A semiparametric model for wearable sensor-based physical activity monitoring data with informative device wear

JAEJOON SONG*, MICHAEL D. SWARTZ, KELLEY PETTEE GABRIEL,
KAREN BASEN-ENGQUIST

*Department of Biostatistics, The University of Texas MD Anderson Cancer Center,
Pickens Academic Tower, 1400 Pressler Street, Floor 4, FCT4.6007, Houston, TX 77030, USA*
jjsong2@mdanderson.org

SUMMARY

Wearable sensors provide an exceptional opportunity in collecting real-time behavioral data in free living conditions. However, wearable sensor data from observational studies often suffer from information bias, since participants' willingness to wear the monitoring devices may be associated with the underlying behavior of interest. The aim of this study was to introduce a semiparametric statistical approach for modeling wearable sensor-based physical activity monitoring data with informative device wear. Our simulation study indicated that estimates from the generalized estimating equations showed ignorable bias when device wear patterns were independent of the participants physical activity process, but incrementally more biased when the patterns of device non-wear times were increasingly associated with the physical activity process. The estimates from the proposed semiparametric modeling approach were unbiased both when the device wear patterns were (i) independent or (ii) dependent to the underlying physical activity process. We demonstrate an application of this method using data from the 2003–2004 National Health and Nutrition Examination Survey ($N = 4518$), to examine gender differences in physical activity measured using accelerometers. The semiparametric model can be implemented using our R package *acc*, free software developed for reading, processing, simulating, visualizing, and analyzing accelerometer data, publicly available at the Comprehensive R Archive Network.

Keywords: Accelerometry; Augmented estimating equations; Information bias; Semiparametric regression model.

1. INTRODUCTION

Physical activity monitors (e.g., accelerometers) are one of the most widely used wearable sensors both for personal health tracking (Evenson *and others*, 2015) and research purposes (Troiano *and others*, 2014; Troiano, 2006). The devices consist of at least three components: a sensor to detect movement, memory to store the data, and a microprocessor to coordinate between the sensor and the memory. Physical activity data from wearable sensors are reported to have: (i) higher reliability (Evenson *and others*, 2012) and validity (Perry *and others*, 2010) compared to self-report and (ii) less administrative difficulties in

*To whom correspondence should be addressed.

measuring real-life activities compared to article based surveys or diary based methods (Robertson *and others*, 2011). In addition, wearable sensors provide an opportunity to collect human physical activity data throughout the 24-h period. Data from the physical activity monitors consist of a long time-series data, with data collection frequencies ranging from sub-seconds to minutes. For example, the physical activity monitor used in the National Health and Nutrition Examination Survey (NHANES) 2003–2004 (AM-7164; ActiGraph, LLC, Pensacola, FL, USA) recorded activity signals every minute, in units of activity counts. The activity count is a metric developed by Actigraph (ActiGraph, LLC, Pensacola, FL, USA) to quantify the amount of movement experienced by the device wearer, over a short period of time. The measure, expressed as an integer value ranging from zero to infinity, are often translated into some meaningful levels of physical activity based on thresholds proposed in the literature (Troiano *and others*, 2014; Freedson *and others*, 1998; Hall *and others*, 2013). One such quantity is the time spent in moderate to vigorous physical activity (MVPA) or activities exceeding Metabolic Equivalent of Task value of 3. Applied scientists are often interested in obtaining minutes of MVPA in a sustained length of time (e.g., in 10 min bouts) (Troiano *and others*, 2014; Freedson *and others*, 1998; Hall *and others*, 2013), since many health agencies such as the Centers for Disease Control and Prevention (CDC) or the World Health Organization (WHO) report physical activity guidelines in this quantity.

1.1. *Statistical considerations for physical activity monitoring*

While accelerometers provide a unique opportunity in measuring human activity in free living conditions, some caveats for analyzing the wearable sensor monitoring have been noted (Troiano *and others*, 2014; Bai *and others*, 2014; Mâsse *and others*, 2005). The concerns, arising from the nature of wearable sensor data collection, can be summarized as the following. First, in observational studies, we rely on the study participants to wear the sensor devices. Consequently, wear times can be highly variable within- and between-individuals over the scheduled measurement days (e.g., 7-day physical activity monitoring). The observation patterns for the NHANES are illustrated in Figure 1, for 100 randomly selected participants. Even for a relatively short observation window of 7 days, accelerometer wear patterns were quite variable between scheduled measurement days and across individuals. Some of the widely reported reasons for the variability in device wear time were forgetting to wear (Robertson *and others*, 2011), discomfort (Perry *and others*, 2010), disability (Evenson *and others*, 2012), and occupational factors (Perry *and others*, 2010). Second, participants are cognizant of their behavior as well as the sensor monitoring. Therefore, their willingness to wear the monitoring devices can be associated with the outcome of interest. The association between observations patterns (e.g., device wear times) and the underlying phenomenon of interest (e.g., physical activity behavior in free living conditions) in longitudinal studies are referred to as informative observation times in the statistical literature. Third, study participants may stop wearing the device from a certain measurement day and onwards (i.e., censored observations). In addition, the early termination of the wearable sensor monitoring may be related to the outcome of interest, and this phenomenon is called informative censoring in the statistical literature. Issues associated with accelerometer wear have triggered a proliferation of research on how to process accelerometer data (Troiano *and others*, 2014). Currently, two approaches are widely used in response to wear time variability. Both approaches start with removing noise signals from the accelerometer data and identifying device wear times to distinguish inactive periods (e.g., sedentary or sleeping times) from non-wear. A predominant approach is to set a minimum wear time per day (Robertson *and others*, 2011; Troiano, 2006; Mailey *and others*, 2014) and discard data for any day the participant did not meet the minimum wear time. The majority of studies set the minimum wear time per day to 10 h (Troiano, 2006), in response to the literature that data from days with shorter device wear time may be a less valid estimate of inactive and active periods during waking hours (Mâsse *and others*, 2005). Another approach is to normalize engagement in different types of physical activity with respect to total wear time per day (e.g., percent moderate to vigorous physical activity per day or percent

Fig. 1. Figure represents device wear times for 100 randomly selected participants in the NHANES 2003–2004 cycle, who were at least 18 years of age and have participated in the physical activity monitoring component ($N = 4518$). Each horizontal line in the figure represents levels of wear time per day for one participant, over the seven scheduled measurement days. Higher wear time per day is represented by a darker color. Wear time per day was calculated following the accelerometer data processing rules in the literature. Details of the data processing procedure can be found in Appendix S1 of [supplementary material](#) available at *Biostatistics* online.

sedentary time per day) (Choi and others, 2011), among days meeting the minimum wear time criteria. This approach implicitly assumes that the amount of time spent in each activity is proportional to the daily wear time.

1.2. Statistical methods for irregularly observed data

In contrast to the extensive list of publications focused on data processing rules to address issues associated with device wear time (Troiano and others, 2014; Freedson and others, 1998; Hall and others, 2013; Bai and others, 2014), limited literature exist on statistical methods to account for the caveats in the wearable sensor monitoring. Namely, the information bias arising from the associations between the device wear patterns and the underlying physical activity process should be considered in statistical models analyzing factors associated with physical activity. In the statistical literature for time-to-event analysis, a family of modeling techniques has been developed to account for these factors (Huang and others, 2006). The methods have been used predominantly to model the number of recurring medical events (such as tumor recurrence) observed at random visits to the doctor's office (Huang and others, 2006). Whereas the classical model for recurrent events assume that the observation times (e.g., visits to the doctor's office) are independent of the observed phenomenon, a recently proposed approach, known as the semiparametric regression approach with augmented estimating equations (Wang and others, 2013), allows for possible associations between the observation pattern and the event process. In Wang and others' approach (2013), parameter estimation is achieved using augmented estimating equations with a missing data perspective. In simulation studies for time to event processes, Wang and others (2013) showed that their model had extended flexibility, less bias, and increased precision compared to the classical model, especially when the observation patterns were notably associated with the event process. In this article, we first present a simulation study to quantify the bias from the conventional statistical method (generalized estimating equations) in estimating the effect of a predictor on irregularly observed accelerometer data. Second,

we frame the semiparametric statistical approach proposed by [Wang and others \(2013\)](#) in modeling physical activity data from wearable sensors. We show that the estimates from the semiparametric modeling approach were unbiased in both conditions, when the device wear patterns were (i) independent or (ii) dependent to the underlying physical activity process. We demonstrate an application of this method using data from the 2003–2004 NHANES ($N = 4518$). The semiparametric modeling approach presented in this article can be implemented using our R package `acc`. The R package is comprehensive software application for reading, processing, simulating, visualizing, and analyzing accelerometer data, publicly available at the Comprehensive R Archive Network ([Song and Cox, 2015](#)). Latest versions of the `acc` package are also available through GitHub: <https://github.com/github-js/acc>. We illustrate the function to fit the semiparametric model using the R package in Appendix S2 of [supplementary material](#) available at *Biostatistics* online.

2. METHODS

2.1. Motivating data

Our study was motivated by the physical activity monitoring component of the NHANES. The NHANES is a nationally representative survey research program conducted by the National Center for Health Statistics (NCHS), Center for Disease Control and Prevention (CDC). Since the NHANES in 2003, accelerometers were used to collect physical activity information (for cycles 2003–2004, 2005–2006, 2011–2012, and 2013–2014). In NHANES 2003–2004 cycle, participants wore uni-axial accelerometers (physical activity monitors with a uni-directional motion sensor) on their waist in free living conditions, for seven consecutive days. Participants were instructed to wear accelerometers all day, except at bedtime and when in contact with water (i.e., removed when swimming or bathing because the devices were not waterproof). In this study, our outcome of interest was the minutes of moderate to vigorous physical activity (MVPA). The MVPA is a widely used activity classification in the epidemiologic literature ([Troiano and others, 2014](#); [Troiano, 2006](#)), often used as a standard unit in national physical activity guidelines (<https://health.gov/paguidelines/guidelines/>). The MVPA is defined as activities with at least three times the energy cost compared to resting state (i.e., exceeding Metabolic Equivalent of Task of 3). In the NHANES 2003–2004, the wearable devices recorded activity intensities every minute, in units of activity counts. For our data analysis, we classified activity intensities into MVPA or below MVPA, and obtained minutes of MVPA in 10 min bouts (i.e., MVPA that persist for a sustained length in a moving time window of 10 min, with allowance of up to 2 min epochs below the MVPA threshold) following the accelerometer data processing rules in the literature ([Freedson and others, 1998](#); [Mâsse and others, 2005](#)). Details of the data processing procedure can be found in Appendix S1 of [supplementary material](#) available at *Biostatistics* online.

2.2. Simulation study

We evaluated the impact of informative observation times in a simulation study. The objective of the simulation study was to assess bias that may occur when estimating associations between a fixed predictor and minutes of MVPA in 10 min bouts, measured longitudinally using wearable sensors. We simulated data for a 7-day physical activity monitoring period (total of $60 \times 24 \times 7 = 10\,080$ min) under two scenarios when: (i) scenario 1: the observation patterns were independent of the physical activity process and (ii) scenario 2: the observation patterns were dependent on the physical activity process. Data were generated to mimic realistic measurements of minutes of MVPA occurring in 10 min bouts, using models from the statistical literature ([Huang and others, 2006](#); [Wang and others, 2013](#)). These models consist of a fixed predictor X_i , generated from a Bernoulli distribution with success probability of 0.5. The model also included a subject-specific random variable Z_i , generated from a Gamma distribution with shape

and rate equal to 5 ($E(Z) = 1$). We also generated a random number of observed MVPA bouts for each simulated individual. This accounts for the fact that individuals may have different number of MVPA bouts in irregular patterns and may be differentially lost to follow-up. For simulation scenario 1 (observation patterns independent of physical activity process), the number of observed MVPA bouts were generated from a Poisson distribution with mean 6. A constant of 1 was added to the generated counts, to avoid having zero number of MVPA bouts. This resulted in an average of 7 observed MVPA bouts over the 7 day monitoring period, consistent with that we found in the real data from the NHANES 2003–2004. The observation times between the MVPA bouts were generated from an exponential distribution with rate equal to $7/10\,080$. Such specification of the observation times allowed to generate random occurrence times for the MVPA bouts with an average of $10\,080/7 = 1440$ min between each observations. The minutes of MVPA in each bouts were generated from a Poisson distribution with intensity $12Z_i\exp(X_i\beta)$, where the true value of β was set to -0.4 . The formulation would generate an average of 12 min of MVPA when $X = 0$ and 8 min of MVPA when $X = 1$. We specified such parameters to be consistent with the real data from the NHANES 2003–2004. In the NHANES 2003–2004, males had an average of 12 min, and females had an average of 8 min of MVPA per bout (when evaluated in a moving window size of 10 min for determining persistent and sustained physical activity of at least 8 min, i.e., 10 min bouts). For simulation scenario 2 (observation patterns associated with physical activity process), the number of observed MVPA bouts were generated differentially by X_i and Z_i . Specifically, the number of observed MVPA bouts were generated (i) from a Poisson distribution with mean $\delta_1 = 6$, when $X_i = 1$ and $Z_i \leq 1$ and (ii) from a Poisson distribution with mean δ_2 less than 6, when $X_i = 0$ or $Z_i > 1$. We varied this parameter δ_2 from 1 to 5, in order to assess the bias and coverage probability under increasing amount of information from X_i and Z_i . Such data generating procedure imposed observation patterns dependent on the observed predictor (X_i), as well as on the unobserved subject-specific variable (Z_i). A constant of 1 was added to the generated counts, to avoid having zero number of MVPA bouts. This resulted in an average of 7 observed MVPA bouts, when $X_i = 1$ and $Z_i \leq 1$, and an average of 2 to 6 observed MVPA bouts $X_i = 0$ or $Z_i > 1$. The observation times between the MVPA bouts were generated from an exponential distribution with rate equal to $7/10\,080$ when $X_i = 1$ and $Z_i \leq 1$. When $X_i = 0$ or $Z_i > 1$, the observation times between the MVPA bout were generated from exponential distribution with rate equal to $(\delta_2 + 1)/10\,080$. Such specification of the observation times allowed to generate random occurrence times for the MVPA bouts with an average of $10\,080/7 = 1440$ or $10\,080/(\delta_2 + 1)$ min between each observations. The minutes of MVPA in each bouts were generated from a Poisson distribution with intensity $12Z_i\exp(X_i\beta)$, where the true value of β was set to -0.4 . The data generating procedure is illustrated in Figure 2.

We compared the performance of our semiparametric approach to the generalized estimating equations through average bias and coverage probability. The Monte Carlo simulation study evaluated model performance over varying sample sizes ($n = 50, 100, 200$), each with 1000 simulated data sets. The simulation study is reproducible through our R code available on GitHub: <https://github.com/github-js/semiparametric>.

2.3. Semiparametric model

Semiparametric models have been found to be useful in modeling events with irregular observation patterns (Huang and others, 2006). In this study, we frame a semiparametric model with augmented estimating equations (Wang and others, 2013) to model minutes MVPA in 10 min bouts as the following.

1. Let the cumulative minutes of MVPA for individual i , from the start of the wearable sensor monitoring up to time t be $N_i(t)$, for subjects $i = 1, \dots, n$.
2. We assume that engagements in MVPA are observed at subject-specific, random observation times. In other words, the minutes of MVPA for individual i are observed for a random number of bouts,

Fig. 2. Data for 200 individuals were generated under simulation scenario 2 (observation patterns associated with physical activity process). The number of observation days were generated differentially by X_i and Z_i . For example, the number of observed MVPA bouts were generated (i) from a Poisson distribution with mean 6, when $X_i = 1$ and $Z_i \leq 1$ and (ii) from a Poisson distribution with mean 1, when $X_i = 0$ or $Z_i > 1$. A constant of 1 was added to the generated counts, to avoid having zero number of MVPA bouts. The charts illustrate the time points in which the events (MVPA bouts) occur, along with the cumulative minutes of MVPA. It can be observed that MVPA bouts are more frequent when $X_i = 1$ and $Z_i \leq 1$ (average number of MVPA bouts would be 7), compared to otherwise (average number of MVPA bouts would be 2). Such data generating process allowed for the observation and censoring times to be informative through a covariate X_i and a latent variable Z_i .

denoted M_i . The actual times observed are indexed as $0 < T_{(i,1)} < \dots < T_{(i,M_i)}$, specific to the subject i . Note that this formulation accounts for the fact that individuals may have different numbers of MVPA bouts in irregular patterns, and may be differentially lost to follow-up.

3. In the semiparametric regression model, the cumulative minutes of MVPA for individual i , from the start of the wearable sensor monitoring up to time t (i.e. $N_i(t)$), are modeled in association to a set of predictors $\mathbf{X}_i$ as the following.

$$\lambda(t; \mathbf{X}_i, Z_i) = E[N_i(t) | \mathbf{X}_i, Z_i] = \Lambda_{0i}(t) Z_i \exp(\mathbf{X}_i^T \boldsymbol{\beta}), \quad (2.1)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression type coefficients, $\mathbf{X}_i^T$ is a $1 \times p$ vector of fixed covariate values for the i th individual, and $\Lambda_{0i}(\cdot)$ is an unspecified baseline mean function.

4. Parameters are estimated using an augmented estimating equations approach, as described below. We denote all observation and censoring times from the data as $s_1, \dots, s_m$, in the scheduled measurement period from 1 to τ . The union of the observation and censoring times $s_1, \dots, s_m$ are used to form a data-dependent grid $G = 0 = s_0 < s_1 < \dots < s_m = \tau$. On the aggregated grid G , we denote individuals not lost to follow-up (censored) as $r_{ij} = I(s_j \leq T_{(i,M_i)})$, for individual i on time j . We express the minutes of MVPA for individual i on time j as a difference between two adjacent

cumulative quantities as $\mathbb{N}_{ij} = N_{i(s_j)} - N_{i(s_{j-1})}$. Also, we let the baseline mean minutes of MVPA occurring as $\lambda_j = \Lambda_{0(s_j)} - \Lambda_{0(s_{j-1})}$, for times $j = 1, \dots, m$.

5. Estimation is obtained through an iterative approach called the Expectation–Solution (ES) algorithm (Elashoff and Ryan, 2004). The ES algorithm starts with a set of estimating equations for parameters $\boldsymbol{\beta}$ and $\lambda_j, j = 1, \dots, m$, as the following.

$$\mathbf{U}(\boldsymbol{\beta}) = \left[\begin{array}{c} \sum_{i=1}^n \sum_{j=1}^m [\mathbb{N}_{ij} - \lambda_j \exp(\mathbf{X}_i^T \boldsymbol{\beta})] r_{ij} \\ \sum_{i=1}^n \sum_{k=1}^m [\mathbb{N}_{ik} - \lambda_k \exp(\mathbf{X}_i^T \boldsymbol{\beta})] \mathbf{X}_i^T r_{ik} \end{array} \right]. \quad (2.2)$$

In order to obtain estimates of $\boldsymbol{\beta}$ and $\lambda_j, j = 1, \dots, m$, with incomplete $\mathbb{N}_{ij}$, the ES algorithm iterates between two steps to (i) first take the conditional expectation given the observed data and (ii) second to solve the conditionally expected estimating equations. Specifically, the first step calculates the conditional expectation as the following:

$$e_{ij} = E[\mathbb{N}_{ij} | \mathbf{X}_i, N_i(T_{i,1}), \dots, N_i(T_{i,M_i})] = \frac{\lambda_j N_i(T_{i,M_i})}{\Lambda_0(T_{i,M_i})}, \quad (2.3)$$

and the second step replaces $\mathbb{N}_{ij}$ with e_{ij} , and solves for $\boldsymbol{\beta}$ using the Newton–Raphson algorithm. The estimate for the baseline mean minutes λ_j for time j in the second step is obtained as $\hat{\lambda}_j = \hat{\lambda}_j(\hat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^n e_{ij} r_{ij}}{\sum_{i=1}^n \exp(\mathbf{X}_i^T \hat{\boldsymbol{\beta}}) r_{ij}}, j = 1, \dots, m$. The parameter estimation is achieved in an iterative manner until a pre-determined convergence criteria was satisfied. The convergence criteria is satisfied if: (i) the maximum difference in the estimated $\boldsymbol{\beta}$ or (ii) the largest relative difference in the estimated $\boldsymbol{\beta}$ (i.e., the difference relative to the estimate from the previous iteration) from the two most recent iterations is smaller than $10^{(-6)}$. Standard errors (SE) of the parameter estimates are obtained by sandwich variance estimation (White, 1982).

3. RESULTS

3.1. Simulation study

A simulation study was performed to evaluate bias in estimating associations between a fixed predictor and levels of physical activity, measured using wearable sensors. For simulation scenario 1, when the observation patterns were independent of the physical activity process, the Wang and others (2013) semiparametric approach were unbiased (Table 1, Figure 3 panel *a* when $\delta_2 = 6$). The GEE approach showed a small magnitude of bias even in this independent scenario (Table 1, Figure 3 panel *a* when $\delta_2 = 6$). Under simulation scenario 2: when the observation patterns were dependent on the physical activity process, estimates from the GEE had substantial amounts of bias (Table 1, Figure 3 panel *a* when $\delta_2 = 1, \dots, 5$ with smaller δ_2 imposing stronger information through the fixed covariate X and latent variable Z). The level of bias increased for GEE when there was stronger association between the observation pattern and the physical activity process (Table 1, Figure 3 panel *a*). The direction of bias was against the null for GEE. Our proposed semiparametric approach was unbiased in both simulation scenarios. In terms of the confidence intervals containing the true parameter value, all models (GEE and the semiparametric approach) showed acceptable coverage probability (close to the nominal level of 0.95 or greater), for simulation scenario 1 (Table 1, Figure 3 panel *b* when $\delta_2 = 6$). However, when observation patterns were associated with the physical activity process (scenario 2), confidence intervals from the GEE contained the true value of the parameter fewer times than the expected level of .95 (Table 1, Figure 3 panel *b* when

Table 1. Results from a simulation study to evaluate bias and coverage probability in estimating associations between a fixed predictor and levels of physical activity, measured using wearable sensors. Average bias across 1000 simulated data sets are presented for simulation scenarios 1 and 2 when the observation patterns were independent or dependent on the physical activity process. In simulation scenario 1, the number of observed bouts (M_i) were generated independent of X_i and Z_i . Specifically, the number of observed MVPA bouts were generated from a Poisson distribution with mean 6 ($\delta_1 = \delta_2 = 6$). In simulation scenario 2, the number of observed bouts (M_i) were generated differentially by X_i and Z_i . Specifically, the number of observed MVPA bouts were generated (i) from a Poisson distribution with mean 6, when $X_i = 1$ and $Z_i \leq 1$ and (ii) from a Poisson distribution with mean less than 6, when $X_i = 0$ or $Z_i > 1$. In order to assess the bias and coverage probability under increasing amount of information from X_i and Z_i , we varied this parameter from $\delta_2 = 1, \dots, 5$. The observation times between the MVPA bouts were generated from an exponential distribution with rate equal to $(\delta_2 + 1)/10\,080$ when $X_i = 0$ or $Z_i > 1$. When $X_i = 1$ and $Z_i \leq 1$, the observation times between the MVPA bout were generated from exponential distribution with rate equal to $7/10\,080$. This generates observation patterns dependent on the observed predictor (X_i), as well as on an unobserved (Z_i) subject-specific variable

	Observation pattern			Semiparametric			GEE		
	Type	δ_1	δ_2	$\bar{\hat{\beta}}$	Bias	Coverage	$\bar{\hat{\beta}}$	Bias	Coverage
$n = 50$	Independent (Scenario 1)	6	6	-0.4154	-0.0154	0.9880	-0.4268	-0.0268	0.9260
$n = 100$				-0.4036	0.0003	0.9890	-0.4196	-0.0178	0.9360
$n = 200$				-0.3997	0.0003	0.9870	-0.4178	-0.0178	0.9400
$n = 50$	Informative (Scenario 2)	6	5	-0.4133	-0.0133	0.9920	-0.4289	-0.0289	0.8970
$n = 100$				-0.4009	-0.0009	0.9920	-0.4241	-0.0241	0.9370
$n = 200$				-0.4060	-0.0060	0.9800	-0.4226	-0.0226	0.9370
$n = 50$	Informative (Scenario 2)	6	4	-0.4173	-0.0173	0.9850	-0.4385	-0.0385	0.9080
$n = 100$				-0.4068	-0.0068	0.9870	-0.4271	-0.0271	0.9330
$n = 200$				-0.4072	-0.0072	0.9850	-0.4300	-0.0300	0.9290
$n = 50$	Informative (Scenario 2)	6	3	-0.4110	-0.0110	0.9810	-0.4457	-0.0457	0.9170
$n = 100$				-0.4037	-0.0037	0.9810	-0.4370	-0.0370	0.9230
$n = 200$				-0.4074	-0.0074	0.9740	-0.4378	-0.0378	0.9180
$n = 50$	Informative (Scenario 2)	6	2	-0.3979	0.0021	0.9830	-0.4597	-0.0597	0.8950
$n = 100$				-0.3981	0.0019	0.984	-0.4496	-0.0496	0.9140
$n = 200$				-0.4073	-0.0073	0.983	-0.4551	-0.0551	0.8880
$n = 50$	Informative (Scenario 2)	6	1	-0.3746	0.0254	0.9670	-0.4827	-0.0827	0.8580
$n = 100$				-0.3881	0.0119	0.9760	-0.4795	-0.0795	0.8630
$n = 200$				-0.3904	0.0096	0.9800	-0.4809	-0.0809	0.7950

$\delta_2 = 1, \dots, 5$ with smaller δ_2 imposing stronger information through the fixed covariate X and latent variable Z). The coverage probability reached 0.795 for GEE when the association between the wear pattern and the physical activity process was the strongest, meaning that only 79.5% of the confidence intervals from the 1000 simulated data sets contained the true coefficient value. In contrast, our proposed semiparametric approach showed coverage of at least 95% regardless of the degree of association between the wear pattern and the physical activity process. The coverage probability in the semiparametric approach

Fig. 3. Results from a simulation study to evaluate bias and coverage probability in estimating associations between a fixed predictor and levels of physical activity, measured using wearable sensors. Average bias across 1000 simulated data sets is presented for simulation scenarios when the observation patterns were independent or dependent on the physical activity process. The bias and coverage estimates are from simulation scenario 1 when $\delta_2 = 6$, and from simulation scenario 2 when $\delta_2 = 1, \dots, 5$.

exceeded the nominal level of 0.95 due to the use of sandwich variance estimator. Although the sandwich variance estimator is known to provide a consistent estimate of the covariance matrix for the parameter estimates, it is known to be more variable than the standard parametric variance estimate (Kauermann and Carroll, 2001).

3.2. Real data analysis

Data for 4518 participants in NHANES 2003–2004 who were at least 18 years of age and have participated in the physical activity monitoring component were included for analysis. Age of the participants in this subset ranged from 18 to 85 years old, with an average of 47.09 (SD = 20.57). Male participants consisted of 47.9% of the sample. Over the 7 days period, average cumulative minutes of MVPA was 34.12 min (SD = 75.17). Males had an average of 44.11 min MVPA throughout the week monitored using wearable sensors compared to 24.93 min in females.

Two different models were fitted [Wang and others, 2013, semiparametric approach and generalized Poisson estimating equations with exchangeable working correlation matrix (GEE)] to estimate associations between gender and levels of MVPA. The outcome of interest were the minutes of MVPA in 10 min bouts. The minute-level wearable sensor data from NHANES were processed using widely used data procedures (defined in Appendix S1 of supplementary material available at *Biostatistics* online) to calculate minutes of MVPA in 10 min bouts. The estimates from the semiparametric and GEE model suggest that females had significantly lower engagement in MVPA, compared to males (Table 2). Specifically, males had higher minutes of MVPA per day by 22.4% ($1/\exp(-0.2018)$) and 32.7% ($1/\exp(-0.2830)$), respectively, for the semiparametric and GEE models. The larger absolute coefficient value from the GEE, compared to the semiparametric model suggests a possible bias against the null, which coincides with what we have found in our simulation study. The R code for data analysis is available on GitHub along with a sample data set: <https://github.com/github-js/semiparametric>.

Table 2. *A semiparametric model and the generalized Poisson estimating equations were fitted with exchangeable working correlation matrix for within-subject observations, as described in the methods section. Standard errors (SE) of the parameter estimates are obtained by sandwich variance estimation. Data for 4518 participants in NHANES 2003–2004, who were at least 18 years of age and have participated in the physical activity monitoring component were considered for analyses. The outcome of interest was the minutes of MVPA per day. The minute-level wearable sensor data from NHANES 2003–2004 were processed using widely used data procedures (defined in Appendix S1 of [supplementary material](#) available at [Biostatistics online](#)) to calculate minutes of MVPA in 10 min bouts. The predictor of interest was gender (reference category set as male)*

	Estimate	Exp (estimate)	SE	Z	P
Semiparametric female	−0.2018	0.8172	0.0311	−6.49	< 0.001
GEE female	−0.2830	0.7535	0.0426	−6.65	< 0.001

4. DISCUSSION

In this study, we outlined some of the key statistical challenges for analyzing accelerometer data collected under real life and observational study conditions. The within- and between-person variability in the device wear times, and its potential associations to the physical activity behavior can result in information bias. In addition, removing individuals with insufficient wear days from analysis can give rise to selection bias. Based on this understanding, we introduced a semiparametric approach proposed by Wang and colleagues, originally developed for modeling recurrent medical events ([Wang and others, 2013](#)), to model accelerometer data with irregular wear time. This method properly addresses the information bias arising from the associations between the device wear patterns and the underlying physical activity process. Our simulation study indicated that estimates from the GEE showed ignorable bias when device wear patterns were independent of the participants physical activity process, but incrementally more biased when the patterns of device non-wear times were increasingly associated with the physical activity process. The estimates from the GEE models were biased against the null. This is due to the fact that the GEE model is a marginal model ([Liang and Zeger, 1986](#)) and the semiparametric approach introduced a random effect to account for the dependence between the physical activity process and the observation process. In physical activity monitoring, it has been discussed that there may be multiple unknown factors that simultaneously contribute to the physical activity process and the monitor wear pattern ([Troiano and others, 2014](#); [Bai and others, 2014](#); [Mâsse and others, 2005](#)). Our simulation study was designed to mimic such real data, so that a fixed known covariate X and a latent variable Z were related to the physical activity and the observation process. Therefore, it is understandable that the GEE, when not accounting for such latent effect would result in fixed effect estimates that are biased against the null. In addition, the small magnitude of bias from the GEE under intermittent observation patterns (even with non-informative observation and censoring times) have been noted previously ([Toledano and Gatsonis, 1999](#)). The magnitude of bias in the GEE model was between -0.02 and -0.08 when the observation and censoring times were informative. This means that the association between the fixed predictor X and the minutes of MVPA would have been estimated between -0.42 and -0.48 , when the true effect was -0.4 . Therefore, using GEE would result in estimating excess risk of 3% ($1/\exp(-0.42) - 1/\exp(0.4) = 0.03$) to 12% ($1/\exp(-0.48) - 1/\exp(0.4) = 0.12$) due to the factor X . Such differences in the estimated risk can be important in making public health decisions, such as determining which factors to be discussed in

public health messages for increasing physical activity. Unbiased estimates can be also important when the objective is to derive effect size estimates, for designing an intervention studies for increasing physical activity. The estimates from our proposed semiparametric modeling approach were unbiased both when the device wear patterns were (i) independent or (ii) dependent to the underlying physical activity process. Application of this method was demonstrated using data from the NHANES 2003–2004. Findings from the analyses agree with existing study, which also reported higher minutes of MVPA in males compared to females in 2003–2004 NHANES data (Troiano and others, 2008). In summary, we have demonstrated that a semiparametric modeling approach can be used to estimate unbiased associations between a fixed covariate and wearable sensor measured physical activity, especially in the presence of informative wear time and censoring. The methods can be generalized to the data collected using the state-of-the-art physical activity monitors (e.g., tri-axial accelerometers (GT3X+; ActiGraph, LLC, Pensacola, FL, USA) used in the NHANES 2013–2014 cycle).

5. SOFTWARE

We note that we have developed an R package that can implement the semiparametric modeling approach presented in this article. The R package `acc` is comprehensive software application for reading, processing, simulating, visualizing, and analyzing accelerometer data, publicly available at the Comprehensive R Archive Network (Song and Cox, 2015). We illustrate the function to fit the semiparametric model using the R package in Appendix S2 of [supplementary material](#) available at *Biostatistics* online.

6. SUPPLEMENTARY MATERIAL

[Supplementary material](http://biostatistics.oxfordjournals.org) is available online at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

National Institutes of Health/National Cancer Institute through R01 CA 109919, R25T CA057730, R25E CA056452, P30 CA016672 (M. D. Anderson's Cancer Center Support Grant and PROSPR Shared Resource), and the Center for Energy Balance in Cancer Prevention and Survivorship, Duncan Family Institute for Cancer Prevention and Risk Assessment.

REFERENCES

- BAI, J., HE, B., SHOU, H., ZIPUNNIKOV, V., GLASS, T. A. AND CRAINICEANU, C. M. (2014). Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics* **15**, 102–116.
- CHOI, L., LIU, Z., MATTHEWS, C. E. AND BUCHOWSKI, M. S. (2011). Validation of accelerometer wear and nonwear time classification algorithm. *Medicine & Science in Sports & Exercise* **43**, 357–364.
- ELASHOFF, M. AND RYAN, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13**, 48–65.
- EVENSON, K. R., BUCHNER, D. M. AND MORLAND, K. B. (2012). Objective measurement of physical activity and sedentary behavior among US adults aged 60 years or older. *Preventing Chronic Disease* **9**, E26.
- EVENSON, K. R., GOTO, M. M. AND FURBERG, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity* **12**, 159.

- FREEDSON, P., MELANSON, E. AND SIRARD, J. (1998). Calibration of the Computer Sciences and Applications, Inc. accelerometer. *Medicine & Science in Sports & Exercise* **30**, 777–781.
- HALL, K. S., HOWE, C. A., RANA, S. R., MARTIN, C. L. AND MOREY, M. C. (2013). METs and accelerometry of walking in older adults: standard versus measured energy cost. *Medicine & Science in Sports & Exercise* **45**, 574–582.
- HUANG, C. Y., WANG, M. C. AND ZHANG, Y. (2006). Analysing panel count data with informative observation times. *Biometrika* **93**, 763–775.
- KAUERMANN, G. AND CARROLL, R. J. (2001). A note on the efficiency of sandwich covariance estimation. *Journal of the American Statistical Association* **96**, 1387–1396.
- LIANG, K.-Y. AND ZEGER, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
- MAILEY, E. L., GOTHE, N. P., WOJCICKI, T. R., SZABO, A. N., OLSON, E. A., MULLEN, S. P., FANNING, J. T., MOTL, R. W. AND MCAULEY, E. (2014). Influence of allowable interruption period on estimates of accelerometer wear time and sedentary time in older adults. *Journal of Aging and Physical Activity* **22**, 255–260.
- MÄSSE, L. C., FUEMMELER, B. F., ANDERSON, C. B., MATTHEWS, C. E., TROST, S. G., CATELLIER, D. J. AND TREUTH, M. (2005). Accelerometer data reduction: a comparison of four reduction algorithms on select outcome variables. *Medicine & Science in Sports & Exercise* **37**, S544–554.
- PERRY, M. A., HENDRICK, P. A., HALE, L., BAXTER, G. D., MILOSAVLJEVIC, S., DEAN, S. G., McDONOUGH, S. M. AND HURLEY, D. A. (2010). Utility of the RT3 triaxial accelerometer in free living: an investigation of adherence and data loss. *Applied Ergonomics* **41**, 469–476.
- ROBERTSON, W., STEWART-BROWN, S., WILCOCK, E., OLDFIELD, M. AND THOROGOOD, M. (2011). Utility of accelerometers to measure physical activity in children attending an obesity treatment intervention. *Journal of Obesity* **2011**, Article ID 398918, 8 pages.
- SONG, J. AND COX, M. G. (2015). acc: an r package to process accelerometer data. <http://cran.r-project.org/web/packages/acc/>.
- TOLEDANO, A. Y. AND GATSONIS, C. (1999). Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* **55**, 488–496.
- TROIANO, R. P. (2006). Translating accelerometer counts into energy expenditure: advancing the quest. *Journal of Applied Physiology* **100**, 1107–1108.
- TROIANO, R. P., BERRIGAN, D., DODD, K. W., MASSE, L. C., TILERT, T. AND McDOWELL, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine & Science in Sports & Exercise* **40**, 181–188.
- TROIANO, R. P., MCCLAIN, J. J., BRYCHTA, R. J. AND CHEN, K. Y. (2014). Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine* **48**, 1019–1023.
- WANG, X., MA, S. AND YAN, J. (2013). Augmented estimating equations for semiparametric panel count regression with informative observation times and censoring time. *Statistica Sinica* **23**, 359–381.
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

[Received December 22, 2016; revised December 4, 2017; accepted for publication December 10, 2017]