analyzed using common and MI methods for handling nonwear. Bias, relative standard error, relative mean squared error, and coverage probabilities were compared across methods.

Results: MI approaches were superior to commonly applied methods, with bias that ranged from −0.001 to −0.028 that was considerably lower than that of discard-based methods (ranging from −0.050 to −0.057) and SI methods (ranging from −0.061 to −0.081). We also reported substantial variation among MI strategies, with coverage probabilities ranging from .04 to .96.

Conclusion: Our findings demonstrate the benefit of applying MI methods over more commonly applied discard- and SI-based approaches. Additionally, we show that how you apply MI matters, where including data from previously observed acceleration measurements in the imputation model when using MI improves model performance.

corresponding author. Desai (manisha.desai@stanford.edu).

Keywords

Accelerometer devices are frequently used in medical and public health research to assess physical activity (PA), sedentary behavior (SB), and sleep. These devices rely on sensors to convert acceleration into electrical signals, which can be quantified by algorithms and interpreted as objective measurements of movement (Chen & Bassett, 2005; John & Freedson, 2012).

Some currently used accelerometers provide raw output in the form of high-frequency acceleration measurements in gravity units (g-units) that represent changes in velocity in three orthogonal directions. Prior to analysis, these measurements are often aggregated into epochs (distinct, equally sized windows of time) within individuals over their observation period. Data at a given epoch may be summarized as a count that represents the acceleration level for the epoch. These aggregated quantities may be further summarized into measures like mean counts per minute (CPM), or proportion of time spent in moderate or vigorous PA. Analysis of accelerometer-derived activity measures is complicated by missing data that arise from periods when the study participant does not wear the device as requested —referred to as nonwear periods. As signal recorded during nonwear periods can be misclassified as the participant's activity (e.g., sleep or SB), a number of algorithms have been developed to distinguish wear periods from nonwear periods (Choi et al., 2011; Evenson et al., 2008; Hecht et al., 2009; Syed et al., 2020; Tackney et al., 2021; Troiano et al., 2008). For example, Troiano and others used a 60-min window of zero counts with an allowance of up to 2 min of counts between0 and 100 (Troiano et al., 2008). Assuming nonwear periods are identifiable and known (as assumed for the purposes of our study), methods for handling nonwear periods need to be considered carefully, as the approach could lead to biased and or inefficient estimates of PA, SB, sleep, and relationships of interest that involve such measures.

A diverse set of approaches have emerged for handling such periods in the analysis (Alhassan et al., 2008). These vary from easy-to-apply discard-based approaches like complete case (CC) analyses to more complex approaches such as single imputation (SI) and multiple imputation (MI). The most commonly applied approaches for the analysis of accelerometer data, however, are discard-based approaches that eliminate participants or days with incomplete data (Alhassan et al., 2008; Esliger et al., 2005; Mâsse et al., 2005). For example, CC analyses involve removing all nonwear epochs and analyzing the remaining periods measured during wear time. Variations on discard-based approaches involve eliminating entire days from the analysis if a minimum number of hours of wear time is not met and/or excluding individuals from the analysis if they do not contribute a minimum number of days (Evenson & Terry, 2009). For example, Mâsse and others evaluated various discard-based approaches common in the literature that include a range of definitions for a day to be eligible for inclusion in the analysis: 60% of recorded data during wake time; 80% of data recorded during a standard day, where the latter is defined by the length of time in which 70% of sample wore the monitor; or 12 hr of wear time during

a 24-hr period (Mâsse et al., 2005). Furthermore, it is common to require an individual to have at least 4 days of data for inclusion in the analysis. Mâsse and others demonstrate how even among various discard-based approaches the interpretation of findings can be greatly affected (Mâsse et al., 2005).

Importantly, discard-based approaches are less efficient and produce biased estimates of relationships of interest unless the mechanism driving nonwear is completely random, known as missing completely at random (i.e., if nonwear periods are not related to any observed or unobserved variables of interest, such as an exposure or outcome). In reality, however, data are rarely missing completely at random. More typically, data are either missing at random (MAR) or not missing at random (NMAR; Little & Rubin, 2014). In the former scenario, missingness is related to observed variables only (e.g., if PA is more likely to be missing in boys, and conditional on gender, unobserved PA values are no different than observed PA values). In the latter scenario, missingness may also be related to unobserved variables (e.g., if among boys, PA data are only available during active hours or if among girls, PA data are only available during sedentary periods). It is in these situations (when the data are not missing completely at random), where likelihood-based methods or MI-based methods should be considered over CC approaches in order to characterize relationships of interest efficiently and without bias.

In contrast to discard-based methods, imputation methods, enable use of the whole data set by filling in missing data. SI methods are those that replace missing values with imputed values once, resulting in one full data set (Donders et al., 2006). For example, Alhassan et al. (2008) applied SI to accelerometer data via a composite method that used wear time data to produce participant-specific data stratified by weekday or weekend, which were then combined to generate summary measures that outperformed those generated by a number of discard-based approaches. While they have the potential to address loss of efficiency of discard-based approaches, SI approaches may still be biased. Furthermore, they may still yield incorrect estimates of the standard error *(SE)*, as it has been well established that SI methods do not appropriately account for the uncertainty of the imputation process nor of the imputed value (Little & Rubin, 2014).

In contrast, MI techniques produce valid estimators under less stringent assumptions regarding missingness (MAR) than discard-based approaches, and can be described in three steps: (a) generate multiple imputed data sets from a plausible distribution specified for the data using an imputation algorithm; (b) fit scientific model to each imputed data set; and (c) pool estimates from models (d) in a way that incorporates variation, both within and across imputated data sets, to appropriately account for all sources of uncertainty (Little & Rubin, 2014; Rubin, 1987; Van Buuren, 2007).

We are not the first to consider MI strategies for analyzing accelerometer studies. However, most of the literature does not specifically focus on handling missing data when characterizing a relationship between accelerometer-based measures and other covariates. For example, Lee and Gill's (2018) goal was to impute agnostically to any prespecified analysis plan as part of data preprocessing. To that end, they targeted imputation efforts directly at the epoch-level count data and developed an imputation model that accounts

for autocorrelation and over/under dispersion based on a zero-inflated Poisson lognormal mixture model. Their approach was evaluated via root mean squared error (MSE) and was superior to several other models in the ability to predict actual count data. The method's performance in estimating relationships involving accelerometer data have yet to be characterized. Similar to Lee and Gill, Liu et al. also imputed on the epoch level, using an additive regression, bootstrapping, and predictive mean matching approach on National Health and Nutrition Examination Survey data to address systematic missing step data across individuals and even whole municipalities (Liu et al., 2016). As they were not interested in assessing the performance of imputation in characterizing relationships between accelerometer-based PA measures and a covariate, evaluation of the imputation procedure was similarly done by comparing imputed and actual distributions. Similarly, Catellier et al. evaluated MI and SI approaches for missing summary measures of accelerometer data. They found the two methods comparable with respect to bias, where MI resulted in superior precision (Catellier et al., 2005).

The literature on MI to handle missing data when characterizing relationships between an accelerometer-based measure and covariate is less prevalent. Tackney et al. (2021) provide a framework for the specific context for imputing an aggregated summary measure—daily steps—when daily steps can be considered right censored and used the MOVE-IT trial to illustrate ideas. In a novel study by Butera et al. (2019), a hot deck-based MI approach that involved imputing counts on the epoch level demonstrated comparable or superior performance relative to discard-based approaches. While interest was centered on bivariable relationships and the study was simulation-based, they did not evaluate their methods based on a known (simulated) bivariable relationship. Instead properties were assessed relative to a full data set without missing data, which was taken as the gold standard. Futhermore, the evaluation was limited to one MAR scenario under a limited range of fraction of missingness (from 1% to 10%). Borgundvaag et al. (2017) compared estimation of the association of minutes of moderate or vigorous PA with cardiovascular fitness measures between a discard-based analysis and epoch-level MI approaches in a limited simulation study with a low percentage of missing data (7%) under one MAR condition. Based on their findings that imputation- and discard-based analyses were comparable, they concluded that discard-based analyses were sufficient to address missing data issues.

Despite the prevalence in the literature of different approaches for handling nonwear time, there has yet to be a thorough, simulation-based exploration of how epoch-level imputation methods compare in their ability to characterize relationships between summarized accelerometer-based data and outcome. More specifically, our focus is in assessing the relationship between a summarized measure of PA based on accelerometer data (CPM) and a clinical outcome when imputing at the epoch level. Aside from work by Butera and others, none of the above studies have thoroughly explored performance of methods for handling non-wear periods when addressing this question. Furthermore, none use a simulation approach that enables comparison of statistical properties under a known relationship, critical for a valid comparison. In response to the current gap in the literature, using real data as the basis, we developed a novel, nonparametric pseudo-bootstrap approach to simulate a new set of complete accelerometer data with known relationships between a summarized PA measure (CPM) and outcome (body mass index [BMI]). Through this approach, we

simulated relationships between CPM and BMI, under a wide range of scenarios, and compared the performance of several MI-based and commonly applied analytic approaches for dealing with missingness.

Methods

Participant Pool

Our study is motivated by the Stanford GOALS study, a randomized controlled clinical trial to evaluate a multifaceted intervention to reduce obesity in children conducted at Stanford University (Robinson et al., 2021) that was part of the larger National Heart Lung Blood Institute-sponsored Childhood Obesity Prevention and Treatment Research Consortium (Pratt et al., 2013; Robinson et al., 2013).

Preparation of the Accelerometer Data Sampling Pool

For simplicity and without loss of generality to other univariable measures such as the triaxial-based vector of magnitude, we generated data based on the vertical axis accelerometer data from participants of Stanford GOALS. Using 1-s epoch accelerometer count data from the vertical axis, we aggregated data over a 60-s epoch length by summing across seconds. We then used default arguments in the wearingMarking() function from the PhysicalActivity R package, which implements the nonwear classification algorithm described by Choi et al. (2021), to detect periods of nonwear, and removed all whole number (e.g., 0:00–0:59, 1:00–1:59, etc.) hours that contained any minutes classified as nonwear. The remaining data were used as the sampling pool for the generation of count data as part of the simulated data sets.

Sampling Process

To construct simulated data, we labeled the count data by their corresponding day (e.g., Monday, Tuesday, etc.) and hour (e.g., 0:00–0:59, 1:00–1:59, etc.). Figure 1 shows an abbreviated illustration of the process. We took all wear epochs present in Stanfod GOALS data (a), split them by hour of the day (b), and collected these hour-long pieces of data into a sampling pool (c). For each hour of the day for each of 7 days for each simulated subject, we randomly sampled with replacement from the sampling pool (d). Each sampled hour was kept intact and the order of epochs were maintained. These samples were then combined across hours and days chronologically to construct data for each subject (e). We used this process of sampling 1-hr blocks to generate 100 data sets where each data set was composed of 100 individuals with 7 days (Sunday–Saturday), with longitudinal count measurements composed of underlying data measured within 1-min length epochs.

Assessing PA Level

The PA level for each minute epoch was classified using classification thresholds defined by Evenson et al. (2008) into either sedentary, light, moderate, or vigorous activity.

Our primary PA measure of interest for this study is mean CPM, which is calculated for each individual by summing up their weartime counts—defined as counts (levels of acceleration for a given epoch) during the time when the individual is wearing the device

—and dividing by their number of weartime minutes (the total number of minutes when the individual is wearing the device). For a simulated participant with complete data, there are $7 \text{ days} \times 24 \frac{\text{hours}}{\text{day}} \times 60 \frac{\text{minutes}}{\text{hour}} = 10,800$ min of wear time. For individuals who do not wear their device the entire time, the number of wear time minutes is reduced. If an individual only wore the device half the time, their wear time minutes would be 5,400 min. CPM would be calculated by summing up the counts during these 5,400 min and dividing by 5,400.

Generating Demographic and Baseline Characteristics

We generated race, age, gender, and season, where race and age were generated based on patterns present in the Stanford GOALS data set. Specifically, we assigned race to each member of our simulated cohort based on the marginal probability distribution that matched the Stanford GOALS study (98% Latinx, 2% Black, and <1% other). Age was generated as a sum of the mean GOALS cohort age (9.53 years), within-subject standardized CPM, and random noise from $N(0, \sigma_{\text{age}}^2)$, where σ_{age}^2 was set as the observed variance of age in the GOALS data. Gender was generated as a probabilistic function of a subject's CPM, so that individuals with higher CPMs were more likely to be girls and individuals with lower CPMs were more likely to be boys. In a typical simulated cohort, this produced a 3:2 ratio of females to males. The relationship induced between CPM and gender was not based on any findings in the literature. Instead the decision to have a relationship between a covariate (gender) and outcome (CPM) was motivated by the goal that the relationship could be leveraged in an MI strategy. Relationships between gender and PA, however, are seen in the literature, although not consistently. In a 2-year longitudinal study of Swedish 8-year-olds, Lahti et al. (2019) found that compared to boys, girls had significantly lower PA 2 years after baseline. On the other hand, Schwarzfischer et al. studied children from the Childhood Obesity Project and found no difference in overall PA between boys and girls, but did find a difference in time spent at different PA intensities (Koletzo et al., 2009; Schwarzfischer et al., 2019). The relationships we induced in our simulations were not intended to reflect findings in the literature. Finally, season (school year vs. summer vacation) was randomly assigned to each subject producing a 3:1 ratio of school year days to summer vacation days.

Generating BMI

Our true data-generating model defines the relationship between BMI and CPM as a simple linear model. We generated BMI as a linear function of CPM under three effect sizes indexed by j ($j = 1, 2,$ and 3), using the general form: $\text{BMI}_{ji} = \alpha_j + \beta_j \times \text{CPM}_i + \epsilon_i$, where individuals are indexed by $i = 1, 2, \ldots, 100$, ϵ_i represents random error that is assumed to be normally distributed such that $\epsilon_i \sim N(0, 3)$, and effect sizes that describe the relationship between CPM and BMI took on three values (no relationship, modest relationship, and strong relationship) corresponding to $\beta = 0, 0.05,$ and 0.1 for $j = 1, 2,$ and 3. Intercepts, represented by αs, were chosen so that the mean of each BMI variable was 25. Note that these simulated relationships are not reflective of those found in the literature, where there exists strong evidence for a negative relationship between PA and BMI (Chen et al., 2021). We chose these associations to illustrate principles, and we consider a range of strengths (no relationship, moderate, and strong).

Generating Auxiliary Counts

We generated a single auxiliary variable that was intentionally strongly related to the epoch-level count data (Pearson correlation with counts of 0.9 and never missing).

Inducing Nonwear for Hours and Days

The missing data mechanisms (MDMs) for hours and days were specified separately. We imposed nonwear time in hour-long blocks using the following four MDMs for missing a given hour; model coefficients are explained below:

1. Missing at random (MAR1): The probability that any particular hour is missing is a function of a constant probability and whether or not the previous hour was nonwear. This corresponds to the following equation:

$$P(\text{nonwear}_{ih} = 1) = L\big[\eta_1 + \gamma_{\text{lag1}}\text{nonwear}_{i,h-1} + O(h)\big],$$

where $\text{nonwear}_{ih} = 1$ if hour hr for subject i is nonwear and 0 otherwise. Here, missingness at the particular hour is a function of missingness in the previous hour.

2. Missing related to covariates (MAR2): For each subject, the probability that any particular hour is nonwear is a function of whether the previous hour was nonwear and the following subject-specific covariates: season, time of day, day of the week, sex, race, age, and BMI. This corresponds to the following equation:

$$\begin{aligned} P(\text{nonwear}_{ih} = 1) \\ = L\big[\eta_2 + X_{2ih}\beta_2 + \gamma_{lag_2}\text{nonwear}_{i,h-1} + O(h)\big], \end{aligned}$$

where nonwear_{ih} is defined as above, X_{2ih} contains covariates for subject i at hour hr and β_2 is a vector of corresponding weights.

3. Missing related to covariates and true overall activity (NMAR1): For each subject, the probability that any particular hour is nonwear is a function of whether the previous hour was nonwear and the following subject-specific covariates: season, time of day, day of the week, sex, race, age, BMI, and a subject's true overall activity levels. As a subject's true overall activity level is not always fully observed, this is considered a NMAR mechanism.

$$\begin{aligned} P(nonwear_{ih} = 1) \\ = L\big(\eta_3 + X_{3ih}\beta_3 + \gamma_{lag_3}nonwear_{i,h-1} + O(h)\big) \end{aligned}$$

where nonwear_{ih} is defined as above, X_{3ih} contains covariates for subject i at hour hr and β_3 is a vector of corresponding weights.

4. Missing related to covariates and true epoch-level activity (NMAR2): For each subject, the probability that any particular hour is nonwear is an arbitrarily prespecified function of whether the previous hour was nonwear and the following subject-specific covariates: season, time of day, day of the week,

sex, race, age, BMI, and a subject's true hour-specific activity level, as often missingness may be more or less likely depending on the actual level of activity. Imposing missingness on the hour level as a function of an hour-level resolution variable produces nonrandom missingness.

$$P(\text{nonwear}_{ih} = 1) = L\big[\eta_4 + X_{4ih}\beta_4 + \gamma_{\text{lag4}}\text{nonwear}_{i,h-1} + O(h)\big],$$

where nonwear_{ih} is defined as above, X_{4ih} contains covariates for subject i at hour hr and β_4 is a vector of corresponding weights.

$L()$ is the standard logistic function. Coefficients describing the relationship between each variable and missingness were set according to empirical relationships observed from the GOALS study. For example, for MAR2, we fit a logistic regression model regressing any nonwear on the hour as a function of the previous hour's nonwear status, season, time of day, day of the week, sex, race, age, and BMI to the Stanford GOALS data. We then used the resulting coefficients to calculate a probability of nonwear for each hour in the simulated data sets. An offset term, $O(h)$, was included in the calculations to enable control over the prevalence of nonwear time in our simulated data sets. Specifically, to more closely mimic the real data, the offset term was defined as a periodic function of time of day that increased the probability of nonwear during the evening using the following expression.

$$O(h) = -3 \times \cos\left(\frac{h-12}{12 \times \pi} \cdot\right)$$

Nonwear was then induced by comparing probabilities of missingness to uniformly distributed random variables such that if $p < P(\text{nonwear}_{ih} = 1)$, $p \in U(0,1)$, the hour was set to nonwear.

Additionally, we induced missing data for an entire day under an NMAR mechanism. Specifically, we assumed the probability that a day was missing was a function of gender and true overall activity level. As a subject's true overall activity level may not be fully observed, this is considered an NMAR mechanism.

Methods Used to Handle Nonwear Periods in Analyses

We implemented three general approaches to dealing with nonwear time, which comprised the specific strategies outlined in this section and shown in Table 1.

Discard

Complete Case—The CC approach discards all nonwear data from the analysis.

Subject and/or Day Exclusion (Exclude Days/Subjects)—Entire days of data are discarded if nonwear time is too extensive, and entire subjects are discarded if they are missing too many days. Specifically, we removed days that had fewer than 15 hr of wear time and subjects who had fewer than five (out of a possible seven) days of data.

Single Imputation

Imputation of Nonwear as Sedentary (Nonwear Sedentary Imputation)—
Nonwear time was recategorized as sedentary wear time and corresponding counts were set to zero. This approach, which may or may not be reasonable, represents a strong NMAR assumption about the nature of nonwear—specifically that all nonwear epochs correspond to SB.

Imputation of Cohort-Wide Regression Mean (Regression Imputation)—We used wear time data to regress counts on time of day, an indicator for whether the count occurred on the weekend, sex, race, age, and BMI across the whole cohort and imputed counts falling in the nonwear periods using the model intercept.

Within-Person MI—We split each subject's wear time data into weekday (Monday–Friday) and weekend (Saturday–Sunday) components and the mean number of counts for each minute of wear time was calculated. For nonwear epochs, the subject-minute-specific mean count for the weekday or weekend was used to impute data that occurred on the weekday or weekend, respectively, as studied by Alhassan et al. (2008).

Multiple Imputation by Chained Equations

We used R's Multiple Imputation by Chained Equations package to multiply impute nonwear counts at the epoch level using Multiple Imputation by Chained Equations's built-in Bayesian Linear Regression method (Van Buuren & Groothuis-Oudshoorn, 2011). We calculated CPM in each imputed data set, regressed BMI on CPM, and pooled our results across each data set using Rubin's Rules (Little & Rubin, 2014). More specifically, for each MI strategy, we imputed five data sets, each with five iterations using the Bayesian linear regression ("norm") imputation model (see example code in Supplementary Code [available online]). These MI strategies have three stages. The first stage (imputation) involves imputing count data for nonwear epochs to create a complete data set with minute-long epochs. In the second stage (analysis), for each imputed data, counts are aggregated into CPM, and each participant's BMI is regressed on their CPM. In the third stage (summarization), results from all regression models are pooled to produce a single estimate and confidence interval describing the relationship between CPM and BMI.

We considered MI models that included the following variables:

- MI1: BMI only

- MI2: BMI and demographic variables (hour, weekend, sex, race, age, and BMI).

- MI3: BMI, demographic variables, and observed CPM.

- MI4: BMI, demographic variables, observed CPM, and auxiliary count.

- MI5: BMI, demographic variables, and lag (this matches the true missing data mechanism specified by MAR1).

- MI6: BMI, demographic variables, lag, and auxiliary count where lag at time t is the observed or missing count value from time $t - 1$.

Metrics Used to Compare Performance

These strategies are designed to capture the known relationship we imposed between BMI and CPM. With that goal in mind, we focus on the following metrics which correspond to measures of fidelity between the regression results produced by these approaches and the known relationship.

- Bias: The mean difference between the true value of a parameter and the estimated value across all simulation replications.

- Mean relative *SE*: The mean ratio of model *SE* to the *SE* obtained by fitting the True Model (true data generating model fit to the data with no nonwear data).

- Relative MSE: The ratio of a model's MSE to the MSE obtained by fitting the True Model.

- Coverage: The proportion of simulated data sets in a scenario for which the model-estimated 95% confidence interval contained the true value.

Results

Accelerometer Data Simulation

Subjects in the Stanford GOALS study ranged in number of nonwear minutes from 0 to 15,986 (percentage of nonwear time: 0–78.3). We removed 6,814 hr (14.1%) that included any nonwear periods as identified by the algorithm. This produced a sample pool with 41,635 hr of count data. Summary statistics were nearly equal between simulated sets and real sets—wear time CPM (*SD*) were 349 (802) and 352 (818), respectively, for observed Stanford GOALS data and a randomly selected simulated data set.

We observed negligible within-person clustering at the count level in the GOALS data (intraclass correlation coefficient = .01), and our sampling procedure preserved this characteristic. In the Stanford GOALS study, 12.9% of the epochs were classified as nonwear. In order to better examine the effects of extensive nonwear data, our simulated data had a higher level of nonwear than GOALS data. Overall, the number of imposed nonwear time epochs in simulated data had a median (interquartile range) of 44% (43%–45%) and ranged from 38% to 53%.

Differences in Effect Sizes Between CPM and BMI

Table 2 shows summary statistics across all scenarios shown in Figures 2–4. For simplicity, we present results for β = 0.10 only and include results for other values of β in Supplementary Table S1 (available online), as patterns remained the same under these conditions.

Comparative Performance Among Discard-, SI-, and MI-Based Approaches

Figure 2 shows mean relative *SE* on the *y*-axis and absolute bias on the *x*-axis for β = 0.1 for discard approaches (CC and exclude days/subjects), and SI and the MI1 (simplest MI strategy) approaches. The results for the True Model fall at the intersection of the two red dashed lines. All approaches fall to the left of the vertical red dashed line, indicating

that they underestimate the true effect of CPM on BMI. With the exception of exclude days/subjects under MAR1 and MAR2, all approaches underestimate *SE*. MI1 has substantially less bias than all of discard and SI approaches, with bias ranging from –0.018 to –0.01 BMI units per CPM (or BMI/CPM). The closest approach to MI1 is exclude days/subjects, which has 2.8–5 times the bias of MI1.

Figure 3 shows relative MSEs on the log_{10} scale for discard, SI, and MI1 approaches. Here too, MI1 shows results more consistent with the True Model, with relative MSE ranging from 1.9 to 3.8.

Comparative Performance Across MI Approaches

Considerable variation in bias and efficiency was observed across MI strategies (Table 2, Figure 4). Across all MDMs, bias across the MI strategies ranged from 0.001 to 0.028 BMI/CPM in magnitude. Aside from MAR1 missing mechanisms, the approach with the auxiliary and lag terms in addition to BMI and demographic variables included in the MI model (MI6) had superior performance among MI approaches. For example in MAR2, MI6 yielded bias of 0.002 BMI/CPM with proportional *SE* of 1.01. Among methods that did not include a strong auxiliary variable, MI1 fared better than other MI approaches in terms of bias, with biases ranging from 0.010 to 0.018 BMI/CPM in magnitude. However, this approach underestimated the *SE* for the CPM effect, especially for non-MAR1 approaches. MI5 was superior to MI1. MI that included demographics and a lag variable performed only slightly worse than the MI1 with respect to bias, but more accurately captured the *SE*, with mean relative model *SE*s ranging from 0.96 to 1.06 for non-MAR1 approaches.

Relative MSE improved with the inclusion of a strong auxiliary variable as observed in performance of MI4 and MI6 (Figure 5). Absent a strong auxiliary variable, the simpler MI1 model had slightly lower relative MSE than the remaining MI methods.

Figure 6 shows coverage for MI approaches. Coverage for auxiliary-including models was highest, ranging across MDMs from 0.82 to 0.96, while coverage for other approaches was smaller and variable, ranging from 0.04 to 0.84.

Discussion

Overview

Our study is the first to quantitatively compare the most common missing data strategies along with numerous types of simple MI strategies for handling nonwear in accelerometer studies where relationships between PA and an outcome are of primary interest. Importantly, our approach is suitable for linear regression models where the goal is to draw inference on a relationship between a PA measure and outcome. Thus our metrics for evaluation include bias, relative MSE, and coverage and do not rely on comparisons between imputed and actual data, as the latter is not relevant for reflecting the ability to capture relationships of interest. These results demonstrate that there are clear benefits to using MI to impute epoch-level counts in accelerometer data. We found that MI produced less biased and more efficient estimation of the underlying relationship between BMI and CPM than did the non-MI approaches (discard and SI). Among MI approaches, the use of a strong auxiliary

variable provided the best results in terms of bias, *SE* estimation, coverage, and relative MSE. Absent a strong, epoch-level auxiliary variable, we found that it was more beneficial to include the previous epoch's count in the imputation model than to simply include BMI and demographics.

Imputing Autocorrelated Data

Given the time series nature of the data, it is not clear how to best formulate the imputation model when imputing epoch-level count data for the purpose of exploring the relationship between an outcome and a summarized measure of counts. Lee and Gill (2018) proposed a model that appropriately accounts for the autocorrelation in the data, although their method has not been evaluated in the context of capturing a bivariable relationship. However, we had practical issues with fitting the model using their package. Their algorithm uses a log-normal, zero-inflated Poisson mixture model for imputation. This algorithm imputes within days, starts in each day at a user-specified epoch and moves iteratively forward from that epoch. At each subsequent epoch, it uses imputed values from the previous epoch as part of its imputation model. Our data produced singularity errors in the algorithm when all the previous epoch's values were zero for all nonzero, nonmissing observations in the current epoch. We thus considered a simpler approach to account for the time series nature of the data by including a lag term in our imputation models and achieved superior performance over commonly used methods.

Use of Patient-Level Characteristics in the Imputation Model

The inclusion of demographic variables in MI2 yielded an increase in bias of around 50% compared with the BMI-only MI (MI1). The results from including demographics and observed CPM (MI3) are almost identical to those of MI2 in terms of bias, *SE*, coverage, and relative MSE. These results may seem counterintuitive, but make sense since the data were generated so that the patient-level characteristics relate more to a person's aggregate activity levels than to epoch-level activity. Thus they do not contain useful information for imputing epoch-level data.

The addition of lag (MI5) to imputation models did not improve bias, but did improve *SE* estimation, which yielded higher coverage and reduced relative MSE. For MAR1, this produced an overestimate for *SE*, while for more severe MDMs, inclusion of the lag term produced more efficient estimates.

Use of Auxiliary Terms in the Imputation Model

Auxiliary variables are those variables that may be useful to the imputation process in that they are correlated with either the variable with missing data, the MDM, or both (Little & Rubin, 2014). Including a strong auxiliary variable without the lag term (observed PA data in the prior epoch) (MI4) reduced bias by approximately 75%, yielded relative *SE* within 0.93 for all MDMs, increased coverage probabilities above .8, and reduced relative MSEs to less than 1.5. The pattern of these improvements was also present when adding the auxiliary variable to more complex models (MI6 compared to MI5). Our most robust MI model, MI6, achieved close to 95% coverage probabilities across all MDMs. All nonauxiliary variable-inclusive MI approaches failed to achieve 95% coverage in any scenario, although

they yielded improved properties over discard and SI approaches. As expected, non-MI approaches yielded estimates with bias and underestimated *SE*s. Excluding individuals based on nonwear time thresholds provided the least bias, but resulted in relative MSEs that were higher than that of MI strategies.

Summary of MI Choices in Practice

Our study suggests that MI approaches provide a general improvement over more standard non-MI approaches. Once the decision to use MI has been made, practitioners must decide which variables to include in the MI model. Our results suggest that a strong, epoch-level auxiliary variable; the analytic outcome; and a lag variable are important to include. More complex strategies, like that of Lee and Gill, may be able to further improve on the statistical properties yielded by simple MI-analytic frameworks that impute under a separate model at the epoch-level, summarize, and then estimate the summary measure relationship to outcome in a simple linear model. This assessment remains as future work.

Strengths and Limitations of Our Study

Our study draws conclusions about methods for imputing accelerometer data in the specific context of relating an outcome to an accelerometer-based summary measure when the underlying accelerometer data may be missing. To that end, we developed a flexible approach to simulate accelerometer data that can produce data reflective of a target population and also allows for the imposition of tailored statistical relationships between outcome, covariates, and accelerometer data. Previous studies have worked with existing data where true relationships are unknown and thus do not provide the ability to assess the performance of how well these approaches accurately capture true relationships in the data.

Our approach relies on the assumption that nonwear time can be realistically simulated using wear time from the same hour on a different day. Additionally, we assume that nonwear is known and measured without error. In practice, accounting for uncertainty of nonwear measurements would be sensible and is future work. Our findings here are the first step in understanding how MI methods perform when data are known to be observed or missing.

In addition, we generated and evaluated a strong auxiliary variable highly correlated with minute-level epoch counts. In practice, the auxiliary variable may vary in strength. Our intention, however, was to demonstrate the possibility of MI in the presence of a strong auxiliary term or in its absence, providing a range of its performance. The authors have been involved in numerous trials, however, when accelerometer data was accompanied by strong auxiliary terms and thus, we believe these are plausible scenarios. For example, one study generated accelerometer data from two devices, where one was worn at the hip, measured at two discrete time points, considered the gold standard, and would serve as the outcome, and the second was generated from a Fitbit worn continuously on the wrist that served as both the intervention and to generate secondary outcome data. Due to battery issues, primary outcome data were missing in a large percentage of participants at one of the three time points. In this example study, the Fitbit data could serve as a strong auxiliary variable when imputing gold standard accelerometer data for the primary outcome.

We recognize that the generalizability of our results is limited by our source data. For example, while we summarized PA only using the vertical axis, our results generalize to use of any axis or even the vector of magnitude, which summarizes across the three axes at a given window. The issue of how to handle missingness at the count level when trying to characterize a summarized PA measure and outcome remain if using uniaxial data or a univariable measure like vector of magnitude that summarizes across the three axes. We therefore do not anticipate that CPM aggregated over the vector of magnitude would yield different findings. Another approach that we did not explore here, however, would be to impute the individual axes prior to deriving vector of magnitude rather than imputing vector of magnitude at a given window. Our results do not generalize to this specific scenario and would need further exploration. Furthermore, only one MI strategy in our study matched the true MDM. Thus, most of the MI strategies we considered did not match the more complex MDMs. We see this as a strength because it allows the evaluation of simple MI methods when the true MDM is unknown. We used a readily available imputation method with an underlying normal parametric assumption that is not consistent with the distribution of our simulated data that were nonparametrically generated. However, given that our ultimate goal was to evaluate performance of an accessible method, we felt this was the most useful contribution to the literature. We also did not evaluate the algorithm developed by Lee and Gill because we could not get it to converge using our data, but this may also reflect the practical difficulties in adopting more complex algorithms. Finally, our findings do not necessarily generalize to handling missing accelerometer data through MI under goals that differ from relating an outcome to a PA summary measure. For example, if the goal is to identity different patterns of activity, other considerations of how to best impute would have to be made that depend on the analytic tool used to identify patterns and other covariates. How to apply MI under different types of goals that employ other analytic tools would involve future studies that evaluate the performance of MI in those specific contexts.

Concluding Remarks

We demonstrated compelling evidence that MI produces results superior to commonly applied methods that discard/ignore missing data or use SI to fill in count values for nonwear time when estimating associations between accelerometry-derived summary variables and individual-level outcomes. How the imputation models are derived matters, and we observed considerable variation in performance across MI strategies. In addition to including terms that are related to PA as auxiliary variables, including PA in a previous epoch was advantageous. Among MI models, we noted that the inclusion of patient-level characteristics did not improve model performance as much as the inclusion of a lag variable and strong auxiliary variables.

While MI-based approaches are slightly more difficult to implement than discard or SI methods, software is accessible and easy to adopt. Furthermore, they make up for the increase in complexity with their improved ability to accurately capture true relationships in the underlying data. In fact, we show that discard- and SI-based approaches yielded estimates with considerably more bias and inefficiency to the extent that these techniques should be avoided with accelerometer data when possible.

Future research can leverage our approach to simulating data to evaluate other methods for the analysis of accelerometry data as well as to evaluate additional missing data strategies. Future research on missing data strategies should include the assessment of more complex research questions. These may include relating patterns of PA to outcome, where patterns of PA may be summarized to better describe PA changes over time. Additionally, performance and ease of adoption of more complex MI strategies should be evaluated that include imputing under one scientific model that can characterize more sophisticated relationships between PA and outcome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Alhassan S, Sirard JR, Spencer TR, Varady A, & Robinson TN (2008). Estimating physical activity from incomplete accelerometer data in field studies. Journal of Physical Activity and Health, 5(Suppl. 1), S112–S125. 10.1123/jpah.5.s1.s112 [PubMed: 18364516]

Borgundvaag E, McIsaac M, Borghese MM, & Janssen I. (2017). Imputing accelerometer nonwear time when assessing moderate to vigorous physical activity in children. Journal of Physical Activity and Health, 14(11):852–860. 10.1123/jpah.2016-0706 [PubMed: 28556671]

Butera NM, Li S, Evenson KR, et al. (2019). Hot deck multiple imputation for handling missing accelerometer data. Statistics in Biosciences, 11(2), 422–448. 10.1007/s12561-018-9225-4 [PubMed: 31447952]

Catellier DJ, Hannan PJ, Murray DM, et al. (2005). Imputation of missing data when measuring physical activity by accelerometry. Medicine & Science in Sports & Exercise, 37(11, Suppl. 1), S555–S562. 10.1249/01.mss.0000185651.59486.4e [PubMed: 16294118]

Chen KY, & Bassett DR Jr. (2005). The technology of accelerometry-based activity monitors: Current and future. Medicine & Science in Sports & Exercise, 37(11, Suppl. 1), S490–S500. 10.1249/01.mss.0000185571.49104.82 [PubMed: 16294112]

Chen T, Lin J, Lin Y, Xu L, Lu D, Li F, Hou L, & Yu C. (2021). Effects of aerobic exercise and resistance exercise on physical indexes and cardiovascular risk factors in obese and overweight school-age children: A systematic review and meta-analysis. PLoS One, 16(9), Article e0257150. 10.1371/journal.pone.0257150

Choi L, Beck C, Liu Z, Moore R, Matthews CE, & Buchowski MS (2021). Physical activity: Process accelerometer data for physical activity measurement. R package version 0.2–4. https://CRAN.R-project.org/package=PhysicalActivity

Choi L, Liu Z, Matthews CE, & Buchowski MS (2011). Validation of accelerometer wear and nonwear time classification algorithm. Medicine & Science in Sports & Exercise, 43(2), 357–364. 10.1249/MSS.0b013e3181ed61a3 [PubMed: 20581716]

Donders AR, van der Heijden GJ, Stijnen T, & Moons KG (2006). Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087–1091. 10.1016/j.jclinepi.2006.01.014 [PubMed: 16980149]

Esliger DW, Copeland JL, Barnes JD, & Tremblay MS (2005). Standardizing and optimizing the use of accelerometer data for free-living physical activity monitoring. Journal of Physical Activity and Health, 3(3), 366–383. 10.1123/jpah.2.3.366

Evenson KR, Catellier DJ, Gill K, Ondrak KS, & McMurray RG (2008). Calibration of two objective measures of physical activity for children. Journal of Sports Sciences, 26(14), 1557–1565. 10.1080/02640410802334196 [PubMed: 18949660]

Evenson KR, & Terry JW Jr. (2009). Assessment of differing definitions of accelerometer nonwear time. Research Quarterly for Exercise in Sport, 80(2), 355–362. 10.1080/02701367.2009.10599570

Hecht A, Ma S, Porszasz J, & Casaburi R. (2009). Methodology for using long-term accelerometry monitoring to describe daily activity patterns in COPD. COPD: Journal of Chronic Obstructive Pulmonary Disease, 6(2), 121–129, 10.1080/15412550902755044 [PubMed: 19378225]

John D, & Freedson P. (2012). ActiGraph and actical physical activity monitors: A peek under the hood. Medicine & Science in Sports & Exercise, 44(1, Suppl. 1), S86–S89. 10.1249/MSS.0b013e3182399f5e [PubMed: 22157779]

Koletzko B, von Kries R, Closa R, Escribano J, Scaglioni S, Giovannini M, Beyer J, Demmelmair H, Gruszfeld D, Dobrzanska A, Sengier A, Langhendries JP, Rolland Cachera MF, & Grote V. (2009). European childhood obesity trial study group. Lower protein in infant formula is associated with lower weight up to age 2 y: A randomized clinical trial. The American Journal of Clinical Nutrition, 89(6), 1836–1845. 10.3945/ajcn.2008.27091 [PubMed: 19386747]

Lahti A, Rosengren BE, Dencker M, Nilsson JÅ, & Karlsson MK (2019). Socioecological and biological associations of lower levels of physical activity in 8-year-old children: A 2-year prospective study. BMJ Open Sport & Exercise Medicine, 5(1), Article e000597. 10.1136/bmjsem-2019-000597

Lee JA, & Gill J. (2018). Missing value imputation for physical activity data measured by accelerometer. Statistical Methods in Medical Research, 27(2), 490–506. 10.1177/0962280216633248 [PubMed: 26994215]

Little RJ, & Rubin DB (2014). Statistical analysis with missing data. John Wiley & Sons.

Liu B, Yu M, Graubard BI, Troiano RP, & 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. 10.1002/sim.7049 [PubMed: 27488606]

Mâsse LC, Fuemmeler BF, Anderson CB, et al. (2005). Accelerometer data reduction: A comparison of four reduction algorithms on select outcome variables. Medicine & Science in Sports & Exercise, 37(11, Suppl. 1), S544–S554. 10.1249/01.mss.0000185674.09066.8a [PubMed: 16294117]

Pratt CA, Boyington J, Esposito L, et al. (2013). Childhood obesity prevention and treatment research (COPTR): Interventions addressing multiple influences in childhood and adolescent obesity. Contemporary Clinical Trials, 36(2), 406–413. 10.1016/j.cct.2013.08.010 [PubMed: 23999502]

Robinson TN, Matheson D, Desai M, et al. (2013). Family, community and clinic collaboration to treat overweight and obese children: Stanford GOALS-A randomized controlled trial of a three-year, multi-component, multi-level, multi-setting intervention. Contemporary Clinical Trials, 36(2), 421–35. 10.1016/j.cct.2013.09.001 [PubMed: 24028942]

Robinson TN, Matheson D, Wilson DM, et al. (2021). A community-based, multi-level, multi-setting, multi-component intervention to reduce weight gain among low socioeconomic status Latinx children with overweight or obesity: The stanford GOALS randomised controlled trial. Lancet Diabetes and Endocrinology, 9(6), 336–349. 10.1016/S2213-8587(21)00084-X [PubMed: 33933181]

Rubin DB (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.

Schwarzfischer P, Gruszfeld D, Stolarczyk A, Ferre N, Escribano J, Rousseaux D, Moretti M, Mariani B, Verduci E, Koletzko B, & Grote V. (2019). Physical activity and sedentary behavior from 6 to 11 Years. Pediatrics, 143(1), Article e20180994. 10.1542/peds.2018-0994

Syed S, Morseth B, Hopstock L, & Horsch A. (2020). Evaluating the performance of raw and epoch non-wear algorithms using multiple accelerometers and electrocardiogram recordings. Scientific Reports, 10(1), 5866. 10.1038/s41598-020-62821-2 [PubMed: 32246080]

Tackney MS, Cook DG, Stahl D, et al. (2021). A framework for handling missing accelerometer outcome data in trials. Trials, 22, 379. 10.1186/s13063-021-05284-8 [PubMed: 34090494]

Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, & McDowell M. (2008). Physical activity in the United States measured by accelerometer. Medicine & Science in Sports & Exercise, 40(1):181–188. 10.1249/mss.0b013e31815a51b3

Van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16(3), 219–242. 10.1177/0962280206074463 [PubMed: 17621469]

Van Buuren S, & Groothuis-Oudshoorn K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. 10.18637/jss.v045.i03
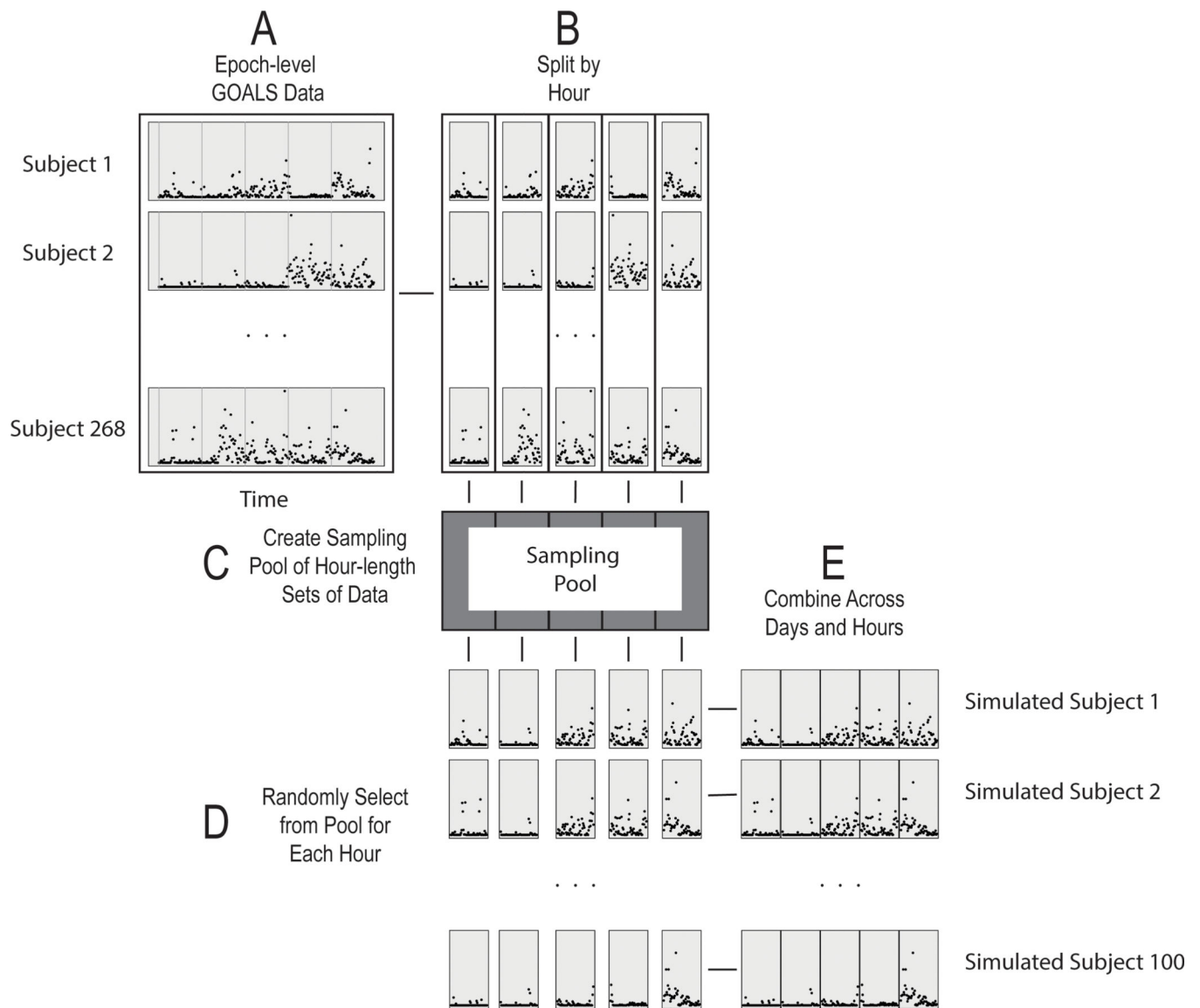
Figure 1 —.
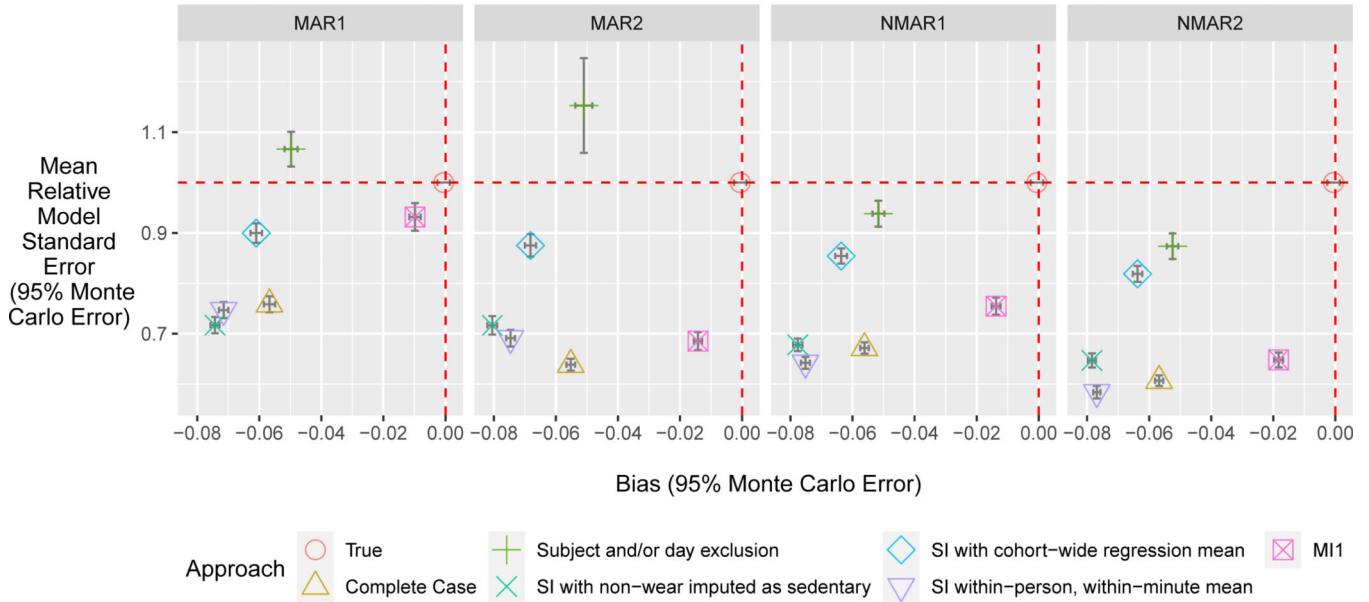Diagram illustrating sampling process used to generate simulated data.

Figure 2 —.

Mean relative *SE* and absolute bias for discard, SI, and MI1 approaches for beta = 0.1. BMI = body mass index; CC = complete case; MAR1 = missing at random; MAR2 = missing related to covariates; NMAR1 = missing related to covariates and true overall activity; NMAR2 = missing related to covariates and true epoch-level activity; SI = single imputation; MI1 = multiple imputation, BMI only.

Figure 3 —.

Relative MSEs for discard, SI, and MI1 approaches for beta = 0.1. CC = complete case; MAR1 = missing at random; MAR2 = missing related to covariates; NMAR1 = missing related to covariates and true overall activity; NMAR2 = missing related to covariates and true epoch-level activity; SI = single imputation; MI1 = multiple imputation, body mass index only; MSE = mean squared error.
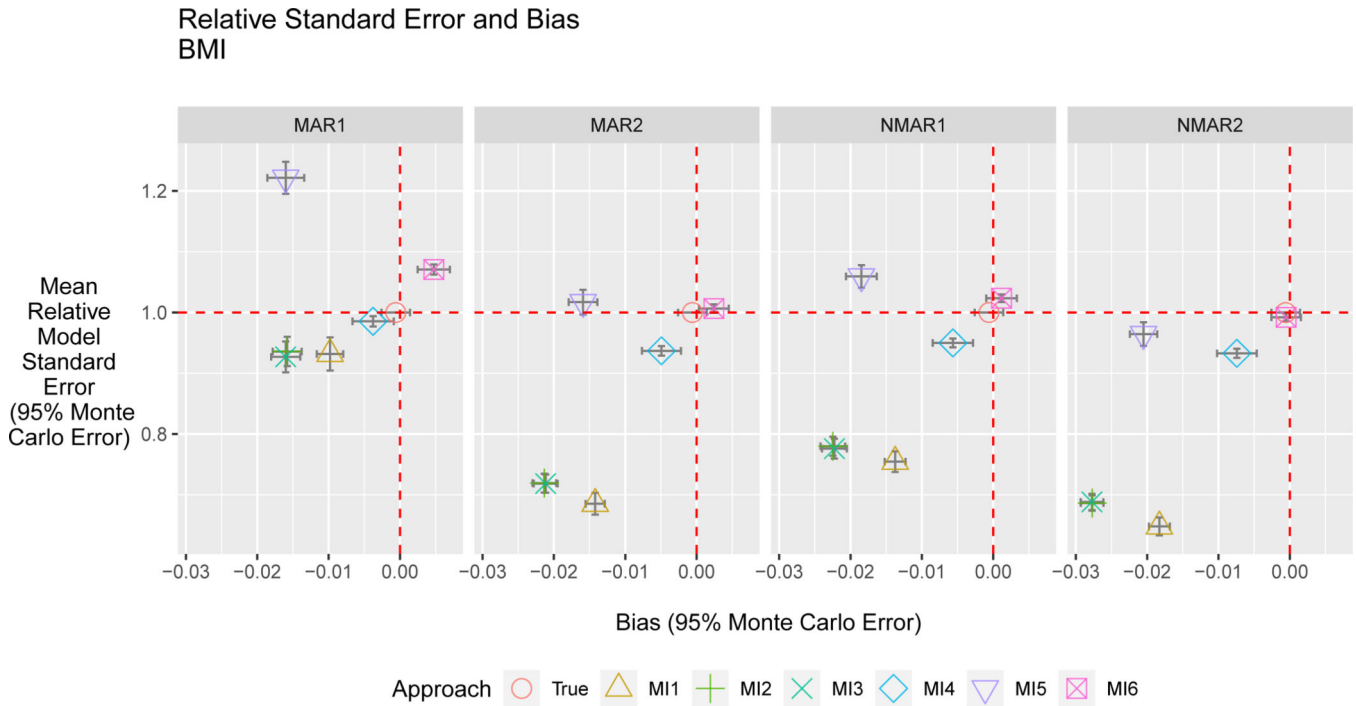
Figure 4 —.

Mean relative *SE* and absolute bias for MI approaches for beta = 0.1. BMI = body mass index; CPM = mean counts per minute; MAR1 = missing at random; MAR2 = missing related to covariates; NMAR1 = missing related to covariates and true overall activity; NMAR2 = missing related to covariates and true epoch-level activity; MI = multiple imputation; MI1 = BMI only; MI2 = BMI and demographic variables (hour, weekend, sex, race, age, and BMI); MI3 = BMI, demographic variables, and observed CPM; MI4 = BMI, demographic variables, observed CPM, and auxiliary count; MI5 = BMI, demographic variables, and lag (this matches the true missing data mechanism specified by missing at random); MI6 = BMI, demographic variables, lag, and auxiliary count.
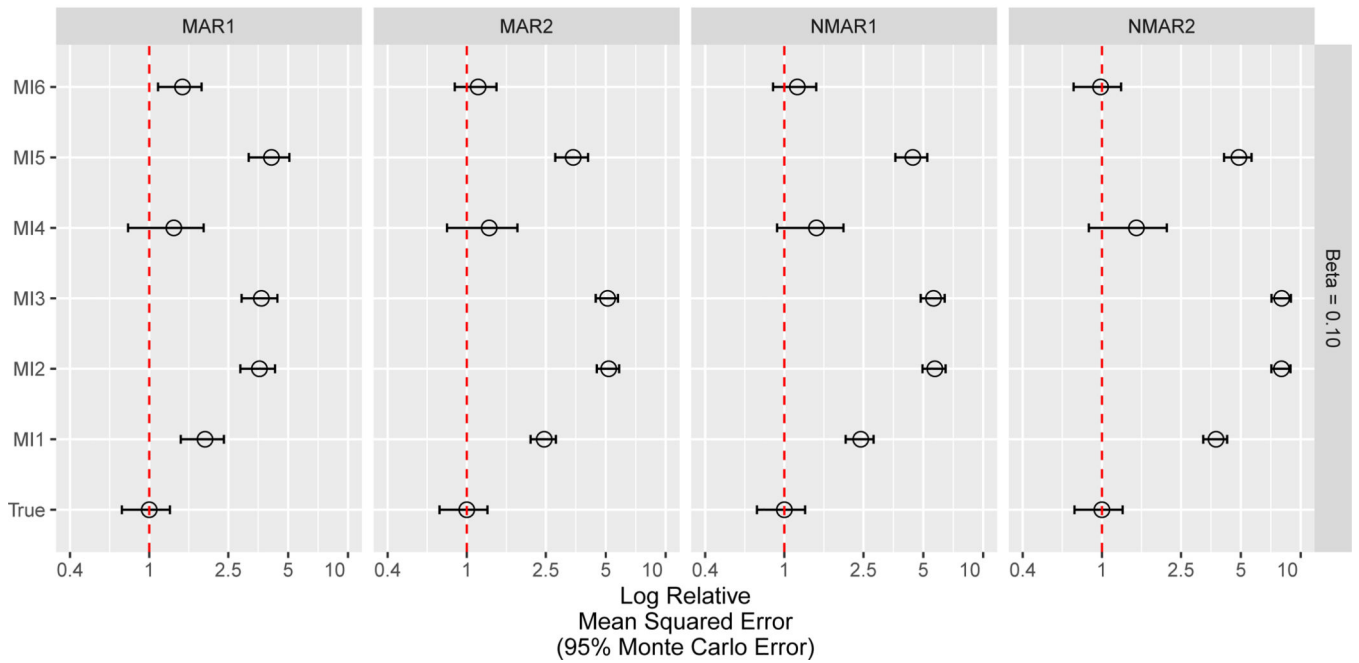
Relative Mean Squared Error

Figure 5 —.
Relative MSEs for MI approaches for beta = 0.1. BMI = body mass index; CPM = mean counts per minute; MAR1 = missing at random; MAR2 = missing related to covariates; NMAR1 = missing related to covariates and true overall activity; NMAR2 = missing related to covariates and true epoch-level activity; MI = multiple imputation; MI1 = BMI only; MI2 = BMI and demographic variables (hour, weekend, sex, race, age, and BMI); MI3 = BMI, demographic variables, and observed CPM; MI4 = BMI, demographic variables, observed CPM, and auxiliary count; MI5 = BMI, demographic variables, and lag (this matches the true missing data mechanism specified by missing at random); MI6 = BMI, demographic variables, lag, and auxiliary count; MSE = mean squared error.

Coverage Probability
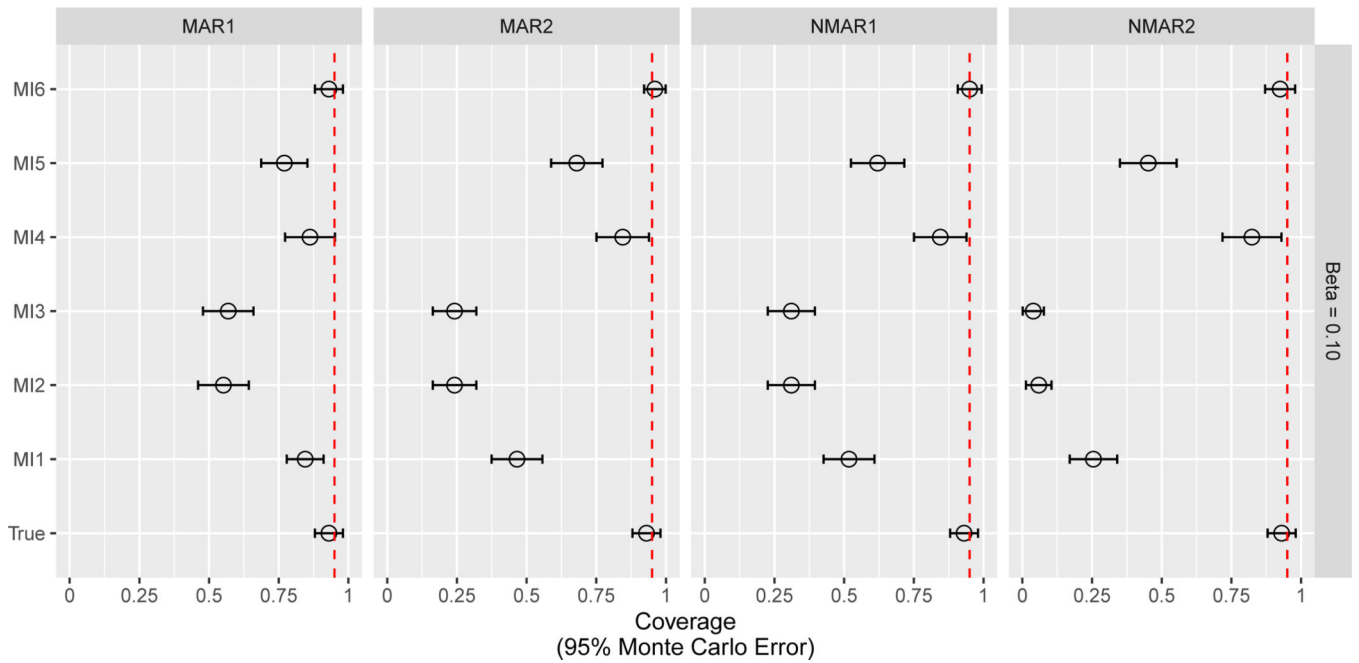
Figure 6 —.

Coverage probabilities for MI approaches for beta = 0.1. BMI = body mass index; CPM = mean counts per minute; MAR1 = missing at random; MAR2 = missing related to covariates; NMAR1 = missing related to covariates and true overall activity; NMAR2 = missing related to covariates and true epoch-level activity; MI = multiple imputation; MI1 = BMI only; MI2 = BMI and demographic variables (hour, weekend, sex, race, age, and BMI); MI3 = BMI, demographic variables, and observed CPM; MI4 = BMI, demographic variables, observed CPM, and auxiliary count; MI5 = BMI, demographic variables, and lag (this matches the true missing data mechanism specified by missing at random); MI6 = BMI, demographic variables, lag, and auxiliary count; MSE = mean squared error.

Table 1:

Non-wear processing approaches

Discard Non-wear	
Complete Case (CC)	Discard all non-wear observations
Exclude Days/Subjects	Discard all days and subjects with insufficient data
Single Imputation (SI)	
Non-Wear Sedentary Imputation	Impute 0 counts for all non-wear observations
Regression Imputation	Impute with a cohort-wide, regression-derived mean count value
Within Person Mean Imputation	Within each subject, impute for each non-wear minute the mean value across that same minute from subject's remaining data separately for weekdays and weekends
Multiple Imputation (MI)	
MI1	Impute counts as a function of subject-specific BMI
MI2	Impute counts as a function of subject-specific BMI and demographic variables
MI3	Impute counts as a function of subject-specific BMI, demographic variables, and observed CPM
MI4	Impute counts as a function of subject-specific BMI, demographic variables, observed CPM, and auxiliary count variable
MI5	Impute counts as a function of subject-specific BMI, demographic variables, and lag
MI6	Impute counts as a function of subject-specific BMI, demographic variables, lag, and auxiliary count variable

Table 2:

Performance statistics for non-wear processing approaches

Approach	Mechanism	Bias	Mean Relative Standard Error	Coverage Probability	Relative Mean Squared Error
True/Complete	MAR1	−0.001	1	0.93	1
True/Complete	MAR2	−0.001	1	0.93	1
True/Complete	NMAR1	−0.001	1	0.93	1
True/Complete	NMAR2	−0.001	1	0.93	1
Complete Case	MAR1	−0.057	0.76	0	31.7
Complete Case	MAR2	−0.055	0.64	0	29.8
Complete Case	NMAR1	−0.056	0.67	0	30.8
Complete Case	NMAR2	−0.057	0.61	0	31.4
Subject and/or day exclusion	MAR1	−0.05	1.07	0	24.9
Subject and/or day exclusion	MAR2	−0.051	1.15	0.03	26.8
Subject and/or day exclusion	NMAR1	−0.052	0.94	0	26.5
Subject and/or day exclusion	NMAR2	−0.052	0.87	0	27.2
SI with non-wear imputed as sedentary	MAR1	−0.074	0.72	0	53.5
SI with non-wear imputed as sedentary	MAR2	−0.081	0.72	0	62.9
SI with non-wear imputed as sedentary	NMAR1	−0.078	0.68	0	58.5
SI with non-wear imputed as sedentary	NMAR2	−0.078	0.65	0	59.5
SI with cohort-wide regression mean	MAR1	−0.061	0.9	0	36.5
SI with cohort-wide regression mean	MAR2	−0.068	0.88	0	45.5
SI with cohort-wide regression mean	NMAR1	−0.064	0.85	0	39.8
SI with cohort-wide regression mean	NMAR2	−0.064	0.82	0	39.5
SI within-person, within-minute mean	MAR1	−0.071	0.75	0	49.6
SI within-person, within-minute mean	MAR2	−0.075	0.69	0	54
SI within-person, within-minute mean	NMAR1	−0.075	0.64	0	54.8
SI within-person, within-minute mean	NMAR2	−0.077	0.58	0	57
MI1	MAR1	−0.01	0.93	0.84	1.9
MI1	MAR2	−0.014	0.69	0.47	2.5
MI1	NMAR1	−0.014	0.75	0.52	2.4
MI1	NMAR2	−0.018	0.65	0.25	3.8
MI2	MAR1	−0.016	0.94	0.55	3.6
MI2	MAR2	−0.021	0.72	0.24	5.2
MI2	NMAR1	−0.022	0.78	0.31	5.7
MI2	NMAR2	−0.028	0.69	0.06	8
MI3	MAR1	−0.016	0.93	0.57	3.7
MI3	MAR2	−0.021	0.72	0.24	5.1
MI3	NMAR1	−0.022	0.78	0.31	5.6
MI3	NMAR2	−0.028	0.69	0.04	8

Approach	Mechanism	Bias	Mean Relative Standard Error	Coverage Probability	Relative Mean Squared Error
MI4	MAR1	−0.004	0.99	0.86	1.3
MI4	MAR2	−0.005	0.94	0.84	1.3
MI4	NMAR1	−0.006	0.95	0.84	1.5
MI4	NMAR2	−0.007	0.93	0.82	1.5
MI5	MAR1	−0.016	1.22	0.77	4.1
MI5	MAR2	−0.016	1.02	0.68	3.4
MI5	NMAR1	−0.018	1.06	0.62	4.4
MI5	NMAR2	−0.021	0.96	0.45	4.9
MI6	MAR1	0.005	1.07	0.93	1.5
MI6	MAR2	0.002	1.01	0.96	1.1
MI6	NMAR1	0.001	1.02	0.95	1.2
MI6	NMAR2	−0.001	0.99	0.92	1