# Adapted Sojourn Models to Estimate Activity Intensity in Youth: A Suite of Tools

PAUL R. HIBBING[1,2], LAURA D. ELLINGSON[1], PHILIP M. DIXON[3], and GREGORY J. WELK[1]

*[1]Department of Kinesiology, Iowa State University, Ames, IA; [2]Department of Kinesiology, Recreation, and Sport Studies, University of Tennessee, Knoxville, TN; and [3]Department of Statistics, Iowa State University, Ames, IA*

## ABSTRACT

HIBBING, P. R., L. D. ELLINGSON, P. M. DIXON, and G. J. WELK. Adapted Sojourn Models to Estimate Activity Intensity in Youth: A Suite of Tools. *Med. Sci. Sports Exerc.*, Vol. 50, No. 4, pp. 846–854, 2018. The challenges of using physical activity data from accelerometers have been compounded with the recent focus on wrist-worn monitors and raw acceleration (as opposed to activity counts). **Purpose**: This study developed and systematically evaluated a suite of new accelerometer processing models for youth. **Methods**: Four adaptations of the Sojourn method were developed using data from a laboratory-based experiment in which youth ($N = 54$) performed structured activity routines. The adaptations corresponded to all possible pairings of hip or wrist attachment with activity counts (AC) or raw acceleration (RA), and they estimated time in sedentary behavior, light activity, and moderate-to-vigorous physical activity. Criterion validity was assessed using direct observation in an independent free-living sample ($N = 27$). Monitors were worn on both wrists to evaluate the effect of handedness on accuracy, and status quo methods for each configuration were also evaluated as benchmarks for comparison. Tests of classification accuracy (percent accuracy, $\kappa$ statistics, and sensitivity and specificity) were used to summarize utility. **Results**: In the development sample, percent accuracy ranged from 68.5% (wrist-worn AC, $\kappa = 0.42$) to 71.6% (hip-worn RA, $\kappa = 0.50$). Accuracy was lower in the free-living evaluation, with values ranging from 49.3% (hip-worn RA, $\kappa = 0.25$) to 56.7% (hip-worn AC, $\kappa = 0.36$). Collectively, the suite predicted moderate-to-vigorous physical activity well, with the models averaging 96.5% sensitivity and 67.5% specificity. However, in terms of overall accuracy, the new models performed similarly to the status quo methods. There were no meaningful differences in performance at either wrist. **Conclusions**: The new models offered minimal improvements over existing methods, but a major advantage is that further tuning of the models is possible with continued research. **Key Words:** ACCELEROMETRY, CHILDREN AND ADOLESCENTS, MACHINE LEARNING, ATTACHMENT SITE, ACTIVITY COUNTS, RAW ACCELERATION

Accelerometer-based activity monitors are an accepted standard for assessing physical activity (PA), but methods for acquiring and processing data vary widely. The range of methods has steadily widened, being driven by several factors that include the following: increased sophistication of the sensors (1); increased use of pattern recognition methods to process sensor data (2); increased interest in processing raw accelerometer data, as opposed to activity counts (3); and increased interest in using data from wrist-worn monitors (4). Although these changes have moved accelerometry research forward, they have also made it challenging to understand the strengths and limitations of the various data processing approaches.

A recent study by Lyden et al. (5) presented a novel way to predict energy expenditure (EE) from activity counts by focusing on how accelerometer data are segmented. Traditionally, these data are segmented into epochs, which can lead to aliasing (5). Their method (called Sojourn) attempts to overcome this limitation by segmenting data into bouts of data-defined activity, rather than arbitrary epochs. This allows for predicting EE for a single bout of activity, rather than averaging across a generic time period (e.g., 60 s) that could contain multiple activities of differing intensities. Although the Sojourn method has shown a compelling degree of accuracy in adults (5,6), it is unclear whether it has reasonable accuracy when applied in youth, and whether it can be adjusted to accommodate wrist-worn data and raw acceleration.

The present study addresses these issues by developing and evaluating a suite of youth-specific Sojourn adaptations. The new methodology includes major modifications that account for the unique patterns and characteristics of movement in children and adolescents, and also builds in flexibility for using monitors with different wear locations (hip or wrist) and data types (activity counts or raw acceleration). This enables researchers to compare how the overall method performs when applied in different ways. Thus, this study is also intended to facilitate efforts to promote standardization as accelerometer methods mature.

Address for correspondence: Paul R. Hibbing, M.S., 307 HPER Building, 1914 Andy Holt Ave, Knoxville, TN 37996; E-mail: phibbing@vols.utk.edu.

Submitted for publication June 2017.
Accepted for publication November 2017.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/18/5004-0846/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000001486

## METHODS

The original Sojourn method uses a complex R function to examine second-by-second data and identify periods of differing activity patterns (i.e., "Sojourns"). Periods of higher activity are fed into an artificial neural network (ANN) to estimate EE directly, whereas periods of lower activity undergo more general estimation of EE, which is ultimately based on a decision tree (5). The specifics of the process are slightly different for the uniaxial (soj-1×) and triaxial (soj-3×) versions. Soj-3× showed comparatively better criterion validity in the original study, so the present study adapts only soj-3× and not soj-1×.

The suite of Sojourn adaptations developed in the present study includes four models, which correspond to hip or wrist data, paired with activity counts or raw acceleration as the data input. For clarity, we refer to an individual Sojourn adaptation as a "model," whereas we refer to the whole suite as a "method," because the same underlying technique (i.e., the Sojourn method) is present in each of the four models.

The present study involved two distinct phases to develop and evaluate the models. In the first phase, existing data from a laboratory-based validation study in youth were used to create the four Sojourn-based models. The second phase entailed a free-living criterion validation of the new models in an independent sample of youth. For reference, the new models were also compared with current status quo methods. Details from each phase are described separately below. All parents and youth participants in both phases provided written consent and assent, respectively, before participation, and the study protocols for both phases were approved by the Iowa State University Institutional Review Board.

### Phase 1: Development and Preliminary Evaluation of Youth Sojourn Models

**Description of calibration data set.** Data for the development of the adapted Sojourn models were obtained from a previous project (7,8). The sample used for the present study consisted of 54 youth (ages 7–13 yr), each of whom performed 12 structured activities that were randomly selected from a pool of 24. The 12-activity routine was performed twice, during two separate visits to the laboratory. The pool of 24 activities included 5 sedentary, 3 light, 11 moderate, and 5 vigorous activities, as classified by the Compendium of Physical Activity (9,10), and the intent of randomly selecting 12 activities for each participant was to generate realistic sequences of youth activity. During both visits, participants wore an Oxycon Mobile (OM) portable gas analyzer and two ActiGraph (AG) devices (model GT3X+; ActiGraph, LLC, Pensacola, FL). One AG was worn on the right hip, and the other was worn on the nondominant wrist. Each activity was performed for 5 min, with a 1-min break between activities.

Complete data were not available in all cases. Data from 108 visits were expected (two visits each for 54 participants). Four follow-up visits were excluded because of participant withdrawal, reducing the total to 104. OM data in 15-s epochs were available for all 104 visits. For the AG, counts data in 1-s epochs were available from 94 (hip) and 84 (wrist) visits, and raw acceleration data were available from 86 (hip) and 81 (wrist) visits. Missing data were attributable to operator error during data download and storage, and all available data were used for analysis. The AG monitors were initialized to sample at 100 Hz and were downloaded with the normal filter applied. Activity counts were stored in 1-s epochs, and the vector magnitude was used for analysis. Raw acceleration data were reduced to 1-s epochs by taking the mean Euclidian Norm Minus One (ENMO) within each second. ENMO is a vector magnitude with gravitational acceleration subtracted (11).

**Initial steps for developing adapted Sojourn models.** The new models were designed to classify PA intensity (i.e., sedentary, light, or moderate-to-vigorous PA (MVPA)), whereas the original method estimated METs and metrics derived from METs (i.e., MET-hours and time spent in sedentary behavior, light PA, and MVPA). To convert the OM data to intensity, the mean oxygen consumption ($\dot{V}O_2$) during minute 4 and the first half of minute 5 for each activity was converted to METs. To calculate METs, measured $\dot{V}O_2$ was divided by the predicted resting $\dot{V}O_2$ from Schofield's equations (12). MET values were then coded as sedentary (≤1.5 METs), MVPA (≥3 METs), or light (between both ranges) on the basis of recommended cutoffs (13,14). Thus, each activity a participant performed was assigned a single intensity.

The second-by-second accelerometer data were annotated to link each 1-s epoch (vector magnitude counts or ENMO) to the activity being performed at that time. This was the last step before the formal development of the new models.

There were three primary steps included in developing each of the four models. The first involved training an ANN to replace the original method's "lab-nnet" (5,15). The second involved redefining and reoptimizing the bout identification criteria. The third involved defining a new decision tree to produce final estimates. Each step is described in more detail hereinafter.

**Development of ANN.** Data cleaning and feature extraction were performed before training each ANN. For cleaning, data from the 1-min transition periods were first removed. Next, outliers were removed following the procedure used by Staudenmayer et al. (15). This required calculating the accelerometer data's coefficient of variation for each participant and activity. The mean coefficient of variation was then calculated for each activity, and individual cases that differed from that value by more than 90% were removed from analysis. Finally, data from minute 4 and the first half of minute 5 were isolated for feature extraction, which ensured that the most consistent movement patterns were included in the data set for training the ANN.

Features were calculated in 15-s, nonoverlapping windows, and the feature set included demographic variables (sex, age in years, and body mass index) and accelerometer features (percentiles (10th, 25th, 50th, 75th, and 90th) and lag-one autocorrelation). The accelerometer features were the same as the original lab-nnet (15), and the demographic variables were included to account for the higher degree of variability among children compared with the variability among adults. Each ANN used the same feature set, so that it would equally resemble the original method, and one ANN was trained for each of the four adapted Sojourn models (i.e., for hip- or wrist-worn devices, using activity counts or raw acceleration).

The ANN training process involved using the R packages caret and nnet (16,17). Caret provides flexible structures for training machine learning models and was first used to split the data set into equally balanced partitions, with 70% of the data used for training, and the remaining 30% held out for cross-validating the ANN. It then called nnet to train models that were optimized through grid search. This involved testing various settings for hidden layer size (5, 10, 15, 20, and 25 nodes tested) and decay rate (0, 0.1, and 0.001 tested), returning the optimal performer.

**Bout identification algorithm.** To update the bout identification algorithm, code from the original soj-3× function was modified to accommodate youth behavior patterns and different monitor usages (i.e., different wear locations and data types). This entailed a strategy that differed slightly from the original. In the original soj-3×, two parameters affected the way in which new Sojourns were defined. The first required that consecutive epochs had to differ by 15 or more counts, and the second required that each Sojourn was a minimum of 30 s. The values for both parameters were tuned using grid search.

The present study's adapted bout identification algorithm used both of these parameters plus a new parameter. The new parameter added a requirement to the comparison of consecutive pairs of epochs, whereby the second value in a given pair had to be below a certain amount, similar to a sedentary cut point. All three parameters in the adapted models were optimized through grid search, with 5 settings tested for each parameter and a total of 125 combinations tested. (See Document, Supplemental Digital Content 1, Summary of model tuning with grid search, http://links.lww.com/MSS/B101.) The optimal parameter values were defined as the combination that achieved the highest $\kappa$ score when the Sojourn models were applied to the whole phase 1 data set, including non–steady-state activity.

Altogether, the bout identification code identifies the boundaries between Sojourns by first comparing consecutive pairs of count values. A transition is identified if two criteria are met: 1) the second value is lower than the first by a predefined amount, and 2) the second value itself is lower than a separate predefined threshold. After tagging transitions, the algorithm calculates the time between each transition. If the span between two transitions is shorter than a predefined amount, it is combined with a neighboring span. The combining process is different for transitions in the middle of the file compared with those at the beginning or end. For those in the middle, the span is combined with either the preceding or succeeding span, whichever is shorter. For the ends of the file, this process can only flow in one direction, so all spans that occur before the first full-length Sojourn are combined, and all spans that occur after the last full-length Sojourn are combined. Once all spans have been combined into blocks of a minimum length, they officially become "Sojourns."

**Decision tree.** The final step in development was to establish the overall decision tree structure needed to synthesize information from the preceding steps and produce a final estimate of activity intensity. The decision tree took three inputs: 1) second-by-second accelerometer data (vector magnitude counts or ENMO), 2) ANN estimates from features calculated in Sojourn-defined windows, and 3) ANN estimates from features calculated in 15-s, nonoverlapping windows. Before discussing the specific branches of the decision tree, it is important to note two things.

First, the inclusion of the third input (on the basis of 15-s windows) served two purposes. It was intended to reduce the effect of misclassifications assigned to long Sojourns. For example, if one Sojourn lasting 60 min was labeled sedentary but the actual activity was a low-intensity, intermittent lifestyle task, that whole hour would be misclassified. By including 15-s estimates, it would be possible to recover some misclassified periods within that hour. Second, it was intended to account for potential differences in features when they are calculated in longer windows than the original training set (i.e., 15 s). In a longer window, the data could hypothetically be less variable than 15-s data, which could affect the ANN's prediction.

The second important note about the decision tree is that the input (accelerometer data and ANN predictions) and output (intensity predictions) were second-by-second data. To obtain second-by-second estimates from the ANN, each prediction was replicated once for each second in the period it represented. For example, a prediction made from features in a 15-s window would be replicated 15 times and aligned with the corresponding second-by-second accelerometer data. This approach does not affect the total time predicted in any intensity category, nor does it go beyond the assumptions of the prediction period, that is, that activity within the prediction period is homogenous.

The decision tree's specific characteristics differ substantially from the decision tree in the original Sojourn method. The original decision tree took only accelerometer data as input and only from Sojourns that were classified as inactivity (using a secondary ANN). The decision tree for the new, youth Sojourn method is not dependent on a preclassification step. Rather, it is applied to data from all Sojourns and takes ANN predictions as input, in addition to accelerometer data. The nodes of the decision tree are detailed in the flowchart contained in Figure, Supplemental Digital Content 2, Decision tree for the adapted Sojourn models for youth, http://links.lww.com/MSS/B102. The materials and code for using the youth Sojourn method are available from the corresponding author, along with a tutorial for using them. Further information is available here: http://www.physicalactivitylab.org/research-links-and-news.html.

**Preliminary evaluation of adapted Sojourn models.** Percent accuracy and $\kappa$ statistics were used to assess overall classification accuracy, and sensitivity and specificity were used to assess intensity-specific classification accuracy. Each ANN was tested on the cleaned 30% holdout sample of 15-s windows, while each of the youth Sojourn models was tested on the whole data set, including non–steady-state activity.

## Phase 2: Free-living Validation of New Youth Sojourn Models

The goals of phase 2 were to systematically evaluate the free-living performance of the new youth Sojourn models and to compare this performance with existing status quo methods applicable to the different monitor configurations. All of the methods classified accelerometer data as either sedentary, light, or MVPA, and the criterion measure was direct observation (DO), which is a validated technique for capturing activity intensity (18).

**Participants and materials.** A sample of 27 youth participated in the independent validation phase of the study. Although the sample was half the size of the sample from phase 1, the key sample characteristics were similar, except that the sex split was more even in phase 2 (see Table 1). Each participant's height and weight were measured using a portable stadiometer and scale, respectively. Other demographics (i.e., sex, age, and handedness) were also recorded. Finally, each participant was fitted with three AG monitors (model wGT3X-BT): one on the right hip and one on each wrist.

**Study protocol and criterion measure.** Participants were each observed for 1 h of simulated free-living during which they were allowed the freedom to choose the type, intensity, and duration of activities performed. Participants were provided with examples and suggestions for possible activities, but the only imposed constraint was that they had to perform at least five different tasks during the hour.

An original R program was developed specifically for this protocol to collect and process DO data with PA intensity as the outcome. (See Document, Supplemental Digital Content 3, Background on direct observation R program, http://links.lww.com/MSS/B103.) The code and materials are available from the corresponding author, along with a tutorial for using them. Further information is available here: http://www.physicalactivitylab.org/research-links-and-news.html. The program's general operation consisted of the researcher (P.H.) documenting transitions between activities and identifying the activity performed (e.g., walking). More specifically, whenever a new activity was indicated, a system timestamp was issued and the researcher was prompted to describe the activity, and then to indicate additional characteristics of the activity, which were primarily related to posture and muscle contraction. This process was similar to the method used by Lyden et al. (5,18) to validate the original Sojourn method. In cases where the intensity could not be determined on the basis of the general characteristics of the activity, the intensity was coded *post hoc* by cross-referencing the Compendium of Physical Activities (9,10) using a youth-specific coding scheme (see Document, Supplemental Digital Content 3, Background on direct observation R program, http://links.lww.com/MSS/B103, and Saint-Maurice et al. [19]).

**Data processing.** For both activity counts (vector magnitude) and raw acceleration (ENMO), data were downloaded and processed in 1-s epochs using the AG software and an R function, respectively. To evaluate the comparative performance of the new suite of models with existing methods, other commonly used, status quo methods were tested. For hip-worn activity counts, the Freedson (20) method was used, and for wrist-worn activity counts, the Crouter regression (21) and Chandler (22) methods were used. At both locations, the original adult Sojourn method was also applied. For raw acceleration at the hip and wrist, the site-specific equations of Hildebrand et al. (23) were used. When appropriate, counts data were reintegrated in R to use the alternative methods (e.g., by applying a 5-s cut point to data in 5-s epochs). Before use, the reintegration code demonstrated 100% agreement with a sample of pilot files that were reintegrated in the AG software. Up to 59 s of data were discarded at the beginning and end of each trial, so that the data began and ended on an exact minute, which is essential for accurately applying and comparing methods in different epoch lengths. The epoch lengths for the Freedson (60 s), Crouter (5 s), and Chandler (5 s) methods were longer than 1 s. Thus, aligning them with second-by-second predictions from the adapted Sojourn models required the same replication technique described previously for the Sojourn decision tree. The specific implementation of each method is detailed in the following paragraphs.

The Freedson method takes counts per minute (i.e., 60-s epoch data) as input to estimate youth METs on the basis of multiples of estimated resting $\dot{V}O_2$ (20). Up to 59 s of data were discarded at the beginning and end of each trial, so that the data began and ended on an exact minute, which is essential for accurately applying and comparing methods in different epoch lengths. For the present study, an age-specific cut point for MVPA was calculated for each participant, on the basis of predicted MET values ≥ 3. A standard sedentary cut point (i.e., ≤100 vertical axis counts per minute) was used to classify sedentary behavior. Lastly, count values between the sedentary and MVPA cut points were classified as light PA.

TABLE 1. Participant demographics from both samples.

| | Phase 1 | | | Phase 2 | | |
|---|---|---|---|---|---|---|
| | Female | Male | Total | Female | Male | Total |
| n | 33 | 21 | 54 | 13 | 14 | 27 |
| Age, yr | 9.9 ± 2.1 | 10.0 ± 2.4 | 10.0 ± 2.2 | 8.6 ± 1.8 | 10.1 ± 2.1 | 9.4 ± 2.1 |
| BMI, kg·m$^{-2}$ | 16.7 ± 2.6 | 18.3 ± 5.9 | 17.3 ± 4.2 | 15.6 ± 2.7 | 18.2 ± 4.0 | 16.9 ± 3.6 |

Values are mean ± SD.
BMI, body mass index.

TABLE 2. Performance of the ANN and adapted Sojourn models in the phase 1 development testing.

| | Activity Counts | | Raw Acceleration | |
|---|---|---|---|---|
| | Hip | Wrist | Hip | Wrist |
| ANN | | | | |
| Sedentary (Se, Sp) | 68.7%, 97.0% | 59.2%, 95.5% | 85.4%, 94.5% | 73.6%, 95.4% |
| Light (Se, Sp) | 43.3%, 94.1% | 22.1%, 92.2% | 36.5%, 94.1% | 19.2%, 96.1% |
| MVPA (Se, Sp) | 95.4%, 66.6% | 93.5%, 54.3% | 93.4%, 73.8% | 96.3%, 55.7% |
| Total (% accuracy, $\kappa$) | 82.4%, 0.59 | 74.9%, 0.43 | 82.1%, 0.62 | 78.3%, 0.50 |
| Sojourn methods | | | | |
| Sedentary (Se, Sp) | 45.7%, 96.6% | 45.9%, 94.3% | 50.8%, 97.4% | 50.5%, 95.9% |
| Light (Se, Sp) | 46.9%, 79.9% | 44.2%, 81.2% | 47.6%, 82.6% | 44.7%, 79.9% |
| MVPA (Se, Sp) | 86.1%, 73.1% | 83.3%, 69.6% | 89.6%, 71.4% | 81.5%, 73.8% |
| Total (% accuracy, $\kappa$) | 69.7%, 0.46 | 68.5%, 0.42 | 71.6%, 0.50 | 70.3%, 0.44 |

The ANN models were evaluated on a 30% holdout sample of steady-state data, and the Sojourn models were evaluated on the whole data set, including non–steady-state activity.
Se, sensitivity; Sp, specificity.

The Crouter regression–based cut points (21) and Chandler cut points (22) both predict intensity categorically using data in 5-s epochs. The original Sojourn method for adults was also applied to counts data at both the hip and the wrist. This method estimates METs on the basis of an assumed resting $\dot{V}O_2$ of 3.5 mL·kg$^{-1}$·min$^{-1}$. To account for this, the MET estimates from the original Sojourn method were converted back to $\dot{V}O_2$ (by multiplying by 3.5 mL·kg$^{-1}$·min$^{-1}$), then divided by the estimated resting $\dot{V}O_2$ from Schofield's equations (12), to obtain youth METs. Intensity was then coded as sedentary behavior (METs < 1.5), MVPA (METs ≥ 3), or light PA (1.5–2.9 METs).

The raw data for the wrist and hip were processed using the attachment-specific $\dot{V}O_2$ prediction equations developed by Hildebrand et al. (23) for AG monitors. The $\dot{V}O_2$ estimates were first converted to METs by dividing by the estimated resting $\dot{V}O_2$ from Schofield's equations (12), then coded into intensities using the same MET ranges as the original Sojourn method.

**Statistical analysis.** The data analyses focused on applying the adapted Sojourn models to each participant's data and comparing the estimates with the criterion measure using the same metrics as phase 1 (percent accuracy, $\kappa$ statistics, sensitivity, and specificity). The same comparisons were also made on the status quo methods, to provide performance benchmarks specific to each attachment site and data type. Wrist data were processed separately for the dominant and nondominant limbs, adding a level of comparison to address the question of whether handedness matters with the use of these methods.

## RESULTS

Table 2 shows results from phase 1, for each ANN itself, as well as for the adapted youth Sojourn models. The ANN models trained for hip-worn monitors were 82.1%–82.4% accurate ($\kappa$ = 0.59–0.62), whereas the wrist-worn ANN models were 74.9%–78.3% accurate ($\kappa$ = 0.43–0.50). Accuracy was lower for all wear locations and data types when the adapted Sojourn models were applied to the entire data set, with an average decrease of 9.4% ($\kappa$ = 0.08). The models had similar overall accuracy, falling within a 3.1% range. Results were also similar within individual activity categories, with sensitivities separated by no more than 8.1% and specificities separated by no more than 4.2%.

For the phase 2 analysis, results for the wrist-worn methods are reported for whichever wrist performed better, due to a high degree of similarity between the wrists for all methods. Table 3 gives a detailed report on the minor differences observed between outcomes from the two wrists. The absolute difference in overall accuracy between the wrists was 0.7%–2.2% ($\kappa$ = 0.01–0.03) using the original Sojourn method, and both adapted Sojourn models. For the other methods, the absolute difference did not exceed 0.8% ($\kappa$ = 0.01). Similarly, when comparing individual intensity categories between the wrists, the absolute difference in classification accuracy (i.e., sensitivity or specificity) for all methods averaged 1.2% (standard deviation = 1.4%).

Confusion matrices and classification accuracy for all methods applied to the Phase 2 data are presented in Table 4. For MVPA classification, all four Sojourn models had at least 2.8% higher sensitivity than in Phase 1, with a mean sensitivity for all four models of 96.5%, compared to 85.1% in Phase 1. Specificities were all ≥58.3%, and the counts-based model for hip-worn monitors had a specificity of 76.2%. The hip-specific status quo methods tended to have lower sensitivity and higher specificity, except for the Freedson method (sensitivity of 94.5% and specificity of 79.9%). The results were similar for the wrist-worn methods.

For sedentary behavior, the Freedson method (using a common generic vertical axis cut point of 100 counts per

TABLE 3. Differences (dominant minus nondominant) between outcomes from both wrists, for each wrist method assessed in phase 2.

| | Sedentary (Se, Sp) | Light (Se, Sp) | MVPA (Se, Sp) | Total (% Accuracy, $\kappa$) |
|---|---|---|---|---|
| Adapted Sojourns, counts | −4.5%, 0.7% | −0.4%, 0.9% | 0.3%, −3.8% | −1.7%, −0.02 |
| Adapted Sojourns, raw | −0.4%, 0.2% | 2.3%, 0.9% | 0.1%, 0.0% | 0.7%, 0.01 |
| Crouter | −0.3%, −1.1% | −2.3%, 0.0% | 0.4%, −0.3% | −0.8%, −0.01 |
| Chandler | −0.2%, −0.5% | −2.9%, 0.6% | 2.0%, −0.8% | −0.6%, −0.01 |
| Hildebrand | 0.0%, 0.0% | 1.0%, −1.4% | −1.3%, 1.3% | 0.0%, 0.00 |
| Original Sojourn | −6.5%, −0.4% | −0.3%, −1.6% | 0.8%, −1.3% | −2.2%, −0.03 |

The highest-magnitude differences for all outcomes are shown in bold.
Se, sensitivity; Sp, specificity.

TABLE 4. Confusion matrices and classification accuracy for the methods tested in the phase 2 study.

| | | | Reference | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | | | Sedentary | Light | MVPA | Se, % | Sp, % | Accuracy, % | κ |
| Hip | Freedson | Sedentary | 410.6 | 263.7 | 0.2 | 72.8 | 73.8 | 59.6 | 0.40 |
| | | Light | 119.0 | 105.9 | 24.2 | 18.9 | 85.8 | | |
| | | MVPA | 34.7 | 191.9 | 420.0 | 94.5 | 79.9 | | |
| | Original Sojourn | Sedentary | 539.3 | 412.6 | 48.8 | 95.6 | 54.1 | 59.0 | 0.38 |
| | | Light | 17.8 | 61.4 | 69.3 | 10.9 | 91.4 | | |
| | | MVPA | 7.1 | 87.5 | 326.2 | 73.4 | 91.6 | | |
| | Adapted Sojourn | Sedentary | 305.0 | 190.5 | 0.3 | 54.1 | 81.0 | 56.7 | 0.36 |
| | | Light | 209.3 | 152.9 | 12.6 | 27.2 | 78.0 | | |
| | | MVPA | 49.9 | 218.0 | 431.4 | 97.1 | 76.2 | | |
| | Hildebrand[a] | Sedentary | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 49.9 | 0.24 |
| | | Light | 557.3 | 463.7 | 124.9 | 82.6 | 32.3 | | |
| | | MVPA | 6.8 | 97.7 | 319.4 | 71.9 | 90.7 | | |
| | Adapted Sojourn[a] | Sedentary | 179.3 | 116.9 | 0.2 | 31.8 | 88.4 | 49.3 | 0.25 |
| | | Light | 269.3 | 183.9 | 33.6 | 32.8 | 70.0 | | |
| | | MVPA | 115.6 | 260.6 | 410.5 | 92.4 | 66.6 | | |
| Wrist | Crouter[b] | Sedentary | 291.3 | 223.8 | 2.9 | 51.6 | 77.5 | 48.4 | 0.23 |
| | | Light | 212.0 | 156.4 | 129.7 | 27.9 | 66.1 | | |
| | | MVPA | 60.9 | 181.2 | 311.8 | 70.2 | 78.5 | | |
| | Chandler[b] | Sedentary | 410.3 | 297.4 | 37.2 | 72.7 | 66.7 | 50.7 | 0.25 |
| | | Light | 125.2 | 127.8 | 149.5 | 22.8 | 72.8 | | |
| | | MVPA | 28.7 | 136.2 | 257.5 | 58.0 | 85.3 | | |
| | Original Sojourn[b] | Sedentary | 443.9 | 302.7 | 26.6 | 78.7 | 67.3 | 53.5 | 0.31 |
| | | Light | 73.1 | 65.2 | 86.3 | 11.6 | 84.2 | | |
| | | MVPA | 47.2 | 193.5 | 331.4 | 74.6 | 78.6 | | |
| | Adapted Sojourn[b] | Sedentary | 253.8 | 202.0 | 0.6 | 45.0 | 79.9 | 50.7 | 0.27 |
| | | Light | 214.5 | 106.6 | 8.8 | 19.0 | 77.9 | | |
| | | MVPA | 95.9 | 252.8 | 434.9 | 97.9 | 69.0 | | |
| | Hildebrand[a,c] | Sedentary | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 47.1 | 0.21 |
| | | Light | 517.1 | 393.0 | 98.2 | 70.0 | 39.0 | | |
| | | MVPA | 47.1 | 168.4 | 346.1 | 77.9 | 80.9 | | |
| | Adapted Sojourn[a,c] | Sedentary | 197.9 | 128.0 | 1.3 | 35.1 | 87.1 | 49.4 | 0.26 |
| | | Light | 191.3 | 139.0 | 4.8 | 24.8 | 80.6 | | |
| | | MVPA | 175.0 | 294.4 | 438.2 | 98.6 | 58.3 | | |

Units in the confusion matrices are minutes.
[a] Method uses raw acceleration.
[b] Results taken from the nondominant wrist.
[c] Results taken from the dominant wrist.
Se, sensitivity; Sp, specificity.

minute) showed the best performance at the hip, with sensitivity and specificity reaching 72.8% and 73.8%, respectively. Performance of wrist methods for predicting sedentary behavior was generally poor, and the original Sojourn method showed the highest classification accuracy with a sensitivity of 78.7% and specificity of 67.3%.

When looking at overall accuracy across the whole period, the new models were not noticeably better than the highest-performing comparison methods. At the hip, the Freedson method ($\kappa$ = 0.40) showed the greatest agreement, whereas the $\kappa$ scores for the original Sojourn method and the counts-based adapted Sojourn model were within 0.04 of that value. For the wrist, the original Sojourn method performed the best ($\kappa$ = 0.31), whereas both the counts-based and raw acceleration Sojourn models were within 0.05.

## DISCUSSION

The present study developed and evaluated a suite of youth-specific prediction models on the basis of adaptations to the Sojourn method (5). This process enabled direct comparisons of MVPA and other outcomes for different attachment sites (hip or wrist) and device outputs (activity counts or raw acceleration). Although the new suite was effective for estimating MVPA, it was not appreciably better than status quo methods in terms of overall performance. Outcomes from the dominant and nondominant wrists were essentially identical at the aggregate level, which was true for the new youth Sojourn models as well as the wrist-based status quo methods. This suggests that dominant versus nondominant attachment may be inconsequential over time in youth.

Interpreting the findings of this study requires careful attention to overall patterns and trends. In particular, the Hildebrand equations (23) did not estimate any sedentary behavior (because of the high intercepts in the equations), whereas the original Sojourn method (5) vastly overestimated sedentary behavior. These trends are problematic, because when one class is systematically underestimated or overestimated, the balance of activities and intensities performed in the study becomes the driving factor behind the observed accuracy metrics (e.g., $\kappa$). Thus, if this study had more sedentary activity, the original Sojourn method would seem more accurate, but if it had more light activity, the Hildebrand equations would seem more accurate. In light of this, it is important to consider the tendencies of each method and the balance of activities and intensities performed by the participants in this study. A clear issue on the basis of the findings of this study is that most methods cannot consistently distinguish light activity from sedentary behavior. In this respect, an advantage of the adapted youth Sojourn method is that there are multiple entry points for improving the models (i.e., improving the ANN, bout identification code, decision tree, or a combination). In contrast, the other methods (except the original Sojourn method intended for adults) are fixed and cannot be readily improved.

As the use and function of accelerometers has evolved, researchers have focused more on the issue of attachment site (23–32) than output type (33). As a result, the former issue has come nearer to resolution than the latter. That is, although advanced models have been able to produce more similar results between the hip and wrist (32), little research has investigated the difference between models trained for activity counts versus raw acceleration. The present study takes this into account by training models specific to each output, which is an important approach to consider amid the rapid changes in the technology itself. The value of a migration away from activity counts remains a largely open question, although it potentially provides a way to harmonize output from different devices and develop monitor-agnostic models. The results of this study do not clearly suggest whether raw acceleration is an improvement compared with activity counts. Both Sojourn models that used activity counts performed better than did those that used raw acceleration, but the differences were fairly small (7.4% for the hip and 1.3% for the wrist).

With wrist-worn monitors, the effect of handedness remains an open question (2). A commonly expressed concern has been that certain activities are performed almost exclusively with one hand or the other (34,35), which could make hand dominance a key variable to consider when selecting models and deploying monitors. Existing evidence (2,36,37) suggests fairly consistently that wrist-specific models are comparable with each other, but most evidence has come from adult populations wearing GENEActiv accelerometers. Mackintosh et al. (38) sampled youth wearing an AG GT3X+ on both wrists and formed machine learning models separately for each wrist. Bias and root mean square error were similar between wrists with all of the models, providing some evidence that the previously observed similarities between wrists may also exist for youth-wearing AG devices. The findings of the present study are consistent with this pattern, but it is important to note that the wrist Sojourn models in this study were not side or handedness specific. Rather, nondominant wrist data were used to develop a single model that was applied to both wrists in the independent validation.

A strength of this study is the systematic evaluation of the models under free-living conditions with criterion data from DO. Other studies in adults (39) and older adults (40) have shown similar findings to the present study (i.e., poorer performance under free-living conditions than in the original development context). Although awareness of this issue is increasing, these types of studies remain rare, potentially because they can be difficult to conduct. The methodological features of this study can make such assessments more feasible in the future, both quantitatively and qualitatively.

Quantitatively, a major advance of the present study is the use of classification accuracy to test the methods. Previous studies (5,18) have reported aggregated indicators of agreement with DO such as total activity time and whole-trial root mean square error. Although these may reflect a method's tendency to converge over time toward accurate totals, they are unable to indicate whether the minutes estimated actually correspond to the minutes accumulated (i.e., whether the temporal trends are linked). In addition, the commonly used metrics do not meaningfully account for the compositional nature of the data (i.e., the reality that intensity categories are not independent). More time in one category means less in another, and thus, the observed accuracies in each category are interdependent. Classification accuracy, as used in this study, assesses agreement at the 1-s level, which has three main benefits. First, it increases the resolution of the measure compared with analyzing total time spent in each category. Second, it gives indicators of temporally linked agreement, which is more informative about the monitor's true validity. Third, classification accuracy accounts, to some degree, for the interdependence of the categories, which is reflected in the relationship between sensitivity and specificity.

Qualitatively speaking, the present study introduces a method to streamline future free-living data collections, primarily through the development of a new R program for DO. This tool collects data in a manner designed for convenient merging with accelerometer data at the epoch level, and it is a cost-effective way to both improve and simplify the process of combining DO and accelerometer data.

Future studies should build on this suite of open-source models to promote standardization and refinement over time. A key advantage of the models presented is their hybrid nature (i.e., inclusion of ANN models, bout identification algorithms, and decision trees). In addition to the previously stated benefit of allowing for multiple entry points for further innovation, hybrid approaches can minimize the effect of limitations specific to a given component. For example, two potential limitations of the ANN models are that they may be overfitted and that they may benefit from having unique feature sets, as opposed to the universal feature set used for this study. In both cases, the hybrid character of the models can offer protection, because the other components of the Sojourn models may compensate for limitations in the ANN. This sort of protection is also present within the decision tree, where window-specific (i.e., 15-s) and bout-specific (i.e., Sojourn) ANN predictions are both considered. Customized features for the wrist were tested while developing the ANN models for this study. However, no relevant improvements were detected. Thus, the advantage of using one universal feature set (which was comparable with the original Sojourn method's feature set) was considered more salient than an extensive search for better features.

In summary, this study provides a flexible suite of data processing models on the basis of the established Sojourn method. A potential limitation is the use of the original compendium of physical activities (9,10) to encode activity intensity. The compendium was developed from studies conducted in adults, which could limit its applicability to youth. However, it is important to note that the compendium was only used in this study when it was not possible to determine activity intensity based solely on the characteristics of the movement. Furthermore, the differences between youth and adults are greater when looking at EE than categorical activity intensity, especially when MVPA is a single category, as it was in this study. Most importantly, compendium-based intensity was encoded using youth-specific MET cutoffs (19). Another limitation of the study is that phase 2 used an AG model (wGT3X-BT) that had an accelerometer with a wider dynamic range (±8 gravitational units) than the monitor in Phase 1 (GT3X+, ±6 gravitational units). Despite this limitation, AG models are often used interchangeably in practice (e.g., by applying cut points for the 7164 model to data from newer models), and the models presented in this study were still able to perform comparably to other methods tested in phase 2. However, they may have performed better if the same model was used in both phases. The collective nature of this suite of models represents a novel methodological improvement for evaluating accelerometer data from youth. We encourage and support further open-source refinement and efforts to promote standardization in methods.

The results of this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. This study does not constitute endorsement by the American College of Sports Medicine.

The authors declare no conflicts of interest.

## REFERENCES

1. John D, Freedson P. ActiGraph and Actical physical activity monitors: a peek under the hood. *Med Sci Sports Exerc*. 2012;44(Suppl):S86–9.
2. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. *Br J Sports Med*. 2014;48(13):1019–23.
3. Freedson P, Bowles HR, Troiano R, Haskell W. Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field. *Med Sci Sports Exerc*. 2012;44(Suppl):S1–4.
4. Freedson PS, John D. Comment on "Estimating Activity and Sedentary Behavior from an Accelerometer on the Hip and Wrist." *Med Sci Sports Exerc*. 2013;45(5):962–3.
5. Lyden K, Keadle SK, Staudenmayer J, Freedson PS. A method to estimate free-living active and sedentary behavior from an accelerometer. *Med Sci Sports Exerc*. 2014;46(2):386–97.
6. Ellingson LD, Schwabacher IJ, Kim Y, Welk GJ, Cook DB. Validity of an integrative method for processing physical activity data. *Med Sci Sports Exerc*. 2016;48(8):1629–38.
7. Kim Y, Crouter SE, Lee JM, Dixon PM, Gaesser GA, Welk GJ. Comparisons of prediction equations for estimating energy expenditure in youth. *J Sci Med Sport*. 2016;19(1):35–40.
8. Kim Y, Lee J-M, Peters BP, Gaesser GA, Welk GJ. Examination of different accelerometer cut-points for assessing sedentary behaviors in children. *PLoS One*. 2014;9(4) [Internet] [cited 29 Sept 2017]. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090630.
9. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of physical activities: a second update of codes and MET values. *Med Sci Sports Exerc*. 2011;43(8):1575–81.
10. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc*. 2000;32(9 Suppl):S498–504.
11. Van Hees VT, Gorzelniak L, León EC, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One*. 2013;8(4) [Internet] [cited 29 Sept 2017]. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061691.
12. Schofield WN. Predicting basal metabolic rate: new standards and review of previous work. *Hum Nutr Clin Nutr*. 1984;39:5–41.
13. Pate RR, O'Neill JR, Lobelo F. The evolving definition of "sedentary." *Exerc Sport Sci Rev*. 2008;36(4):173–8.
14. Pate RR, Pratt M, Blair SN, et al. Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *JAMA*. 1995;273(5):402–7.
15. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *J Appl Physiol (1985)*. 2009;107(4):1300–7.
16. Kuhn M, Wing J, Weston S, et al. *caret: Classification and Regression Training*. 2016 [cited 29 Sept 2017]. Available from: https://CRAN.R-project.org/package=caret.
17. Venables WN, Ripley BD. *Modern Applied Statistics with S. Fourth Edition*. New York: Springer; 2002. pp. 1–498.
18. Lyden K, Petruski N, Mix S, Staudenmayer J, Freedson P. Direct observation is a valid criterion for estimating physical activity and sedentary behavior. *J Phys Act Health*. 2014;11(4):860–3.
19. Saint-Maurice PF, Kim Y, Welk GJ, Gaesser GA. Kids are not little adults: what MET threshold captures sedentary behavior in children? *Eur J Appl Physiol*. 2016;116(1):29–38.
20. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Med Sci Sports Exerc*. 2005;37(Suppl):S523–30.
21. Crouter SE, Flynn JI, Bassett DR. Estimating physical activity in youth using a wrist accelerometer. *Med Sci Sports Exerc*. 2015;47(5):944–51.
22. Chandler JL, Brazendale K, Beets MW, Mealing BA. Classification of physical activity intensities using a wrist-worn accelerometer in 8-12-year-old children. *Pediatr Obes*. 2016;11(2):120–7.
23. Hildebrand M, Van Hees VT, Hansen BH, Ekelund U. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. *Med Sci Sports Exerc*. 2014;46(9):1816–24.
24. Bao L, Intille SS. *Activity Recognition from User-Annotated Acceleration Data*. In: *Proceedings of the 2nd International Conference on Pervasive Computing*; 2004 Apr 18–23: Linz/Vienna (Austria). Johannes Kepler Universität. 2004. p. 1–17.
25. Chen KY, Acra SA, Majchrzak K, et al. Predicting energy expenditure of physical activity using hip- and wrist-worn accelerometers. *Diabetes Technol Ther*. 2003;5(6):1023–33.
26. Cleland I, Kikhia B, Nugent C, et al. Optimal placement of accelerometers for the detection of everyday activities. *Sensors (Basel)*. 2013;13(7):9183–200.
27. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiol Meas*. 2014;35(11):2191–203.
28. Mannini A, Intille SS, Rosenberger M, Sabatini AM, Haskell W. Activity recognition using a single accelerometer placed at the wrist or ankle. *Med Sci Sports Exerc*. 2013;45(11):2193–203.
29. Puyau MR, Adolph AL, Vohra FA, Butte NF. Validation and calibration of physical activity monitors in children. *Obes Res*. 2002;10(3):150–7.
30. Rosenberger ME, Haskell WL, Albinali F, Mota S, Nawyn J, Intille S. Estimating activity and sedentary behavior from an accelerometer on the hip or wrist. *Med Sci Sports Exerc*. 2013;45(5):964–75.

31. Swartz AM, Strath SJ, Bassett DR, O'Brien WL, King GA, Ainsworth BE. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc*. 2000;32(9 Suppl):S450–6.
32. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: hip versus wrist data. *Physiol Meas*. 2014;35(11):2183–9.
33. Rowlands AV, Stiles VH. Accelerometer counts and raw acceleration output in relation to mechanical loading. *J Biomech*. 2012;45(3):448–54.
34. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA accelerometer. *Med Sci Sports Exerc*. 2011;43(6):1085–93.
35. Hibbing PR, Kim Y, Saint-Maurice PF, Welk GJ. Impact of activity outcome and measurement instrument on estimates of youth compliance with physical activity guidelines: a cross-sectional study. *BMC Public Health*. 2016;16(223) [Internet] [cited 29 Sept 2017]. Available from: http://www.biomedcentral.com/1471-2458/16/223.
36. Montoye AH, Mudd LM, Biswas S, Pfeiffer KA. Energy expenditure prediction using raw accelerometer data in simulated free living. *Med Sci Sports Exerc*. 2015;47(8):1735–46.
37. Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA. Wrist-independent energy expenditure prediction models from raw accelerometer data. *Physiol Meas*. 2016;37(10):1770–84.
38. Mackintosh KA, Montoye AH, Pfeiffer KA, McNarry MA. Investigating optimal accelerometer placement for energy expenditure prediction in children using a machine learning approach. *Physiol Meas*. 2016;37(10):1728–40.
39. Gyllensten IC, Bonomi AG. Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *IEEE Trans Biomed Eng*. 2011;58(9):2656–63.
40. Sasaki JE, Hickey AM, Staudenmayer JW, John D, Kent JA, Freedson PS. Performance of activity classification algorithms in free-living older adults. *Med Sci Sports Exerc*. 2016;48(5):941–50.