2023

# Statistical Learning Methods to Identify Nonwear Periods From Accelerometer Data

Sahrej Randhawa

Manoj Sharma

Madalina Fiterau

Jorge A. Banda

Farish Haydel

*See next page for additional authors*

## Authors

Sahrej Randhawa, Manoj Sharma, Madalina Fiterau, Jorge A. Banda, Farish Haydel, Kristopher Kapphahn, Donna Matheson, Hyatt Moore IV, Robyn L Ball, Clete Kushida, Scott Delp, Dennis P. Wall, Thomas Robinson, and Manisha Desai

*Journal for the Measurement of Physical Behaviour,* 2023, 6, 124-133
https://doi.org/10.1123/jmpb.2022-0034

Human Kinetics
ORIGINAL RESEARCH

# Statistical Learning Methods to Identify Nonwear Periods From Accelerometer Data

**Sahej Randhawa,[1] Manoj Sharma,[2] Madalina Fiterau,[3] Jorge A. Banda,[4] Farish Haydel,[5] Kristopher Kapphahn,[6] Donna Matheson,[5] Hyatt Moore IV,[6] Robyn L. Ball,[7] Clete Kushida,[8] Scott Delp,[9] Dennis P. Wall,[10] Thomas Robinson,[5] and Manisha Desai[6]**

[1]Department of Orthopedic Surgery, University of California at Davis, CA, USA; [2]GRAIL Inc., Menlo Park, CA, USA; [3]Manning College of Information & Computer Sciences, University of Massachusetts, MA, USA; [4]College of Health and Human Sciences, Department of Public Health, Purdue University, IN, USA; [5]Stanford Solutions Science Lab, Departments of Pediatrics and Medicine, Stanford University, CA, USA; [6]Quantitative Sciences Unit, Department of Medicine, Stanford University, CA, USA; [7]The Jackson Laboratory, Bar Harbor, ME, USA; [8]Sleep Medicine Division, Department of Psychiatry and Behavioral Sciences, Stanford University, CA, USA; [9]Departments of Bioengineering and Mechanical Engineering, Stanford University, CA, USA; [10]Division of Systems Medicine, Departments of Pediatrics and Medicine, Stanford University, CA, USA

***Background:*** Accelerometers are used to objectively measure movement in free-living individuals. Distinguishing nonwear from sleep and sedentary behavior is important to derive accurate measures of physical activity, sedentary behavior, and sleep. We applied statistical learning approaches to examine their promise in detecting nonwear time and compared the results with commonly used wear time (WT) algorithms. ***Methods:*** Fifteen children, aged 4–17, wore an ActiGraph wGT3X-BT monitor on their hip during overnight polysomnography. We applied Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) to classify states of nonwear and wear in triaxial acceleration data. Performance of methods was compared with WT algorithms across two conditions with differing amounts of consecutive nonwear. Clinical scoring of polysomnography served as the gold standard. ***Results:*** When the length of nonwear was less than or equal to WT algorithms' predefined thresholds for consecutive nonwear time, GMM methods yielded improved classification error, specificity, positive predictive value, and negative predictive value over commonly used algorithms. HMM was superior to one algorithm for sensitivity and negative predictive value. When the length of nonwear was longer, results were mixed, with the commonly used algorithms performing better on some parameters but GMM with the greatest specificity. However, all approached the upper limits of performance for almost all metrics. ***Conclusions:*** GMM and HMM demonstrated robust, consistently strong performance across multiple conditions, surpassing or remaining competitive with commonly used WT algorithms which had marked inaccuracy when nonwear time periods were shorter. Of the two statistical learning algorithms, GMM was superior to HMM.

***Keywords:*** polysomnography (PSG), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), activity, classification

Physical activity, sedentary behavior, and sleep behavioral patterns are associated with multiple clinical outcomes, including obesity, diabetes, and cardiovascular disease (https://www.nhlbi.nih.gov/health/health-topics/topics/phys/benefits). Accurate measurement is essential to understanding the determinants of these behaviors and their relationships to health and disease, public health surveillance of these behaviors, and designing and evaluating interventions to change them. Accelerometers have long been used to measure physical activity, sedentary behavior and sleep, and have been widely adopted as a device-based means to measure movement in free-living humans (Sadeh, 2011; Troiano, 2007; Trost, 2007). Accelerometer measures have great potential to provide more accurate assessments of movement and sleep than self-reported measures, which are prone to errors in recall (Trost, 2001).

Modern accelerometers continue to record data even when users are not wearing the monitor, and data recorded during nonwear periods can appear similar to data captured when participants are wearing the monitor during sleep or sedentary periods. Thus, distinguishing nonwear periods from those of sleep and sedentary behavior is challenging but essential to accurately

Sharma https://orcid.org/0000-0002-3393-4825
Fiterau https://orcid.org/0000-0003-4179-7274
Banda https://orcid.org/0000-0002-5472-7526
Kapphahn https://orcid.org/0000-0002-5017-8697
Matheson https://orcid.org/0000-0001-6279-6786
Moore IV https://orcid.org/0000-0001-7866-2280
Ball https://orcid.org/0000-0002-7335-3339
Kushida https://orcid.org/0000-0002-9430-3752
Delp https://orcid.org/0000-0002-9643-7551
Wall https://orcid.org/0000-0002-7889-9146
Robinson https://orcid.org/0000-0002-2367-0774
Desai (manishad@stanford.edu) is corresponding author, https://orcid.org/0000-0002-6949-2651

characterize physical activity, sedentary behavior, and sleep. Misclassification of sleep or sedentary periods as nonwear time (NWT) or of NWT as sleep or sedentary periods can create substantial and misleading errors in estimates of physical activity levels, sedentary behavior, and sleep. Importantly, because NWT can mimic sleep or sedentary behavior and vice versa, misclassification of NWT is a particular threat to the validity of the estimates of sleep and sedentary behavior, which are often key quantities of interest.

In practice, several algorithms are commonly applied to distinguish between periods of wear time (WT) and NWT (Choi et al., 2011, 2012; Troiano et al., 2008). These WT algorithms are typically based on the number of epochs—defined periods of time—with zero count values for acceleration (Banda et al., 2016; Cain et al., 2013). WT algorithms are attractive for their simplicity, and while they provide good accuracy in specific settings (Banda et al., 2016; Cain et al., 2013), they may lack generalizability; specifically, they are sensitive to intra- and inter-individual variations and to the precise placement of the sensors, resulting in suboptimal classification accuracy. Furthermore, the most commonly used WT algorithms typically do not make use of all available data from contemporary multiaxial accelerometers; they utilize only uniaxial count and/or summarized measurements such as vector of magnitude (VM)—the square root of the sum of the "counts" (derived from the accelerations in multiple axes) squared for each axis. In addition, they are often applied to varying epoch lengths, resulting in estimates that are not comparable across studies, potentially leading to disparate results, interpretations, conclusions, and/or misleading findings (Banda et al., 2016).

To address some of these challenges, we investigated the potential of statistical learning methods to improve discrimination between WT and NWT periods, particularly nonwear periods where the device would have signal that is similar to signal recorded when sedentary or sleeping (e.g., device is on a dresser) in contrast to when a device is not worn and mobilized (e.g., being thrown up and down or in a moving vehicle). More specifically, we applied statistical learning methods to triaxial accelerometer data to develop and evaluate a new approach that distinguishes NW periods from those of sleep and sedentary behavior. We are not the first to consider statistical learning methods for application to accelerometer data. For example, Gaussian Mixture Models (GMMs), *k*-Nearest Neighbors (*k*-NN), and Hidden Markov Models (HMMs) have all been applied in accelerometer studies for activity classification (Mannini & Sabatini, 2010). However, none of these approaches have been applied and evaluated for NW detection. For example, HMMs, which assume the accelerometer data are generated according to a Markov process and that activities are unobserved, "hidden" states, have been used to classify an individual's type of activity based upon uniaxial acceleration data and VM for biometric gait recognition (Nickel & Busch, 2013). In another study, HMMs were applied to accelerometry data obtained using a smartphone application available for Google Android devices for action and activity recognition (Lee & Cho, 2011). The underlying assumption of Markovian dependence is particularly attractive in this setting, as people tend to remain in the same state (WT or NWT) for an extended period (Rabiner, 1989). GMMs, which assume that an observed sequence belongs to a weighted sum of multiple Gaussian distributions, have been adaptively employed for classification of three postures (sitting, standing, and lying) and five movements (sit-to-stand, stand-to-sit, lie-to-stand, stand-to-lie, and walking) in a home-based multiple days study with a limited number of subjects (Allen et al., 2006; Reynolds et al., 2000). Given that we anticipate movement belongs to one of numerous states, GMMs are a natural choice for classification of these data.

Importantly, in contrast to the uniaxial or summarized vector WT algorithms, which do not utilize the covariance structure of the triaxial accelerations, we were motivated to incorporate data from all three axes in our development of statistical learning approaches, as their covariance may contribute relevant information about movement, potentially increasing classification accuracy. In addition, we considered nonprocessed triaxial acceleration data (measured in gravity units), as opposed to processed counts (a unit distinct to the manufacturer), to increase algorithmic transparency and allow others to fully understand and possibly even enhance our approach. In addition, it allows for translation across different device brands.

To pursue these aims that especially relate to distinguishing nonwear signal from signal recorded during sleep or sedentary behavior, we used accelerometer data collected during clinical sleep studies with overnight polysomnography (PSG) in 15 children. We evaluated the performance of HMM and GMM approaches with metrics such as classification error, sensitivity, specificity, and positive (PPV) and negative predictive values (NPVs) for identifying NWT.

# Methods

In a study of 15 children, we compared performance of HMM and GMM results against two commonly used WT algorithms under two conditions: (a) with 1-hr periods of NW on each side of a WT period and (b) with 5-hr NW periods flanking the WT period (See Figure 1 below), where the WT period takes place before, after, and during a sleep study. The condition with longer NW intervals was used to highlight the WT algorithms' "best-case" performance, while the shorter period mimicked a situation expected to be more common and difficult to detect, where users remove their accelerometers for periods of activities that are shorter than 5 hr in length (e.g., swimming, showering, sports, etc., where the accelerometer may be removed for protection, safety, or convenience).

Eligible participants were children 2–17 years of age when completing an overnight PSG at our institution's Sleep Medicine Center. Children were ineligible if they (a) were participating in a treatment sleep assessment; (b) had a condition interfering with normal sleep body movements or wearing the monitors; (c) were deemed inappropriate for study participation in the opinion of their sleep center clinician; or (d) they or their parents were unable to read, understand, or complete informed consent or assent (for children ≥7 years) in English or Spanish. Our institution's Administrative Panel on Human Subjects in Medical Research approved the study.

## Data Collection

The ActiGraph wGT3X-BT monitor (ActiGraph) was used to measure movement at the hip during an overnight sleep study for approximately 12 hr. Although accelerometers are often worn on the wrist for sleep research, we used the hip placement as is typical of studies of movement in children over full 24-hr days. The monitor measures acceleration in three individual axes, has a dynamic range of ±8 units of gravity, and was set to record at a frequency of 40 Hz, as this frequency was recommended to enable a battery life of 1 week of continuous use in field studies. ActiLife (version 6.10.2) was used to download these data from the monitor (ActiGraph GT3X+ and wGT3X+ Device Manual, n.d.). Participants wore the

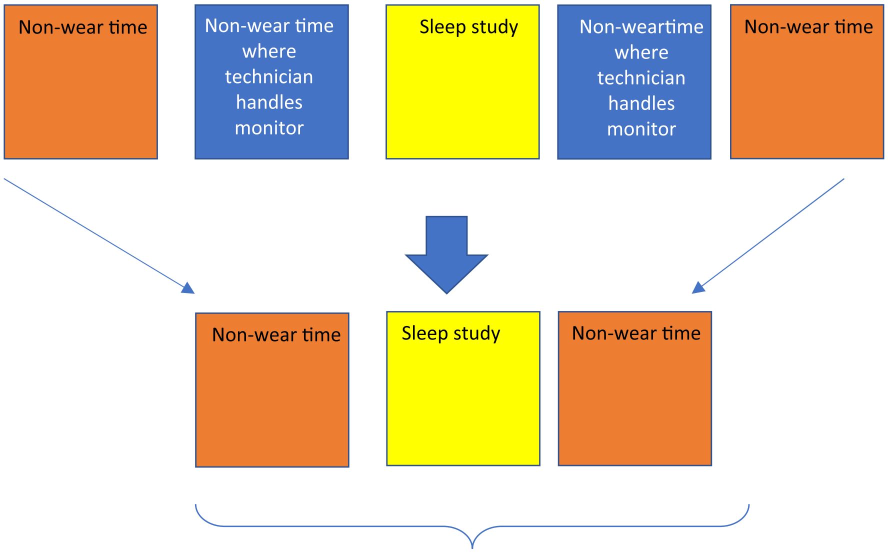
Non-wear time
Non-wear time where technician handles monitor
Sleep study
Non-weartime where technician handles monitor
Non-wear time
Non-wear time
Sleep study
Non-wear time

**Figure 1** — Data included in our study to develop statistical learning procedures.

monitor on a belt on their right hip, placed by registered sleep technologists prior to attaching PSG recording electrodes and removed at the completion of the sleep study, prior to electrode detachment. Study staff documented the time each monitor was placed on and removed from participants. Because the 30-min periods prior to placement and after removal of the monitor included periods when technologists transported and handled the monitors, these periods were not considered nonwear for purposes of this study because they do not capture typical nonwear signals, and thus excluded from the analysis data (see Figure 1). From the remaining data, we constructed two data sets to represent two scenarios of shorter and longer periods of NWT: the first consisting of 1-hr periods of NW time before and after the WT and the second consisting of 5-hr periods of NW time before and after WT.

Sleep technologists scored PSG data using clinical software (Sandman Elite™ Sleep Diagnostic Software, Covidien) and following American Academy of Sleep Medicine 2015 guidelines, categorizing 30-s intervals as nonrapid eye movement sleep (Stages N1, N2, and N3), rapid eye movement sleep and wake, and scoring other sleep and respiratory events (Berry, 2015 and Covidien, n.d.). This served as the ground truth for our study. Participants were awake and monitored for periods immediately after donning the PSG sleep equipment and again at the end of the sleep study, just prior to removing the sleep equipment, providing periods of both wake and sleep during the wear periods.

## Statistical Analyses

Let $Y_t$ represent the triaxial acceleration data around three axes ($x$-axis, $y$-axis, and $z$-axis) at time $t$, and let $s_t$ be the categorical variable representing the three activity states (nonwear, sleep, and wake) at time $t$. Below, we describe the application of HMM and GMM to process and analyze the triaxial acceleration data ($Y_t$) for classification of activity states ($s_t$) from the *Sleep Study*, where NW and wear states are observed and known.

### *Activity Classification Algorithms*

**Feature Extraction.** We considered features that could be derived universally using raw data. We extracted such features from nonoverlapping windows of 30-s intervals for classification for the HMMs and GMMs. Windows of 30-s intervals were utilized as they coincided with the frequency of the recordings of true states, and the median vector $m$, and covariance matrix $S$ were computed for each window. Derivations of these features were considered, such as the determinant of variance–covariance matrices, which reduces the multidimensional matrix to a scalar, and log-transformations to reduce the large variation observed in the signal. The following features were therefore derived for each window and considered in the activity classification algorithm:

- $m$: The median vector of data corresponding to each of the three axes ($x$, $y$, and $z$) recorded at 40 Hz
- $L_2$-norm = $\sqrt{x^2 + y^2 + z^2}$: The VM
- Logarithm of determinant of covariance matrix (logDetS): Log transformation of the determinant of the sample covariance matrix
- $\theta_x = \cos^{-1}(x)/L_2$-norm: Description of tilt using data from $x$-axis
- $\theta_y = \cos^{-1}(y)/L_2$-norm: Description of tilt using data from $y$-axis
- $\theta_z = \cos^{-1}(z)/L_2$-norm: Description of tilt using data from $z$-axis
- logDetT: The logarithm of the determinant of $T$, the variance covariance matrix of ($\theta_x$, $\theta_y$, and $\theta_z$)
- mSm-norm1 = $m \times S \times m'$
- mSm-norm2 = $m \times S^{-1} \times m'$

Figure 2 depicts the primary features (centroid represented as median vector of three axes and logarithm of determinant of covariance matrix for 30-s interval along with activity state) for one subject. Strong correlation among the features is observed with considerably different patterns corresponding to each of the three states.

**Hidden Markov Model.** Using the *mhsmm* package in R (O'Connell & Højsgaard, 2011; R Core Team, 2015), we relied on multiple sequence training to develop the HMM, using a leave-one-out approach, where the model was trained on training

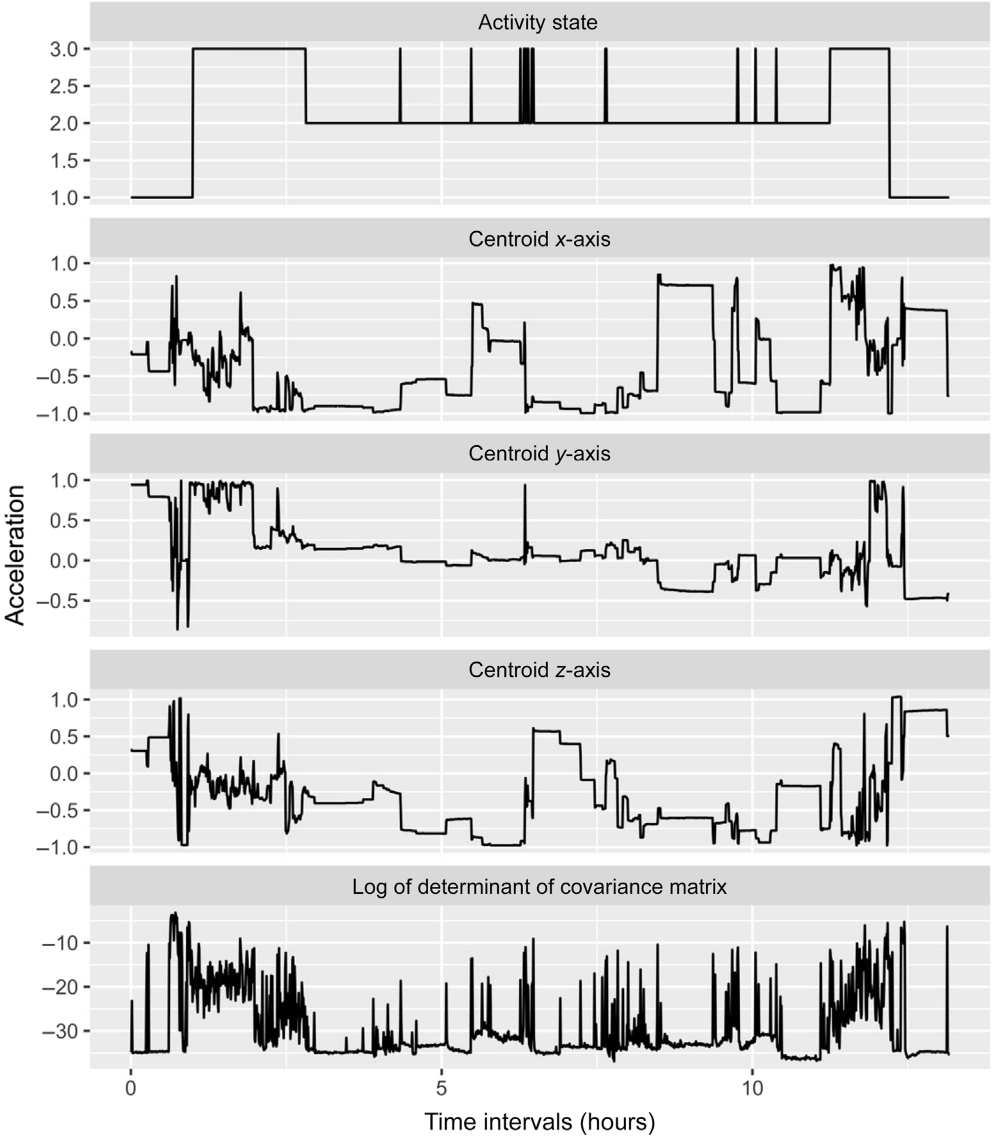

**Figure 2** — Activity states, triaxial acceleration data centroids, and logarithm of determinant of covariance matrix for one participant for a series of 30-s intervals. (Centroids for three axes and the determinant of the sample covariance matrix represent acceleration data in gravity units.)

sequences that remained after leaving out one of the 15 individual sequences for testing. Multiple sequence training enabled an alternative to providing only one training sequence derived as 14 concatenated sequences, which would provide an artificially incorrect sequence. We trained on all possible iterations among the 15 choose 14 possible ways to construct a training data set, and for each iteration, we tested on the left out individual. The initial state of the HMM model was set as the first known (true labeled) state. Transition probabilities were estimated from the training data along with parameters for conditional Gaussian densities for each state at time *t*, given the observed data. Evaluation of model performance involved testing on the sequences of the participants not used in the training set, predicting states for each time interval *t* of a given participant's sequence. Our publicly available code demonstrates how to replicate this approach for any giving training size.

Alternate training strategies were also considered but yielded poorer performance. Chief among these, we considered training and testing within a given participant's sequence, which required partitioning the first x% to form the training set and the remaining intervals into the testing set for all participants. Other training strategies included randomly concatenating the data from all 15 participants before training and testing on a partitioned version of this larger data set.

Prior to fitting the HMM models, we applied smoothing to the reference labels using the *smooth.discrete* function from *mhsmm* package in R (O'Connell & Højsgaard, 2011; R Core Team, 2015). Similarly, we smoothed the predicted activity states post prediction and before assessing the performance of these models described below in subsection "Performance Evaluation Metrics".

Through iteratively comparing the performance of the HMM on various sets of the aforementioned extracted features, the optimal set of features for this model consisted of each of the three axes of the acceleration vector *m* as well as the logDetS feature.

**Gaussian Mixture Model.** Data from all 15 participants were combined to estimate the Gaussian mixture densities for different clusters based upon the triaxial data features for 30-s intervals (as in the HMM) using the *mclust* package in R (R Core Team, 2015; Fraley & Raftery, 2007). From this combined pool, the training set was constructed by randomly sampling 80% of the time intervals labeled as WT and randomly sampling the same number of intervals from the time intervals labeled as NW. The remaining data were used to create the test set. From the training set, the maximum likelihood estimates were derived for weights, mean vectors, and variance–covariance matrices for nine Gaussian distributions assumed. These estimates were then applied to the observed sequence data to determine the posterior probability for generating the resulting state sequence. Each cluster was then mapped to one of two activity states (NW or wear) based upon the modal frequency observed for a particular state in each cluster. For example, if Cluster one resulted in mostly wear (sleep or wake), we assigned this cluster to wear. Thus, the GMM's assignment of a given interval to a cluster could be translated to an assignment to either wear or NW. We compared the estimated state sequence to the true state sequence labels. Other attempted training strategies included training the model on a random sampling of 80% of the pooled data without balancing the amount of wear and nonwear, since the total data set has an uneven amount of wear and nonwear in it and such a method would obtain a training set of an expected composition more similar to the data set. Additionally, using a method similar to training the HMM (training the GMM on the data of 14 subjects pooled together and testing on the remaining one) was also tried. However, these methods yielded inferior results.

Through comparing the performance of the GMM on various sets of the aforementioned extracted features, the optimal set of features for this model consisted of each of the three axes of the acceleration vector *m*, logDetS, the three $\theta$ features for tilt, and logDetT.

### Commonly Applied Nonwear Classification Methods Considered

We compared the performance of HMMs and GMMs with that of two commonly applied WT algorithms: Choi et al. (2011) and Troiano et al. (2008). These algorithms are based on *counts* of consecutive zero values, where a nonzero *count* refers to a proprietary unit derived by ActiGraph to represent a level of activity indicating accelerometer signal that exceeds some threshold—enough to "count" as some activity and where the number of counts is some indication of the level of intensity of activity (ActiGraph GT3X+ and wGT3X+ Device Manual, n.d.). Importantly, the count-based WT algorithms serve as the current standards for determining activity state from accelerometer data. For example, they were applied to data from the National Health and Nutrition Examination Survey and several population-based studies (Tudor-Locke et al., 2012).

**Choi et al. Algorithm.** The Choi et al. algorithm is applied to counts for 1-min epochs and classifies periods of time with consecutive zero counts of at least 90 min as NWT, with an allowance for nonzero counts lasting up to 2 min as long as the 30-min windows flanking this allowance period contain only nonzero counts (Choi et al., 2011). The algorithm was previously validated for 24-hr periods including sleep on several subjects (adults and youth) in a controlled environment and assessed on data during a 7-day study in free-living environment in older female adults (Choi et al., 2011, 2012). For the purposes of comparison, we applied this algorithm for NW detection using the VM computed from count data over three axes, as recommended by the authors.

**Troiano et al. Algorithm.** This algorithm is also based upon counts within 1-min epochs and was developed for uniaxial data. Periods with an interval of at least 60 consecutive minutes of zero activity intensity counts are classified as NW, with an allowance of 1–2 min of counts between 0 and 100 (Tudor-Locke et al., 2012). As it was developed for uniaxial data, we applied this algorithm for NW detection using data from each of the three axes independently, represented as *TroianoX*, *TroianoY*, and *TroianoZ*. Note that the original algorithm was developed and validated for the vertical axis (*TroianoY*). However, because it is possible to extend the algorithm to other axes and because people sometimes place the device in different orientations, we decided to consider its performance extended to the other two axes.

### Comparative Performance Evaluation Metrics

We used the following metrics to compare the performance of the methods, assessing their predictions on the intervals comprising test set against the true states/labels for those intervals:

- Classification error: The proportion of intervals in the test set that were misclassified by their predicted states.
- Sensitivity: The proportion of correctly predicted nonwear among all nonwear intervals in the test set.
- Specificity: The proportion of correctly predicted *wear* intervals among all wear intervals in the test set.
- PPV or precision: The proportion of predicted nonwear intervals in the test set that was truly nonwear.
- Negative predictive value: The proportion of predicted *wear* intervals in the test set that was truly wear.

All of our code has been annotated and are available online at www.github.com/qsuProjects/r-nonwear-methods.

## Results

The 15 participants had median and mean ages of 8 and 9.5 years, respectively, with a range from 4.0 to 17.5 years, and roughly half of the participants (46.7%) were female. The proportion of time spent in each of three activity states (nonwear, sleep, and wake) across all time periods for the 15 participants were 8.1%, 63.1%, and 28.7%, respectively.

For the first condition, with 1-hr periods of nonwear flanking each side of the wear interval, the GMM outperformed the commonly applied WT algorithms for four of the five metrics: classification error, specificity, PPV, and NPV (Figure 3, Table 1). The Choi et al. method exhibited the best sensitivity. HMM was superior to each axis for the Troiano method for sensitivity and NPV but did not fare well relative to them or Choi et al. otherwise. For instance, the HMM had a mean classification error of 40.2% (*SD* of 24.5%) relative to a classification error of only 13.9% from Choi et al. (*SD* of 8.7%) and to a mean classification error of 32.6% (*SD* of 10.9%) for the best-performing version of the Troiano et al. algorithm (TroianoY).

The second condition, with 5-hr periods of nonwear flanking each side of the wear interval, resulted in the performance of the HMM and GMM being comparable but often poorer performing compared with the commonly applied WT algorithms (Figure 4, Table 2). The HMM had a mean classification error of 22.6% (*SD* of 16.0%). With six iterations, the GMM results yielded a mean classification error of 16.2% (*SD* of 9.4%). The Choi et al. algorithm showed superior performance with a mean classification error of 11.2% (*SD* of 10.1%), and the best-performing version of

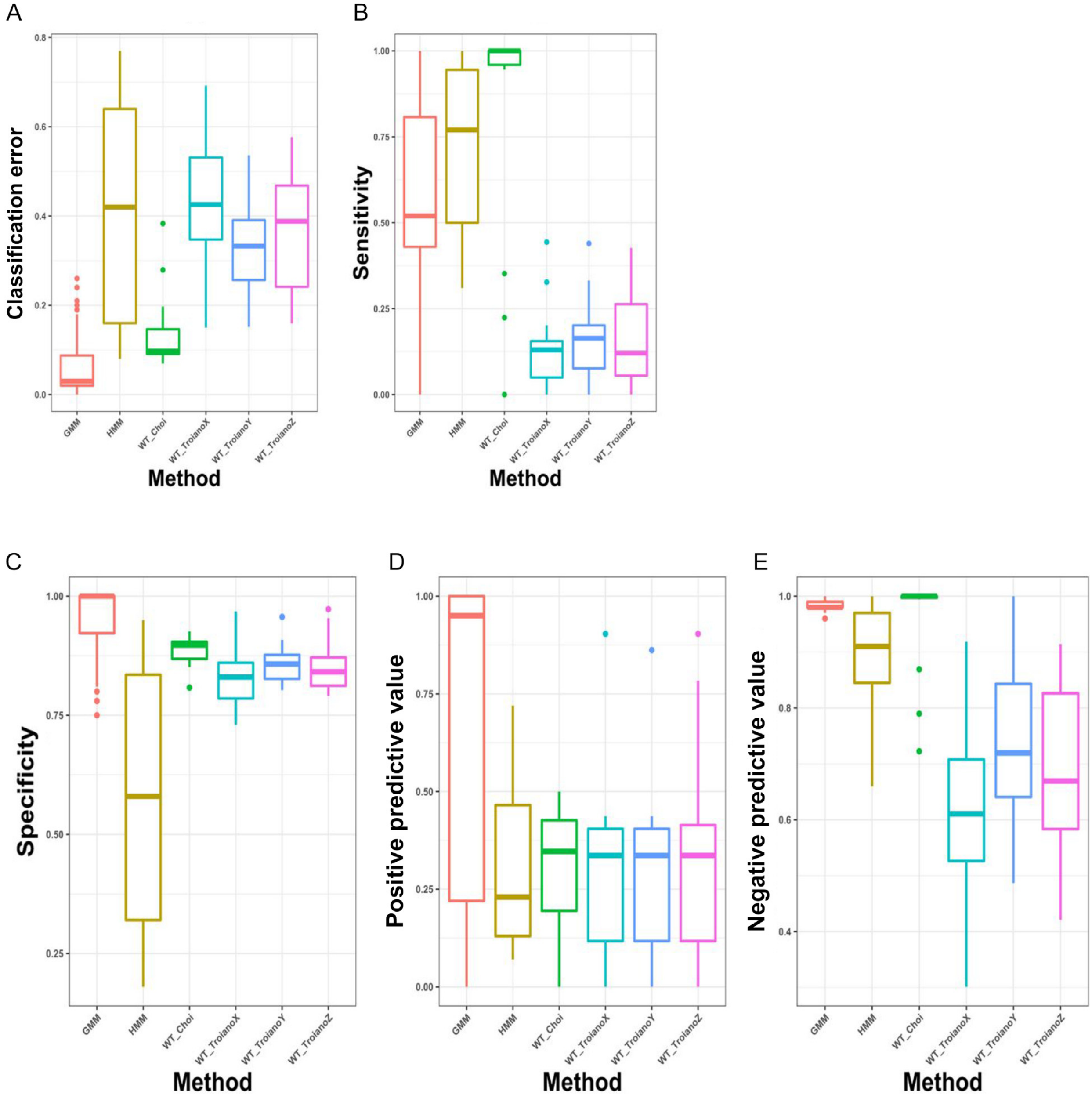

**Figure 3** — Box-and-whisker plot comparison of algorithms' performance for 1-hr condition. Individual points represent outliers at least 1.5 × (interquartile distance) less than the first quartile or greater than the third quartile.

**Table 1 Mean Values (as Decimals) and *SD*s (in Parentheses) for Each Algorithm's Results per Performance Metrics for 1-hr Condition**

| Algorithm | Classification error | Sensitivity | Specificity | Positive predictive value | Negative predictive value |
|---|---|---|---|---|---|
| Choi et al. | 0.139 (0.087) | 0.823* (0.349) | 0.887 (0.031) | 0.301 (0.178) | 0.958 (0.090) |
| *TroianoX* | 0.429 (0.148) | 0.131 (0.122) | 0.828 (0.064) | 0.288 (0.233) | 0.620 (0.172) |
| *TroianoY* | 0.326 (0.109) | 0.164* (0.115) | 0.857 (0.042) | 0.285 (0.225) | 0.742 (0.142) |
| *TroianoZ* | 0.363 (0.136) | 0.167 (0.156) | 0.852 (0.154) | 0.316 (0.266) | 0.694  0.148) |
| Hidden Markov Models | 0.402 (0.245) | 0.706 (0.244) | 0.577 (0.270) | 0.303 (0.206) | 0.896 (0.103) |
| Gaussian Mixture Models | 0.059 (0.063) | 0.577 (0.249) | 0.956 (0.067) | 0.671* (0.383) | 0.983 (0.010) |

*Data contained one missing value that was removed when calculating the mean.

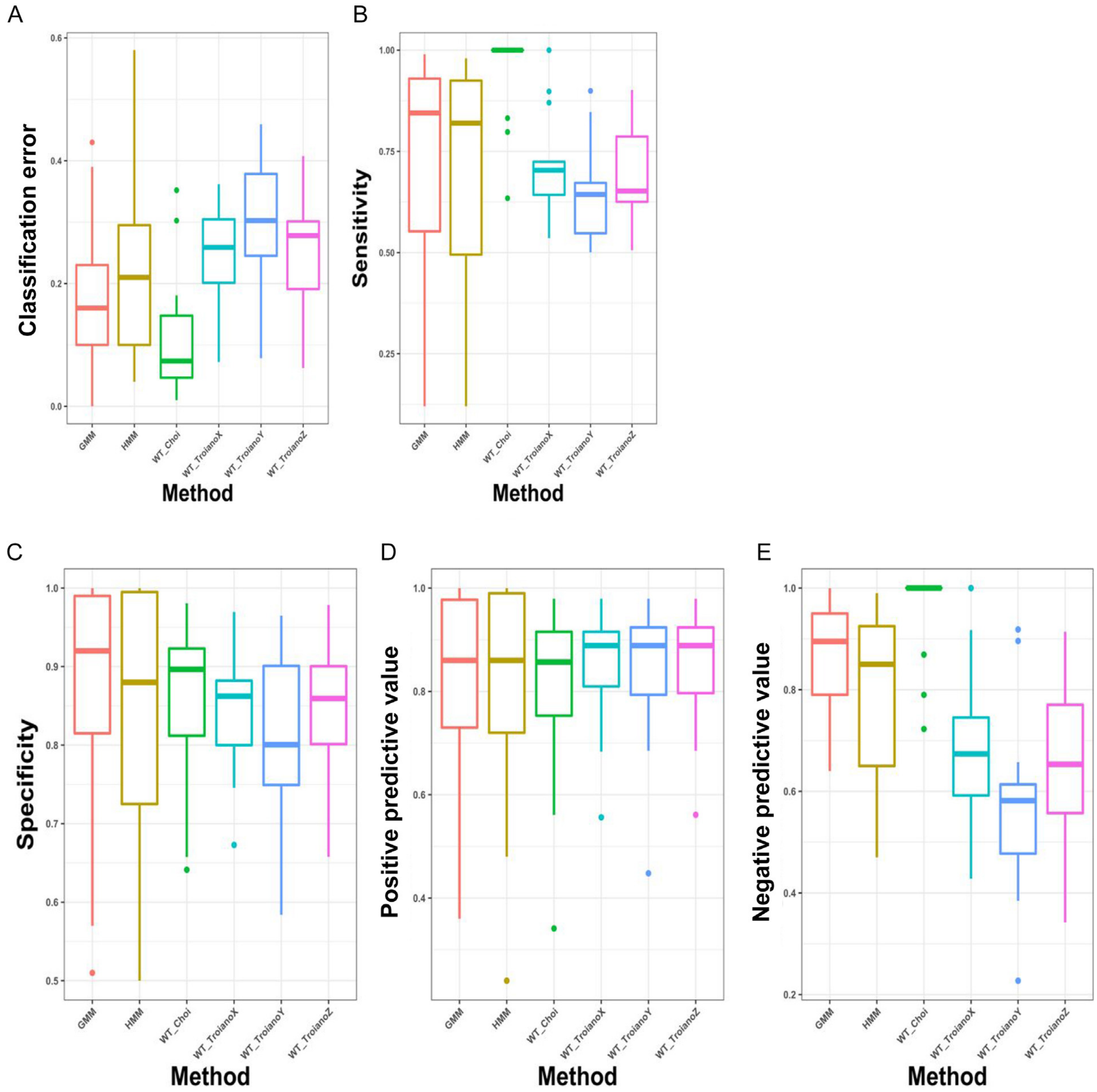

**Figure 4** — Box-and-whisker comparison of algorithms' performance for 5-hr condition.

**Table 2 Mean Values (as Decimals) and *SD*s (in Parentheses) for Each Algorithm's Results per Performance Metrics for 5-hr Condition**

| Algorithm | Classification error | Sensitivity | Specificity | Positive predictive value | Negative predictive value |
|---|---|---|---|---|---|
| Choi | 0.112 (0.101) | 0.951 (0.109) | 0.860 (0.103) | 0.806 (0.171) | 0.959 (0.090) |
| *TroianoX* | 0.242 (0.086) | 0.716 (0.122) | 0.845 (0.079) | 0.845 (0.114) | 0.682 (0.168) |
| *TroianoY* | 0.304 (0.101) | 0.643 (0.114) | 0.814 (0.106) | 0.841 (0.134) | 0.569 (0.179) |
| *TroianoZ* | 0.256 (0.097) | 0.697 (0.122) | 0.840 (0.087) | 0.848 (0.112) | 0.657 (0.174) |
| Hidden Markov Models | 0.226 (0.160) | 0.688 (0.281) | 0.849 (0.160) | 0.807 (0.225) | 0.785 (0.168) |
| Gaussian Mixture Models | 0.162 (0.094) | 0.757 (0.218) | 0.884 (0.125) | 0.820 (0.161) | 0.877 (0.093) |

Troiano et al. (*TroianoX*) had a mean classification error of 24.2% (*SD* of 8.6%). As depicted below in Figure 4 and Table 2, Choi et al. performed best in three out of five performance metrics, while the GMM performed best with regard to specificity. The GMM and HMM algorithms were able to perform better than any Troiano et al. version for classification error, specificity, and NPV, with GMM also displaying improved sensitivity over Troiano et al.

## Discussion

The performance of GMM and HMM relative to the commonly applied WT algorithms varied between the two settings considered. For the 1-hr data set, the GMM methods had more favorable properties than all the Troiano et al. versions and yielded improved classification error, specificity, PPV, and NPV over Choi et al. HMM was superior to each axis for the Troiano method for sensitivity and NPV. However, Choi et al. had a much improved sensitivity over both the GMM and HMM (i.e., Choi et al. was less likely to miss NWT). In contrast, for the 5-hr data set, the WT algorithms performed better than they did for the 1-hr data set, as anticipated, and had substantially improved performance than the GMM and HMM with much increased sensitivity and PPV, although the GMM and HMM approaches remained competitive. In general, the commonly applied WT algorithms—when confronted with shorter NW periods—are more likely to misclassify WT as NW, whereas the GMM and HMM approaches may miss NW periods and misclassify as WT. When confronted with longer NW periods, increased sensitivity is observed across the board, with the algorithms less likely to misclassify NW as WT. These comparisons demonstrate that the Choi et al. and Troiano et al. algorithms are less suitable for scenarios with shorter nonwear intervals, particularly those shorter than their dictated thresholds, which are preselected and thereby require particular, initial knowledge of how the nonwear data will be distributed. Conversely, the HMM and GMM models appear to be flexible, largely accurate, and useable for a variety of scenarios and with limited advance knowledge of the distributions of nonwear in the data. Of the two statistical learning algorithms, the GMM exhibited stronger performance for most metrics across the two conditions.

Importantly, while all approaches performed better when there were long continuous blocks of NW, the 1-hr data set may be more representative of common real-world scenarios that are most difficult to predict or classify, where the wearable technology is removed for relatively shorter periods of time for specific activities like showering, swimming, or sports as discussed by Vert and others who mention that the most common lengths of NWTs are shown to be under 60 min of length (Vert et al., 2022). While long periods of nonwear can exist if users potentially remove them during sleep at night or forget to wear them for the day, the regularity, duration, and timing of these instances may already yield key fingerprints for detection and classification.

### Study Strengths

This study's use of statistical learning methods provides algorithms of not only superior or comparable performance to commonly applied WT algorithms across a variety of scenarios but also ones that do not rely on predetermined, hard-coded thresholds and decision rules. As a result, they are likely to be more adaptable to different profiles of data and patterns. In addition, this study demonstrates that this classification task can be accomplished with a few given variables, namely acceleration in each of the three spatial axes, and without other types of data, such as ambient light or time of day. Furthermore, this study set and met a rigorous threshold for success of its classification task. Specifically, while the value of machine learning-based algorithms are often compared against simple naïve baseline algorithms, our study compared the GMM and HMM performance with currently used algorithms with more complex logic, demonstrating the GMM and HMM as not only effective but competitive and practically applicable. The data used in this study originated from a sleep study, which allowed us to evaluate our algorithms for an activity type that may arguably be the most similar to nonwear and thus most challenging. The variety of performance metrics used in this study also enabled several options for comparing algorithms and defining "superior performance." The data set itself contributes to further study on this topic through its PSG labels, allowing for potential work on classification of various sleep stages from the given data.

### Study Limitations

Our study is limited in terms of representation as it only includes 15 participants 4.0–17.5 years of age and only three activity states (nonwear, and wear consisting of sleep and wake) in a limited setting of a sleep study. However, the classification results indicate that these participants were heterogeneous in the signal measured. Importantly, our study does not attempt to draw inference based on these 15 individuals. Instead, our intent is to examine the potential of statistical learning methods to discriminate between WT and NWT particularly when signal during NWT may appear similar to that during sedentary behavior or sleep. For this purpose, we presented summary statistics such as sensitivity and specificity along with the variation in such measures across 15 individuals in box plots. In addition, while using data from the controlled environment of a sleep study is a strength, providing a rigorous ground truth measure of sleep, wear, and nonwear, it is also a limitation in terms of generalizability. This could potentially affect children's typical movements in sleep, compromising PSG interpretation itself and contributing noise or measurement error into what we refer to as true labels (Van De Water et al., 2011). This may contribute to some of the observed subject-to-subject variation. It also raises issues with generalizability to the free-living environment and the ability to extend the results to daily wear in children and to adults. For example, our findings are based on sleep studies where sleep was the most common activity. Thus, performance of these approaches may differ in the context when sleep is not the predominant activity state. For example, the Choi and Troiano algorithms were not validated for detecting nonwear during sleep periods and may contribute to some of the variation in performance observed for these algorithms (Choi et al., 2011, 2012; Troiano et al., 2008). In addition, the quality of the sleep may differ from sleep in a noncontrolled environment, and it may be that nonwear detection differs in these two settings. These findings therefore need to be replicated in a context where activity states are more varied in both order, and, length of occurrence, similar to, and ideally in proportions of how they would arise in free-living settings. Our work is relevant for distinguishing a specific type of nonwear—one where the signal would mimic that of sleep or sedentary behavior (e.g., when the device is placed on a dresser) in contrast to not wearing the device and having the device carried by an individual from one location to another in order to place on a participant. This latter scenario is one that we attempted to remove from our studies by excluding the half hour of time just prior to putting the device on the participant. We wanted to include the common type of NWT we find in free-living settings where an individual removes the device while showering, for example. An additional limitation is the fixed sequence of states of nonwear to wear to nonwear, which may further compromise generalizability to scenarios with different sequences of states particularly those where NW is interspersed throughout WT periods. Also, in our experiment where true activity state labels were available for 30-s intervals, we compared performance of features extracted from various window sizes and determined the 30-s intervals to perform comparatively well. However, we recognize that this may not generalize to other studies and may be a function of our study design having true labels measured with the same window size. Furthermore, our results may differ for consumer wrist-worn and smartphone-based accelerometers, as the inertial measurement units are likely to differ between such devices and those studied here. In addition, note that the sampling frequency of 40 Hz in our data set may differ from the sampling frequency of other studies. In work by Clevenger et al. (2019), the authors evaluated differences in interpretation resulting from different sampling rates of 30 and 100 Hz. While largely no significant differences were observed in mean acceleration, the authors did note that the more intensively sampled data resulted in significantly more total counts, particularly for higher intensity activities. Importantly, however, the percent agreement between the two types of data was high and ranged from 97.4% to 99.7% when machine learning algorithms were applied to the data (Clevenger et al., 2019). Thus, we anticipate that our findings here would generalize across a range of sampling rates. It is important, however, that the data set to which the machine learning tools are applied is comprised of consistently sampled data across individuals within the data set. The groundwork for applying these methods is now established but extending them to additional technologies needs further exploration. Finally, further study is needed to pinpoint the reasons behind why the clustering approach of GMM generally performed better than the Markovian chain approach of HMM for these data sets.

## Conclusions

Statistical learning methods hold promise over commonly used algorithms for detecting NWT periods. GMM and HMM demonstrated robust, consistently strong performance across multiple conditions, surpassing or remaining competitive with the commonly used WT algorithms, which had marked inaccuracy when NWT periods were shorter. Of the two statistical learning algorithms, the GMM was superior to the HMM.

## References

Actigraph GT3X+ and wGT3X+ Device Manual. (n.d.). http://dl.theactigraph.com/GT3Xp_wGT3Xp_Device_Manual.pdf

Actilife 6 User's Manual. (n.d.). http://actigraphcorp.com/support/manuals/actilife-6-manual/

Allen, F.R., Ambikairajah, E., Lovell, N.H., & Celler, B.G. (2006). An adapted Gaussian mixture model approach to accelerometry-based movement classification using time-domain features [Paper presentation]. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society.

Banda, J.A., Haydel, K.F., Davila, T., Desai, M., Bryson, S., Haskell, W.L., Matheson, D., & Robinson, T.N. (2016). Effects of varying epoch lengths, wear time algorithms, and activity cut-points on estimates of child sedentary behavior and physical activity from Accelerometer Data. *PLoS One, 11*(3), Article 0150534. https://doi.org/10.1371/journal.pone.0150534

Berry, R.B., Gamaldo, C.E., Harding, S.M., Brooks, R., Lloyd, R.M., Vaughn, B.V., & Marcus, C.L. (2015). AASM scoring manual version 2.2 updates: New Chapters for scoring infant sleep staging and home sleep apnea testing. *Journal of Clinical Sleep Medicine, 11*(11), 1253–1254. https://doi.org/10.5664/jcsm.5176

Cain, K.L., Sallis, J.F., Conway, T.L., Van Dyck, D., & Calhoon, L. (2013). Using accelerometers in Youth Physical Activity Studies: A review of methods. *Journal of Physical Activity and Health, 10*(3), 437–450. https://doi.org/10.1123/jpah.10.3.437

Choi, L., Liu, Z., Matthews, C.E., & Buchowski, M.S. (2011). Validation of accelerometer wear and non-wear time classification algorithm. *Medicine & Science in Sports & Exercise, 43*(2), 357–364.

Choi, L., Ward, S.C., Schnelle, J.F., & Buchowski, M.S. (2012). Assessment of wear/non-wear time classification algorithms for tri-axial accelerometers. *Medicine & Science in Sports & Exercise, 44*(10), 2009–2016.

Clevenger, K.A., Pfeiffer, K.A., Mackintosh, K.A., McNarry, M.A., Brønd, J., Arvidsson, D., & Montoye, A.H. (2019). Effect of sampling rate on acceleration and counts of hip-and wrist-worn ActiGraph accelerometers in children. *Physiological Measurement, 40*(9), Article 095008.

Covidien. (n.d.). *Sandman Elite™ Sleep Diagnostic Software*.

Fraley, C., & Raftery, A. (2007). Model-based methods of classification: Using the mclust software in Chemometrics. *Journal of Statistical Software, 18*(6).

Lee, Y.-S., & Cho, S.-B. (2011). Activity recognition using hierarchical hidden Markov models on a smartphone with 3D accelerometer [Conference session]. HAIS'11 Proceedings of the 6th International Conference on Hybrid Artificial Intelligent, Part I, 460–467.

Mannini, A., & Sabatini, A.M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors, 10*(2), 1154–1175. https://doi.org/10.3390/s100201154

Nickel, C., & Busch, C. (2013). Classifying accelerometer data via Hidden Markov models to authenticate people by the way they walk. *IEEE Aerospace and Electronic Systems Magazine, 28*(10), 29–35.

O'Connell, J., & Højsgaard, S. (2011). Hidden semi Markov models for multiple observation sequences: The mhsmm package for R. *Journal of Statistical Software, 39*(4), 1–22. https://doi.org/10.18637/jss.v039.i04

Rabiner, L.R. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. https://doi.org/10.1109/5.18626

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reynolds, D.A., Quatieri, T.F., & Dunn, R.B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing, 10*(1), 19–41.

Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Reviews, 15*(4), 259–267. https://doi.org/10.1016/j.smrv.2010.10.001

Troiano, R.P. (2007). Large-scale applications of accelerometers. *Medicine & Science in Sports & Exercise, 39*(9), Article 1501. https://doi.org/10.1097/mss.0b013e318150d42e

Troiano, R.P., Berrigan, D., Dodd, K.W., Masse, L.C., Tilert, T., & McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine & Science in Sports & Exercise, 40,* 181–188.

Trost, S.G. (2001). Objective measurement of physical activity in youth: Current issues, future directions. *Exercise and Sport Sciences Reviews, 29*(1), 32–36. https://doi.org/10.1097/00003677-200101000-00007

Trost, S.G. (2007). State of the art reviews: Measurement of physical activity in children and adolescents. *American Journal of Lifestyle Medicine, 1*(4), 299–314. https://doi.org/10.1177/1559827607301686

Tudor-Locke, C., Camhi, S.M., & Troiano, R.P. (2012). A catalog of rules, variables, and definitions applied to accelerometer data in the National Health and Nutrition Examination Survey, 2003–2006. *Preventing Chronic Disease, 9,* Article E113.

Van De Water, A.T.M., Holmes, A, & Hurley, D.A. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—a systematic review. *Journal of Sleep Research, 20,* 183–200.

Vert, A., Weber, K.S., Thai, V., Turner, E., Beyer, K.B., Cornish, B.F., Godkin, F.E., Wong, C., McIlroy, W.E., & Van Ooteghem, K. (2022). Detecting accelerometer non-wear periods using change in acceleration combined with rate-of-change in temperature. *BMC Medical Research Methodology, 22*(1), Article 147. https://doi.org/10.1186/s12874-022-01633-6