

**CALIBRATION AND VALIDATION OF GYROSCOPE INCLUSIVE
YOUTH SOJOURN MODELS**

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

Paul Robert Hibbing

May 2020

DEDICATION

For Sarah – not that you want to read it.

ACKNOWLEDGMENTS

I want to start by thanking everyone who has helped me in unseen ways over the past four years. I wish I could name you all, but that would take another dissertation, and given my verbose tendencies, I think we can all agree it is better that I settle for a quick and utterly inadequate word of thanks. Here it is: Good work needs great support, and you have buoyed me with every hug, word of encouragement, act of service, and so on. Special thanks to those who have feigned interest in my work.

Now, on to the verbose part: Notwithstanding the preceding laconicism, I shall still salute several specific supporters, starting with my wife and kids. In seriousness, I cannot thank you enough for keeping me sane, especially when I have had the opposite effect on you. Sarah, thank you for helping me pursue balance in life, and thank you for your patience when I get it wrong. Kids, thank you for reminding me to love life for the right reasons.

On a technical level, I would like to thank the Applied Physiology Lab staff for making this work possible. Special thanks to Samuel LaMunion for doing everything well, especially the things I do not want to do. Thank you to my advisor, Dr. Scott Crouter, for providing guidance and opportunities over the years, and for always keeping my best interests at heart. And thank you to the other members of my committee, Drs. David Bassett, Dawn Coe, and Haileab Hilafu, for stretching me as a student and researcher. I have learned priceless lessons from each of you.

Lastly, I would like to thank technology for being like a good fast food burger: economical, well-oiled, and guaranteed to have the last word. Without good tech, I could not have written this dissertation. I hope this statement wins me favor with the robots so they will spare my life when they take over the world.

The work in this dissertation was funded under NIH R01HD083431.

ABSTRACT

Youth Sojourn models are sensor-based physical activity (PA) assessment methods that are used to predict sedentary behavior (SB), light PA (LPA), and moderate-to-vigorous PA (MVPA). The existing models are accelerometer-based and rely on non-statistical segmentation. Improvements may be possible through the incorporation of gyroscope data and statistical change point detection. **PURPOSE:** To calibrate and validate new youth Sojourn models (i.e., Soj_{recall} and Soj_{precision}) that incorporate gyroscope data and change point detection. **METHODS:** Subsets of two data sets were used, one from a study conducted in a semi-structured laboratory setting (n = 86), and the other from a study conducted in free-living environments with an independent sample (n = 31). Throughout both studies, participants wore ActiGraph GT9X devices (each containing an accelerometer and gyroscope) on the hip and both wrists. Energy expenditure was also assessed (via indirect calorimetry), and direct observation was performed to document activity and postural transitions. The latter data streams were used to obtain criterion values for the timing of activity transitions, as well as intensity (SB, LPA, and MVPA). Predictions from Soj_{recall} and Soj_{precision} and the original accelerometer-based youth Sojourn models (Soj_{original}) were then compared to the criterion values, in terms of segmentation and intensity classification. **RESULTS:** When replicating the original youth Sojourn calibration with gyroscope data instead of accelerometer data, all gyroscope-based components outperformed their accelerometer-based counterparts. For segmentation of laboratory data, Soj_{recall}, Soj_{precision}, and Soj_{original} all had peak mean aggregated performance of 56.8%-64.2% (across all models and attachment sites). However, for the free-living data the values were lower for Soj_{recall} and Soj_{precision} (34.5%-46.3%) than Soj_{original} (53.2%-53.6 %). For intensity classification across all attachment sites, Soj_{recall} and Soj_{precision} had mean percent correct classification of 71.2%-77.6%

(laboratory data) and 57.5%-71.3% (free-living data), whereas Soj_{original} had mean percent correct classification of 56.9%-69.5% (laboratory data) and 28.4%-52.7% (free-living data).

CONCLUSION: Change point detection was not effective for segmentation, but gyroscope data led to improved intensity classification. Future work should emphasize development of models that capitalize on the unique strengths of accelerometer and gyroscope data.

TABLE OF CONTENTS

Chapter I: Introduction.....	1
Statement of the Problem.....	4
Statement of the Purpose	5
Research Questions and Hypotheses	5
Overview and Significance of the Dissertation	6
Chapter II: Review of Literature	7
Background	8
Scope and Terminology	8
Theoretical Basis for Sensor-Based PA Assessment	8
History of Sensors in PA Assessment.....	9
Summary and Transition.....	12
Procedural Factors in Sensor-Based PA Assessment	12
Sensor Placement	12
Data Format	14
Monitor Brand.....	15
Sensor Fusion.....	16
Summary and Transition.....	35
Computational Factors in Sensor-Based PA Assessment.....	36
Segmentation.....	36
Feature Generation.....	42
Model Application	44
Novel Methods.....	50
Summary and Transition.....	60
Recap and Synthesis: Gaps with Respect to Segmentation and Sensor Fusion.....	61
Chapter III: Description of Laboratory Data Set.....	62
Participants.....	63
Protocol	63
Equipment	65
ActiGraph GT9X	65
Noldus Pocket Observer	66

Cosmed K4b ²	68
Initial Data Processing	69
ActiGraph GT9X	69
Noldus Pocket Observer	69
Cosmed K4b ²	70
Preview of Further Processing.....	70
Further Processing: Resting Metabolic Rate Data	71
Further Processing: Steady State Data	71
Processing Criterion Data	71
Processing Sensor Data.....	72
Assembling the Final Data Set.....	72
Further Processing: Continuous Time Course Data.....	74
Data Loss	74
Summary of Final Data Set.....	78
Chapter IV: Research Question 1: Methods and Results	81
Overview	82
Scope.....	82
Methods.....	84
Bout Detection Algorithm.....	85
Neural Network.....	90
Decision Tree	93
Results.....	97
Bout Detection Algorithm.....	97
Neural Network.....	99
Decision Tree	100
Summary and Transition.....	100
Chapter V: Research Question 2: Methods and Results	102
Overview	103
Methods.....	104
Change Point Detection Method.....	104
Soj _{recall} Algorithm.....	105
Soj _{precision} Algorithm.....	105

Analysis.....	109
Results.....	113
Analysis of Segmentation	113
Analysis of Activity Intensity Classification	115
Summary and Transition.....	119
Chapter VI: Description of Free-Living Data Set.....	120
Participants.....	121
Protocol	122
Equipment	122
ActiGraph GT9X	122
Noldus Observer XT	123
Cosmed K4b ² and K5.....	127
Initial Data Processing	128
ActiGraph GT9X	128
Noldus Observer XT	128
Cosmed K4b ² and K5.....	129
Assembling the Final Data Set.....	130
Data Loss	131
Summary of Final Data Set.....	134
Chapter VII: Research Question 3: Methods and Results.....	136
Overview.....	137
Methods.....	137
Analysis of Model Runtime	137
Analysis of Segmentation	138
Analysis of Activity Intensity Classification	139
Analysis of Activity Time.....	140
Results.....	140
Analysis of Model Runtime	140
Analysis of Segmentation	141
Analysis of Activity Intensity Classification	145
Analysis of Activity Time.....	150
Summary and Transition.....	153

Chapter VIII: Discussion and Conclusion.....	154
Summary of Findings.....	155
Implications of Model Design	156
Implications of Study Design.....	159
Core Implications.....	161
Sensor Fusion.....	162
Change Point Detection	166
Combining Sensor Fusion and Change Point Detection.....	170
Model Generalizability	171
Technical and Practical Implications	172
Strengths & Limitations.....	174
Conclusion	175
References	177
Appendices.....	210
Appendix A: Recruitment and Enrollment Materials for the Laboratory Study	211
Recruitment Flyer	211
Health History Questionnaire	212
Parent Informed Consent Form.....	214
Youth Assent Form	216
Appendix B: Recruitment and Enrollment Materials for the Free-Living Study	219
Recruitment Flyer	219
Parent Informed Consent Form.....	220
Youth Assent Form	223
Vita	225

LIST OF TABLES

Table 1. Summary of sensor fusion studies described in the literature review.....	17
Table 2. Activities performed by participants in the laboratory study	64
Table 3. Layout of the steady state data.....	73
Table 4. Layout of the continuous time course data	75
Table 5. Summary of activity level data loss for steady state activities	77
Table 6. Participant characteristics for the laboratory study.....	79
Table 7. Summary of development and cross-validation for gyroscope-based youth Sojourn bout detection algorithms.....	86
Table 8. Distribution of key participant characteristics within the subsamples used for development and cross-validation of gyroscope-based bout detection algorithms	87
Table 9. Summary of development and cross-validation for gyroscope-based youth Sojourn neural networks	91
Table 10. Distribution of key participant characteristics within the subsamples used for training, validation, and cross-validation of gyroscope-based neural networks	92
Table 11. Summary of development and cross-validation for gyroscope-based youth Sojourn decision trees.....	95
Table 12. Cross-validation results for each of the accelerometer-based and gyroscope-based youth Sojourn model components	98
Table 13. Distribution of key participant characteristics within the subsamples used for developing and testing the segmentation components in Soj _{recall} and Soj _{precision}	110
Table 14. Activity selections in the coding scheme for video direct observation	124
Table 15. Layout of the free-living data	132
Table 16. Participant demographics for the free-living study.....	135
Table 17. Summary of data loss (participant level) for each analysis of segmentation using the free-living data set.....	142
Table 18. Summary of data loss (participant level) for each analysis of classification agreement using the free-living data set	146

LIST OF FIGURES

Figure 1. Noldus Pocket Observer mobile interface for performing real-time direct observation	67
Figure 2. Summary of participant level data loss and retention for the laboratory data set	76
Figure 3. Depiction of the process by which the original youth Sojourn models classify second-by-second data	83
Figure 4. Visual depiction of the Soj _{recall} algorithm.....	106
Figure 5. Visual depiction of the Soj _{precision} algorithm.....	108
Figure 6. Performance of segmentation algorithms when applied to the testing subset.....	114
Figure 7. Classification agreement of the original youth Sojourn models and the newly developed, gyroscope inclusive Soj _{recall} and Soj _{precision} models when applied to the laboratory data set	117
Figure 8. Summary of participant level data loss and retention for the free-living data set.....	133
Figure 9. Performance of segmentation algorithms when applied to the free-living data set	144
Figure 10. Classification agreement of the original youth Sojourn models and the newly developed, gyroscope inclusive Soj _{recall} and Soj _{precision} models when applied to the free-living data set	148
Figure 11. Measured and predicted percent of free-living time spent in sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA).....	151

CHAPTER I
INTRODUCTION

Physical activity (PA) is an essential part of a healthy lifestyle (1). The general benefits of PA have been known since ancient times (2), but modern empirical research has provided more and more detail about the health benefits and benefit-producing mechanisms of PA (3). One of the first modern empirical discoveries was the connection between physical inactivity and coronary heart disease, as reported by Morris et al. in 1953 (4). The insight was obtained by comparing rates of coronary heart disease among those with inactive occupations (i.e., double-decker bus drivers and postal office workers) and those with active occupations in the same industries (i.e., double-decker bus conductors and postal deliverers).

Since the study by Morris et al. (4), PA has been linked to many aspects of health and wellbeing (3, 5). To establish such links, it has often been necessary to assess PA in ways that provide more detail than the occupation-focused method of Morris et al. That is, information has often been needed about specific aspects of PA (6, 7), such as frequency, intensity, type, and duration, as well as accumulation of PA in various domains (e.g. occupational versus leisure time). That level of complexity has led to the development of many PA assessment methods (7–9), including a wide range of sensor-based methods, i.e., methods involving wearable devices that contain one or more sensors, particularly accelerometers (10).

Sensor-based methods have gained widespread use largely because they are not limited by recall bias and related factors that affect self-report methods, e.g. questionnaires and diaries (10, 11). Furthermore, most wearable devices are reusable, fairly inexpensive, and able to collect movement data continuously. Thus, in addition to their objectivity, they are also scalable for large studies, and able to provide a movement profile with unprecedented detail (12). However, despite the favorable qualities of wearable devices, there are also limitations that have prompted continual innovation on both the devices and the ways they are used (13, 14).

Early innovations were geared toward reducing the physical size of wearable devices, while also increasing their onboard memory. Such changes enabled the collection and storage of more data (e.g. dozens of triaxial samples per second, instead of one uniaxial sample per minute) over periods longer than one week (10). As the technology improved, there were also rapid developments in the area of computation, wherein the focus shifted away from simple regression-based modeling (e.g. for predicting energy expenditure (EE) or activity intensity) and onto more sophisticated approaches, many of which involved machine learning models (15–18).

Recently, several unique avenues have been explored for continuing to improve sensor-based PA assessment. One such avenue is sensor fusion, which is the combined usage of data from multiple types of concurrently-worn sensors, such as accelerometers and gyroscopes (19). Sensor fusion allows models to draw on multiple types of information (e.g. both linear and angular aspects of movement), which creates unique potential compared to the status quo of using only linear acceleration data (20). But despite the novelty of sensor fusion, it is understudied because it has only recently become technologically feasible on a large scale. That is, miniaturization has only recently reached a point where different types of sensors can be housed in a single unobtrusive and inexpensive unit (21). Early work involving sensor fusion has shown that it can be used effectively for predicting EE and activity type (22–24). However, there is room for continued innovation, with one clear gap being the lack of methods that exist for commercially available devices that are also scalable for health research (25, 26).

Another recent focus has been segmentation, which is a pre-processing step in which sensor data are grouped into subunits for further analysis (27). Historically, the convention has been to separate data into epochs of equal duration (27, 28), allowing predictions to be made at regular intervals, e.g. once per minute. While the epoch method is both intuitive in terms of analysis and

expedient in terms of data storage (29, 30), it is also subject to limitations. Previous work has noted that multiple activities can be performed in a given epoch if the intervals are too long, making it difficult to represent activity for the whole epoch using a single prediction (31–34). Alternatively, a short epoch may not have enough data to fully characterize a motion, leading to compromised predictions (35). Therefore, some have attempted to improve sensor-based PA estimates by analyzing the data one activity at a time, where a specialized segmentation algorithm is used to automatically identify the start and stop times for each activity (36–38). The latter approach has been applied most notably in several iterations of the Sojourn accelerometer method (32–34), wherein segmentation is generally designed to capture transitions into and out of stationary behavior. More recently, some work has begun to explore change point detection as a way of performing automated segmentation (39, 40). Change point detection algorithms search for transitions into and out of distinct data distributions, thereby taking a statistical approach to segmentation, rather than a theory-driven approach like Sojourn.

Statement of the Problem

Sensor fusion and segmentation are both fairly new areas of investigation, and the available evidence suggests that both are promising. In particular, several studies have demonstrated the value of using both accelerometer and gyroscope data (24, 25, 41), and others have demonstrated the utility of change point detection for improved segmentation of sensor data (42–44). Despite the promising findings, nearly all studies have used data from small samples of adult participants. Thus, there is scant evidence to suggest whether accelerometer data, gyroscope data, and change point detection can complement one another when applied to data from youth and larger samples.

Statement of the Purpose

The purpose of this dissertation is to calibrate and validate youth Sojourn models (34) that use novel approaches for segmentation (via sensor fusion and change point detection) and activity intensity classification (via gyroscope-based neural networks). The youth Sojourn framework is a logical starting point for this investigation, because of both its existing emphasis on innovative segmentation techniques, and its ability to be iteratively refined over time. Three research questions will be addressed to accomplish the purpose of this dissertation.

Research Questions and Hypotheses

1. **Question:** In a laboratory setting, do gyroscope-based youth Sojourn model components (bout detection algorithm, neural network, and decision tree) perform better than the existing accelerometer-based components?

Hypothesis: Gyroscope-based youth Sojourn model components will achieve better performance than the existing accelerometer-based components when applied to a laboratory data set.

2. **Question:** In a laboratory setting, does the performance of youth Sojourn models improve when using sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification?

Hypothesis: When applied to laboratory data, youth Sojourn models that use sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification will have improved performance, compared to existing youth Sojourn models.

3. **Question:** In free-living settings, does the performance of youth Sojourn models improve when using sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification?

Hypothesis: When applied to free-living data, youth Sojourn models that use sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification will have improved performance, compared to existing youth Sojourn models.

Overview and Significance of the Dissertation

In this dissertation, the research questions are addressed using methods that overlap in some places, while remaining crucially distinct in others. To prevent the distinctions from becoming muddled by the similarities, the document has been strategically arranged, and thus an overview is warranted. In Chapter 2, relevant literature is reviewed. In Chapter 3, a laboratory data set is described, which is then used in Chapter 4 and Chapter 5 to address the first and second research questions, respectively. In Chapter 6, a free-living data set is described, which is then used in Chapter 7 to address the third research question. Finally, an integrated discussion and conclusion is given in Chapter 8.

This dissertation makes an important contribution to PA assessment by providing youth-specific insights into the importance of sensor fusion and change point detection for segmentation, as well as the importance of gyroscope data for intensity classification. This dissertation is also one of the first works (in youth or adults) to show how sensor fusion and segmentation can interact and contribute to the performance of intensity classification models in a large sample. Ultimately, this dissertation benefits sensor-based PA assessment by providing new and improved models that can be deployed in future research involving youth participants.

CHAPTER II
REVIEW OF LITERATURE

Background

Scope and Terminology

Diverse methods are available for assessing physical activity (PA), but each method has limitations (9). Sensor-based PA assessments have become prominent (10), and there is currently a heavy emphasis on refining sensor technology and usage to mitigate their limitations. This review will discuss current issues in sensor-based PA assessment, with a specific focus on methods that use inertial sensors, e.g. accelerometers and gyroscopes (45, 46). Methods that are not sensor-based – such as self-report methods and doubly labeled water (11, 47) – will not be addressed. Unless otherwise noted, the term “sensor” will refer hereafter to an inertial sensor, which excludes physiological and contextual sensors (48). In this review, sensor-based PA assessment will be examined from numerous perspectives including theoretical, historical, technological, procedural, and computational perspectives.

Theoretical Basis for Sensor-Based PA Assessment

PA is fundamentally characterized by bodily movement and the corresponding physiological processes (1, 6, 49, 50). Thus, PA is closely related to the inertia of the body (21), which can be monitored using inertial sensors (48). Such sensors provide a movement profile that can be used to predict several PA variables (45). But even though the sensors exist and there is a clear basis for using them to assess PA, numerous difficulties make such assessments imperfect in practice (51). Some of the difficulties reflect inherent limitations of sensor-based PA assessment, examples of which include the following:

- susceptibility to non-PA-related movements, e.g. from riding in a vehicle (52, 53)
- limited usefulness for activities like resistance training, in which the inertia of the body is altered by design (11)

- differences between localized movement (i.e., sensor data gathered from a single body segment) and the overall movement of the whole body (11, 21, 54, 55).

Other difficulties do not reflect inherent limitations of sensor-based PA assessment, and the latter difficulties have been mitigable (to varying degrees) through continual innovations in both the technology and usage of sensors.

Before exploring current directions in sensor-based PA assessment, it is helpful to survey how the state of the art has developed over time. Thus, the following section will briefly address the history of sensor-based PA assessment, as a primer for the more detailed discussion to follow concerning contemporary issues.

History of Sensors in PA Assessment

SENSOR TECHNOLOGY. In 1961, Cavagna et al. (56) became the first to use an inertial sensor (i.e., a triaxial strain gauge accelerometer) to track human movement. However, it was not until the 1980s that sensor-based PA assessment began to gain momentum (13), with Wong et al. (57) and Montoye et al. (58) presenting the seminal works. They used piezoelectric accelerometers, which measure acceleration via components that deform (i.e., bend or compress) when accelerated. The deformation is transduced into a voltage signal that changes in proportion to the acceleration, and the voltage can then be digitized and converted to units of acceleration (13, 21, 48). Piezoelectric accelerometers are better suited to measuring dynamic accelerations than static accelerations (21), whereas piezoresistive and variable capacitance accelerometers can measure both by transducing deformation into changes in resistance and capacitance, respectively (13, 45). Thus, the latter technologies eventually came to be used more widely than piezoelectrical technologies.

Early devices had to be assembled by hand (48), and were somewhat cumbersome to wear. Advances in miniaturization – especially those achieved through microelectrical mechanical systems (MEMS) technology – caused a dramatic shift in sensor-based PA assessment by enabling devices to become simultaneously cheaper, easier to manufacture, and less obtrusive to wear (45, 48). The origins of MEMS technology are complex, with the earliest developments occurring as early as the 1950s, while major developments began to emerge in the 1970s (59). By 1979, the first MEMS accelerometers had been developed (59–61), but they did not enter mainstream use for PA assessment until after the turn of the century (48). It is important to note that MEMS technology is not exclusive to accelerometers, and the broader applications include some sensors (notably gyroscopes) that are relatively uncommon in PA assessments despite their potential usefulness, which will be discussed later in this review.

As sensing components were simultaneously shrinking and improving, so were other device components, e.g. onboard memory (10). Thus, as monitors were decreasing in size and cost, they were also developing higher capacities for collecting and storing data (10, 48). It even became possible to remove certain components (e.g. analog signal filters) by digitizing their functions and applying them through onboard software known as firmware (13). Altogether, the rapid technological improvements made sensor-based PA assessments increasingly feasible, while also contributing to the evolution of sensor usage.

SENSOR USAGE. Accelerometers have been used more widely in PA assessments than any other sensor (48, 62), and their use has increased steadily in recent years (10, 12, 63). The earliest accelerometers were usually worn on the waist or low back (64–66), which were placements intended to be as close to the center of body mass as possible (67, 68). Data were stored in integrative summaries known as “activity counts” or simply “counts”, which reflected

cumulative movement within regularly-spaced intervals called epochs (52, 58, 68). Memory limitations prevented early devices from storing large amounts of raw data (i.e., acceleration samples collected several times each second), so the epoch length was selected during monitor initialization, and raw data were reduced to count values in real time during data collection, thereby preserving memory (29, 69). Additional memory preservation was achieved by using long epochs, often 60 s (28, 30).

Once a device stopped recording, the count values were extracted and typically converted to either categorical estimates of activity intensity – e.g. sedentary behavior (SB), light PA, or moderate-to-vigorous PA (MVPA) – or continuous estimates of energy expenditure (EE) (70). In most cases, the two were closely related. Specifically, regression equations were developed to predict EE in metabolic equivalents (METs) (70–72), and the predictions could optionally be encoded to intensity using established cutoffs, namely ≤ 1.5 METs for SB, 1.51–2.99 METs for light PA, and ≥ 3.0 METs for MVPA (73, 74). By rearranging the regression equations and solving for fixed MET values, it was possible to define count ranges (i.e., cut-points) that corresponded to each intensity level. Over time, receiver operating characteristic analysis emerged as another popular way to develop cut-points, which did not require the intermediate prediction of EE via regression (75–78).

Virtually every aspect of sensor usage has changed since commercially-available accelerometers were first validated in the 1990s (29, 79). Monitors are now commonly worn on the wrist (10, 67, 68), and analyses increasingly rely on raw acceleration values rather than activity counts (62, 80–82); most analyses use epoch lengths much shorter than 60 s (83), and some analyses are performed on an event-by-event basis, rather than an epoch-by-epoch basis (27, 84); regression and receiver operating characteristic analyses have been downplayed in

favor of complex machine learning models (10, 14, 15, 85, 86), and sensors are now used increasingly to predict diverse outcomes beyond intensity and EE (62, 87–89); sensor fusion – e.g. combining data from accelerometers and gyroscopes – has also emerged as an increasingly viable path to making better predictions (90, 91).

Summary and Transition

In this section, theoretical, historical, and technological perspectives have been used to introduce fundamental concepts in sensor-based PA assessment. In the remainder of this review, those concepts are expanded upon by examining recent developments from procedural and computational perspectives. Afterwards, a synthesis is provided, wherein the evidence is recapitulated to highlight the rationale for this dissertation.

Procedural Factors in Sensor-Based PA Assessment

Sensor Placement

Although sensors have traditionally been placed near the center of mass, alternative placements have been discussed and explored for some time (68, 92–94). As noted previously, wrist-worn monitors are becoming increasingly common. The latter shift will be the primary focus of this section. However, it is important to note that other placement sites (including the thigh and ankle) are sometimes used, instead of the more common hip and wrist placements (95–100).

In the past, there was a twofold critique of wrist-worn monitors. First, it was noted that handedness and small wrist movements (e.g. from gesturing and eating) could complicate the interpretation of wrist-based data (55, 76, 81). Second, in lab-based studies there was typically worse performance for wrist-worn monitors than for hip-worn monitors (92, 101). More recent evidence has shown that wrist-worn monitors can achieve similar performance as hip-worn

monitors, particularly if innovative modeling approaches are used (25, 67, 76, 82, 102). The latter findings have helped to diminish the concerns about wrist-worn monitors, thereby contributing to their increased use. At the same time, precedent has been set for wrist-worn monitor usage, through its adoption in population-level surveillance studies, particularly the United Kingdom Biobank (103) and the National Health and Nutrition Examination Survey (NHANES) (10).

In the case of the NHANES, the decision to use wrist-worn monitors in the cycles from 2011-2014 contrasted with previous use of hip-worn monitors in the cycles from 2003-2006 (10). The switch was made for several reasons, including the ability to assess sleep via wrist actigraphy (104, 105) and the potential for increased participant compliance when asked to wear wrist-worn versus hip-worn monitors (10, 106). It is reasonable to suggest that the increased use of wrist-worn monitors in general practice has been driven by the NHANES precedent, since NHANES accelerometer protocols are often used as a reference for field studies. For example, non-wear screening methods derived from NHANES procedures (107, 108) have gained widespread use, including applications to data from wrist-worn monitors despite being originally developed for data from the hip (109).

Overall, the drawbacks of wrist-worn monitors have been reduced enough to warrant increased use, especially considering the potential for concurrent sleep assessment and better compliance. The next section will discuss similar developments in the area of data format, wherein the drawbacks of raw acceleration data have been reduced enough to warrant increased use compared to activity counts.

Data Format

As discussed previously, activity counts were historically used with long epoch lengths, as a way of conserving device memory (28–30, 69). Today, increased onboard memory allows most sensor-based monitors to store raw acceleration data at high resolution (e.g. 30 Hz) for extended periods (e.g. a week or more) on a single device and battery charge (110, 111). Thus, it is feasible to use raw acceleration data in place of activity counts, which offers two main benefits: First, the raw data stream is potentially richer than the reduced activity counts data stream (69, 112); and second, raw data increases the comparability of output across different device brands (113, 114).

Given the potential benefits of raw acceleration, it was explicitly recommended in 2012 that sensor-based PA assessments move away from using proprietary activity counts (85). At that time, some work had already demonstrated the potential merits of raw acceleration data in activity classification (115) and EE prediction (116). Since then, there has been a gradual transition toward widespread use of raw acceleration data, wherein Hildebrand et al. (117) presented a seminal work in youth and adults wearing multiple accelerometers on the hip and wrist. Numerous processing metrics have been explored for raw acceleration models, and several open-source software tools (most notably the R package GGIR) have been developed to promote field-wide engagement with raw acceleration data processing (118–122).

Although raw acceleration data have made important contributions to sensor-based PA assessments, the science is still nascent, and several challenges have emerged (51, 123). Thus, further investigation is warranted. In the next section, the discussion is focused on a particularly important aspect of raw acceleration data, i.e., the potential for increasing the comparability of values from different monitor brands.

Monitor Brand

There has been a proliferation of commercially-available devices for assessing PA (124, 125), which are generally marketed separately to researchers (29, 81, 93, 103, 126–128) and consumers (129, 130). With so many options, it is difficult to make comprehensive comparisons between different monitors, outside of general usage information, which has been given (for research monitors) by Wijndaele et al. (131). Furthermore, new generations of existing monitors are frequently released in both the research and consumer markets, making comparisons even more difficult (132–135). A full treatment of between-brand differences (especially for consumer monitors) is outside the scope of this review, but there are two notable points to raise.

First, there are always questions about the comparability of sensor outputs from different monitor brands (124). It has long been known that activity count values differ between monitor brands (65, 136), which was largely attributable to differences in how acceleration data were pre-processed before integration (21, 137). As noted previously, raw acceleration data may promote greater comparability of outputs from different monitor brands, which could have several benefits (47, 113, 114, 117, 123). However, some work has shown that even raw acceleration values from different devices may not be interchangeable (138–140).

Second, the issue of device brand (and generation) sometimes intersects with the concept of sensor fusion (91). More than ever, it is now possible for one small device to house multiple sensors (137), which can potentially play a role in setting monitor brands apart from one another (125). For researchers, the best-known example of a multi-sensor monitor may be the SenseWear Armband (no longer sold), which is a device that contains an array of inertial and physiological sensors to measure acceleration, galvanic skin response, heat flux, skin temperature, and near-body temperature (141, 142). For consumers, monitors like the Fitbit and Apple Watch

commonly combine accelerometers with heart rate sensing (via photoplethysmography), and in some cases other sensors such as altimeters (129, 143). Recently, several studies have been published using the ActiGraph GT9X (25, 26, 144–146), which is a research monitor that includes a backward-compatible triaxial accelerometer, as well as an inertial measurement unit (IMU) that houses a secondary triaxial accelerometer, a triaxial gyroscope, a triaxial magnetometer, and a sensor thermometer (25, 26). It was previously noted that sensor fusion is becoming an increasingly important tool for improving sensor-based PA assessments. Thus, the following section will explore previous applications of sensor fusion.

Sensor Fusion

Table 1 presents a summary of key studies involving sensor fusion. For the purpose of this review, the term sensor fusion will refer to any usage of different types of sensors, with no distinction between sensor integration and sensor fusion (19). Furthermore, there will not be an emphasis on distinguishing between competitive, complementary, and cooperative sensor fusion (91, 147). The primary focus in this section is on fusion of different types of inertial sensors, meaning multi-accelerometer approaches (i.e., involving multiple accelerometers worn on different limbs) will not be discussed, with a key example being the Intelligent Device for Energy Expenditure and Activity (IDEEA) (126, 148). Additionally, methods will not be discussed that revolve primarily around fusion of inertial and physiological data (149–151). In this section, another focus is on transparent use of sensor fusion, meaning studies have been excluded if they relied on the “black box” implementations of some devices, notably the SenseWear Armband (152, 153), most consumer monitors (130, 154), and the Xsens Link system (155). In spite of the restrictions that have just been described, some exceptions have been made for especially noteworthy studies.

Table 1. Summary of sensor fusion studies described in the literature review.

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
<i>Activity Recognition Studies</i>						
Pärkkä et al., 2006 (22)	16 adults	DTr (hand built) DTr (standard) ANN	X		X	Overall classification accuracy was between 82% and 86%, depending on classifier.
Ermes et al., 2008 (23)	12 adults	DTr (hand built) DTr (standard) ANN EM	X		X	EM had better classification accuracy (89%) when calibrating with both laboratory and free-living data, than when using only laboratory data (72%).
Lau et al., 2008 (156)	3 adults	SVM ANN RBFN BBN	X	X		SVM had 100% classification accuracy for distinguishing walking up stairs, walking down stairs, and other kinds of walking; other models had lower accuracy, but still > 90%.
Boughorbel et al., 2010 (157)	1 child	QDA	X	X	X	Mean $\pm$ SD classification accuracy was lower when using only accel. data (79.9% $\pm$ 1.5%) than when using all sensors (97.8% $\pm$ 0.2%).
Zhang et al., 2011 (20)	10 adults	DTr	X	X		DTr had sensitivities $\geq$ 96% for all postures and activities.

Table 1 (Continued)

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
Reiss and Stricker, 2012 (158)	9 adults	DTr DTr (bagged) DTr (boosted) NB KNN	X	X	X	Provided a benchmark data set; KNN had highest activity recognition accuracy (89.3%).
Baños et al., 2012 (159)	17 unknown	NCC KNN DTr	X	X	X	Provided a benchmark data set; KNN had highest classification accuracy (> 90%)
Nam and Park, 2013 (147)	10 children	NB BN SVM KNN DTa DTr MP LR	X		X	When using only accel. data, classification accuracy was lower for SVM (86.2%) and DTr (88.3%) than when using accel. and barometer data (98.4% for SVM and 96.3% for decision tree).
Liu et al., 2013 (160)	30 adults	SVM KNN	X		X	Mean $\pm$ SD classification accuracy was lower when using only accel. data (73.9% $\pm$ 14.0% for SVM; 68.7% $\pm$ 13.1% for KNN) than when using data from all sensors (85.1% $\pm$ 6.8% for SVM and 81.4% $\pm$ 9.6% for KNN).

Table 1 (Continued)

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
Moncada-Torres et al., 2014 (161)	6 adults	KNN	X	X	X	Mean classification accuracy was higher when using wrist data (96.1%) than other data (85.7%-89.1%). Gyro. features were algorithmically excluded from all wrist models, suggesting gyro. may have limited value for activity recognition.
Shoaib et al., 2013 (41)	4 adults	NB SVM ANN LR KNN DTa DTr	X	X	X	Gyro. and accel. each had classification accuracy between 73% and 99% when used individually, whereas combining them led to classification accuracy between 87% and 99%. Magnetometer data did not meaningfully contribute.
Shoaib et al., 2014 (162)	10 adults	BN NB SVM LR KNN RBC (version one) RBC (version two) DTr RF	X	X	X	Time and frequency domain features drove detection of stair ascent and descent recognition for both accel. and gyro.; magnetometer data may be useful if a direction-independent feature set is selected.

Table 1 (Continued)

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
Reyes-Ortiz et al., 2016 (163)	3 data sets (9-30 adults)	PSVM	X	X	X	TAHAR was most effective when including transitional data in the training process; TAHAR outperformed reference methods across data sets.
Zebin et al., 2016 (164)	12 adults	SVM CNN	X	X		CNN was more computationally efficient than SVM with no loss in classification accuracy.
Zebin et al., 2017 (165)	5 adults	ANN SVM KNN	X	X		Provided evidence that accel. has greater value in classification than gyro.; prompted additional research regarding combination of the two.
Zebin et al., 2017 (166)	20 adults	DTr (complex) DTr (medium) LDA QDA SVM KNN EM	X	X		Gyro. data contributed less meaningful information than accel., but was still useful in certain scenarios.
Zebin et al., 2018 (167)	20 adults	LSTM ANN SVM Deep ANN	X	X		LSTM ANN achieved similar classification accuracy to SVM, despite using less training data.

Table 1 (Continued)

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
D'souza and R, 2017 (168)	10 unknown	KNN NB SVM CIT DTr RF	X	X		Combining accelerometer and gyroscope data showed promise for achieving higher classification agreement than either can achieve alone.
Other Studies						
Najafi et al., 2003 (169)	44 adults	Custom	X	X		Classifier recognized various types of postural transitions with > 90% sensitivity across laboratory, clinical, and free-living settings
Pärkkä et al., 2007 (170)	11 adults	OLSR	X	X		Within attachment sites (hip, wrist, ankle), accel. models had lower root mean squared error than gyro. models. The effectiveness of accel. versus gyro. data was dependent on the specific activity in question.
Vathsangam et al., 2010 (24)	8 adults	BLR	X	X		Combining accel. and gyro. data led to error reductions up to 58.8% for predictions of oxygen consumption.

Table 1 (Continued)

	Sample	Model(s)	Sensors Used			Key Findings
			Accel.	Gyro.	Other	
John et al., 2011 (171) + Liu et al., 2012 (172)	50 adults	SVM (Gaussian) SVM (linear) KNN LDA NB SVR	X		X	Combining accel. and respiration data led to correct intensity classification of $73.7\% \pm 13.5$ and root mean squared error of 0.4 METs for EE assessment
Reiss and Stricker, 2012 (158)	9 adults	DTr DTr (bagged) DTr (boosted) NB KNN	X	X	X	Provided a benchmark dataset; KNN had highest intensity classification accuracy (96.7%).
Hibbing et al., 2018 (25)	30 adults	2RM	X	X	X	Gyro. effectively distinguished between sedentary behavior and locomotor versus intermittent activities. Across attachment sites (hip, wrists, and ankles), root mean squared error was reduced by up to 14.4% when combining accel. and gyro. data. Magnetometer data did not meaningfully contribute to any models.

Accel.- accelerometer; gyro.- gyroscope; DTr- decision tree; ANN- artificial neural network; EM- ensemble method; SVM- support vector machine; RBFN- radial basis function network; BBN- Bayesian belief network; QDA- quadratic discriminant analysis; SD- standard deviation; NB- Naïve Bayes classifier; KNN- k nearest neighbors; NCC- nearest class center; BN- Bayes net; DTa- decision table; MP- multilayer perceptron; LR- logistic regression; RBC- rule-based classifier; RF- random forest; PSVM- probabilistic support vector machine; TAHAR- transition aware human activity recognition; CNN- convolutional neural network; LDA- linear discriminant

analysis; LSTM- long-short term memory; CIT- conditional interference tree; OLSR- ordinary least squares regression; BLR- Bayesian linear regression; SVR- support vector regression; METs- metabolic equivalents; EE- energy expenditure; 2RM- two-regression model.

Below, studies are presented in two subsections, according to whether the focus was on activity type classification (which was true in most cases) or a different outcome such as EE (which was true in only a few cases). Within each subsection, studies are presented chronologically, with exceptions made in cases where methodological links between studies warrant non-chronological sequence. Before proceeding, it is important to note that the vast majority of evidence has originated from adult participants. The lack of youth-specific evidence will be addressed in the synthesis at the end of this review.

STUDIES ON ACTIVITY TYPE CLASSIFICATION. Pärkkä et al. (22) were among the first to explore sensor fusion. They recruited 16 adult volunteers (three females) to performed a 2-h, semi-structured routine involving tasks in various contexts (e.g. home, transportation, public spaces). Throughout the protocol, participants wore an extensive system that provided 35 data streams from sensors worn on the wrist, chest, finger, forehead, shoulder, and upper back. The sensors ranged from wrist- and chest-worn accelerometers to a microphone (for recording ambient noise) and global positioning system (GPS; for tracking location). Features were calculated in the time and frequency domains, and visually compared to determine which were most effective for distinguishing between lying down, sitting or standing, walking, running, Nordic walking, cycling, and rowing. Three classification models were developed, namely a hand-built decision tree, standard decision tree, and artificial neural network (ANN). Model performance (i.e., classification accuracy) was compared using leave-one-subject-out cross-validation. For individual activities, classification accuracy varied widely, from 22% (ANN, running) to 97% (both decision trees, running). For all activities combined, classification accuracy was highest (86%) using the standard decision tree, whereas the hand-built decision tree and ANN (both 82%) were slightly lower.

In a follow-up to the study by Pärkkä et al. (22), Ermes et al. (23) presented a fusion-based (accelerometer and GPS) classifier that was developed from data collected on 12 adults who were each observed for 6 h. Accelerometers were placed on the hip and wrist, and the GPS was worn in a rucksack. The observation protocol included structured and unstructured activities, and a focus of the study was on changes in classification accuracy when training and cross-validating the models using different subsets of the data. The classification categories were the same as in the study by Pärkkä et al. (22), with the addition of overground cycling and football. Furthermore, the classifiers themselves were the same (hand-built decision tree, standard decision tree, and ANN), with the addition of a hybrid model combining the hand-built decision tree and the ANN. When structured and unstructured activity were included in both the training and cross-validation phases, percent correct classification for the hybrid model was 89%, whereas it dropped to 72% when using the structured data for training and the unstructured data for cross-validation. Thus, the findings underscored the importance of free-living data during model development.

Lau et al. (156) developed a support vector machine (SVM) to predict walking activities using data from shank- and foot-worn devices that each contained an accelerometer and gyroscope. Participants ($N = 3$) wore the devices while performing five walking activities, and the SVM was able to achieve 100% classification accuracy when distinguishing between stair ascent, stair descent, and other forms of walking. The authors also examined other classification models (ANN, radial basis function network, Bayesian belief network) that all had lower classification accuracy than the SVM, yet still $> 90\%$. In some of those cases, modest gains were attained by combining data from the shank and foot.

Boughorbel et al. (157) observed a two-year-old child during indoor activities while wearing a multi-sensor unit (triaxial accelerometer, barometer, and triaxial gyroscope) in the back pocket. They developed a quadratic discriminant analysis (QDA) model to predict seven activities, using video direct observation as a criterion measure. When using only accelerometer data, mean classification accuracy was 79.9% (SD: 1.5%), whereas it increased to 97.8% (SD: 0.2%) when using all sensors.

Zhang et al. (20) developed a decision tree to classify postures, locomotor activities, and transitional activities from tilt angles that were obtained from a combination of accelerometer and gyroscope data. The algorithm was developed using mechanical data (e.g. obtained from an oscillator), after which it was tested on ten adult participants (five females) who wore three multi-sensor units on the chest and both thighs. Each unit contained a triaxial accelerometer and triaxial gyroscope. The participants performed short activity scripts involving postural transitions (e.g. sit-to-stand) and interchanges between standing and walking at different speeds. Specifically, the participants executed 10 sit-to-stand transitions, 10 stand-to-sit transitions, a sequence of stand-to-sit-to-lie-to-sit-to-stand transitions, and sequences of stand-to-walk-to-stand transitions (five repetitions each at preferred, slow, and fast walking speeds). The algorithm had 100% sensitivity for locomotor activities and transitional activities, whereas for individual postures (i.e., sitting, standing, or lying) it had sensitivity of 96%-100%.

Reiss and Stricker (158) and Baños et al. (159) presented benchmark data sets to enable ongoing research involving sensor fusion. Both data sets included primarily data from IMUs (each containing an accelerometer, gyroscope, and magnetometer) worn at various body locations. For the data set of Reiss and Stricker (158), nine adults (one female) wore three IMUs (chest, dominant wrist, and ankle on the same side as the wrist) and a heart rate strap while

performing up to 18 activities. Five modeling approaches were examined to provide benchmark performance metrics, with k nearest neighbors (KNN) achieving the highest overall classification accuracy (89.3%). For the data set of Baños et al. (159), 17 participants (no descriptions given) wore IMUs (nine total) on the back, arms, and legs. The participants performed 33 activities while wearing the sensors in three different ways (potentially differing in both exact placement and orientation) to examine the effect of incorrect sensor placement. Three models were examined to provide benchmark performance metrics, with KNN achieving the highest classification accuracy (> 90%; reported visually only).

Nam and Park (147) studied ten children (16-29 months old; five females) who wore a waist-mounted device containing a triaxial accelerometer, barometer, thermometer, and radio frequency identification chip. Direct observation was used as a criterion measure to code engagement in 11 behaviors (e.g. crawling, rolling), and models were trained to predict activity using eight approaches (e.g. naïve Bayes classifier and KNN). Sensitivity was assessed for individual activities, and averaged across activities as a measure of overall performance. When using only accelerometer data, the best-performing models were SVM, logistic regression, and decision tree (overall sensitivities of 86.2%, 86.9%, and 88.3%, respectively). When adding barometer data, overall sensitivity increased to 98.4% (SVM) and 96.3% (decision tree), with no data reported for logistic regression.

Liu et al. (160) presented a primarily classification-focused application of the “integrated PA measurement system” (IMS) (171, 172), which was an array that included accelerometers (worn on the hip and wrist), respiration sensors (worn on the chest and abdomen), and an ultraviolet radiation sensor housed in the wrist accelerometer enclosure. While wearing the IMS, participants (N = 30, 14 males) performed a treadmill test and one of two activity routines

comprising a variety of intensities from SB to MVPA. Activity classification models were developed using SVM and KNN, and classification accuracy for both models was higher when using data from all sensors ($85.1\% \pm 6.8\%$ for SVM and $81.4\% \pm 9.6\%$ for KNN), compared to accelerometers only ($73.9\% \pm 14.0\%$ and $68.7\% \pm 13.1\%$, respectively).

Moncada-Torres et al. (161) developed models for classifying activities of daily living using data from ReSense devices (173) worn on the chest, both wrists, and both ankles. Each ReSense device contained an IMU (containing a triaxial accelerometer and triaxial gyroscope) and a barometric pressure sensor. Six adults wore the monitors while performing 16 activities ranging from SB to MVPA, and including upper limb-dominant activities such as food preparation and teeth brushing. Time and frequency domain features were calculated, and a specialized feature selection algorithm was used to determine which features should be retained in various KNN models. Classification accuracy was higher for the wrist-worn sensor models ($96.2\% \pm 2.3\%$) than the all other models (mean classification accuracies of 85.7%-89.1%), which included a model that used data from all sensors. Thus, further investigations were focused only on the wrist, wherein a notable finding was that the feature selection algorithm removed all gyroscope features regardless of whether barometer data were included in the model or not. The authors concluded that the gyroscope may lack value for activity recognition, particularly in light of its disproportionate battery consumption compared to other sensors.

Numerous studies have attempted to perform activity recognition using data from smartphone sensors, namely accelerometers, gyroscopes, and magnetometers (174–177). For the present work, smartphone applications have limited relevance, since there are fundamental differences between wearable devices (which are meant to be worn in a fixed location) and smartphones (which are meant to be transferred between various locations, particularly hands,

pockets, and holsters). However, there have been relevant insights in some cases, particularly from the studies by Shoaib et al. (41, 162). In one of their studies (41), participants (N = 4, all males) performed six activities while wearing smartphones on the right arm, right wrist, belt, and right jeans pocket. Seven classification models (e.g. SVM and ANN) were tested using features from individual sensors, as well as combinations. The main finding was that, for all attachment sites and models (except decision trees at the wrist and belt locations), classification accuracy was higher when combining accelerometer and gyroscope data (true positive rates of 87%-99%) than when using either alone (true positive rates of 73%-99%). Models using only gyroscope data tended to have worse performance than those using only accelerometer data, but for certain specific activities (i.e., walking up stairs and down stairs) the gyroscope models performed much better than the accelerometer models. Magnetometer models had the worst performance of all (e.g. true positive rate < 50% for naïve Bayes classifier on the belt), suggesting magnetometers have limited use for activity recognition.

The second study by Shoaib et al. (162) addressed the individual contributions available from each sensor, including which features provided the greatest contribution to the observed performance. Ten participants (all males) wore five smartphones (both jeans pockets, plus right upper arm, right wrist, and belt) while performing the same tasks from the previous study (41), plus cycling. The analysis was robust, involving five attachment sites, nine classification models, three feature sets, three evaluation scenarios, and implementations for both individual and combined sensor use. The findings expanded on what was previously shown (41) by demonstrating that walking detection was primarily driven by time domain features (for accelerometer) and frequency domain features (for gyroscope), whereas for stair ascent and descent, both feature domains were important for both sensors. A notable finding was that it was

particularly useful to combine accelerometer and gyroscope data in situations where classification accuracy was low for each of them individually, and some evidence was provided that magnetometer data can be useful if a direction-independent feature set is used.

Reyes-Ortiz et al. (163) used accelerometer and gyroscope data to perform activity recognition within a novel SVM-based framework called Transition-Aware Human Activity Recognition (TAHAR). Instead of using standard SVM models to predict activity type directly, TAHAR relied on probabilistic SVM models that generated probability estimates. That is, for each data window the models assigned a probability value to each of the possible activities. The final activity predictions were then made based on temporal analysis of probability trends, wherein unique emphasis was given to the recognition of transitional activities. The authors presented two approaches to classifying transitional activities, depending on whether transitional data were excluded during training or not. When excluding transitional data, TAHAR would infer transitional activities based on periods when the probabilistic SVM output gave low probabilities for all activities. When including transitional data, TAHAR could infer transitional activities from affirmative probability evidence. TAHAR was intended for use with accelerometer and gyroscope data from smartphones. However, the authors noted that additional sensors could be readily added, and demonstrations were also given using non-smartphone data. Three data sets were used to evaluate TAHAR using leave-one-participant-out cross-validation, meaning TAHAR was re-trained several times for each data set. For the first data set (178), the key finding was that TAHAR was much less effective at identifying transitional activities when excluding versus including transitional data in model training (classification errors of $7.8\% \pm 8.6\%$ and $0.2\% \pm 0.7\%$, respectively), whereas overall classification error was similar for both approaches ($3.6\% \pm 4.4\%$ and $3.2\% \pm 4.3\%$, respectively). The second data set (158) did not

include transitional data, but the key finding was that classification error was reduced by nearly half when using TAHAR (mean of 5.7%) versus a reference KNN model (10.8%). The third data set (159) also had no transitional data, but classification error was again much lower using TAHAR (mean of 0.5%) than a reference KNN model (4.0%).

In a series of studies, Zebin et al. (164–167) examined activity recognition accuracy using multi-sensor units (each having a triaxial accelerometer and triaxial gyroscope) worn on the waist, both thighs, and both shanks. In one study (164), classification accuracy was compared between a SVM and a convolutional neural network. Data were obtained from 12 adult participants who performed six activities. The main finding was that the convolutional neural network was able to achieve slightly better classification accuracy than the SVM (97.0% versus 96.4%) with a dramatically reduced computational cost (3.5 ms versus 10.6 ms). Another study (165) showed performance of the accelerometer and gyroscope separately. Five volunteers performed six ambulatory activities 10 times each. Several accelerometer-only models (ANN and SVM) were developed to establish proof-of-concept for feature extraction methods, and KNN models were set up to compare the effectiveness of accelerometer versus gyroscope data. Across all attachment sites (individually and combined), findings showed stronger classification accuracy for accelerometer data (83.1%-91.3% precision; 51.3%-62.4% recall) than gyroscope data (52.0%-85.2% precision; 19.4%-46.4% recall). In the next study by Zebin et al. (166), the focus was on combining accelerometer and gyroscope data and comparing classification accuracy across numerous modeling structures (e.g. SVM with quadratic kernel, versus ensemble classifier with bagging and boosting applied). Although results were not presented comparing accelerometer and gyroscope data, the authors noted that gyroscope data tended to contribute fewer meaningful features than accelerometer data. However, they also noted that gyroscope

features were able to improve classification accuracy for activities that involve substantial rotation. In their final study (167), hip-worn data from 20 participants were used to fit a long-short term memory deep recurrent neural network, which was compared to other reference modeling approaches (e.g. SVM). The former model was able to obtain similar classification accuracy to the reference methods, while requiring substantially less training data to do so. Altogether, the studies of Zebin et al. (164–167) provided strong evidence that there is value in using both accelerometer and gyroscope data, even if the applications are more specialized for gyroscopes than accelerometers.

D'souza and R (168) used a publicly-available data set (179) to examine activity recognition. Ten participants (characteristics unspecified) performed 12 activities (ranging from SB to MVPA) while wearing electrocardiography sensors along with accelerometers, gyroscopes, and magnetometers (all triaxial) in various locations. The authors used only accelerometer and gyroscope data from the right wrist and left ankle, with both sensors being worn at both attachment sites. Time domain features were extracted, and activity recognition was performed using various modeling structures (e.g. KNN, SVM). The focus was on comparing classification agreement when using data from single sensors, versus data from sensors grouped by attachment site (i.e., accelerometer and gyroscope combined, separately for the wrist and ankle), and versus data from sensors grouped by type (i.e., wrist and ankle combined, separately for the accelerometer and gyroscope). When using a portion of data from each participant for both training and testing, individual sensors. achieved correct classification of 81.0%-98.6% (wrist) and 75.8%-93.4% (ankle). The gyroscope tended to outperform the accelerometer on the wrist (depending on the modeling structure), while the sensors performed similarly to one another on the ankle. When combining accelerometer and gyroscope data, correct classification increased to

84.4%-100.0% (wrist) and 87.7%-94.3% (ankle). When using accelerometer data alone (wrist and ankle combined), correct classification was 91.9%-98.6%, whereas it was 85.3%-98.1% when using gyroscope data alone. The study provided evidence that gyroscope data can complement accelerometer data, and that combining the two can potentially lead to improved classification compared to using either of them alone.

STUDIES ON OTHER OUTCOMES. Occasionally, sensor fusion studies have focused on outcomes other than activity type classification. For example, Najafi et al. (169) presented a system for detecting postural transitions in elderly participants. The device was a chest-mounted kinematic unit containing two accelerometers and a gyroscope, and assessments were made in three contexts, i.e., a laboratory setting (n = 11 older adults), a clinical setting (n = 24 older adults), and a free-living setting (n = 9 older adults). In each setting, the classifier had sensitivity > 90% for various types of transitions, e.g. sit-to-stand or stand-to-sit.

Pärkkä et al. (170) examined how accelerometer and gyroscope data from the hip, non-dominant wrist, and ankle (same side as wrist) related with EE during common everyday tasks. Specifically, 11 adults (six males) performed activities ranging from light PA (e.g. folding laundry) to vigorous PA (e.g. treadmill running at 85% maximal effort). The analysis was focused on assessing the strength of linear relationships between signal and EE, where indirect calorimetry was used as a criterion measure. Thus, simple linear regression models were fitted, which showed lower root mean squared error for ankle-worn data (1.2 METs for the accelerometer model and 1.3 METs for the gyroscope model) than for hip- and wrist-worn data (1.4 METs for the accelerometer models and 1.7-2.1 METs for the gyroscope models). Generally, accelerometer models performed better than gyroscope models, but there were certain activities (notably stair walking for all three attachment sites) for which the reverse was true.

Vathsangam et al. (24) combined accelerometer and gyroscope data to predict oxygen consumption (VO_2) during walking. Adult participants ($N = 8$, four females) performed treadmill walking at five speeds between 2.5 and 3.5 miles per hour. Throughout the trials, participants wore an IMU (containing an accelerometer and gyroscope) on the hip, as well as a facemask for criterion measurement of VO_2 via a metabolic cart. Data were isolated from steady state periods, and Bayesian linear regression was used to form different predictive models based on different combinations of sensor data. The main finding was that prediction error was lower by up to 58.8% when combining accelerometer and gyroscope data than when using accelerometer or gyroscope data alone.

John et al. (171) and Liu et al. (172) used the IMS to predict a variety of outcomes (i.e., minute ventilation, breathing rate, activity intensity, and EE) beyond activity classification. For the calibration procedures, participants ($N = 50$) performed one of two indoor activity routines involving activities that ranged from SB to MVPA. A rich modeling approach was taken in both studies, whereby different sensor combinations were explored, along with different modeling approaches including SVM, k-nearest neighbors, naïve Bayes, linear discriminant analysis, and support vector regression. The IMS showed promise for predicting activity intensity, with correct classification reaching 73.7% (SD: 13.5%) for the SVM when combining accelerometer and respiration data. Similarly for EE, the best performance (root mean squared error 0.4 METs; mean bias -0.0 METs; mean absolute percent error 6.3%) occurred when using combined data from the accelerometers and the abdominal respiration sensor.

In the previously described benchmark data set of Reiss and Stricker (158), activity intensity (light, moderate, or vigorous) was also examined. The compendium of physical activities (180) was used to assign an EE value for each activity, which was then encoded to intensity. Five

models were examined to obtain benchmark values for correct classification of activity intensity, with KNN achieving the highest value (96.7%).

Hibbing et al. (25) presented two-regression algorithms for GT9X devices worn on the hip, both wrists, and both ankles. In their study, adult participants (N = 30, 10 females) performed 10 activities for seven minutes each while wearing the GT9X devices and a portable indirect calorimeter (for criterion measurement of EE). Three algorithms were developed for each attachment site, the first including only accelerometer data, the second including accelerometer and gyroscope data, and the third including accelerometer, gyroscope, and magnetometer data. During the initial classification stages of the two-regression algorithms (i.e., detection of SB and locomotion), gyroscope data produced better discrimination than accelerometer data, including 100% sensitivity and specificity for SB classification at all attachment sites and $\geq 93.1\%$ sensitivity and specificity for locomotion at all attachment sites. For the hip and ankles, combining gyroscope and accelerometer data led to improvements in root mean squared error (reductions of 0.1-0.2 METs, or 12.3%-14.4%) and mean absolute percent error (reductions of 2.6%-6.0%), whereas improvements were smaller for the wrists. Adding magnetometer data did not meaningfully affect any of the algorithms, as reflected in mean absolute percent error increases of 0.2% (all sites).

Summary and Transition

Over time, sensor usage has changed dramatically in PA assessments. So far, this review has emphasized changes related to data acquisition (i.e., procedural factors), with a particular focus on sensor fusion. The next section will cover changes related to the way data are handled following data acquisition.

Computational Factors in Sensor-Based PA Assessment

To obtain estimates (e.g. EE) from sensor data, three steps are necessary, namely segmentation, feature generation, and model application (15, 27, 181). Each step is reviewed below, along with relevant innovations that have been presented in previous work. Afterwards, methods are discussed that have involved innovations for multiple steps.

Segmentation

Segmentation refers to the way sensor data are partitioned into analytical subunits. The default approach has long been the epoch method, in which partitioning occurs at fixed time intervals, e.g. every second or every minute. The epoch method is a special instance of sliding window segmentation (27) in which the windows do not overlap. However, there are alternative segmentation approaches others have sometimes taken. Apart from overlapping-window approaches, Preece et al. (27) described event-based and activity-defined segmentation as the key alternatives. Event-based segmentation is an approach designed to create segments based on event sequences (e.g. heel strike and toe off), whereas activity-defined segmentation is an approach designed to create a new segment each time a new activity occurs. Both approaches rely on segmentation algorithms to detect when specific events occur or when transitions between activities occur. The difference is that event-based segmentation may be discontinuous and sequence-based (e.g. retaining data only between heel strike and toe off, while discarding data between toe off and heel strike), while activity-defined segmentation is continuous. Event-based segmentation is more common in biomechanical applications, and thus is not highly relevant to sensor-based PA assessments. However, activity-defined segmentation is becoming increasingly important, and warrants discussion. Below, activity-defined segmentation techniques are presented in two broad categories, i.e., dynamic segmentation and change point

detection. A subsection is devoted to each category, with studies being presented chronologically within each subsection except for cases where methodological links between studies warrant non-chronological sequence. Notably, gait-focused studies will not be discussed (182–184).

DYNAMIC SEGMENTATION. It is easiest to think of dynamic segmentation as an extension of the epoch method, wherein extra processing steps are added that allow each epoch to expand or contract until its start and end correspond to the onset and offset of a single activity. There have been a variety of implementations, often with an emphasis on smart home applications (185–191). For this review, smart homes are not highly relevant, and thus the examples given below are from studies focused on wearable sensors.

Kozina et al. (36) presented one of the first dynamic segmentation algorithms for sensor-based PA data, drawing on work that originated in other areas (192, 193). Their algorithm identified runs of decelerating data points, where the overall deceleration exceeded a dynamically computed threshold. For all such runs, the ending points were used to define breaks between consecutive windows, thereby allowing the windows to have different (and dynamically determined) sizes. In an experimental evaluation, 11 participants performed six types of activity (e.g. lying or sitting) and four types of transitions (e.g. sit-to-stand) while wearing accelerometers on the chest and left thigh. Across all activities and transitions, dynamic segmentation led to the higher classification accuracy (92.9%) than sliding window approaches (89.0% and 89.6% for non-overlapping and overlapping windows, respectively).

Noor et al. (37) took a similar approach to Kozina et al. (36), noting that such approaches have been common across a variety of applications. Their algorithm included three decision trees: one for detecting transitional activities, another for classifying the latter activities (e.g. as sit-to-stand or stand-to-sit), and another for classifying the remaining non-transitional activities

(e.g. as walking or sitting). Whenever a transitional activity was identified, the window expansion routine was initiated, wherein successive expansions were performed and compared in terms of probability fit within a multivariate Gaussian distribution. The finalized window size was selected based on the highest probability. Their models were developed using laboratory data from hip-worn accelerometers worn by six participants (five adults and one child) in a controlled setting. Two public data sets were used for independent validation, wherein their new algorithm was compared to a baseline algorithm based on a sliding window approach. In both cases, the new algorithm had overall recall $> 95\%$, and outperformed the baseline method by 5.5%-5.6%.

Gharghabi et al. (38) presented a multipurpose segmentation algorithm applicable to sensor-based PA assessment. Their algorithm took a geometric approach, wherein transition points were identified when the shape of the time series changed. The changing shapes were identified based on arc curves, which were values that changed in relation to the probability of a transition occurring. Based on changes in the arc curves, their algorithm was able to identify transition points. To demonstrate the multipurpose effectiveness of their algorithm, performance was assessed on a variety of data sets, including a physical activity data set. In the latter case, accelerometer and gyroscope data were used to perform the segmentation. A formal analysis was not presented, but visual examples were given showing excellent alignment between the predicted transitions and the labeled sensor data.

CHANGE POINT DETECTION. In contrast to the theory-driven approach of dynamic segmentation, change point detection entails a statistical approach. Specifically, change point detection algorithms search for changes in the statistical properties of time series data. Many algorithms have been developed, for many types of outcomes, as summarized by

Aminikhanghahi and Cook (39). Fundamentally, the algorithms work by labeling each data point with a change score, where higher scores reflect greater likelihood that a transition has occurred. The change scores are determined by comparing the distribution of data on either side of the data point in question, where greater dissimilarity is reflected in a higher change score. Thresholds can then be determined (e.g. using receiver operating characteristic analysis) such that transitions are predicted whenever the change score exceeds a certain amount. A full treatment of change point detection is outside the scope of this review, but several studies are notable either for their inclusion of PA data, or for their implications in sensor-based PA assessments.

Kawahara and Sugiyama (194) presented a novel approach to change point detection, whereby change scores could be obtained without estimating the probability distributions on either side of each data point. Instead, the ratio of the distributions could be directly estimated, which dramatically simplified the algorithm. Liu et al. (195) extended the density estimation approach by presenting the relative unconstrained least squares importance fitting (RuLSIF) method, which has been shown to be one of the most effective change point detection methods across a range of applications (39). For physical activity specifically, Feuz et al. (42) used RuLSIF to identify activity transitions using smart home data from both structured (N = 15 adults, nine females) and unstructured (N = 8 older adults, six females) activity protocols. For both data sets, RuLSIF had true positive rate > 80% and false positive rate < 15%, both of which were better than baseline methods.

Killick et al. (196) presented the pruned exact linear time (PELT) algorithm. Although it has not been applied to PA data, it bears mentioning from the perspective of computational cost. Runtime is an important consideration for change point detection algorithms, and PELT is one of the fastest (196). It is built on a dynamic programming foundation with machine learning

characteristics. Specifically, the algorithm iteratively adjusts to the data in such a way that a cost function is minimized. Killick et al. (196) demonstrated PELT on a variety of data sets (both simulated and real, none involving PA data), showing a much stronger balance of performance and runtime compared to other methods.

Sprint et al. (43) presented a method called physical activity change detection (PACD), which provided a framework for evaluating the significance of change scores. Thus, transitions could be predicted whenever the change score exceeded a significance threshold, rather than based on receiver operating characteristic curves or similar methods. The PACD method was unsupervised, meaning it provided information about data structure (e.g. in terms of participant habits related to bout length) rather than information about deviation from a criterion measure of activity transitions. To demonstrate PACD, the authors combined it with numerous change point detection methods, and used two data sets (one using simulated data based on a Fitbit Charge HR (143), and the other using data from Fitbit Flex monitors (197) worn by 11 older adults) to compare data structure discoveries from each method. The main finding was that PACD was able to detect changes when paired with any of the change point detection algorithms.

Ni et al. (44) used the multivariate online change detection algorithm (198, 199) in a laboratory-based activity recognition framework. Participants ($N = 10$) wore an accelerometer on the wrist while performing 12 activities, including three stationary activities (e.g. sitting), three dynamic activities (e.g. walking), and six transitional activities (e.g. sit-to-stand). After applying the segmentation algorithm, various classification models were applied, of which random forest performed the best (95.5% correct classification). Agreement decreased when using standard segmentation methods, namely sliding windows with 50% overlap (94.7% correct classification) and no overlap (93.6% correct classification). For recognition of specific activities, the results

generally followed similar trends (especially for transitional activities), with the segmentation algorithm achieving the highest sensitivity, followed by sliding windows with 50% overlap and no overlap. Of studies focused on physical activity data, the study by Ni et al. (44) provided some of the strongest evidence in favor of change point detection.

Recently, Twaites et al. (40) presented an adaptation of a method known as online Bayesian change point detection (OBCPD)(200), which runs quickly and allows multivariate data. Data were obtained in a laboratory setting, as well as in free-living environments. In the laboratory study, 16 participants (characteristics unspecified) wore a GENEActiv accelerometer on the right wrist while performing repeated transitions (sit-to-stand, stand-to-sit, sit-to-walk, walk-to-sit) at specific intervals. Direct observation was the criterion measure of when the transitions occurred. In the free-living study, 47 participants (characteristics unspecified) wore a GENEActiv accelerometer on the right wrist and an activPAL device on the thigh (side unspecified) for one week. The activPAL data were used to identify postural transitions, which were then used as a criterion measure. In both studies, performance comparisons were made between OBCPD and existing dynamic segmentation methods (32, 36, 37). For the laboratory data, all methods were tested using different combinations of accelerometer axes, as well as with and without a refractory period, if applicable. OBCPD had Matthews correlation coefficients (201) of 0.38-0.76 (versus 0.18-0.58 for the existing methods) and root mean squared errors of 3.17-7.99 s (versus 4.42-7.43 s for the existing methods). For the free-living data, methods were tested with and without a refractory period, and only two comparison methods were tested (32, 36). The analytical emphasis was placed on sensitivity and mean minimum distance as performance metrics, as well as ratios of those metrics to the values obtained when predicting the same number of transitions at equally spaced intervals. OBCD had sensitivity of 80.2%-83.1%

(compared to 63.2%-87.2% for the existing methods) and mean minimum distance of 3.54-3.67 (compared to 5.16-8.73 for the existing methods). OBCD had ratio metric values from 2.40-3.21 (compared to 1.74-3.02 for the existing methods). The study presented one of the first potentially usable change point detection methods for sensor-based PA assessment.

Feature Generation

Feature generation refers to segment-level data aggregation that prepares the data for input into the actual predictive model (15, 27, 202). An intuitive example is the computation of activity counts, wherein raw data are first gathered within each segment (i.e., each epoch), then filtered, rectified, and summed for each segment (21). The latter process is often treated as a downsampling step that occurs prior to feature generation, but it is actually per se a simple example of feature generation.

In machine learning settings, a wide range of features are used, some of which require much more complicated computations than activity counts. Generally, features are classified as heuristic, time domain, frequency domain, or time-frequency domain (27, 203). Apart from the features themselves, there are methods that can guide the decision of which features to use. Such feature selection methods are categorized as filter methods, wrapper methods, and embedded methods (204, 205). An exhaustive treatment of feature generation is outside the scope of this review, and many of the important feature-related trends in sensor-based PA assessment will come up in the following section on model application. Nevertheless, several studies have made feature-specific contributions that warrant mention here.

Preece et al. (206) compared classification accuracy for a KNN model when using 14 different feature sets, seven using time-frequency domain (i.e., wavelet) features, and seven using other features from the time and frequency domains. Data were obtained from 20 adult

participants (10 males) who wore accelerometers on the hip, thigh, and ankle while performing eight activities that ranged from walking to running and jumping. Video direct observation was used as a criterion measure of behavior timing, and participants were given a high level of autonomy to determine the pace of each activity and the timing of transitions. When using hip, thigh, and ankle data in the KNN model, classification accuracy was $87.3\% \pm 12.1\%$ for the time and frequency domain feature sets, compared to $77.4\% \pm 13.6\%$ for the wavelet feature sets. Thus, the study provided evidence that wavelet features may have limited utility for recognizing MVPA.

Kate et al. (207) compared feature sets and classifiers using hip-worn accelerometer data collected from 146 adults (67 males). Participants completed six treadmill walking tasks (speed from 1.5 to 4.0 miles per hour) and five activities of daily living (e.g. computer work, moving boxes) while wearing the accelerometer and a portable indirect calorimeter for criterion EE assessment. Nine predictive models (e.g. SVM, random forest, and linear regression) were fitted to predict activity type or EE. Many features were calculated and compared, being broadly categorized as basic statistical features (e.g. percentiles and lag-1 autocorrelation), all statistical features (e.g. coefficient of variation, skewness, and zero crossings), distance based features (e.g. dynamic time warping), and ngram features (which involve specific techniques from document analysis). Classification accuracy was highest when using the SVM, so it was used as the basis for feature comparisons, which showed that it was more effective to combine several categories of features, rather than using any category alone. Specifically, the highest classification accuracy (86.7%) was obtained when using all statistical features and derivative dynamic time warping, whereas classification accuracy for individual feature categories ranged from 65.8% (derivative dynamic time warping) to 81.8% (ngram features). For EE estimates, there was limited

distinction between models and feature sets, which prompted the authors to advocate for predictive frameworks that first classify activities, then use specialized models to predict EE based on the activity classification.

Kheirkhahan et al. (208) compared features using hip-worn accelerometer data collected from 1135 older adults (778 females) in free living environments. The study objective was to identify the feature (or features) that most effectively enabled objective detection of mobility impairment, using clinical assessments as a criterion measure. Features were calculated based on individual axes, vector magnitude, and steps. The initial total was 67 features, of which 37 were removed due to multicollinearity. Based on five feature selection approaches, models were able to correctly classify 76.7%-79.3% of participants, with a variety of features (e.g. magnitude-based, variability-based, timing-based, and step-based) contributing to the performance. Thus, the study provided evidence that several features are necessary to identify mobility impairment.

Apart from studies that have directly examined specific features, some studies have adopted feature selection methods that allow a model to self-identify useful features. Although feature selection methods are not the primary focus in this section, notable applications have been slow feature learning as presented by Koch et al. (209), and various methods (both supervised and unsupervised) presented by Kheirkhahan et al. (210). Such methods have not been widely adopted in general practice for sensor-based PA assessment, but they may become increasingly important as sensor-based PA assessment continues to evolve.

Model Application

Once features have been calculated for each segment, the features can be fed into a model that predicts a specific outcome variable, e.g. EE. In the preceding sections, there have been references to many different modeling structures, along with several different outcome variables.

In this section, the objective is not to provide an exhaustive overview of different models and how they work, but rather to highlight key studies that have pioneered the use of complex models (especially those involving machine learning) in sensor-based PA assessment.

Linear regression and receiver operating characteristic cut-points have already been discussed as the traditional approaches to producing estimates from sensor data. Pober et al. (18) were among the first to demonstrate alternative approaches, namely QDA and hidden Markov model (HMM), both of which they used to classify various activities from simple features (mean and standard deviation). Six adults (four males) performed four activities while wearing a hip-worn uniaxial accelerometer. On average, QDA correctly recognized activities for 70.9% of time points, whereas HMM correctly recognized 80.8%. The study provided a strong proof of concept for the use of machine learning in sensor-based PA assessment.

Rothney et al. (116) were among the first to take a machine learning approach for EE prediction. Furthermore, they were among the first to create a model using raw acceleration data. Data were collected from 102 adults (46 males) who were assessed in a room calorimeter for approximately 24 hours, while wearing a hip-worn ActiGraph accelerometer and a modified IDEEA array. During the protocol, participants performed two scripted activity routines that elicited a range of intensities from SB to MVPA. For the hip-worn accelerometer, an ANN was developed to predict EE from five minute-by-minute features (peak value, interquartile interval, minimum 10-s coefficient of variation, sum of signal power < 0.7 Hz, and sum of signal power > 0.7 Hz), plus several demographic features (sex, age, height, body mass, and ethnic background). The model performance was then compared against the criterion measure (i.e., room calorimeter) as well as the proprietary IDEEA algorithm and an existing ActiGraph regression-based model. Mean absolute error was lower for the ANN ($0.3 \pm 0.1 \text{ kcal} \cdot \text{day}^{-1}$) than the IDEEA (0.4 ± 0.1

kcal·day⁻¹) or regression-based model (0.5 ± 0.1 kcal·day⁻¹), which showed that sophisticated modeling could improve EE prediction without the limitations of obtrusive arrays like the IDEEA.

Two more ANN models were presented by Staudenmayer et al. (211), one for predicting EE and another for predicting activity type. Data were obtained from 48 adults (24 females) who wore a hip-mounted accelerometer and portable indirect calorimeter while performing 6-18 activities that ranged from SB to MVPA. The features for both models were five percentiles (10th, 25th, 50th, 75th, and 90th) and the lag-1 autocorrelation. The EE ANN was compared against three existing methods, and shown to have a lower root mean squared error (1.2 METs) than the other methods (1.6-2.1 METs). The activity type ANN correctly classified activities into one of four groups 88.8% of the time. The study had notable impacts on future research, including a replication study (with a larger training sample of 277 adults, plus an independent validation sample of 65 adults) by Freedson et al. (212). In that study, root mean squared error was 1.9 METs, whereas two reference measures had root mean squared error of 2.0 METs and 2.1 METs. For activity type classification, the ANN gave a correct prediction 80.9% of the time. Both the original study (211) and the replication study (212) were instrumental in the development of Sojourn accelerometer models (32–34), which will be a key focus in the following section.

Zhang et al. (82) were among the first to apply machine learning techniques to data from a wrist-worn monitor. Data were obtained from 60 adults (24 males) who wore GENEActiv accelerometers on both wrists and the hip. Participants performed 10-12 activities across a range of intensities, and several classification models (decision tree, naïve Bayes, logistic regression, SVM, and ANN) were fitted to predict engagement in four activity categories. Features were

calculated from the signal vector magnitude (12.8-s sliding windows) and included the mean and standard deviation, plus nine features drawn from fast Fourier transform and discrete wavelet transform. Classification accuracy was similar for all modeling structures, ranging from 95.3%-97.0% for either wrist, compared to 98.2%-99.6% for the hip. Thus, the study provided strong evidence to support the use of frequency and time-frequency domain features, and it was one of the first to show that doing so can narrow the performance gap between hip-worn and wrist-worn monitors.

Ellis et al. (213) developed one random forest for EE prediction and another for prediction of activity type. Data were obtained from 40 adults (21 females) who performed six activities each (from a list of nine) while wearing three ActiGraph accelerometers (left and right hip, plus non-dominant wrist) and a portable indirect calorimeter. For the EE prediction model, root mean squared error (1.0 METs for both hip-worn monitors; 1.1 METs for the wrist) was lower than three comparison methods (1.1-1.5 METs), which included the ANN of Staudenmayer et al. (211). Additional improvements were shown when combining data from various combinations of the three accelerometers, plus heart rate data. For the activity type prediction model, classification accuracies were 70.2% (right hip), 72.5% (left hip), and 80.2% (wrist), with minimal improvements obtained by combining accelerometers and heart rate data (maximum classification accuracy of 83.6%). The study was one of the first to apply random forest modeling in sensor-based PA assessment.

Staudenmayer et al. (16) collected data from 20 healthy adults (10 females) who wore an ActiGraph accelerometer on the dominant wrist while performing a range of activities. A portable indirect calorimeter was used as the criterion measure of EE. Multiple models were developed, including simple models for predicting EE (multiple linear regression) and

classification activity intensity and basic type (decision tree). They also developed more complex models, i.e., ANN, SVM, and random forest. Features were mean and standard deviation, plus acceleration angles and frequency domain features. The primary analyses were focused on EE predictions, wherein root mean squared error was lowest using the random forest model (approximately 1.2 METs) than other models (approximately 1.3 METs for SVM and 1.4 METs for ANN, up to > 1.7 METs for regression-based methods). A key contribution of the study involved dissemination of the new methods, which were able to be subsequently implemented in several studies (214, 215) unlike many other novel methods in sensor-based PA assessment (216).

Chowdhury et al. (217) provided a novel demonstration of ensemble learning, in which predictions are drawn from several models and combined to obtain a final estimate. Wrist-worn accelerometer data were drawn from three data sets that included adult and youth participants. A total of 45 features were calculated (10-s windows) from the time and frequency domains. Four standard classification models were fitted (decision tree, KNN, SVM, and ANN), and six ensemble techniques were compared, of which three were conventional ensemble techniques (i.e., bagged decision tree, random forest, and boosted decision tree) and three were customized (i.e., weighted majority voting, naïve Bayes combination, and behavior knowledge space). The key classification metric was the F1 score, with overall accuracy being represented as an average across all activities. Of the conventional ensemble methods, random forest performed the best for all three data sets, reaching average F1 of 79.7-85.0. When using standard classifiers, SVM was generally the strongest, with average F1 of 73.5-83.2. For the customized ensemble methods, weighted majority voting was the most effective technique, with average F1 of 76.9-86.0. The study made a novel contribution to sensor-based PA assessment by demonstrating ensemble

learning as a potential way to combine the strengths of various classification models while counterbalancing their limitations.

Trost et al. (218, 102, 219) have provided key extensions of previous work by examining youth-specific applications and the importance of attachment site for machine learning model performance. In an initial study (218), ANN models were developed to predict EE and activity type in youth, similar to the approach taken by Staudenmayer et al. (211) in adults (including the use of the same features, calculated in various window sizes). The study included 100 participants who wore a uniaxial ActiGraph accelerometer on the right hip, along with a portable indirect calorimeter for criterion measurement of EE. The participants completed 12 activities, which were grouped into five categories. Overall classification accuracy was > 80%, regardless of window size, with greater agreement in 60-s windows (88.4%) than 10-s windows (81.3%). For EE prediction across window sizes, root mean squared error was 30%-40% lower for the ANN (0.9-1.1 METs) than two established regression equations (1.3-1.5 METs).

A subsequent study (102) was devoted to addressing performance differences between hip and wrist data when classifying activity type. Data were obtained from 52 children and adolescents (28 males) who wore triaxial ActiGraph accelerometers on the hip and non-dominant wrist while performing 12 activities that were grouped into seven categories. Features were the same as Staudenmayer et al. (211), with the addition of several time domain features, e.g. skewness and zero crossings. Regularized logistic regression was used for the classification, and results showed 91.0% correct classification for the hip model, versus 88.4% for the wrist. Thus, the study provided youth-specific evidence that advanced modeling can lead to similar performance at both sites.

Trost et al. (219) presented additional data collected from preschoolers (N = 11, six females) who performed 12 activities while wearing triaxial ActiGraph accelerometers on the hip and non-dominant wrist. The activities were grouped into five categories, and two classification models (random forest and SVM) were developed using similar features (time and frequency domain) to previous work (15, 102). The mean percent correct classification was 78.1%-81.8% for the hip and wrist individually or combined. Similarly for the SVM, percent correct classification was 80.4%-85.2%. Weighted kappa scores ranged from 0.76-0.84, which was considerably higher than comparison methods ($\kappa = 0.49-0.65$) that were developed using traditional receiver operating characteristic analyses.

Novel Methods

Traditional analyses (e.g. regression equations and receiver operating characteristic cut-points) took primitive approaches to segmentation, feature generation, and model application. Segmentation and feature extraction were accomplished by calculating epoch-by-epoch activity counts, and model application was accomplished by plugging the count values into the corresponding model. The previous sections have shown that sensor-based analyses have evolved in all three areas. The evolution has made it possible to address more outcomes beyond activity intensity and EE. Intille et al. (62) listed nine outcomes of interest in PA assessments, many of which have been approached from a sensor-based perspective. One key example is identification of individual activities, which can potentially lead to subsequent EE and intensity predictions (via activity-specific models) that are more effective than predictions from a catchall model. The latter paradigm has been referred to as “divide and conquer” (220, 221), and it is one of two major innovative categories, with the other being Sojourn methods. Both categories are discussed below.

“DIVIDE AND CONQUER” PARADIGM. Two-regression models (2RMs) are the best-known example of the divide and conquer paradigm. In a 2RM, there are two regression equations that correspond to different activity categories. Simple signal features are used to determine which equation to use on an epoch-by-epoch basis. Additionally, low-magnitude signal is assigned a static resting EE value. One of the earliest 2RMs was presented by Heil et al. (93), who calibrated 2RMs for Actical accelerometers worn on the hip, non-dominant wrist, and ankle (same side as wrist monitor). The calibration procedures involved 24 children (14 males) and 24 adults (12 males) performing 10 activities (ranging from supine rest to treadmill jogging) while wearing the accelerometers and a portable indirect calorimeter. Their algorithms involved two count-based cut-points that were visually determined. Below the first cut-point, the static resting EE value was assigned. Between the two cut-points, a linear regression was invoked based on data from sitting and cleaning activities. Above the second cut-point, a linear regression was invoked based on data from locomotor activities. For youth, mean total activity EE estimates differed from the criterion by 12.0 kcal (10.0%) for the ankle and ≤ 2.8 kcal (2.3%) for the hip and wrist. For adults, the differences were 27.4 kcal (18.1%) for the ankle, 7.1 kcal for the hip (4.7%), and 31.1 kcal (20.5%) for the wrist. Crouter and Bassett (222) presented another 2RM for hip-worn Actical devices, which was subsequently refined and compared to the Heil models by Crouter et al. (223). In the latter study, a 6-h free-living protocol was used, wherein 29 adults (17 females) performed self-selected behaviors while wearing a hip-worn Actical and a portable indirect calorimeter (which was used as the criterion measure of EE). Mean EE was within 0.1-0.2 METs of the criterion for both the refined Crouter 2RM and the Heil 2RM. However, activity time estimates were more favorable for the former model. Specifically, estimated time spent in SB and light PA was within 0.7%-5.3% of the criterion for the refined Crouter 2RM, versus

7.2%-29.7% for the Heil 2RM. Errors were larger for time spent in moderate or vigorous PA, but again the refined Crouter 2RM was substantially better (within 28.4%-86.4% of the criterion) than the Heil 2RM (within 100.0%-108.5%).

Following the study of Heil et al. (93), Crouter et al. (17) presented the first of several 2RMs for hip-worn ActiGraph accelerometers. The key novelty of the Crouter 2RMs was that, instead of using count-based cut-points to select from different linear EE prediction equations, a variation-based cut-point (i.e., signal coefficient of variation) was used to determine whether a log-linear equation (for locomotion) or cubic equation (for lifestyle activities) should be used. Data were obtained from 48 adults (24 females) who wore a hip-worn uniaxial accelerometer and portable indirect calorimeter (the criterion measure of EE) while performing up to 17 tasks ranging from SB to MVPA. Across all activities (mean criterion EE of 4.2 ± 2.7 METs) except cycling, the 2RM predicted 4.2 ± 2.6 METs, which was substantially closer to the criterion measure than three standard regression equations (means of 3.5, 4.0, and 4.4 METs). The 2RM only examined the coefficient of variation at the start of each minute, meaning it could not detect changes in variability that occurred within the minute. To overcome the latter limitation, a refined 2RM was published by Crouter et al. (31), in which coefficient of variation was determined on a rolling basis, rather than the original approach that used non-overlapping windows. On the same data set, the refined 2RM estimated 4.1 ± 2.6 METs and had a root mean squared error of 0.6 METs (compared to 0.7 METs for the original 2RM). Crouter et al. (224) performed an independent validation in a free-living sample of 29 adults (17 females) who were observed in 1-2, 6-h sessions. Compared to the criterion measure (indirect calorimetry), the refined 2RM produced higher mean estimates of light, moderate, and vigorous PA (by 9.6%-

63.6%) and a lower mean estimate of SB (by 20.8%). When considering baseline methods, there were equivocal findings concerning which performed the best.

A youth-specific 2RM was also developed by Crouter et al. (225). Data were obtained from 109 youth (57 males) who wore hip-mounted triaxial ActiGraph accelerometers and a portable indirect calorimeter while performing one of three activity routines, each of which involved six activities across a range of intensities. Two 2RMs were developed, one using vertical axis counts, and the other using vector magnitude counts. Both 2RMs were developed using the same approach as the refined 2RM (31), except the units of EE were youth metabolic equivalents (notated MET_{RMR}). For all activities combined, the mean criterion EE was $3.1 \pm 2.0 \text{ MET}_{\text{RMR}}$. For both 2RMs, mean EE was within $0.2 \text{ MET}_{\text{RMR}}$ of the criterion, whereas for four reference methods the means differed by $0.3\text{-}0.4 \text{ MET}_{\text{RMR}}$. Thus, the study showed that 2RMs could be similarly effective in youth as adults. Crouter et al. (226) performed an independent validation in a sample of 72 youth (40 males) who performed structured activities or an approximately 2-h free-living assessment. For the structured activities, both youth 2RMs had root mean squared error between 0.1 MET_y (supine rest for the vector magnitude 2RM) and 2.6 MET_y (running for both 2RMs). When averaged across all activities, root mean squared error was $1.3\text{-}1.4 \text{ MET}_y$ for the youth 2RMs, and $1.3\text{-}1.6 \text{ MET}_y$ for baseline methods. In the free-living investigation, root mean squared error was $1.5\text{-}1.7 \text{ MET}_y$ for all methods.

Dongwoo and Kim (227) presented data from one adult male who performed five activities while wearing a portable indirect calorimeter (used as a criterion measure of EE) and accelerometers on the hip, wrist, and ankle (sides not specified). Results were presented for regression equations developed under three conditions. For the first condition, four regression equations (one for each sensor, plus another for all sensors combined) were fitted across all

activities combined. For the second condition, two regression equations (both using data from all sensors combined) were fitted, one for arm-dominant activities and the other for leg-dominant activities. For the third condition, four regression equations (each using data from all sensors combined) were fitted, again for arm-dominant and leg-dominant activities, but with an additional layer of subdivision based on high versus low intensity. The authors did not present a method for classifying data into any of the categories, and thus the analysis was strictly focused on comparing error with and without the use of specific regression equations for activity and intensity subgroups. For the first condition, standard error of the estimate was much lower when using all sensors combined (1.2 kcal) than any of the individual sensors (1.6-2.2 kcal). Standard error of the estimate decreased even further in the second (0.8 kcal) and third conditions (0.7). Although the study had extremely limited scope, it is notable for being one of the first to show the potential benefits of the divide and conquer paradigm outside of the 2RM framework.

Bonomi et al. (228) used a divide and conquer approach to evaluate EE estimates in 15 free living adults (nine males). Doubly labeled water was the criterion measure of EE, and participants wore a triaxial Tracmor accelerometer on the lower back for five days. A previously-developed decision tree was used to classify accelerometer signal into one of six activity categories (lie, sit-stand, active standing, cycle, walk, run) based on primarily time domain features (means and standard deviations of different axes, peak-to-peak distance) and one time domain feature (peak frequency). The compendium of physical activities (180) was used to assign a MET value for each activity, and daily aggregates were compared to doubly labeled water in terms of physical activity level (PAL, i.e., the ratio of total EE to basal metabolic rate), total EE, and activity EE. For the PAL analysis, simple linear regression was used to examine variability accounted for by the accelerometer-derived daily MET estimates. Additionally, PAL

was regressed against total daily activity counts. For total EE and activity EE, multiple regression was used to fit models where the key predictors were total activity counts or accelerometer-derived METs. Additional covariates were added from three categories, namely basal metabolic rate, basic physical characteristics (e.g. body mass), or advanced physical characteristics (e.g. fat-free mass). Separate models were formed for covariates in each category, and stepwise variable selection was used to determine which variables had the greatest explanatory value. For the PAL analysis, accelerometer-derived MET estimates explained 51% of the variability, compared to 46% explained by total activity counts. For total EE and activity EE, the percentage of explained variability always increased when using the accelerometer-derived MET estimates compared to using total activity counts, with increases of 9%-13% except when predicting total EE with basal metabolic rate as a covariate (increase of 2%). Although none of the models explained more than 76% of the variance for PAL, total EE, or activity EE, the study provided some free-living evidence in support of the divide and conquer paradigm.

Albinali et al. (72) presented a method designed to extend the Crouter 2RM framework. Data were obtained from 24 adults (10 males) who performed up to 22 activities while wearing a portable indirect calorimeter (criterion measure of EE), a hip-mounted uniaxial ActiGraph accelerometer, and additional triaxial accelerometers (not made by ActiGraph) on the hip, thigh, and upper arm (all on the dominant side). The algorithm used a decision tree to initially classify activity type using features from the time domain (axis distances, variability, and correlations) and frequency domain (entropy and spectral peaks). Regression equations were then invoked, which were specifically fitted using data from the activity predicted by the decision tree. The authors also presented algorithms that used the compendium of physical activities (180) in different ways (rather than regression equations), but the latter approaches will not be reviewed

here. Assuming perfect classification of the activities, the regression equations had mean error of 1.5% across participant sessions. When using the decision tree to classify activity, mean error increased to 3.9%. Four reference methods were also evaluated, having mean error of 29.2%-38.5%. Thus, the study demonstrated the merit of using many activity-specific regression equations, thereby providing a key extension of the Crouter 2RM while also confirming the value of the divide and conquer paradigm.

Ruch et al. (229) examined EE predictions from an ANN versus activity-specific regression equations. Data were obtained from 43 youth (22 females) who performed eight activities while wearing a portable indirect calorimeter (criterion measure of EE) and hip-mounted triaxial ActiGraph accelerometer. Features (1-s epochs) included individual axis counts, vector magnitude, steps, and posture (obtained from the inclinometer function). Demographic variables were also included in the models. The activity-specific regression equations were fitted using backward selection of individual variables. Additional models were also tested, including reference methods (e.g. 2 RMs) and a variant of the activity-specific equations in which a decision tree was first used to classify activity and thereby determine which equation to use. The classification accuracy of the decision tree was also compared to a classification focused ANN. For all activities except crawling, absolute mean bias was lower using the ANN than the activity-specific equations, with the margins ranging from 0.1 METs to 4.2 METs. When combining the decision tree with the activity-specific equations, absolute mean bias was lower than the ANN (by 0.1-2.0 METs) for all activities except cycling and running. Similarly, the combination of decision tree and activity-specific equations led to lower absolute mean bias (by 0.3-24.1 METs) than the reference methods, except for during walking and crawling using the Crouter vertical axis 2RM. Activities were correctly classified 60.5% of the time for the decision tree, versus

63.8% for the ANN. The study provided evidence that the divide and conquer paradigm may enable better EE predictions than catchall models (even machine learning models), with the added benefits of greater transparency, simplicity, and ease of use.

Mu et al. (230) presented an approach called Bipart, which was broadly similar to the method of Ruch et al. (229). The Bipart algorithm first clustered data using a distance-based and block structured approach. An activity classification was then assigned, which was subsequently used to determine which of several EE prediction equations should be used. Data were obtained from 112 children who performed supine rest and a subset ($n = 6$) of 18 possible activities.

Throughout the trials, the participants wore a triaxial ActiGraph accelerometer on the hip, plus a portable indirect calorimeter (which was used as the criterion measure of EE). For each activity, multiple linear regression was used to develop a prediction equation based on accelerometer percentiles and autocorrelative features. The Bipart method had root mean squared error of 1.4 METs when using leave-one-participant-out cross-validation, whereas comparison methods (in which the initial classification was made using more common models such as KNN) had root mean squared error between 1.4 and 2.6 METs. The study had similar implications to that of Ruch et al. (229), namely that effective EE estimation can be achieved using simple regression if there are multiple activity-specific equations to choose from, and an adequate way of recognizing which activity is being performed.

Hibbing et al. (25) used 2RMs as a way to provide proof of concept for sensor fusion in the estimation of EE. The study has already been summarized (see “Studies on Other Outcomes” under “Sensor Fusion” in the procedural factors section), but it bears mention here as one of the most recent examples of both 2RMs and the broader divide and conquer paradigm.

SOJOURN MODELS. Of the innovative methods in sensor-based PA assessments, Sojourn models represent a minority that have included a distinct focus on dynamic segmentation. The term “Sojourn” refers to the underlying segmentation algorithm, which attempts to identify transitions between periods of non-movement and sojourns of movement (i.e., sustained periods of elevated accelerometer signal). After identifying the sojourns and thereby segmenting the data, predictions are made for each segment using techniques that differ depending on which Sojourn variation is being used. Specifically, there are three variations (32–34), which will be described below.

Lyden et al. (32) presented the original Sojourn model, including separate versions for uniaxial versus triaxial data. Detailed descriptions of the two algorithms are beyond the scope of this review, but in general they both employed an approach whereby bouts of homogenous movement data were first identified using a Sojourn algorithm. Next, a decision tree was used to classify each bout as either activity or one of several types of inactivity. Finally, EE values were assigned using static values for the inactivity categories and ANN predictions for activity bouts. The Staudenmayer et al. (211) neural network was used for the last step, meaning features were the same as previously described (i.e., percentiles and lag-1 autocorrelation). The study used data from 13 adults (eight females) who were each observed in free-living for approximately 10 h while wearing a hip-mounted triaxial ActiGraph accelerometer. Direct observation was used as a criterion measure of behavior and EE. Six participants were randomly selected for model development, and the remaining seven were used for cross-validation, wherein root mean squared error was lowest for the Sojourn models (1.0 METs for both uniaxial and triaxial), compared to reference models for which root mean squared error ranged from 1.4 METs (Crouter refined 2RM) to 5.4 METs (Staudenmayer et al. (211) neural network).

A novel aspect of Sojourn models is their ability to be iteratively refined over time. The study by Ellingson et al. (33) provided an example of such refinement, which was achieved by fusing the existing Sojourn model with activPAL data, enabling more effective discrimination between SB and other forms of low-acceleration activity. Data were obtained from 49 adults (31 females) who wore a hip-mounted triaxial ActiGraph accelerometer, plus a thigh-mounted activPAL accelerometer and a portable indirect calorimeter (for criterion assessment of EE) while performing 15 activities in a laboratory setting. Activities ranged in intensity from SB to MVPA. The new algorithm used activPAL data to identify postural transitions and add them to the list of transitions detected by the Sojourn bout detection algorithm. Furthermore, the activPAL data were used to adjust EE estimates for inactive periods, such that standing activities would have $EE \geq 1.5$ METs and seated activities would have $EE < 1.5$ METs. The EE estimates from the original and updated models were similar (within 0.2 METs of one another for nine of 15 activities; within 0.8 METs of one another for all but one activity), but classification of activity intensities was greater for the new model (overall correct classification of 78.7%) than the original (56.0%). The latter trend was driven primarily by much greater detection of light PA by the new method than the original.

Hibbing et al. (34) presented youth-specific adaptations of the Sojourn model, with a key focus being to provide methods for activity counts and raw acceleration, as well as for hip-worn and non-dominant wrist-worn monitors. Instead of predicting EE, the models predicted intensity, i.e., SB, light PA, or MVPA. Calibration data were obtained from 54 youth (33 females) who wore triaxial ActiGraph accelerometers on the hip and non-dominant wrist while performing 12 activities that were randomly selected from a diverse pool of 24. A portable indirect calorimeter was also worn as a criterion measure of EE, which was then encoded to intensity. The calibration

data were used to develop algorithms wherein bouts were first identified using a modified Sojourn algorithm. Afterwards, a newly-developed ANN was used to make two intensity predictions, one drawn from features calculated across entire bouts, and one drawn from features calculated in 15-s windows. In the final step, the two intensity predictions were fed into a hand-built decision tree along with accelerometer data (i.e., activity counts or raw acceleration, depending on the model being used) to obtain a final estimate. For the ANN, features were primarily the same as the original Sojourn model (i.e., percentiles and lag-1 autocorrelation) based on Staudenmayer et al. (211). However, sex, age, and body mass index were also added as features to account for greater individual differences among youth than adults. An independent validation was carried out in simulated free-living, wherein 27 youth (13 females) completed one hour of unstructured activity. Direct observation was used as the criterion measure of intensity, and performance was compared between the new models and numerous reference methods. In terms of overall accuracy, the new methods (maximum correct classification of 56.7% using activity counts on the hip) did not outperform the reference methods (maximum correct classification of 59.6% when using the Freedson (231) cut-points). However, in terms of MVPA classification the new methods had some strengths, particularly when using activity counts on the hip (MVPA sensitivity and specificity of 97.1% and 76.2%, respectively).

Summary and Transition

In this section, a detailed review has been given of different approaches to analysis and modeling in sensor-based PA assessment. A particular emphasis has been given to two areas of innovation (i.e., the divide and conquer paradigm and Sojourn methods) that have shown exceptional promise in recent years. The following section concludes this review by synthesizing the evidence and bringing out gaps in key areas that are relevant to this dissertation.

Recap and Synthesis: Gaps with Respect to Segmentation and Sensor Fusion

There is a vast body of previous work related to sensor-based PA assessment, and there are ongoing efforts to develop improved methods. Innovation can be approached from many angles, some that are procedural, and others that are analytical. On the procedural side, sensor fusion is a promising pathway that warrants further investigation. On the analytical side, activity-defined segmentation may lead to improvements, particularly through change point detection.

Several studies have examined either sensor fusion or activity-defined segmentation, but few have done so in youth, and most have had small sample sizes. Almost no studies have combined sensor fusion and activity-defined segmentation. One exception was the demonstration given by Gharghabi et al. (38), with its combined use of accelerometer and gyroscope data for transition detection in youth (see “Dynamic Segmentation” under “Segmentation” in the computational factors section). While the former study informally demonstrated the potential value of combining sensor fusion and activity-defined segmentation, its scope was limited, and thus there remains much room for investigation.

Among sensor-based PA assessment methods, the youth Sojourn approach is unique in its ability to be refined over time (34). Furthermore, activity-defined segmentation is a core focus within each youth Sojourn model. Therefore, the youth Sojourn approach represents a logical starting point from which to examine how sensor fusion and activity-defined segmentation interact, and whether their combination can lead to improved model performance. The latter observation forms the basis for the research questions to be addressed in the following chapters.

CHAPTER III

DESCRIPTION OF LABORATORY DATA SET

Data were drawn from a larger study that has been presented elsewhere (154, 232). The present investigation used only a subset of the data, specifically focusing on indirect calorimetry, direct observation (DO), and research-grade multi-sensor devices worn on the hip and both wrists, as described below. Participants also wore other devices on the wrists, ankles, thighs, and shoes, which will not be discussed. In this chapter, the focus is strictly on describing how the data set was generated, whereas subsequent chapters will describe how the data were analyzed. Specifically, the data presented in this chapter will be used in Chapter 4 (pages 81-101) to address the first research question, and in Chapter 5 (pages 102-119) to address the second.

Participants

One hundred youth (age 6-18 y) volunteered to participate in the study. Recruitment was accomplished through word of mouth and publicly posted flyers, and was specifically aimed at obtaining an even balance across sex (i.e., matching numbers of males and females) and age group (i.e., matching numbers of participants between 6-9, 9-12, 12-15, and 15-18 years old). Written parental informed consent and youth assent were obtained prior to data collection, and exclusion criteria (e.g. contraindications to exercise, or recent musculoskeletal injury) were screened based on a health history questionnaire filled out for the participants by a parent or guardian. The study was approved by the University of Tennessee, Knoxville Institutional Review Board. The key materials used for recruitment and enrollment are shown in Appendix A (see page 211).

Protocol

Participants came to the Applied Physiology Lab on two occasions. At the first visit, anthropometric measurements were taken, followed by a 30-min resting metabolic rate (RMR) assessment. Then eight activities were performed, from a list of 16 (see Table 2). At the second

Table 2. Activities performed by participants in the laboratory study.

Operational Definition	
Activities Performed in Applied Physiology Lab	
Supine rest	Lying still on back with no further activity (i.e., no conversation, screen time, etc)
Reading	Reading a book (hard copy) while seated or reclining, optionally at a table or desk
Internet surfing	Using a desktop computer to perform self-selected online activities (e.g. for leisure or educational purposes)
Online gaming	Using a desktop computer to play online games (e.g. Tetris, Pacman)
Reclining	Occupying an office chair in a seated or reclined posture with no further activity (i.e., no conversation, screen time, etc.)
Sweeping	Cleaning debris off the floor using a broom and dustpan
Dusting	Cleaning various surfaces (tables, counters, etc.) using spray cleaner and paper towels
Cycling	Pedaling on an auto-braked cycle ergometer at a self-selected cadence and resistance (minimum 25 W)
Activities Performed in Nearby Recreation Areas	
Basketball	Playing basketball (dribbling, shooting, one-on-one, etc.)
Soccer	Playing soccer (dribbling, passing, shooting, etc.)
Catch	Playing catch with a football (self-selected distance)
Jumping jacks	Performing jumping jacks continuously at a self-selected pace maintainable for 4-5 min
Stair walking	Ascending and descending a flight of stairs at a self-selected pace maintainable for 4-5 min
Overground slow walk [#]	Walking at a self-selected slow pace
Overground brisk walk [#]	Walking at a self-selected brisk pace
Overground run [#]	Running at a self-selected pace maintainable for 4-5 min

[#]Performed in repeated circuits around pre-defined paths, i.e., following the baseline of an indoor basketball court or outdoor tennis court in half laps or full laps.

visit, the other eight activities were performed. Throughout all activities, participants wore activity monitors (ActiGraph GT9X; ActiGraph, LLC, Pensacola, FL, USA), as well as a portable indirect calorimeter (Cosmed K4b²; Cosmed, Rome, Italy) for assessment of energy expenditure (EE). DO was performed throughout the protocol using tablet-based software (Pocket Observer; Noldus Information Technology bv, Wageningen, The Netherlands) for focal sampling, which served as a criterion measure of activity type and transition timing, as well as posture. The sessions were video recorded in most cases (depending on participant opt-in), which allowed review of the DO data when necessary.

Participants performed two bouts of each activity, one short (60-90 s) and the other long (4-5 min). The activities were performed in one of two general locations, namely the Applied Physiology Lab or a nearby recreation area (e.g. indoor gym or outdoor tennis court). Each day, there was an approximately equal split of activities performed in each location. The activity order was restricted such that 1) the research team determined which eight activities would be performed each day, with combinations varying widely by design; 2) the same activity could not be performed twice in a row; and 3) repeated location changes were avoided. Outside of those restrictions, the participants selected the activity order, and a random number generator was used to determine whether the first bout of each activity would be short or long.

Equipment

ActiGraph GT9X

For each participant, data from three GT9X devices were used. One device was worn on the right hip along the anterior axillary line at the level of the iliac crest, and the other two were worn on the wrists (i.e., one each on the dominant and non-dominant sides) immediately superior to the ulnar and radial styloid processes. The GT9X contains a triaxial primary accelerometer

and an inertial measurement unit (IMU). The primary accelerometer has a dynamic range of ± 8 g and is the same component used in ActiGraph GT3X+ models. It can be configured to sample at rates between 30 Hz and 100 Hz, and the data can be extracted in two formats, namely raw acceleration (in g; low pass filtered at a proprietary cutoff) or the integrated proprietary units known as “activity counts”, which are integrated over regular intervals of user-specified length between 1 s and 240 s. The IMU contains a device thermometer and three triaxial sensors, i.e., secondary accelerometer (dynamic range ± 16 g), gyroscope (dynamic range $\pm 2000^{\circ}\cdot\text{s}^{-1}$), and magnetometer (dynamic range ± 4800 μT). The IMU sensors can be selectively enabled, but the sampling rate (100 Hz) is not modifiable. Prior to each session, the GT9X devices were initialized to collect at 90 Hz (primary accelerometer) with all IMU sensors activated.

Timestamps were automatically synced with the Windows system clock during initialization.

Noldus Pocket Observer

Noldus Pocket Observer is a mobile application driven by the Noldus Observer XT desktop application. Users define a coding scheme through the desktop application, and then transfer the coding scheme to a mobile device for real-time data collection. For the present study, Android tablets were used for data collection, and the interface is shown in Figure 1. At the start of the collection, a default behavior called “CommentStartTime” prompted the user to record the start time based on a Windows system clock, for future temporal alignment with other data sources. Afterwards, users tapped two boxes on the screen whenever activity or postural transitions occurred. The first box reflected the activity being performed, with options corresponding to the 16 shown in Table 2, plus “Transition” and “UNKNOWN” (the latter being used for cases such as restroom breaks). The second box reflected the posture being assumed by the participant, with seven possibilities that were either non-weight-bearing (lying, sitting, and reclining) or weight-

Observation

Observed: 00:04:00.24

Suspend

Stop

Marker

Comment

Undo

Time	Behavior	Modifier	Comment
104.0	Sweep	Stand	
093.4	Transition	Stand	
068.2	Dust	Stand	
054.3	Transition	Stand	
037.7	Internet	Sit	
000.0	CommentStartTime		10:00:00

Activity	G - Internet	U - UNKNOWN
A - Basketball	H - Jumping Jacks	N - Transition
B - Book	I - Lying	O - CommentStartTime
C - Catch	J - Reclining	P - Walk_Brisk
D - Cycling	K - Run	Q - Walk_Slow
E - Dust	L - Soccer	R - Walk_Stairs
F - Games	M - Sweep	

Figure 1. Noldus Pocket Observer mobile interface for performing real-time direct observation. A) The activity log showed which activities and postures had been performed (see the “Behavior” and “Modifier” columns, respectively), as well as the timing for each. B) The first activity in each observation was “CommentStartTime”, which prompted users to manually enter the start time as a comment (“10:00:00” in the example) based on a Windows system clock. C) The input area allowed users to indicate that a new behavior had started by tapping the corresponding box, at which point a follow-up prompt (not shown) asked the user to specify the posture being assumed.

bearing, i.e., upright (standing, stepping, ascending stairs, and descending stairs). Simultaneous behaviors were not allowed in the coding scheme, and behaviors were coded continuously (i.e., gaps were not allowed in the observation). Thus, the Pocket Observer data allowed each moment of the data collection to be linked to a single activity and posture. User input was automatically timestamped, thereby establishing when activity and postural transitions occurred. The intervening periods (i.e., periods of continuous engagement in a single activity and posture) will be referred to hereafter as behavioral intervals, which are notably distinct from activity bouts (i.e., periods of continuous engagement in a single activity, which could include one or more postures). Researchers were also able to add comments throughout the observation, which were used to document any issues that occurred (e.g. a monitor came off, or a transition was missed). When necessary, the latter issues were cross-referenced with video data, which allowed the DO records to be corrected.

Cosmed K4b²

The K4b² is a portable indirect calorimetry unit that has previously been validated against the Douglas bag method for assessment of oxygen consumption (VO₂) and carbon dioxide production (VCO₂) across a range of intensities (233). Participants wore a facemask that was connected to a chest-mounted unit containing analyzers for oxygen (O₂) and carbon dioxide (CO₂). Before each test, the K4b² was calibrated using four modules, one for analyzer calibration to room air, one for analyzer calibration to a reference gas mixture (15.98% O₂, 4.008% CO₂, balance nitrogen), one for flowmeter calibration to a known volume (3 L, drawn from a Hans Rudolf syringe), and one for delay calibration (i.e., to account for the time lag between recognition of a breath by the flowmeter, and sampling of the same breath by the analyzers). At

the start of each data collection, the Windows system clock time was manually recorded for future alignment with other data sources.

Initial Data Processing

ActiGraph GT9X

Binary data from the primary accelerometer (90 Hz) and gyroscope (100 Hz) were first read into R using the function ``AGread::read_gt3x``, then averaged each second and merged together using the function ``AGread::AG_merge`` (234). The two commands generally followed the procedures of Hibbing et al. (25). Specifically, for primary accelerometer data (in g) two variables were computed for each sample. The first variable was the Euclidian norm minus one (ENMO, in milli-g), which was defined as $[(\sqrt{X^2 + Y^2 + Z^2} - 1) \cdot 1000]$, with negative values rounded up to zero. The second variable was the vector magnitude (VM, in g), which was defined as $\sqrt{X^2 + Y^2 + Z^2}$. After computing ENMO and VM for each sample, second-by-second averages were calculated for both variables. For the 100-Hz gyroscope data (in $^{\circ}\cdot\text{s}^{-1}$), each axis was low-pass filtered at 35 Hz, after which the gyroscope vector magnitude (GVM, defined as $\sqrt{X^2 + Y^2 + Z^2}$) was calculated and averaged each second. The second-by-second ENMO, VM, and GVM data were then merged.

Noldus Pocket Observer

Pocket Observer data were exported as Excel files in “event log” format, meaning each behavioral interval was represented with two rows of data, one reflecting the start of the interval and the other reflecting the end. When necessary, video recordings were used to manually correct the event logs for issues noted by the researchers, as discussed above on page 68. The finalized Excel files were then processed in R, wherein a sync procedure was first performed to align the data with the Windows system clock. Specifically, the offset was calculated between the initial

Android timestamp and the manually-recorded Windows start time (see Figure 1B), and the offset was then subtracted from all Android timestamps to convert them to Windows timestamps. The remaining R code then reformatted the data such that each behavioral interval was represented with one row of data rather than two. That is, after the R code was complete, each row of data reflected one behavioral interval with an activity label, a posture label, and two Windows timestamps denoting the start and end of the interval.

Cosmed K4b²

Breath-by-breath data were screened for invalid breaths by examining respiratory frequency (acceptable range of 5-80 breaths·min⁻¹), ventilation rate (acceptable range of 0.2-10 L·min⁻¹), fraction of expired O₂ (acceptable range of 10%-20%), and fraction of expired CO₂ (acceptable range of 1%-10%). Breaths were excluded if one or more values fell outside the acceptable range. Timestamps were calculated by adding the K4b² relative timestamps (i.e., elapsed time since the start of the measurement) to the manually recorded Windows start time.

Preview of Further Processing

The previous sections have described how initial processing was carried out for individual data sources (i.e., for GT9X, DO, and K4b² separately). After the individual data sources had been prepared, the ensuing procedures all used data from multiple sources. Specifically, there were three tasks, with the first being to determine RMR for each participant. The RMR values were then used to calculate youth metabolic equivalents ($MET_y = \frac{activity\ VO_2}{RMR}$) in the second and third tasks, which were focused on processing steady state data and continuous time course data (i.e., data collected throughout the whole protocol), respectively. Below, each task is described in a separate subsection. Notably, breath-by-breath VO₂ data were converted to different epoch lengths depending on the task. All such conversions were performed using procedures given by

the manufacturer. Specifically, within each epoch, means were calculated across breaths for key variables (e.g. gas fractions), and then those means were used to calculate the VO_2 value for the whole epoch.

Further Processing: Resting Metabolic Rate Data

DO data were used to identify when each 30-min RMR assessment began and ended. The corresponding breath-by-breath VO_2 data were then extracted and processed. Data were discarded from the first 10 min of each assessment, along with data from the last minute. Using the remaining data, VO_2 was calculated for all possible 5-min epochs (i.e., using a 5-min rolling window), and the lowest value was taken as RMR.

Further Processing: Steady State Data

The steady state data set was created in three broad steps, one for processing criterion data (i.e., data from DO and the K4b²), one for processing sensor data (i.e., data from the GT9X), and one for assembling the final data set. Below, the steps are described in separate subsections.

Processing Criterion Data

DO data were used to define when steady state periods began and ended for each participant and activity. Specifically, activity bouts lasting ≥ 220 s were examined. For all such bouts, the last 10 s were removed, and the steady state period was defined as the last 60 s of what remained. For each steady state period, the corresponding breath-by-breath data were extracted and averaged (i.e., converted to a 60-s epoch) to obtain a steady state VO_2 value, which was subsequently divided by RMR to obtain MET_y . Each MET_y value was then merged with the corresponding DO data, such that all steady state periods had activity and posture labels (from DO) and a MET_y value (from the K4b²). Intensity classification was then determined using accepted definitions that included both posture and EE (73, 235). Specifically, each steady state

period was labeled as sedentary behavior (SB; $\leq 1.5 \text{ MET}_y$ and lying, seated, or reclining posture), light physical activity (LPA; $1.51\text{-}2.99 \text{ MET}_y$, or $\leq 1.50 \text{ MET}_y$ with upright posture), or moderate-to-vigorous physical activity (MVPA; $\geq 3.0 \text{ MET}_y$). After completing the prior steps, criterion values (activity, posture, MET_y , and intensity) were available for each steady state period.

Processing Sensor Data

The GT9X data were converted to summary features using the following three-step process for each steady state period: First, the corresponding 60 s of second-by-second GT9X data were extracted; second, those GT9X data were grouped into four 15-s sequences; and third, features were calculated for each sequence. The features were five percentiles (10th, 25th, 50th, 75th, and 90th) and the lag-1 autocorrelation, each of which was calculated for ENMO, VM, and GVM. The features were selected in keeping with the neural network training procedures in the original youth Sojourn calibration (34). After completing the prior steps, there were four rows of GT9X data per steady state period, with each row containing features that were calculated over a 15-s period.

Assembling the Final Data Set

The GT9X features were used as the starting point for assembling the final data set. Each row of feature data was labeled with the criterion values (activity, posture, MET_y , and intensity) from the corresponding steady state period. Afterwards, participant characteristics (e.g. sex and age) were added to complete the data set. Table 3 shows how the data set was laid out.

Table 3. Layout of the steady state data, as used for model development in the present investigation. Each row of data corresponds to a 15-s window of steady state activity (with a total of 60 s per activity).

A 15-s window of steady-state activity (with a total of 60-s per activity):															
Demographics					ENMO (milli-g)				... [#]	GVM (°·s ⁻¹)				Criterion Labels [#]	
					Percentiles					Percentiles					
PID	Attachment	Sex	Age	... [#]	10th	...	90th	L1	...	10th	...	90th	L1	Activity	Intensity
000	Hip	M	11.9	...	113.7	...	150.5	0.42	...	44.1	...	60.6	0.00	Sweep	MVPA
000	Hip	M	11.9	...	89.2	...	130.2	0.52	...	42.5	...	67.4	0.57	Sweep	MVPA
000	Hip	M	11.9	...	46.9	...	127.7	0.58	...	41.9	...	77.4	0.39	Sweep	MVPA
000	Hip	M	11.9	...	61.2	...	124.2	0.03	...	42.2	...	88.0	0.56	Sweep	MVPA
000	Hip	M	11.9	...	29.6	...	31.4	-0.35	...	0.9	...	1.7	-0.21	Reclining	SB
000	Hip	M	11.9	...	29.7	...	31.1	-0.08	...	1.0	...	2.1	-0.01	Reclining	SB
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

ENMO- Euclidian norm minus one; GVM- gyroscope vector magnitude; PID- participant code; L1- lag-1 autocorrelation; M- male; MVPA- moderate-to-vigorous physical activity; SB- sedentary behavior.

[#]To conserve space, some variables are not shown. Most notably, data are not shown for the accelerometer vector magnitude, which had the same features (i.e., percentiles and lag-1 autocorrelation) as ENMO and GVM. Additionally, criterion values for posture and metabolic equivalents are not shown, having already been combined to determine intensity (i.e., SB, LPA, or MVPA).

Further Processing: Continuous Time Course Data

The continuous time course data consisted of everything collected throughout the whole trial, including transition periods and non-steady-state activity. Prior to assembling the full data set, breath-by-breath VO_2 data were converted to 15-s epochs. That epoch length was selected to allow granular examination of the VO_2 trends while also smoothing the data enough to stabilize the values. VO_2 was then divided by RMR to obtain MET_y .

The data set was assembled by starting with the second-by-second GT9X data. First, each second was labeled with the MET_y value that was occurring at that time, based on the K4b² data (15-s epochs). Next, DO data were added by labeling each second with the activity and posture values from the corresponding behavioral interval. After completing the prior steps, the data set included second-by-second data from the GT9X (ENMO, VM, GVM), K4b² (MET_y), and DO (activity and posture). Data were then excluded from any periods with missing values for any of the variables (e.g. if the K4b² mask was removed for a water break).

For each second of data, posture and MET_y were examined to determine activity intensity (i.e., SB, LPA, or MVPA) using the same classification scheme described above for the steady state data. Additionally, the activity labels were examined to identify transition points, which were labeled using a binary variable (i.e., taking a value of 1.0 whenever a transition occurred and 0.0 otherwise). Finally, participant characteristics (e.g. sex and age) were added to complete the data set. Table 4 shows how the data set was laid out.

Data Loss

Data loss occurred for a variety of reasons, and partial data were used whenever possible. Figure 2 summarizes participant level data loss, and Table 5 summarizes activity level data loss. The causes of data loss were straightforward, with the exception of loss due to data cleaning.

Table 4. Layout of the continuous time course data (second-by-second format), as used for model development in the present investigation.

PID	Attachment	Demographics			Sensor Variables			Criterion Labels [#]		
		Sex	Age	... [#]	ENMO (milli-g)	VM (g)	GVM (°·s ⁻¹)	Activity	Intensity	Transition [^]
000	Hip	M	11.9	...	8.5	1.004	31.173	Lying	SB	0
000	Hip	M	11.9	...	23.3	1.023	18.472	Lying	SB	0
000	Hip	M	11.9	...	27.6	1.027	15.328	Transition	LPA	1
000	Hip	M	11.9	...	17.4	1.014	22.783	Transition	LPA	0
000	Hip	M	11.9	...	25.1	1.024	35.127	Transition	LPA	0
...	...	...	...	...	...	...	...	...	...	...

PID- participant code; ENMO- Euclidian norm minus one; VM- vector magnitude; GVM- gyroscope vector magnitude; M- male; SB- sedentary behavior

[#]To conserve space, some variables are not shown. Most notably, criterion values for posture and metabolic equivalents are not shown, having already been combined to determine intensity (i.e., SB, LPA, or MVPA).

[^]Indicator of whether a transition occurred (1) or not (0) for each second.

Note: VM does not always equal [(ENMO/1000)+1], which is a side effect of converting the raw data (90 Hz) to 1-s averages.

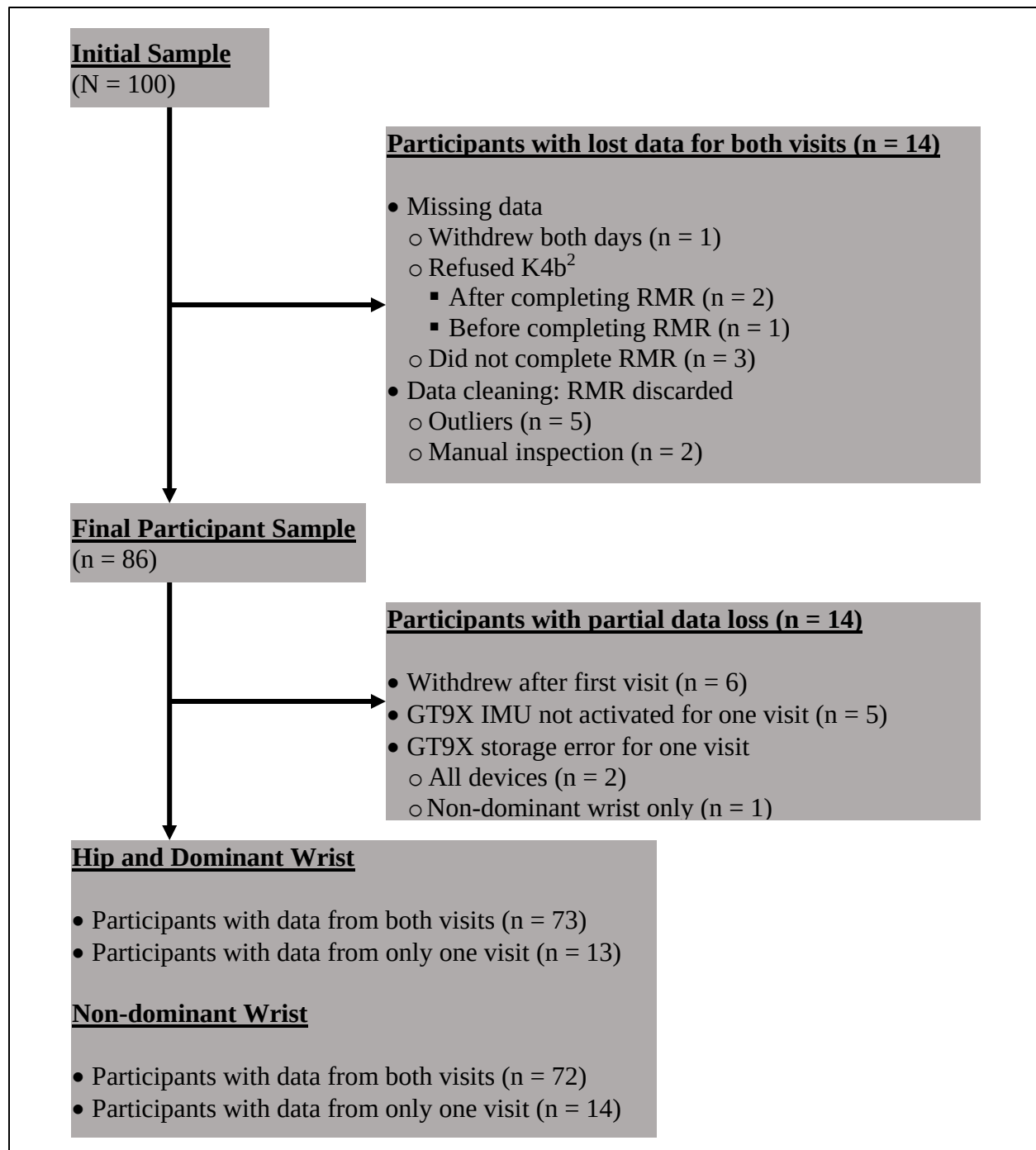


Figure 2. Summary of participant level data loss and retention for the laboratory data set.
RMR- resting metabolic rate; IMU- inertial measurement unit.

Table 5. Summary of activity level data loss for steady state activities.

Description	Activities lost (n)	Remaining Activities (n)
Expectation (86 participants times 16 activities)	--	1376
K4b² data cleaning		
Activity duration < 3 min, 40 s	68	1308
Value > 2 SD's from mean for activity and age group [#]	55	1253
K4b² data missing or incomplete		
Activity interrupted (e.g. monitor came off)	3	1250
Activity not completed	1	1249
Other missing data (e.g. battery died)	17	1232
Miscellaneous data loss		
Basketball court unavailable	8	1224
Failure/refusal to perform an activity	2	1222
Partial data loss at participant level		
Withdrawal after first visit	48	1174
GT9X initialization error (inertial measurement unit not activated)	34	1140
GT9X download or user error	15	1125
Additional participant-level data loss by attachment site		
Hip and dominant wrist	0	1125
Non-dominant wrist (download error)	8	1117

SD- standard deviation.

[#]Age groups were ≤ 12 years old and > 12 years old.

Therefore, the remainder of this section will focus on describing how data cleaning was carried out, specifically for RMR and steady state data.

After obtaining RMR for each participant, the values were cleaned separately for two age groups (i.e., those ≤ 12 years old and > 12 years old), with individual values being removed if they were > 2.0 standard deviations from the mean. Upon manual inspection of the RMR data, two additional participants were excluded due to apparent malfunction of the K4b². Without RMR values, it was impossible to calculate MET_y for steady state or continuous time course data, and thus there was complete loss of all data for participants ($n = 7$) whose RMR values were discarded.

For the steady state data, a similar cleaning approach was taken. Specifically, mean EE was calculated for each activity and age group. Individual values were then removed if they were > 2.0 standard deviations from the corresponding mean.

Summary of Final Data Set

After cleaning the data, participants were included in the final data sets if they had at least one day of activity data. There were 86 such participants, most of whom had data from both visits ($n = 73$ for hip and dominant wrist data; $n = 72$ for non-dominant wrist data). Steady state data were available for 4-16 activities per participant (1125 total activity bouts for the hip and dominant wrist; 1117 total activity bouts for the non-dominant wrist), and all participants had 0.9-2.6 h of continuous time course data (mean $\pm$ SD 2.1 ± 0.4 h per participant for all attachment sites). The total volume of continuous time course data amounted to more than one week (182.5 h for the hip and dominant wrist; 181.7 h for the non-dominant wrist). Table 6 shows participant characteristics for the final sample. The data presented in this chapter will serve as a starting point for analyses in the next two chapters. Specifically, the data will be used

Table 6. Participant characteristics for the laboratory study. Values are n (%), except for age (mean $\pm$ SD).

	Female (n = 45)	Male (n = 41)	Total (N = 86)
Age (yr)	12.5 $\pm$ 3.6	12.8 $\pm$ 3.1	12.6 $\pm$ 3.4
Weight Status[#]			
Underweight (< 5 th percentile)	2 (4%)	1 (2%)	3 (3%)
Normal Weight (5.0-84.9 th percentile)	35 (78%)	34 (83%)	69 (80%)
Overweight (85.0-94.9 th percentile)	6 (13%)	2 (5%)	8 (9%)
Obese ($\geq$ 95 th percentile)	2 (4%)	4 (10%)	6 (7%)
Handedness			
Left	3 (7%)	2 (5%)	5 (6%)
Right	42 (93%)	39 (95%)	81 (94%)
Ethnicity			
Hispanic	3 (7%)	1 (2%)	4 (5%)
Non-Hispanic	42 (93%)	40 (98%)	82 (95%)
Race			
African American	6 (13%)	2 (5%)	8 (9%)
White	39 (87%)	39 (95%)	78 (91%)

[#]Percentiles determined from the Centers for Disease Control and Prevention body-mass-index-for-age growth charts.

in Chapter 4 (pages 81-101) to address the first research question, and in Chapter 5 (pages 102-119) to address the second.

CHAPTER IV

RESEARCH QUESTION 1: METHODS AND RESULTS

Overview

In a previous study, Hibbing et al. (34) developed youth Sojourn models that used accelerometer data to predict activity intensity, namely sedentary behavior (SB), light physical activity (LPA), or moderate-to-vigorous physical activity (MVPA). In that study, four models were developed, each corresponding to one of two attachment sites (hip or non-dominant wrist) and one of two data formats (raw acceleration or activity counts). All four models contained three components, namely a bout detection algorithm, a neural network, and a decision tree. The components were used to produce intensity estimates through the process shown in Figure 3.

In this chapter, laboratory data (both steady state and continuous time course) were used to replicate the calibration methods of Hibbing et al. (34) using gyroscope data instead of accelerometer data. Holdout samples were then used for comparing the new gyroscope-based components to their accelerometer-based counterparts. The procedures were designed to address the first research question, which was formulated as follows:

1. **Question:** In a laboratory setting, do gyroscope-based youth Sojourn model components (bout detection algorithm, neural network, and decision tree) perform better than the existing accelerometer-based components?

Hypothesis: Gyroscope-based youth Sojourn model components will achieve better performance than the existing accelerometer-based components when applied to a laboratory data set.

Scope

For this investigation, the focus was strictly on comparing the performance of individual youth sojourn model components when using accelerometer versus gyroscope data. The focus was not on developing new models, which was a separate research question (see the next

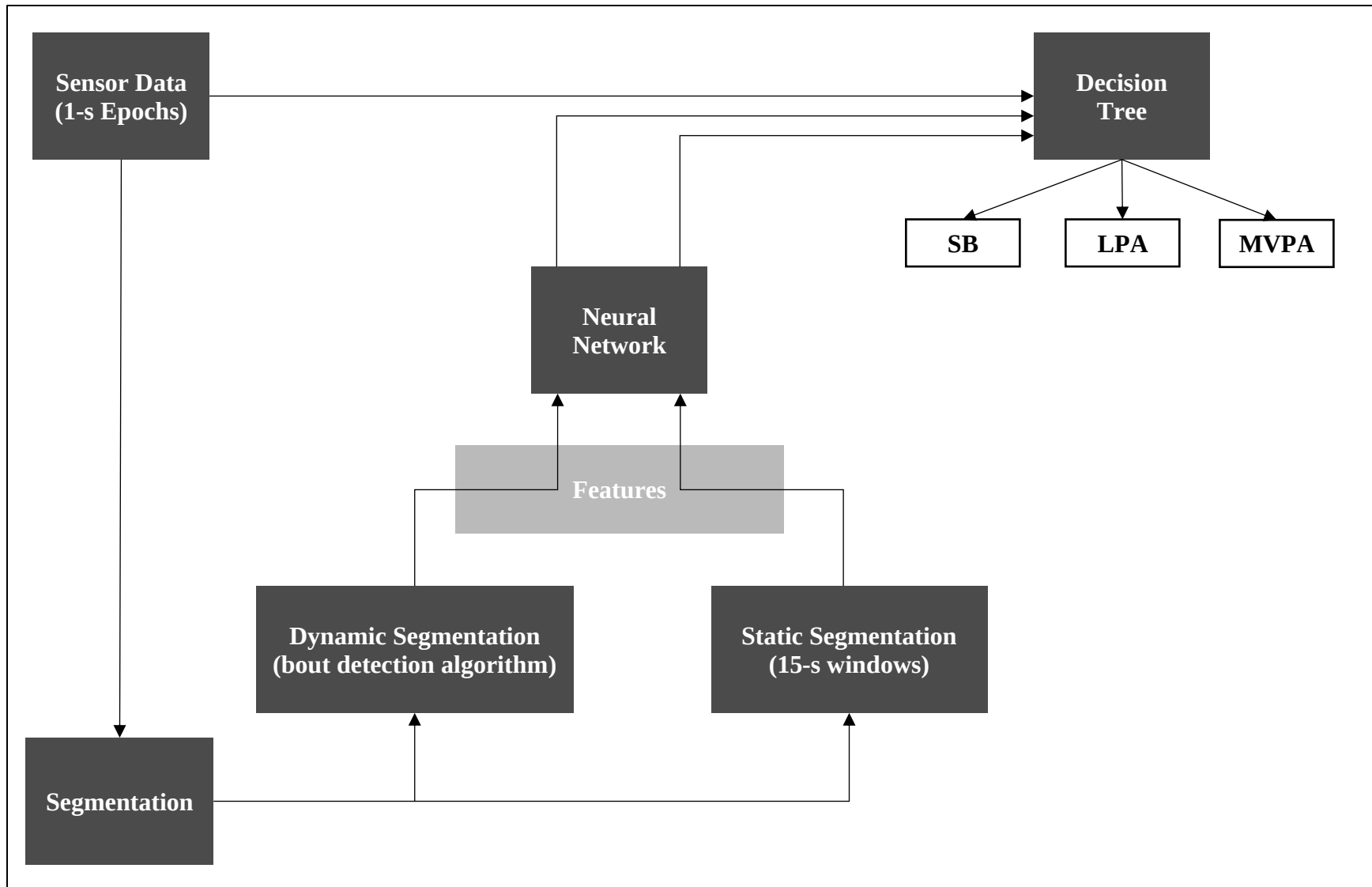


Figure 3. Depiction of the process by which the original youth Sojourn models classify second-by-second data as sedentary behavior (SB), light physical activity (LPA), or moderate-to-vigorous physical activity (MVPA).

chapter, pages 102-119). In this chapter, the new gyroscope-based components were only compared to their raw acceleration-based counterparts, not their activity count-based counterparts. There were four reasons for excluding the count-based components:

1. **Sampling rate.** In the original youth Sojourn calibration, the accelerometer sampling rate was 100 Hz, whereas it was 90 Hz for the present study. Activity counts are influenced by sampling rate (236), whereas raw acceleration data generally are not (237). Therefore, the original count-based components were not applicable to the present data, whereas the raw acceleration components were.
2. **Consistency with best practices.** There has recently been an explicit recommendation against using activity counts in new calibrations (85), which has spurred a broad shift towards raw acceleration data in recent calibration studies (51).
3. **Formatting.** In modeling terms, it was sensible to require that the accelerometer and gyroscope signals had similar format.
4. **Simplicity.** Including activity count data would add complexity to the analysis without commensurate benefits.

Methods

Data from the previous chapter (steady state and continuous time course; $n = 86$) served as a starting point for the methods in this chapter. In the original youth Sojourn calibration study (34), models were only presented for the hip and non-dominant wrist, not the dominant wrist. Thus, dominant wrist data are not used in this chapter. The methods described below were carried out twice each (i.e., separately for both attachment sites).

Each youth Sojourn model component (i.e., bout detection algorithm, neural network, and decision tree) was assessed in two stages. The first stage (development) entailed creation of a

new gyroscope-based component, and the second stage (cross-validation) entailed a performance comparison between the latter component and its existing accelerometer-based counterpart. For development, grid search was used to test different settings for the tuning parameters associated with each component. Performance was evaluated for each grid (i.e., each combination of settings), and the grid yielding the best performance was selected for cross-validation, wherein a holdout sample was used for the accelerometer-versus-gyroscope performance comparison. Notably, performance metrics for all processes ranged from 0% (pessimal performance) to 100% (optimal performance), meaning the analyses were not influenced by whether participants had one day of data (e.g. due to withdrawal) or two. Therefore, all participants were included in all analyses. For participants with data from two visits ($n = 73$ for the hip; $n = 72$ for the non-dominant wrist), the visits were pooled before analysis. Below, separate sub-sections are used to describe the component-specific methods by which development and cross-validation were carried out.

Bout Detection Algorithm

Table 7 outlines the development and cross-validation procedures for the bout detection algorithm. The continuous time course data were used, with data from 75% of participants being used for development, while data from the remaining 25% were held out for cross-validation. Participants were assigned to the development or cross-validation subsamples using a guided random process, such that participant characteristics were similar in both groups (see Table 8). Specifically, random selection was performed repeatedly until a manually specified frequency target was met. By pairing a frequency-based target with random selection, bias was avoided that could arise from hand-selecting specific participants for certain partitions.

Table 7. Summary of development and cross-validation for gyroscope-based youth Sojourn bout detection algorithms (for data from the hip or non-dominant wrist).

Summary	
Development	
<u>Data Used</u>	Continuous time course (75% of participants; n = 65; 139.0 total h of hip data and 138.2 total h of non-dominant wrist data)
<u>Grid Search</u>	125 permutations tested (3 parameters, 5 settings each)
Tuning Parameter 1	<i>Description:</i> Minimum decrease from previous GVM epoch for a new Sojourn to be triggered <i>Values tested:</i> 5, 10, 15, 20, and $25^{\circ}\cdot s^{-1}$
Tuning Parameter 2	<i>Description:</i> Maximum GVM to trigger a new Sojourn <i>Values tested:</i> 0, 5, 10, 25, and $50^{\circ}\cdot s^{-1}$
Tuning Parameter 3	<i>Description:</i> Minimum Sojourn duration. <i>Values tested:</i> 5, 10, 15, 30, and 60 s.
<u>Procedure</u>	For each data file: - Segmentation algorithm iterated for all 125 tuning grids For each participant and tuning grid: - If applicable, true/false positives and false negatives pooled across visits - Aggregated performance calculated (mean of recall, precision, and relative root mean squared error) Analysis: - Aggregated performance averaged across participants for each tuning grid - Grid with highest average selected for cross-validation
Cross-Validation	
<u>Data Used</u>	Continuous time course (remaining 25% of participants; n = 21; 43.5 total h of data for both the hip and non-dominant wrist)
<u>Procedure</u>	For each data file: - Original algorithm (accelerometer) applied - New algorithm (gyroscope) applied For each participant and algorithm: - If applicable, true/false positives and false negatives pooled across visits - Performance metrics calculated (recall, precision, relative root mean squared error, and aggregated performance) Analysis: - Metrics averaged across participants for both algorithms - For each metric, means compared to evaluate which algorithm performed better

GVM- gyroscope vector magnitude.

Table 8. Distribution of key participant characteristics within the subsamples used for development and cross-validation of gyroscope-based bout detection algorithms (for data from the hip or non-dominant wrist). Values are n(%).

	Development (75%; n = 65)	Cross-Validation (25%; n = 21)	Total (n = 86)
Age Group (yr)			
< 9.0	10 (15%)	3 (14%)	13 (15%)
9.0-11.9	20 (31%)	7 (33%)	27 (31%)
12.0-14.9	15 (23%)	4 (19%)	19 (22%)
≥ 15	20 (31%)	7 (33%)	27 (31%)
Weight Status			
Underweight/Normal Weight	54 (83%)	18 (86%)	72 (84%)
Overweight/Obese	11 (17%)	3 (14%)	14 (16%)
Handedness			
Left	4 (6%)	1 (5%)	5 (6%)
Right	61 (94%)	20 (95%)	81 (94%)

Note: The same subsamples were also used to develop and cross-validate gyroscope-based decision trees, as discussed beginning on page 93. For gyroscope-based neural networks, the same cross-validation subsample was used, but additional subdivision was applied when using the development subsample (as discussed beginning on page 90).

DEVELOPMENT. For bout detection algorithm development, the candidate settings for the first and third tuning parameters (see Table 7) were designed to be the same as those used by Hibbing et al. (34), although the first parameter used different units ($^{\circ}\cdot\text{s}^{-1}$, versus milli-g). For the second tuning parameter, the candidate settings were designed to mimic the method of graduated increase used in previous studies (34, 238, 239). The range ($0\text{-}50^{\circ}\cdot\text{s}^{-1}$) was established with reference to the SB cut-points for gyroscope vector magnitude (GVM) in the two-regression models of Hibbing et al. (25), which fell between $11.9^{\circ}\cdot\text{s}^{-1}$ and $37.8^{\circ}\cdot\text{s}^{-1}$ depending on the attachment site. The latter study involved adult participants wearing monitors on the ankles (plus the hip and wrists), which is why the range of candidate settings for the second tuning parameter was widened to $0\text{-}50^{\circ}\cdot\text{s}^{-1}$ for the present investigation.

For each data file and tuning grid, the bout detection algorithm was implemented, and performance was assessed by comparing the predictions to the criterion measure of direct observation (DO) using the transition pairing method (TPM) of Hibbing et al. (35). The TPM identifies true positive predictions by pairing them to criterion transitions using a variant of the Gale-Shapley algorithm (240). A user-specified spurious pairing threshold defines the maximum lag time for a prediction to be considered a true positive. For the present investigation, the spurious pairing threshold was set to 6 s, based on the point at which youth Sojourn algorithms achieved $> 50\%$ recall in the original TPM study (35).

For participants with data from two visits, the true positives, false positives, and false negatives from both visits were pooled before further analysis. The outcome variable was aggregated performance, defined as the mean of recall, precision and relative root mean squared error (RMSE%). Recall and precision are commonly used in many classification settings, but RMSE% is unique to the TPM. Ordinarily, root mean squared error (RMSE) is used on an

absolute scale to represent timing error by a dynamic segmentation algorithm, and lower values reflect higher performance. Within the TPM, the maximum lag time is known (i.e., equal to the spurious pairing threshold value), as is the minimum lag time (i.e., 0 s). Therefore, it is possible to express RMSE on a relative scale (i.e., RMSE%) such that performance is reflected the same way as precision and recall, namely as a percentage between 0% and 100% where higher values reflect better performance. To calculate RMSE%, the following formula can be used: $RMSE\% = 100\% - \left(\frac{RMSE}{SPT} * 100\% \right)$, where SPT is the spurious pairing threshold value. More information is available in the vignette for the R implementation of the TPM (see <https://cran.r-project.org/web/packages/PAutilities/vignettes/tpm.html>).

After implementing the TPM for each participant and tuning grid, all results were screened for missing values (e.g. due to no true positives forming). Listwise deletion was then performed across tuning grids, to ensure the same samples were used for comparing them. Mean aggregated performance was then calculated for each grid, and the grid with the highest mean was selected for cross-validation.

CROSS-VALIDATION. For bout detection algorithm cross-validation, the new gyroscope algorithms (hip and non-dominant wrist) were compared to their accelerometer counterparts when applied to the 25% holdout sample. For each data file, the site-appropriate gyroscope algorithm was applied, and the R function ``Sojourn::apply_youth_sojourn`` was used to apply the site-appropriate accelerometer algorithm (241). Predictions from both algorithms were then compared to the criterion measure using the same methods described for development (i.e., using the TPM with a spurious pairing threshold setting of 6 s, pooling both visits where applicable, and handling missing values via listwise deletion), except that aggregated performance was not

the only metric that was examined. Instead, recall, precision, RMSE%, and aggregated performance were all averaged across participants to allow detailed descriptive comparisons.

Neural Network

The existing youth Sojourn neural networks are accelerometer-based classification models that predict activity intensity. The model inputs include participant characteristics (age, sex and body mass index) and accelerometer features, namely percentiles (10th, 25th, 50th, 75th, and 90th) and lag-1 autocorrelation. For the present investigation, gyroscope-based neural networks were created using steady state data. The development and cross-validation procedures are outlined in Table 9. Overfitting was avoided during development by using data from 50% of participants (n = 43) to train candidate models, then using data from 25% of participants (n = 22) to compare the performance of each candidate model. Hereafter, the 50% and 25% subsamples will be referred to as the training and validation subsamples, respectively. Altogether, development used data from 75% of participants (n = 65), with data for the remaining 25% (n = 21) being held out for cross-validation. Table 10 shows that participant characteristics were balanced across the training, validation, and cross-validation subsamples. The balance was achieved using the same guided random approach that was described for the bout detection algorithm.

DEVELOPMENT. For neural network development (hip and non-dominant wrist), 15 neural networks were developed corresponding to all combinations of the candidate settings (two tuning parameters) shown in Table 9. The candidate settings were the same as originally used by Hibbing et al. (34). The training data were balanced before fitting the models. That is, equal numbers of sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA) observations were selected for training. The selection involved iterative removal of observations from the two largest categories, until they both had the same

Table 9. Summary of development and cross-validation for gyroscope-based youth Sojourn neural networks (for data from the hip or non-dominant wrist).

Summary	
Development	
<u>Data Used</u>	Steady state data
Model Training	50% of participants (n = 43; 584 total activity bouts for hip data and 576 total activity bouts for non-dominant wrist data)
Model Validation	25% of participants (n = 22; 272 total activity bouts for both hip and non-dominant wrist data)
<u>Grid Search</u>	15 permutations tested (2 parameters with 5 and 3 settings, respectively)
Tuning Parameter 1	<i>Description:</i> Number of nodes in hidden layer <i>Values tested:</i> 5, 10, 15, 20, and 25
Tuning Parameter 2	<i>Description:</i> Rate at which coefficients were penalized toward 0 with continued training <i>Values tested:</i> 0, 0.1, and 0.001
<u>Procedure</u>	Training: - Data set balanced to include equal frequencies for SB, LPA, and MVPA - Models fitted for each tuning grid Analysis: - All models applied to the validation subsample - Model with the highest percent correct classification selected for cross-validation
Cross-Validation	
<u>Data Used</u>	Steady state data (remaining 25% of participants; n = 21; 269 total activity bouts for both the hip and non-dominant wrist)
<u>Procedure</u>	- Original model (accelerometer) applied - New model (gyroscope) applied - Performance metrics calculated for both models <i>Overall agreement:</i> Percent correct classification, kappa <i>Category-specific agreement:</i> sensitivity, specificity - For each metric, values compared to evaluate which model performed better

SB- sedentary behavior; LPA- light physical activity; MVPA- moderate-to-vigorous physical activity.

Table 10. Distribution of key participant characteristics within the subsamples used for training, validation, and cross-validation of gyroscope-based neural networks (for data from the hip or non-dominant wrist). Values are n(%).

	Development (75%; n = 65)		Cross-Validation (25%; n = 21)	Total (n = 86)
	Training (50%; n = 43)	Validation (25%; n = 22)		
Age Group (yr)				
< 9.0	6 (14%)	4 (18%)	3 (14%)	13 (15%)
9.0-11.9	14 (33%)	6 (27%)	7 (33%)	27 (31%)
12.0-14.9	9 (21%)	6 (27%)	4 (19%)	19 (22%)
≥ 15	14 (33%)	6 (27%)	7 (33%)	27 (31%)
Weight Status				
Underweight/Normal Weight	36 (84%)	18 (82%)	18 (86%)	72 (84%)
Overweight/Obese	7 (16%)	4 (18%)	3 (14%)	14 (16%)
Handedness				
Left	2 (5%)	2 (9%)	1 (5%)	5 (6%)
Right	41 (95%)	20 (91%)	20 (95%)	81 (94%)

Note: The overall 75% development subsample contains the same participants as the development subsample shown in Table 8, and the cross-validation subsamples are the same in both cases.

number of observations as the third (i.e., smallest) category. To prevent changes in the total number of participants at each intensity, a repeated two-step process was used to determine which observations to remove: First, the participant was identified who had the most observations for the category in question; and second, an observation was randomly selected and removed for that participant before repeating the process (or terminating if no further removals were necessary).

After fitting the neural networks, they were all applied to the validation subsample (25% of participants; $n = 22$). The performance metric was percent correct classification, and the model achieving the highest percent correct classification was selected for cross-validation.

CROSS-VALIDATION. For neural network cross-validation, the new gyroscope neural networks were applied to the 25% holdout sample, as were the original accelerometer neural networks. Accelerometer-versus-gyroscope comparisons were then made descriptively in terms of overall agreement (percent correct classification, kappa statistics) and category-specific agreement (sensitivity and specificity).

Decision Tree

The decision trees in youth Sojourn models take three inputs that are illustrated in Figure 3. The inputs are obtained through a complex process that depends on both the bout detection algorithm and the neural network. Thus, gyroscope-based decision trees could not be developed or cross-validated until appropriate data preparation had been performed to obtain all three inputs. Accordingly, the preparation steps will be described below, before moving on to the development and cross-validation procedures.

DATA PREPARATION. All three decision tree inputs were second-by-second variables. The first input was GVM drawn directly from the continuous time course data without additional

processing. The second and third inputs were each intensity estimates generated by a neural network (i.e., the gyroscope-based neural network described in the previous section; see pages 90-93), with the difference being how the features were calculated. For the second input (referred to hereafter as Intensity_{15}), the second-by-second GVM data were grouped into 15-s sequences, and features were calculated for each sequence. Predictions were then made by the neural network, resulting in one prediction per 15-s sequence. Each prediction was then used to label the corresponding individual seconds. For the third input (referred to hereafter as $\text{Intensity}_{\text{soj}}$), the second-by-second GVM data were grouped into bouts by the gyroscope-based bout detection algorithm (see pages 85-90), and features were calculated for each bout. Predictions were then made by the neural network, resulting in one prediction per bout. Each prediction was then used to label the corresponding individual seconds, at which point preparation was complete for all three decision tree inputs (i.e., GVM, Intensity_{15} , and $\text{Intensity}_{\text{soj}}$).

DEVELOPMENT. Data were used from 75% of participants ($n = 65$). As shown in Table 11, the decision tree had only one tuning parameter, which was tested at five values that were selected with reference to the GVM cut-points for SB in the two-regression models of Hibbing et al. (25), as discussed previously for the bout detection algorithm. The decision tree was implemented according to the following pseudocode, where SB_Threshold refers to the decision tree tuning parameter:

- if ($\text{Intensity}_{15} = \text{MVPA} \ \& \ \text{Intensity}_{\text{soj}} = \text{MVPA}$)
 predict MVPA
- else if ($\text{GVM} < \text{SB_Threshold} \ \& \ \text{SB} \in \{\text{Intensity}_{15}, \text{Intensity}_{\text{soj}}\}$)
 predict SB

Table 11. Summary of development and cross-validation for gyroscope-based youth Sojourn decision trees (for data from the hip or non-dominant wrist).

Summary	
Development	
<u>Data Used</u>	Continuous time course (75% of participants; n = 65; 139.0 total h of hip data and 138.2 total h of non-dominant wrist data)
<u>Grid Search</u>	5 permutations tested (1 parameter, 5 settings)
Tuning Parameter 1	<i>Description:</i> Sedentary threshold <i>Values tested:</i> 10, 20, 30, 40, and 50°·s ⁻¹ .
<u>Procedure</u>	For each data file: - Decision tree iterated for all 5 tuning grids For each participant and tuning grid: - Data pooled across visits (if applicable) - Percent correct classification calculated Analysis: - Percent correct classification averaged across participants for each tuning grid - Grid with highest average selected for cross-validation
Cross-Validation	
<u>Data Used</u>	Continuous time course (remaining 25% of participants; n = 21; 43.5 total h of data for both the hip and non-dominant wrist)
<u>Procedure</u>	For each data file: - Original decision tree (accelerometer) applied - New decision tree (gyroscope) applied For each participant and algorithm: - Data pooled across visits (if applicable) - Performance metrics calculated for both models <i>Overall agreement:</i> percent correct classification, kappa <i>Category-specific agreement:</i> sensitivity, specificity Analysis: - Metrics averaged across participants for both algorithms - For each metric, values compared to evaluate which model performed better

- else if (60+ consecutive upstream/downstream $GVM \leq SB_Threshold$)
predict SB
- else if (60+ consecutive upstream/downstream $MVPA \in \{Intensity_{15}, Intensity_{soj}\}$)
predict MVPA
- else
predict LPA

From the three inputs, the decision tree produced second-by-second intensity estimates, which were compared to the criterion labels in terms of percent correct classification. For participants with data from both visits, the visits were pooled prior to assessing performance, resulting in one percent correct classification value per participant and tuning grid. The grid-wise means were compared, and the best-performing grid was selected for cross-validation.

CROSS-VALIDATION. The gyroscope decision tree was applied to the testing partition, along with the existing accelerometer decision tree, which was invoked using the R function ``Sojourn:::youth_sojourn_tree``. Notably, gyroscope-based values for $Intensity_{15}$ and $Intensity_{soj}$ were used when testing both decision trees. That is, the gyroscope-based bout detection algorithm and neural network were used to obtain $Intensity_{15}$ and $Intensity_{soj}$ for both the accelerometer-based and gyroscope-based decision trees. Doing so ensured that performance differences would exclusively reflect differences between the decision trees, not the other components.

For participants with data from two visits, the data were pooled prior to analysis. Then for each participant, the decision tree estimates were compared to the criterion measure using tests of overall agreement (percent correct classification and kappa statistics) and category-specific

agreement (sensitivity and specificity). The metrics were averaged across participants for each decision tree, then used to make descriptive comparisons.

Results

Development and cross-validation results are presented in separate subsections for each component, with Table 12 showing a combined summary of cross-validation results for all three components. Results are presented as mean $\pm$ SD, except where otherwise noted.

Bout Detection Algorithm

DEVELOPMENT. For both the hip and non-dominant wrist, 25 tuning grids were removed because aggregated performance was missing for all participants (e.g. due to no true positives forming). The affected grids were the same for both sites. Thus, development results are presented for the 100 remaining grids. There were 16 participants (hip) and 45 participants (non-dominant wrist) with complete data across the tuning grids. For the hip, the highest aggregated performance ($51.5\% \pm 2.9\%$) was achieved using settings of $5^\circ \cdot s^{-1}$, $50^\circ \cdot s^{-1}$, and five s for the first, second, and third tuning parameters, respectively. For the non-dominant wrist, the highest aggregated performance ($41.7\% \pm 3.7\%$) was achieved by the same grid as for the hip.

CROSS-VALIDATION. There were no cases of missing performance metric values for either attachment site, meaning participants always had at least one true positive prediction. For the hip, aggregated performance was higher for the gyroscope algorithm ($51.2\% \pm 2.7\%$) than the original accelerometer algorithm ($42.3\% \pm 5.0\%$), which was attributable to 30.6% higher mean recall and 8.9% higher mean RMSE%. Mean precision for the gyroscope algorithm was less than half as high as the gyroscope algorithm, but both were $< 25\%$.

For the non-dominant wrist, aggregated performance was marginally higher for the gyroscope algorithm ($43.3\% \pm 4.5\%$) than the accelerometer algorithm ($41.6\% \pm 3.9\%$), which

Table 12. Cross-validation results for each of the accelerometer-based and gyroscope-based youth Sojourn model components. Values are mean $\pm$ SD, except for the neural network results, which were calculated for all steady state samples at once (i.e., not on a per-participant basis).

	Hip		Non-Dominant Wrist	
	ENMO	GVM	ENMO	GVM
Bout Detection Alg.				
Recall	56.9% $\pm$ 6.8%	87.5% $\pm$ 6.6%	61.0% $\pm$ 7.2%	68.5% $\pm$ 10.5%
Precision	24.7% $\pm$ 5.4%	11.7% $\pm$ 2.3%	18.2% $\pm$ 4.1%	12.1% $\pm$ 3.1%
RMSE%	45.5% $\pm$ 6.8%	54.4% $\pm$ 3.8%	45.6% $\pm$ 6.5%	49.4 $\pm$ 5.1%
AP	42.3% $\pm$ 5.0%	51.2% $\pm$ 2.7%	41.6% $\pm$ 3.9%	43.3% $\pm$ 4.5%
Neural Network				
<u>Overall Agreement</u>				
CC	53.8%	79.6%	70.0%	74.4%
Kappa	0.19	0.68	0.48	0.61
<u>SB</u>				
Sensitivity	19.8%	94.0%	69.8%	89.1%
Specificity	96.7%	98.9%	96.7%	97.8%
<u>LPA</u>				
Sensitivity	8.2%	68.4%	7.1%	63.3%
Specificity	82.2%	82.2%	88.4%	78.2%
<u>MVPA</u>				
Sensitivity	92.9%	74.2%	93.2%	69.0%
Specificity	41.9%	89.9%	63.8%	87.7%
Decision Tree				
<u>Overall Agreement</u>				
CC	71.8% $\pm$ 5.5%	73.3% $\pm$ 6.1%	70.3% $\pm$ 7.3%	71.2% $\pm$ 6.5%
Kappa	0.56 $\pm$ 0.08	0.59 $\pm$ 0.09	0.55 $\pm$ 0.11	0.56 $\pm$ 0.10
<u>SB</u>				
Sensitivity	99.3% $\pm$ 1.1%	92.3% $\pm$ 7.0%	82.5% $\pm$ 17.4%	86.8% $\pm$ 10.8%
Specificity	80.1% $\pm$ 7.1%	88.6% $\pm$ 4.5%	89.2% $\pm$ 5.3%	88.9% $\pm$ 4.7%
<u>LPA</u>				
Sensitivity	40.2% $\pm$ 12.1%	54.2% $\pm$ 12.9%	57.3% $\pm$ 13.2%	54.1% $\pm$ 13.2%
Specificity	85.8% $\pm$ 7.4%	80.5% $\pm$ 6.7%	76.2% $\pm$ 12.2%	77.8% $\pm$ 9.8%
<u>MVPA</u>				
Sensitivity	68.6% $\pm$ 14.4%	68.6% $\pm$ 14.3%	68.8% $\pm$ 13.3%	68.7% $\pm$ 13.3%
Specificity	92.0% $\pm$ 6.5%	92.0% $\pm$ 6.5%	91.4% $\pm$ 5.7%	91.4% $\pm$ 5.6%

ENMO- Euclidian norm minus one; GVM- gyroscope vector magnitude; Alg.- algorithm; RMSE%- relative root mean squared error; AP- aggregated performance (mean of recall, precision, and RMSE%); CC- correct classification; SB- sedentary behavior; LPA- light physical activity; MVPA- moderate-to-vigorous physical activity.

was attributable to 7.5% higher mean recall and 3.8% higher mean RMSE%. Mean precision was lower for the gyroscope algorithm than the accelerometer algorithm by a factor of 1.5, but both were $< 20\%$.

Neural Network

For both development and cross-validation, metrics were calculated after pooling all steady state data across participants, rather than on a per-participant basis. Therefore, metrics are given as single values rather than participant means and standard deviations.

DEVELOPMENT. For the hip, percent correct classification was highest (80.3%) when using five nodes in the hidden layer and a decay setting of 0.0. For the non-dominant wrist, percent correct classification was highest (74.6%) when using 20 nodes in the hidden layer and a decay setting of 0.1.

CROSS-VALIDATION. For the hip, overall agreement was higher for the new gyroscope model (79.6% correct classification, $\kappa = 0.68$) than the existing accelerometer model (53.8% correct classification, $\kappa = 0.19$), and the same was generally true for category-specific agreement. Specifically, the gyroscope model had 60.2%-74.1% higher sensitivity than the accelerometer model for SB and LPA, but 18.6% lower sensitivity for MVPA. Both models had specificity within 2.2% of one another for SB and LPA, whereas for MVPA the gyroscope model had considerably higher specificity than the accelerometer model (89.9% versus 41.9%).

For the non-dominant wrist, overall agreement was higher for the new gyroscope model (74.4% correct classification, $\kappa = 0.60$) than the existing accelerometer model (70.0% correct classification, $\kappa = 0.48$), and the same was generally true for category-specific agreement. Specifically, the gyroscope model had 19.3%-56.1% higher sensitivity than the accelerometer model for SB and LPA, but 24.2% lower sensitivity for MVPA. Both models had specificity

within 10.2% of one another for SB and LPA, whereas for MVPA the gyroscope model had considerably higher specificity than the accelerometer model (87.7% versus 63.8%).

Decision Tree

DEVELOPMENT. For the hip, the highest percent correct classification ($76.9\% \pm 5.3\%$) was achieved using a threshold of $10^\circ \cdot s^{-1}$. For the non-dominant wrist, the same threshold yielded the highest percent correct classification ($75.1\% \pm 6.2\%$).

CROSS-VALIDATION. For the hip, overall agreement was marginally higher for the new gyroscope model ($73.3\% \pm 6.1\%$ correct classification, $\kappa = 0.59 \pm 0.09$) than the existing accelerometer model ($71.8\% \pm 5.5\%$ correct classification, $\kappa = 0.56 \pm 0.08$). For the individual intensity categories, both models had mean sensitivity and specificity within 8.5% of one another in all cases except sensitivity for LPA, wherein the mean was 14.0% higher for the gyroscope model than the accelerometer model.

For the non-dominant wrist, overall agreement was marginally higher for the new gyroscope model ($71.2\% \pm 6.5\%$ correct classification, $\kappa = 0.56 \pm 0.10$) than the existing accelerometer model ($70.3\% \pm 7.3\%$ correct classification, $\kappa = 0.55 \pm 0.11$). For the individual intensity categories, both models had mean sensitivity and specificity within 4.3% of one another in all cases.

Summary and Transition

In this chapter, the original youth Sojourn calibration methods were replicated using gyroscope data instead of accelerometer data. The gyroscope-based and accelerometer-based model components were then compared to one another, as a preliminary way of assessing the potential merits of gyroscope data within the youth Sojourn framework. In terms of segmentation, the findings showed that aggregated performance was higher for gyroscope-based

bout detection algorithms than their accelerometer-based counterparts. In terms of intensity classification, the findings were similar, with gyroscope-based neural networks and decision trees showing higher percent correct classification than their accelerometer-based counterparts, although the differences were small in some cases. For specific intensities, the findings showed that gyroscope data were more effective than accelerometer data for detecting SB and LPA, while accelerometer data were more effective for detecting MVPA. Altogether, there was clear evidence in favor of using gyroscope data within the youth Sojourn framework. Thus, the next chapter will focus on exploratory development of gyroscope inclusive youth Sojourn models.

CHAPTER V

RESEARCH QUESTION 2: METHODS AND RESULTS

Overview

In this chapter, new youth Sojourn models were developed using the continuous time course data that were presented in Chapter 3 (see pages 62-80). The new models were developed without a decision tree component, meaning they included only a segmentation algorithm and neural network. The segmentation algorithms combined sensor fusion and change point detection, and the gyroscope-based neural networks from Chapter 4 (see pages 81-101) were used for the classification. For each attachment site, two models were developed having different segmentation properties. One model ($\text{Soj}_{\text{recall}}$) was designed to favor recall over precision by erring on the side of over-predicting transitions, and the other algorithm ($\text{Soj}_{\text{precision}}$) was designed to favor precision over recall by erring on the side of under-predicting transitions. The contrasts were designed to elucidate the relative importance of recall and precision in terms of eventual model performance (i.e., intensity classification accuracy). Furthermore, $\text{Soj}_{\text{precision}}$ was designed to run more quickly than $\text{Soj}_{\text{recall}}$, to examine the potential tradeoff between runtime and performance.

In the following section, the change point detection method is described, followed by descriptions of $\text{Soj}_{\text{recall}}$ and $\text{Soj}_{\text{precision}}$. Afterwards, the analyses are presented, in which the continuous time course data set was used to conduct performance comparisons between the new youth Sojourn models and the original raw acceleration-based models that were developed by Hibbing et al. (34). The analyses were designed to address the second research question, which was formulated as follows:

2. **Question:** In a laboratory setting, does the performance of youth Sojourn models improve when using sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification?

Hypothesis: When applied to laboratory data, youth Sojourn models that use sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification will have improved performance, compared to existing youth Sojourn models.

Methods

Change Point Detection Method

The pruned exact linear time (PELT) method was developed by Killick et al. (196), and it is an extension of a dynamic programming method called optimal partitioning (242). Both methods identify change points in such a way that a cost function is minimized. A global minimum is reached, meaning both algorithms are exact (as opposed to approximate methods like binary segmentation). The distinction between the two is that PELT adds a pruning step to each iteration of the underlying algorithm, whereby data points are flagged if they cannot contribute to minimizing the cost function. The flagged data points are then excluded from subsequent iterations, which can dramatically reduce the algorithm runtime, in some cases achieving a linear time complexity that contrasts sharply with the polynomial or exponential complexities of other algorithms. In their original work, Killick et al. (196) showed that PELT was able to achieve better performance than binary segmentation, and similar performance (with substantially faster runtime) to optimal partitioning. PELT is available in the R package changepoint (243), and it can be run for detection of changes in the signal mean, variance, or both. Each of the new youth Sojourn algorithms draw on PELT for identifying variance change points in accelerometer data.

Soj_{recall} Algorithm

The Soj_{recall} algorithm takes two inputs, namely second-by-second vector magnitude (VM) and gyroscope vector magnitude (GVM) data (see Figure 4A and Figure 4B, respectively). From the inputs, two main features are calculated, namely point-to-point changes in the VM data (i.e, lag-1 differences; see Figure 4C) and intensity estimates (see Figure 4D) obtained using a rolling 5-s window (GVM neural network).

There are two main procedures in Soj_{recall}. In the first procedure, the standard PELT method is used to detect variance change points in the lag-1 VM differences (see Figure 4E). The default PELT settings are used, except that the minimum segment length is set to 5 s. In the second procedure, the intensity predictions are searched for prolonged (> 60 s) periods of continuous SB (see Figure 4F). A change point is predicted at the start and end of all such periods, and the change points from both processes are then pooled (see Figure 4G).

After both main processes are complete, a final screening process is executed. Two steps are used to screen the change points. First, change points are rejected if they occur at the first or last data points in the time series. Second, PELT change points are rejected if they occur within ± 5 s of the start or end of a bout of SB.

Soj_{precision} Algorithm

The Soj_{precision} algorithm is different from the Soj_{recall} algorithm in two ways. First, it is stricter in terms of rejecting PELT change points that coincide with bouts of SB. Second, it uses data smoothing to reduce the number of data points prior to implementing PELT, thereby accelerating the segmentation process. Both features of Soj_{precision} favor precision over recall, the first by suppressing all predicted transitions that occur during a bout of SB (as opposed to suppressing only those that occur within ± 5 s of the start or end of the bout), and the second by

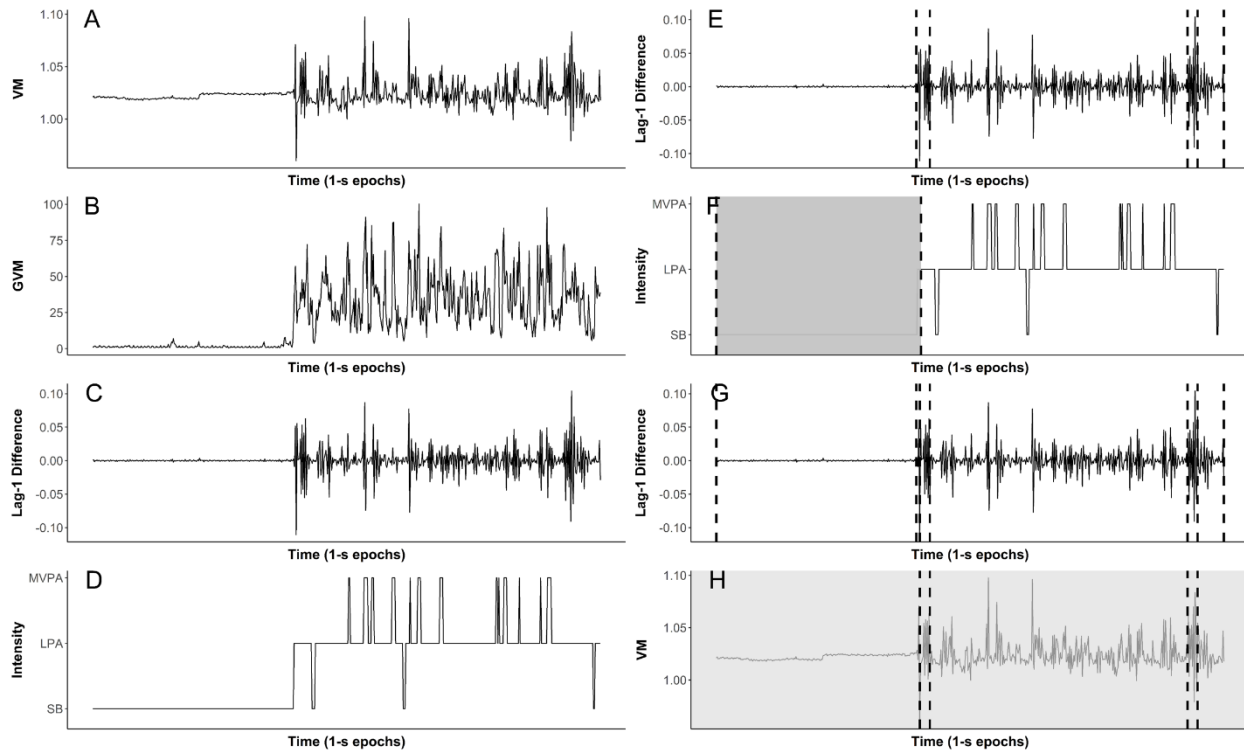


Figure 4. Visual depiction of the Soj_{recall} algorithm, where vertical dashed lines represent transitions detected by the algorithm. Inputs are: A) second-by-second vector magnitude (VM) from the primary accelerometer, and B) second-by-second gyroscope vector magnitude (GVM). Features are: C) lag-1 differences of the VM data, and D) second-by-second GVM neural network output (SB- sedentary behavior; LPA- light physical activity; MVPA- moderate-to-vigorous physical activity), obtained via a rolling window. Procedures are: E) standard PELT algorithm is applied to lag-1 differences; F) prolonged bouts (> 60 s) of SB are identified, and additional change points are predicted at the start and end of each bout; G) all change points are pooled; H) change points are rejected if they occur at the first or last data point, and if PELT change points occur within ± 5 s of the start or end of a bout of SB.

limiting the number of change points that can be predicted. Outside of those two differences, the algorithms are mostly similar. $\text{Soj}_{\text{precision}}$ takes the same inputs as $\text{Soj}_{\text{recall}}$ (i.e., second-by-second VM and GVM; see Figure 5A and Figure 5B, respectively). From the inputs, three main features are calculated. The first feature is a vector of absolute lag-1 differences (see Figure 5C). The second feature is a series of signed lag-1 differences obtained after smoothing the VM data (i.e., averaging every 5 s in non-overlapping windows; see Figure 5D). The third feature is a series of intensity predictions from a GVM neural network, where one prediction is made every 5 s (see Figure 5E).

There are two main procedures in the $\text{Soj}_{\text{precision}}$ algorithm. In the first procedure, the standard PELT method is used to identify variance change points in the smoothed VM data (see Figure 5F). The default PELT settings are used, except that the minimum segment length is set to 5 s. By using smoothed VM data, the PELT method is prevented from treating aberrant changes in second-by-second VM as change points. However, a correction is necessary to account for the loss in data resolution. To perform the correction, second-by-second VM changes are examined within ± 5 s of each change point (see Figure 5G), and the change point is assigned at whichever point the largest change occurs (see Figure 5H). In the second procedure, the neural network predictions are used to identify prolonged periods (> 60 s) of continuous SB. A change point is predicted at the start and end of all such periods (see Figure 5I).

After both main procedures are complete, a final screening process is executed (see Figure 5J). Three steps are used to screen the change points. First, change points are rejected if they occur at the first or last data points in the time series. Second, PELT change points are rejected if they occur during a bout of SB, as determined by the neural network. Third, a refractory period is instituted, whereby 5 s must elapse after a change point before another change point can occur.

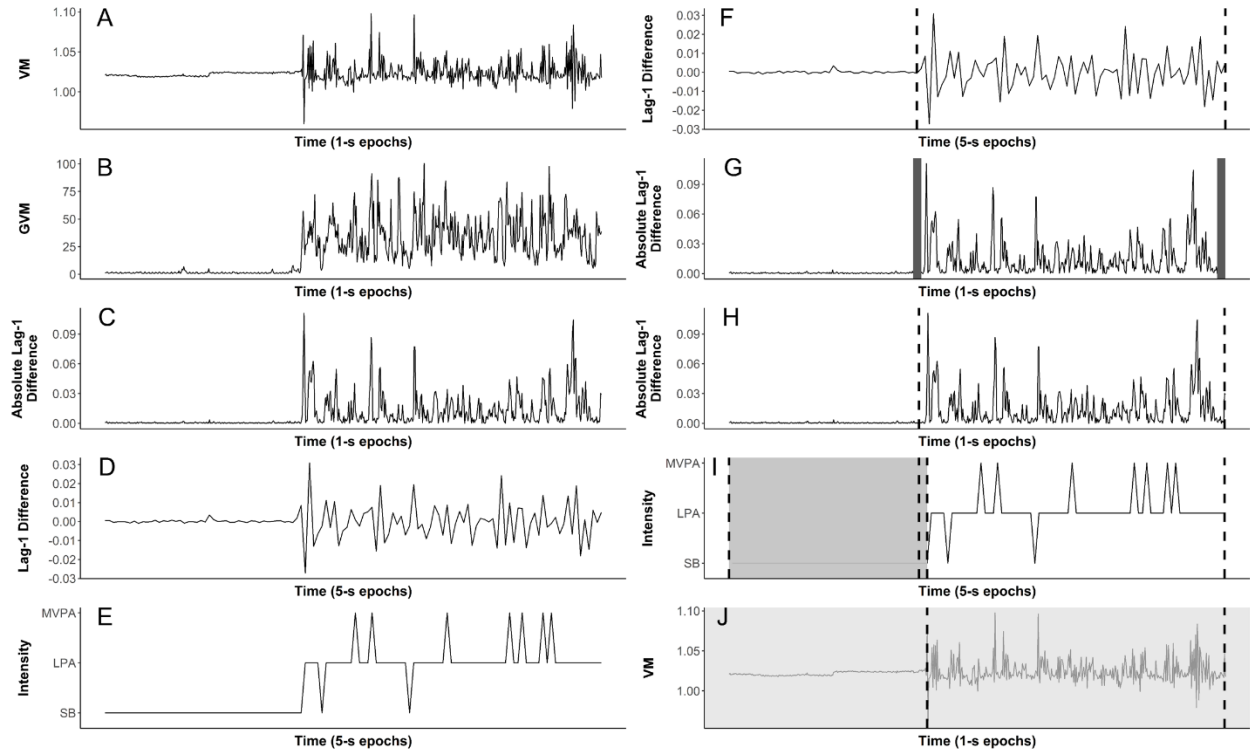


Figure 5. Visual depiction of the Soj_{precision} algorithm, where vertical dashed lines represent transitions detected by the algorithm. Inputs are: A) second-by-second vector magnitude (VM) from the primary accelerometer, and B) second-by-second gyroscope vector magnitude (GVM). Features are: C) absolute lag-1 differences of VM, D) signed lag-1 differences of smoothed VM data, and E) GVM neural network output (SB- sedentary behavior; LPA- light physical activity; MVPA- moderate-to-vigorous physical activity) in 5-s epochs. Procedures are: F) standard PELT algorithm is applied to signed 5-s VM differences to detect variance change points; G) second-by-second magnitude changes are examined within ± 5 s of each predicted change point; H) change points are assigned to the highest magnitude change in each range; I) prolonged bouts (> 60 s) of SB are identified, and additional change points are predicted at the start and end of each bout; J) change points are rejected if they occur at the first or last data point, or if a PELT change point occurs during a prolonged bout of SB. A refractory period also prohibits predictions from occurring within 5 s of one another.

Analysis

Versions of Soj_{recall} and Soj_{precision} were developed for the hip, non-dominant wrist, and dominant wrist. For each attachment site, performance comparisons were made between Soj_{recall}, Soj_{precision}, and the original youth Sojourn models (34). Notably, there was no original model for dominant wrist data, and thus the original non-dominant wrist model was used for both wrists. The same approach was taken in the free-living portion of the original calibration study (34), wherein the findings showed marginal performance differences between the wrists.

For all analyses in the present study, data were excluded from files in which there were missing sensor values (e.g. a monitor came off during the test), since the segmentation algorithms were designed for application to uninterrupted data. There were two main analyses, one focused on segmentation and the other focused on activity intensity classification. Below, the two analyses are described separately.

ANALYSIS OF SEGMENTATION. The segmentation components of Soj_{recall} and Soj_{precision} were developed using an exploratory process that involved data from 75% of participants (n = 65). The analyses described below were then carried out using data from the remaining 25% of participants (n = 21). As shown in Table 13, participant characteristics were similar for the development and analysis subsets. The similarity was achieved using a guided random process. Specifically, random assignments were made repeatedly until a manually specified frequency target was met. By pairing a frequency-based target with random selection, bias was avoided that could arise from hand-selecting specific participants for certain subsets.

For the analyses (using 25% of participants), the transition pairing method (TPM) of Hibbing et al. (35) was used to make site-specific comparisons of segmentation correctness for Soj_{recall}, Soj_{precision}, and the original youth Sojourn model. Comparisons were made for various settings of

Table 13. Distribution of key participant characteristics within the subsamples used for developing and testing the segmentation components in Soj_{recall} and Soj_{precision}. Values are n(%).

	Exploratory Development (75%; n = 65)	Analysis (25%; n = 21)	Total (n = 86)
Age Group (yr)			
< 9.0	10 (15%)	3 (14%)	13 (15%)
9.0-11.9	20 (31%)	7 (33%)	27 (31%)
12.0-14.9	15 (23%)	4 (19%)	19 (22%)
≥ 15	20 (31%)	7 (33%)	27 (31%)
Weight Status			
Underweight/Normal Weight	54 (83%)	18 (86%)	72 (84%)
Overweight/Obese	11 (17%)	3 (14%)	14 (16%)
Handedness			
Left	4 (6%)	1 (5%)	5 (6%)
Right	61 (94%)	20 (95%)	81 (94%)

the spurious pairing threshold, namely 1-20 s in 1-s increments. Predicted and criterion transitions were first obtained for each participant and visit. For those with data from two visits ($n = 17$), the values were then pooled across visits. For each participant, performance metrics were then calculated, namely recall, precision, relative root mean squared error (RMSE%), and aggregated performance. Recall and precision are commonly used in many classification settings, but RMSE% and aggregated performance are unique to the TPM. Ordinarily, root mean squared error (RMSE) is used on an absolute scale to represent timing error by a dynamic segmentation algorithm, and lower values reflect higher performance. Within the TPM, the maximum lag time is known (i.e., equal to the spurious pairing threshold value), as is the minimum lag time (i.e., 0 s). Therefore, it is possible to express RMSE on a relative scale (i.e., RMSE%) such that performance is reflected the same way as precision and recall, namely as a percentage between 0% and 100% where higher values reflect better performance. To calculate RMSE%, the following formula can be used: $RMSE\% = 100\% - \left(\frac{RMSE}{SPT} * 100\% \right)$, where SPT is the spurious pairing threshold value. Aggregated performance is defined as the mean of recall, precision, and RMSE%. More information is available in the vignette for the R implementation of the TPM (see <https://cran.r-project.org/web/packages/PAutilities/vignettes/tpm.html>).

After calculating performance metrics for each participant, the values were averaged for analysis. Listwise deletion was used to ensure that the same cases were used when making comparisons among models for a given metric, attachment site, and spurious pairing threshold setting. Statistical significance was examined using the overlapping confidence interval method. That is, for a given metric there was a statistically significant difference between two methods whenever there was no overlap between their 95% confidence intervals. The latter approach was chosen for simplicity (given the large number of comparisons), and in accord with the secondary

importance of the segmentation analysis, behind the analysis of intensity classification agreement.

ANALYSIS OF ACTIVITY INTENSITY CLASSIFICATION. Tests of classification agreement were used for evaluating the second-by-second intensity classifications from each method. All participants ($n = 86$) were included in the analyses. Predicted and criterion intensities were first obtained for individual visits, and then for participants with data from both visits ($n = 73$ for hip and dominant wrist data; $n = 72$ for non-dominant wrist data) the values were pooled before analysis.

The predictions were compared to the criterion measure using a total of 11 metrics. Two metrics (percent correct classification and kappa statistic) reflected overall agreement, while the remaining nine reflected agreement for individual intensity categories. Specifically, for each of the three intensities (SB, LPA, and MVPA), agreement was assessed with sensitivity, specificity, and F1 (i.e., the harmonic mean of recall and precision). In the event of missing values for a given metric (e.g. undefined F1 due to recall or precision of zero), the values were discarded for all three methods to ensure the same samples were used for group level comparisons.

After obtaining the metrics for each participant, the values were averaged across participants and compared statistically. Specifically, 33 tests were run (11 metrics times three attachment sites) using repeated measures analysis of variance (ANOVA). Pairwise post hoc tests were then used to compare Soj_{recall} , $Soj_{precision}$, and the original youth Sojourn model. The false discovery rate correction (244) was applied to account for multiple comparisons.

Results

Analysis of Segmentation

MISSING PARTICIPANT DATA. Missing data led to exclusion of data from eight visits, including both visits from two participants. Thus, the holdout sample included data from 19 participants, of whom 18 had data for both visits. For the 1-s spurious pairing threshold, RMSE% and aggregated performance data were excluded from one additional participant (hip and non-dominant wrist only) due to one or more methods predicting no true positives. For spurious pairing thresholds ≥ 2 -s, all analyses used the same 19-person sample.

REJECTION OF INDIVIDUAL TRANSITION PAIRINGS. Non-sequential pairings were rare for any of the segmentation methods (i.e., original youth Sojourn, Soj_{recall}, and Soj_{precision}), meaning few pairings were rejected by the TPM. Specifically, there were no rejected pairings for devices worn on the hip or non-dominant wrist. For the dominant wrist, there were no rejections for spurious pairing threshold settings ≤ 13 s, whereas for higher settings 1-2 participants had a rejected pairing (never more than one per participant).

OUTCOMES. Figure 6 shows results of the segmentation analysis. Aggregated performance (see Figure 6D, Figure 6H, and Figure 6L) was similar for all three methods across attachment sites, with no differences exceeding 5.3%. The same was true for RMSE%, (see Figure 6C, Figure 6G, and Figure 6K) with the vast majority of values falling within 5-6% of one another. In contrast, recall and precision were dramatically different for the three segmentation algorithms, as discussed below.

The original youth Sojourn algorithm had the highest mean recall for all wear locations, regardless of the spurious pairing threshold (see Figure 6A, Figure 6E, and Figure 6I). Mean recall for the original youth Sojourn algorithm exceeded 50% for threshold values of 4-5 s, and

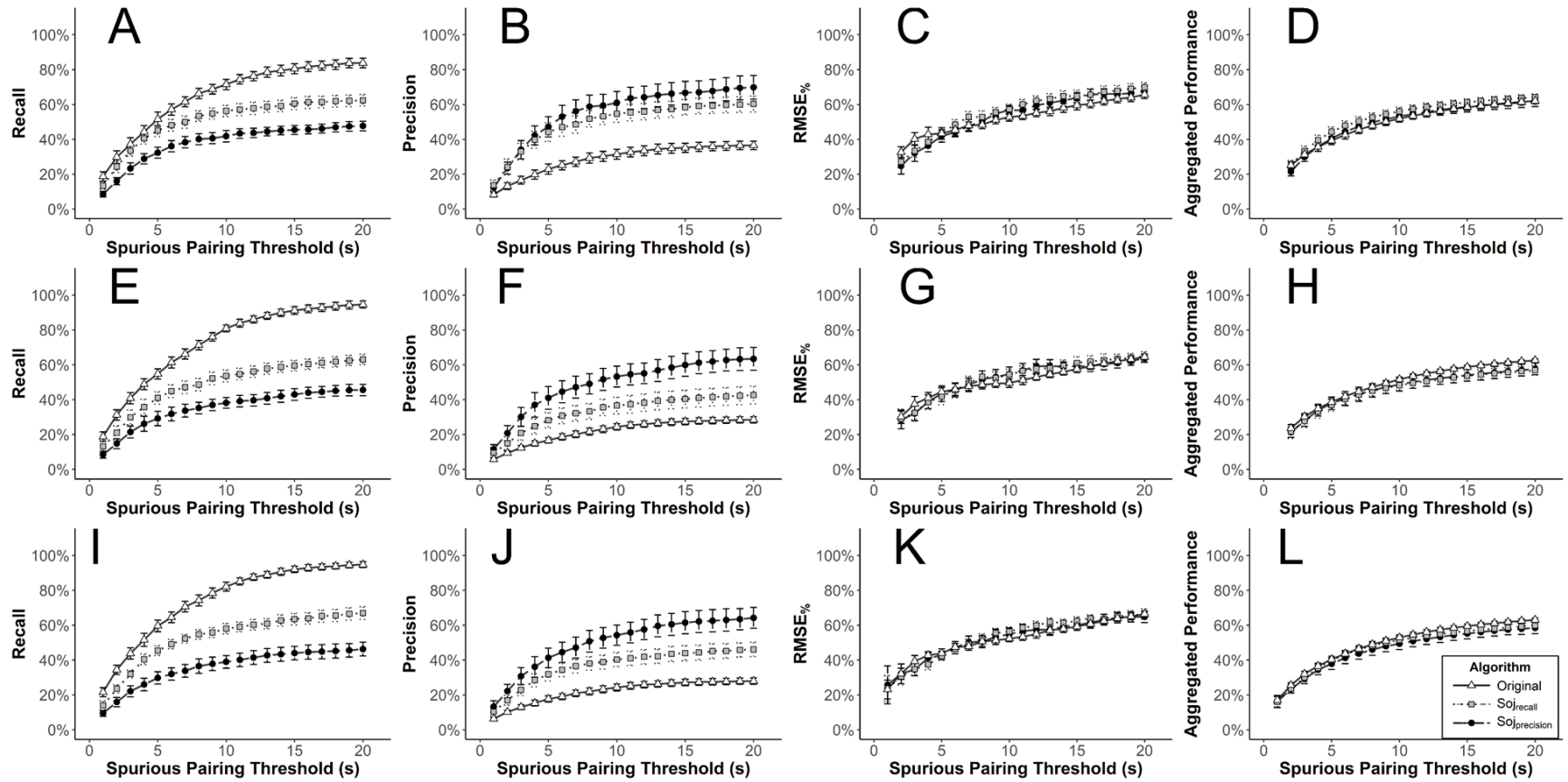


Figure 6. Performance of segmentation algorithms when applied to the testing subset. Results are shown for recall, precision, RMSE%, and aggregated performance on the hip (A-D, respectively), non-dominant wrist (E-H, respectively), and dominant wrist (I-L, respectively). Values are means, and error bars are 95% confidence intervals. The spurious pairing threshold (x-axis) reflects the longest-allowable lag time for true positives.

exceeded 75% for threshold values of 9-12 s. Mean recall was greater for Soj_{recall} than $Soj_{precision}$ (regardless of wear location and spurious pairing threshold), by 4.7%-20.9%. For Soj_{recall} , recall exceeded 50% when the spurious pairing threshold was 7-9 s, and recall $> 75\%$ was never achieved (maximum of 66.9% for the dominant wrist, spurious pairing threshold of 20 s). For $Soj_{precision}$, mean recall did not exceed 50% for any setting of the spurious pairing threshold (maximum of 47.7% for the hip, spurious pairing threshold of 20 s).

The original youth Sojourn algorithm had the lowest mean precision for all wear locations, regardless of the spurious pairing threshold (see Figure 6B, Figure 6F, and Figure 6J). Specifically, mean precision for the original youth Sojourn algorithm never exceeded 36.6% (hip, spurious pairing threshold of 20 s). For the PELT-based algorithms, mean precision was greater for $Soj_{precision}$ than Soj_{recall} (for all but spurious pairing thresholds of 1-2 s on the hip only), by 1.1%-21.0%. For Soj_{recall} , precision only exceeded 50% for the hip (spurious pairing threshold ≥ 8 s; maximum mean precision of 60.3%), whereas it was always $\leq 46.1\%$ for the wrists. For $Soj_{precision}$, mean precision exceeded 50% when the spurious pairing threshold was 6-9 s, but it never exceeded 69.8% (hip, spurious pairing threshold of 20 s).

Analysis of Activity Intensity Classification

Missing data led to exclusion of 31 visits, after which there were 78 participants left for analysis. Most of the participants had data from both visits ($n = 49$ for the hip and dominant wrist; $n = 48$ for the non-dominant wrist). No additional exclusion occurred for monitors worn on either wrist. For the hip, participants were excluded for the analyses of F1 for SB ($n = 7$ participants excluded; $n = 71$ remaining, 45 with data for both visits) and LPA ($n = 4$ participants excluded; $n = 74$ remaining, 48 with data for both visits).

A total of 33 ANOVAs were performed (11 metrics times three attachment sites), and main effects were significant for 31 of the 33 ($p < 0.03$). The exceptions were SB F1 on the non-dominant wrist, and LPA specificity on the dominant wrist ($p = 0.11$ - 0.12). Figure 7 shows the results of the entire classification analysis. Below, results are presented sequentially for overall agreement (percent correct classification and kappa), SB, LPA, and MVPA. Summary statistics are presented as mean $\pm$ standard deviation.

OVERALL AGREEMENT. For the hip, mean percent correct classification was 19.6%-20.7% higher for Soj_{recall} ($76.5\% \pm 7.1\%$) and Soj_{precision} ($77.6\% \pm 6.9\%$) than the original model ($56.9\% \pm 15.4\%$; both $p < 0.001$). The differences were more modest at the wrists, where the greatest difference occurred on the dominant side, between Soj_{recall} ($71.7\% \pm 7.5\%$) and the original model ($68.6\% \pm 7.9\%$; $p = 0.02$). Kappa statistics followed similar trends to percent correct classification, with means ranging from 0.55-0.65 (for Soj_{recall} and Soj_{precision}) versus 0.34-0.53 (for the original models; all $p < 0.05$). Overall agreement was extremely similar between Soj_{recall} and Soj_{precision}, with non-significant differences ($p > 0.12$) for all but the hip.

SEDENTARY BEHAVIOR. For the hip, mean sensitivity was approximately twice as high for Soj_{recall} ($96.7\% \pm 4.3\%$) and Soj_{precision} ($94.8\% \pm 5.7\%$) than the original model ($48.5\% \pm 31.1\%$; both $p < 0.001$), whereas mean specificity for all three methods was between 87.5% and 94.7%. Mean F1 scores were 25.9%-26.3% higher for the new models than the original (both $p < 0.001$). For the wrists, mean sensitivity was $> 90\%$ for Soj_{recall} and Soj_{precision}, whereas it was significantly lower (82.4%-84.9%; all $p < 0.04$) for the original model. Mean specificities for all three models were between 84.3% and 87.3%, and F1 scores were higher for the new models than the original by 2.0%-7.9% ($p = 0.27$ for both comparisons on the non-dominant wrist; $p < 0.001$ for both comparisons on the dominant wrist). When comparing Soj_{recall} and Soj_{precision} in

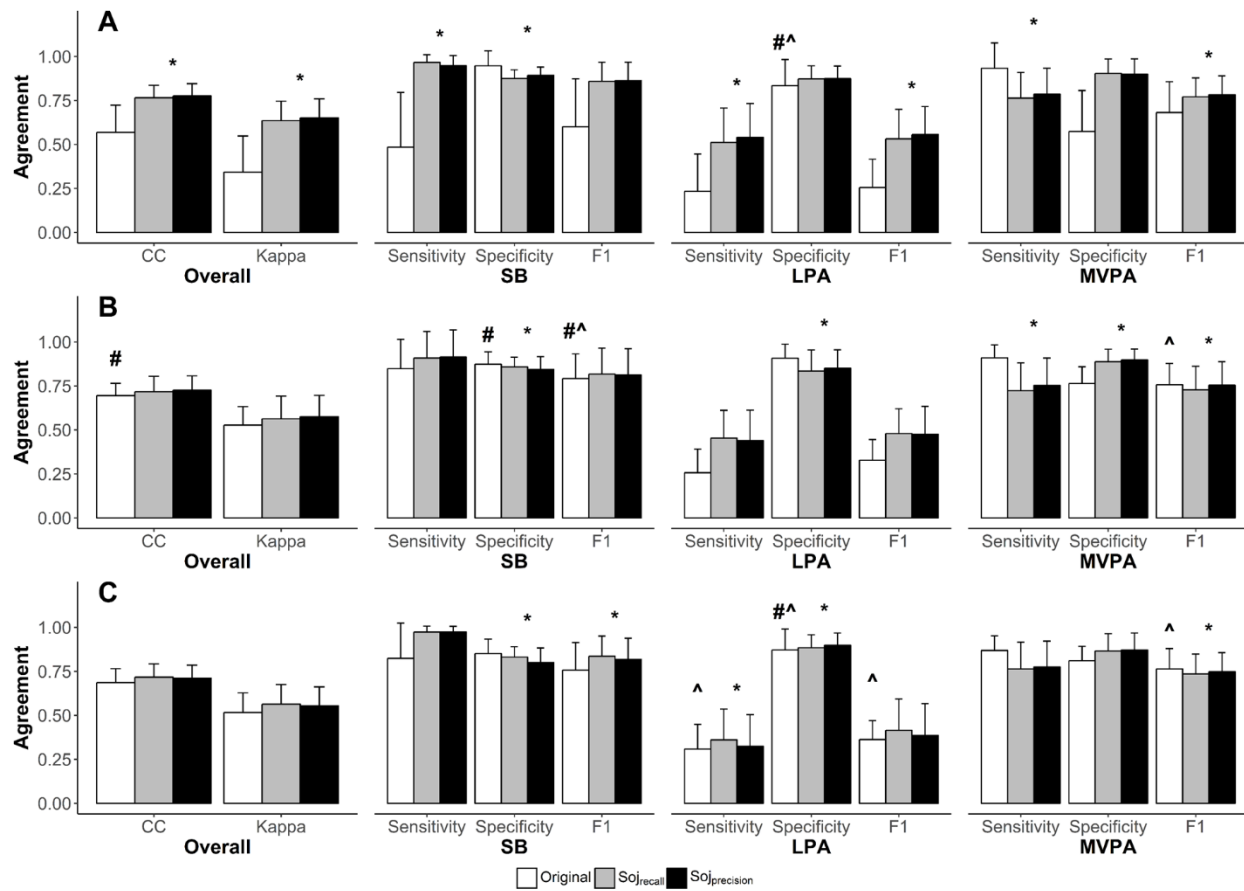


Figure 7. Classification agreement of the original youth Sojourn models and the newly developed, gyroscope inclusive Soj_{recall} and Soj_{precision} models when applied to the laboratory data set. Assessments include overall agreement and intensity-specific agreement for sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA). Attachment sites are (A) hip, (B) non-dominant wrist, and (C) dominant wrist. Error bars are standard deviation.

CC- correct classification

#Non-significant difference ($p \geq 0.05$) between original youth Sojourn and Soj_{recall}

^Non-significant difference ($p \geq 0.05$) between original youth Sojourn and Soj_{precision}

*Significant difference ($p < 0.05$) between Soj_{recall} and Soj_{precision}

All other comparisons involving the original youth Sojourn model were significant ($p < 0.05$).

terms of sensitivity, specificity, and F1, means differed by $\leq 3.1\%$ regardless of attachment site ($p = 0.00-0.70$).

LIGHT PHYSICAL ACTIVITY. For the hip, mean sensitivity was more than twice as high for Soj_{recall} ($51.2\% \pm 19.5\%$) and Soj_{precision} ($54.1\% \pm 19.1\%$) than the original model ($23.3\% \pm 21.2\%$; both $p < 0.001$), and mean specificity was between 83.4% and 87.4% for all three models. F1 scores were 27.7%-30.2% higher for the new models than the original (both $p < 0.001$). For the wrists, the findings differed. On the non-dominant wrist, the patterns were largely the same as for the hip, with mean sensitivity of 44.0%-45.4% (for Soj_{recall} and Soj_{precision}) versus 25.7% (original model; both $p < 0.001$), and mean specificity between 83.4% and 90.8% for all three models. F1 scores were 14.7%-15.1% higher for the new models than the original (both $p < 0.001$). On the dominant wrist, the three methods were much more similar to each other (mean sensitivity of 30.9%-36.0%; mean specificity of 87.2%-89.9%; mean F1 of 36.3%-41.5%; pairwise $p = 0.01-0.49$). When comparing Soj_{recall} and Soj_{precision} in terms of sensitivity, specificity, and F1, means differed by $\leq 3.7\%$ regardless of attachment site ($p = 0.00-0.74$).

MODERATE-TO-VIGOROUS PHYSICAL ACTIVITY. For the hip, mean sensitivity was significantly lower for Soj_{recall} ($76.4\% \pm 14.6\%$) and Soj_{precision} ($78.7\% \pm 14.7\%$) than the original model ($93.3\% \pm 14.5\%$; both $p < 0.001$), whereas mean specificity was significantly higher (90.0%-90.4% for Soj_{recall} and Soj_{precision}, versus 57.4% for the original model; both $p < 0.001$). F1 scores were 8.9%-10.1% higher for the new models than the original (both $p < 0.001$). For the wrists, mean sensitivity was 72.3%-77.6% for Soj_{recall} and Soj_{precision}, whereas it was significantly higher for the original model (86.9%-91.0%; all $p < 0.001$). In contrast, mean specificity was 86.6%-89.8% for Soj_{recall} and Soj_{precision}, whereas it was significantly lower for the original model (76.4%-81.0%; all $p < 0.001$). F1 scores were 0.3%-2.9% lower for the new models than the

original ($p = 0.02-0.82$). When comparing Soj_{recall} and $Soj_{precision}$ in terms of sensitivity, specificity, and F1, means differed by $\leq 3.0\%$ regardless of attachment site ($p = 0.00-0.81$).

Summary and Transition

In this chapter, two gyroscope inclusive youth Sojourn models (i.e., Soj_{recall} and $Soj_{precision}$) were developed, with versions of both models being developed for the hip, non-dominant wrist, and dominant wrist. The models used both accelerometer and gyroscope data for segmentation, and then intensity classifications were obtained from gyroscope data alone. The main finding was that the new models demonstrated higher overall performance than the existing youth Sojourn models, with a greater margin for hip data than for data from either wrist. Although the findings were promising, it is important to note that the data were collected in a semi-structured laboratory setting. Independent validations in free-living settings are essential for understanding the generalizability of model performance. Thus, the ensuing chapters will be focused on an independent validation conducted in free-living settings. In the next chapter, a free-living data set is introduced, and in the following chapter, the independent free-living validation is performed.

CHAPTER VI

DESCRIPTION OF FREE-LIVING DATA SET

Data were drawn from a larger study that was conducted in free-living environments. Similar to the laboratory study (see Chapter 3), the present investigation used only data from indirect calorimetry, direct observation (DO), and research-grade multi-sensor devices worn on the hip and both wrists. Data from other monitors worn by the participants (e.g. on the thighs and ankles) will not be discussed. In this chapter, the focus is strictly on describing how the data set was generated, whereas the next chapter (pages 136-153) will describe how the data were analyzed to address the third research question.

Participants

At the time of this investigation, data were still being collected and processed for the parent study. Thus, data were only available from a subset of the participants ($n = 42$, age 6-18 y), of whom 10 also participated in the laboratory study that was presented in Chapter 3 (pages 62-80). Recruitment was accomplished through word of mouth and publicly posted flyers, and was specifically aimed at obtaining an even balance across sex (i.e., matching numbers of males and females) and age group (i.e., matching numbers of participants between 6-9, 9-12, 12-15, and 15-18 years old). Written parental informed consent and youth assent were obtained prior to data collection, and exclusion criteria (e.g. contraindications to exercise, or recent musculoskeletal injury) were screened based on a health history questionnaire that was filled out for the participants by a parent or guardian. The study was approved by the University of Tennessee, Knoxville Institutional Review Board. The key materials used for recruitment and enrollment are shown in Appendix B (see page 219). Notably, the health history questionnaire matched the questionnaire used for the laboratory study (see page 212), and thus is not shown in Appendix B.

Protocol

Participants were seen on two occasions. The first was conducted in a home setting, and the second was conducted in a location selected by the participant (e.g. a park or gym), with an emphasis on choosing a setting conducive to distinct behaviors from the first visit.

At the first visit, height and weight were assessed using a portable stadiometer and scale, respectively. Participants were then fitted with activity monitors (ActiGraph GT9X; ActiGraph LLC, Pensacola, FL, USA), as well as a portable metabolic unit (Cosmed K4b² or Cosmed K5; Cosmed, Rome, Italy). A resting metabolic rate (RMR) assessment was then conducted, consisting of 30 min of supine rest. The remainder of the visit involved a free-living assessment lasting approximately 2 h, in which participants were allowed to self-select their behaviors while the research team observed.

For the second visit, participants were fitted with the same equipment worn during the first visit. Another free-living assessment was then conducted, again lasting approximately 2 h, and again performed while the research team observed.

All 42 of the participants opted into having the trials video recorded. At the start of each recording, the Windows system clock was briefly filmed for future calculation of Windows timestamps. Video DO was then performed post-hoc using desktop software (Noldus Observer XT; Noldus Information Technology bv, Wageningen, The Netherlands) to document activity type and transition timing, as well as posture.

Equipment

ActiGraph GT9X

For each participant, data from three GT9X devices were used. One device was worn on the right hip along the anterior axillary line at the level of the iliac crest, and the other two were

worn on the wrists (i.e., one each on the dominant and non-dominant sides) immediately superior to the ulnar and radial styloid processes. The GT9X contains a triaxial primary accelerometer and an inertial measurement unit (IMU). The primary accelerometer has a dynamic range of ± 8 g and is the same component used in ActiGraph GT3X+ models. It can be configured to sample at rates between 30 Hz and 100 Hz, and the data can be extracted in two formats, namely raw acceleration (in g; low pass filtered at a proprietary cutoff) or the integrated proprietary units known as “activity counts”, which are integrated over regular intervals of user-specified length between 1 s and 240 s. The IMU contains a device thermometer and three triaxial sensors, i.e., secondary accelerometer (dynamic range ± 16 g), gyroscope (dynamic range $\pm 2000^{\circ}\cdot\text{s}^{-1}$), and magnetometer (dynamic range ± 4800 μT). The IMU sensors can be selectively enabled, but the sampling rate (100 Hz) is not modifiable. Prior to each session, the GT9X devices were initialized to collect at 90 Hz (primary accelerometer) with all IMU sensors activated. Timestamps were automatically synced with the Windows system clock during initialization.

Noldus Observer XT

Video DO was performed using the Noldus Observer XT desktop application. Specifically, videos were imported to the software and annotated using a specialized coding scheme. At the start of each coding session, a default behavior called “CommentStartTime” prompted the user to enter a comment indicating the Windows system clock time that was filmed on the camera. Doing so allowed future temporal alignment with other data sources. Afterwards, users clicked a series of buttons whenever a participant transitioned to a new activity or posture. The first button indicated the activity being performed (see Table 14 for common examples), and the second indicated the posture being assumed, with eight options (lying, sitting, reclining, standing, stepping, squatting/kneeling, mixed posture, or unknown). Simultaneous behaviors were not

Table 14. Activity selections in the coding scheme for video direct observation. To aid interpretability, activities are grouped under common headings, which were not per se part of the coding scheme.

Activity	Operational Definition
Ambulatory Activities	
Running (overground)	Running continuously in an overground manner (e.g. on a sidewalk)
Running (treadmill)	Running continuously on a treadmill
Stair climbing	Traversing stairs (ascending or descending) in a continuous manner (i.e., excluding transitional use of stairs)
Walking (overground)	Walking continuously in an overground manner (e.g. on a sidewalk)
Walking (treadmill)	Walking continuously on a treadmill
Baseball and Softball	
Drills	Performing baseball- or softball-related tasks (e.g. batting, pitching, fielding) outside of a gameplay scenario
Gameplay	Performing baseball- or softball-related tasks (e.g. batting, pitching, fielding) in a gameplay scenario
Basketball	
Drills	Performing basketball-related tasks (e.g. dribbling, shooting) outside of a gameplay scenario
Gameplay	Performing basketball-related tasks (e.g. dribbling, shooting) in a gameplay scenario
Football	
Drills	Performing football-related tasks (e.g. passing, catching, kicking) outside of a gameplay scenario
Gameplay	Performing football-related tasks (e.g. passing, catching, kicking) in a gameplay scenario
Golf	
Drills	Performing golf-related tasks (e.g. driving, putting) outside of a gameplay scenario
Gameplay	Performing golf-related tasks (e.g. driving, putting) in a gameplay scenario
Homework	
Non-screen-based	Performing homework-related tasks not involving a screen (e.g. written assignments)
Screen-based	Performing homework-related tasks involving a screen, including mixed media tasks (e.g. involving hand-written portions with digital prompts or feedback)
Household Chores	
Cleaning or wiping	Clearing grime or debris from a surface (e.g. counter, tabletop) using a cloth, rag, sponge, etc.
Cooking	Performing food production tasks (e.g. chopping, mixing, pouring)

Table 14. (Continued)

Activity	Operational Definition
Household Chores (Continued)	
Dishes	Cleaning flatware or silverware by hand or using a dishwasher
Laundry	Performing wardrobe management tasks (e.g. washing, drying, folding or hanging clothes)
Sweeping or mopping	Cleaning the floor using a handled implement (e.g. broom, Swiffer mop)
Leisure Activities (non-screen-based)	
Arts and crafts	Producing fine or functional art using various tools (e.g. pencil, crayon, scissors, glue)
Blocks	Building structures or models using various materials (e.g. Legos, Lincoln logs)
Book reading	Reading a book (hard copy)
Catch	Throwing a ball or disc back and forth in a manner unlike drills or gameplay for a specific sport
Dog walking	Walking continuously while holding a leash being used to guide a dog
Games	Playing non-screen-based games of various forms (e.g. card games, board games)
Gardening	Performing garden-related tasks (e.g. tilling, planting, weeding)
Tag or hide-and-seek	Engaging in one of two specific games for children, namely tag or hide-and-seek
Leisure Activities (screen-based)	
Cell phone	Cell usage encompassing various purposes (e.g. communication, entertainment)
Computer	Desktop computer usage encompassing various purposes (e.g. communication, entertainment)
Laptop	Laptop computer usage encompassing various purposes (e.g. communication, entertainment)
Tablet	Tablet usage encompassing various purposes (e.g. communication, entertainment)
Video gaming	Playing a screen-based game through the use of a handheld controller
Watching television	Viewing television programming on a traditional television (i.e., excluding other devices such as tablets)
Outdoor Chores	
Mowing (push)	Using a push lawnmower to mow the lawn
Mowing (riding)	Using a riding lawnmower to mow the lawn
Raking	Using a rake for outdoor tasks (e.g. gathering leaves)
Playground Activities	
Climbing	Engaging with playground equipment through climbing motions

Table 14. (Continued)

Activity	Operational Definition
Playground Activities (Continued)	
Slide	Engaging with a playground slide in the intended manner (i.e., sliding from top to bottom)
Swing	Engaging with a swing of any type (e.g. including tree swings) in the intended manner
Other	Playground activity that cannot be described as climbing, sliding, or swinging
Soccer	
Drills	Performing soccer-related tasks (e.g. dribbling, shooting) outside of a gameplay scenario
Gameplay	Performing soccer-related tasks (e.g. dribbling, shooting) in a gameplay scenario
Stationary Activities	
Lying	Lying still with no additional activity (e.g. screen time)
Sitting	Sitting still with no additional activity (e.g. screen time)
Tennis	
Drills	Performing tennis-related tasks (e.g. serving, returning) outside of a gameplay scenario
Gameplay	Performing tennis-related tasks (e.g. serving, returning) in a gameplay scenario
Volleyball	
Drills	Performing volleyball-related tasks (e.g. bumping, setting, spiking) outside of a gameplay scenario
Gameplay	Performing volleyball-related tasks (e.g. bumping, setting, spiking) in a gameplay scenario
Wheeled Activities	
Cycling (overground)	Bicycling continuously in an overground manner (e.g. on a bike path)
Cycling (ergometer)	Pedaling continuously on a stationary bicycle
Roller skating	Moving about while wearing wheeled footwear (e.g. inline skates, roller skates)
Scooter	Continuous use of a scooter for transportation or sport (e.g. tricks)
Skateboarding	Continuous use of a skateboard for transportation or sport (e.g. tricks)
Other Activities	
Transition	Any activity that occurs as a linkage between two other activities, rather than a standalone activity
Unknown activity	Any activity that is not adequately described by the preceding options

allowed in the coding scheme, and behaviors were coded continuously (i.e., gaps were not allowed in the observation). Thus, the Observer XT data allowed each moment of the data collection to be linked to a single activity and posture. User input was automatically timestamped using relative timestamps (i.e., elapsed time since the start of the observation), thereby establishing when activity and postural transitions occurred. The intervening periods (i.e., periods of continuous engagement in a single activity and posture) will be referred to hereafter as behavioral intervals, which are notably distinct from activity bouts (i.e., periods of continuous engagement in a single activity, which could include one or more postures).

Cosmed K4b² and K5

The K4b² (233) and K5 (245) are both portable indirect calorimeters that have previously been validated against the Douglas bag method for assessment of oxygen consumption (VO₂) and carbon dioxide production (VCO₂) across a range of intensities. Each participant wore a facemask connected to a chest-mounted (K4b²) or back-mounted (K5) unit that housed the oxygen (O₂) and carbon dioxide (CO₂) analyzers. The facemasks were interchangeable between the K4b² and the K5, and both systems collected data in breath-by-breath format. Prior to each test, the systems were calibrated in 4-5 steps. First, the analyzers were calibrated to room air. Second, the analyzers were calibrated to a reference gas (16.0% O₂, 5.0% CO₂, balance nitrogen). Third, the flowmeter was calibrated to a known volume (3 L, drawn from a Hans Rudolf syringe). And fourth, a delay calibration was conducted for ensuring temporal alignment between data collected by the flowmeter and the analyzers. The fifth procedure applied only to the K5 during indoor testing, and it was a scrubber calibration that zeroed the CO₂ analyzer to the indoor room air. At the start of each data collection (K4b² and K5), the Windows system clock time was manually recorded for future alignment with other data sources.

Initial Data Processing

ActiGraph GT9X

Binary data from the primary accelerometer (90 Hz) and gyroscope (100 Hz) were first read into R using the function `AGread::read\_gt3x`, then averaged each second and merged together using the function `AGread::AG\_merge` (234). The two commands generally followed the procedures of Hibbing et al. (25). Specifically, for primary accelerometer data (in g) two variables were computed for each sample. The first variable was the Euclidian norm minus one (ENMO, in milli-g), which was defined as $\left[\left(\sqrt{X^2 + Y^2 + Z^2} - 1 \right) \cdot 1000 \right]$, with negative values rounded up to zero. The second variable was the vector magnitude (VM, in g), which was defined as $\sqrt{X^2 + Y^2 + Z^2}$. After computing ENMO and VM for each sample, second-by-second averages were calculated for both variables. For the 100-Hz gyroscope data (in $^{\circ}\cdot\text{s}^{-1}$), each axis was low-pass filtered at 35 Hz, after which the gyroscope vector magnitude (GVM, defined as $\sqrt{X^2 + Y^2 + Z^2}$) was calculated and averaged each second. The second-by-second ENMO, VM, and GVM data were then merged.

Noldus Observer XT

Two reviewers coded each video, and their codes were compared using built-in functionality of the Noldus software. Disagreements were resolved by a senior reviewer, at which point the observation was exported as an Excel file. Each Excel file was in “event log” format, meaning separate rows of data were used to represent the beginning and end of each behavioral interval (i.e., each period of continuous engagement in a single activity and posture). Timestamps were given in relative format (i.e., elapsed time since the start of the observations), and the first line of every file indicated what the filmed Windows timestamp was.

The Excel files were then processed in R, wherein a sync procedure was first run to obtain Windows timestamps for each behavioral interval. Specifically, for each file the Windows timestamps were calculated by adding the relative timestamps (outputted from the Observer software) to the Windows time indicated in the first row. After completing the sync procedure, the R code reformatted the data such that each behavioral interval was represented with a single row of data containing activity and posture labels, plus Windows-formatted start and stop times.

Cosmed K4b² and K5

Breath-by-breath data were first screened by excluding any breaths for which one or more of the following four conditions were met: 1) respiratory frequency outside the range 5-80 breaths·min⁻¹; 2) ventilation rate outside the range 0.2-10 L·min⁻¹; 3) fraction of expired O₂ outside the range 10%-20%; or 4) fraction of expired CO₂ outside the range 1%-10%. After excluding invalid breaths, Windows timestamps were calculated. The K4b² and K5 systems each gave timestamps in relative format (i.e., elapsed time since the start of data collection). Thus, Windows timestamps were calculated by adding the relative timestamps to the manually recorded Windows starting time.

Further processing occurred in two broad steps. The first step involved determining RMR for each participant. Those values were then used to calculate youth metabolic equivalents ($MET_y = \frac{activity\ VO_2}{RMR}$) in the second step, which involved creating a second-by-second time course of MET_y values throughout each trial. Below, the procedures are described separately for the RMR data and the continuous time course data. Notably, breath-by-breath VO₂ data were converted to different epoch lengths for RMR versus continuous time course data. All such conversions were performed using procedures given by the manufacturer. Specifically, within each epoch, means

were calculated across breaths for key variables (e.g. gas fractions), and then those means were used to calculate the VO_2 value for the whole epoch.

RESTING METABOLIC RATE DATA. Start and stop times were manually recorded for the RMR assessments. The corresponding breath-by-breath VO_2 data were then extracted and processed. Data were discarded from the first 10 min of each assessment, along with data from the last minute. Using the remaining data, VO_2 was calculated for all possible 5-min epochs (i.e., using a 5-min rolling window), and the lowest value was taken as RMR.

Outlier screening was performed by determining the mean RMR for participants ≤ 12 years old, and those > 12 years old. Within each age group, values were removed if they differed from the mean by more than 2.0 standard deviations. Notably, the outlier screening was performed using available RMR data from 73 participants, rather than only the 42 who also had DO and GT9X data. By using as large a sample as possible, there was reduced potential for identifying spurious outliers due to small sample size in either age group. Thus, it was important to use data from all 73 participants, even though 31 of them were not eligible for inclusion in any other aspects of this investigation.

CONTINUOUS TIME COURSE. For each participant, VO_2 data were first converted to 15-s epochs. The averages were then converted to MET_y via division by RMR. Lastly, each second of the data collection was labeled with the MET_y value from the corresponding 15-s epoch. Thus, second-by-second MET_y values were available after completing the prior steps.

Assembling the Final Data Set

The data set was assembled by starting with the second-by-second GT9X data. First, the GT9X data were merged with the K4b² and K5 data by labeling each second with the MET_y value from the corresponding 15-s epoch. Next, DO data were added by labeling each second

with the activity and posture values from the corresponding behavioral interval. After completing the prior steps, the data set included second-by-second data from the GT9X (ENMO, VM, GVM), K4b² (METy), and DO (activity and posture).

For each second of data, posture and METy were examined to determine activity intensity based on accepted definitions that included both posture and EE (73, 235). Specifically, each second was classified as sedentary behavior (SB; ≤ 1.5 METy and lying, seated, or reclining posture), light physical activity (LPA; 1.51-2.99 METy, or ≤ 1.5 METy with upright posture), or moderate-to-vigorous physical activity (MVPA; ≥ 3.0 METy, regardless of posture). Activity labels were then examined to identify transition points, which were labeled using a binary variable (i.e., taking a value of 1.0 whenever a transition occurred and 0.0 otherwise). Afterwards, participant characteristics (e.g. sex and age) were added to complete the data set. Table 15 shows how the data set was laid out. After completing the prior steps, there were separate data files for each participant and visit. Pooling of data from both visits will be discussed in the following chapter, since the pooling was specifically related to data analysis.

Data Loss

Figure 8 summarizes participant level data loss. Seven participants were excluded for not having data from two days, and four additional participants were excluded due to various metabolic data issues that affected both days. Attachment site-specific exclusion was also performed since two participants had partially missing GT9X data for one of their visits. Specifically, one participant was missing hip data for one visit (due to researcher error) and another participant was missing dominant wrist data for one visit (due to download error). The affected participants were excluded from analyses that involved the affected attachment site, but they were not excluded from analyses involving the other attachment sites. Thus, the analytical

Table 15. Layout of the free-living data (second-by-second format).

PID	Attachment	Demographics			Sensor Variables			Criterion Labels [#]		
		Sex	Age	... [#]	ENMO (milli-g)	VM (g)	GVM (°·s ⁻¹)	Activity	Intensity	Transition [^]
000	Hip	M	12.8	...	108.9	0.975	71.171	Basketball (drills)	MVPA	0
000	Hip	M	12.8	...	44.8	1.016	28.532	Basketball (drills)	MVPA	0
000	Hip	M	12.8	...	32.0	1.028	18.207	Transition	MVPA	1
000	Hip	M	12.8	...	22.1	1.022	22.550	Transition	MVPA	0
000	Hip	M	12.8	...	18.9	1.018	14.000	Transition	MVPA	0
...	...	...	...	...	...	...	...	...	...	...

PID- participant code; ENMO- Euclidian norm minus one; VM- vector magnitude; GVM- gyroscope vector magnitude; M- male; MVPA- moderate-to-vigorous physical activity

[#]To conserve space, some variables are not shown. Most notably, criterion values for posture and metabolic equivalents are not shown, having already been combined to determine intensity (i.e., SB, LPA, or MVPA).

[^]Indicator of whether a transition occurred (1) or not (0) for each second.

Note: VM does not always equal [(ENMO/1000)+1], which is a side effect of converting the raw data (90 Hz) to 1-s averages.

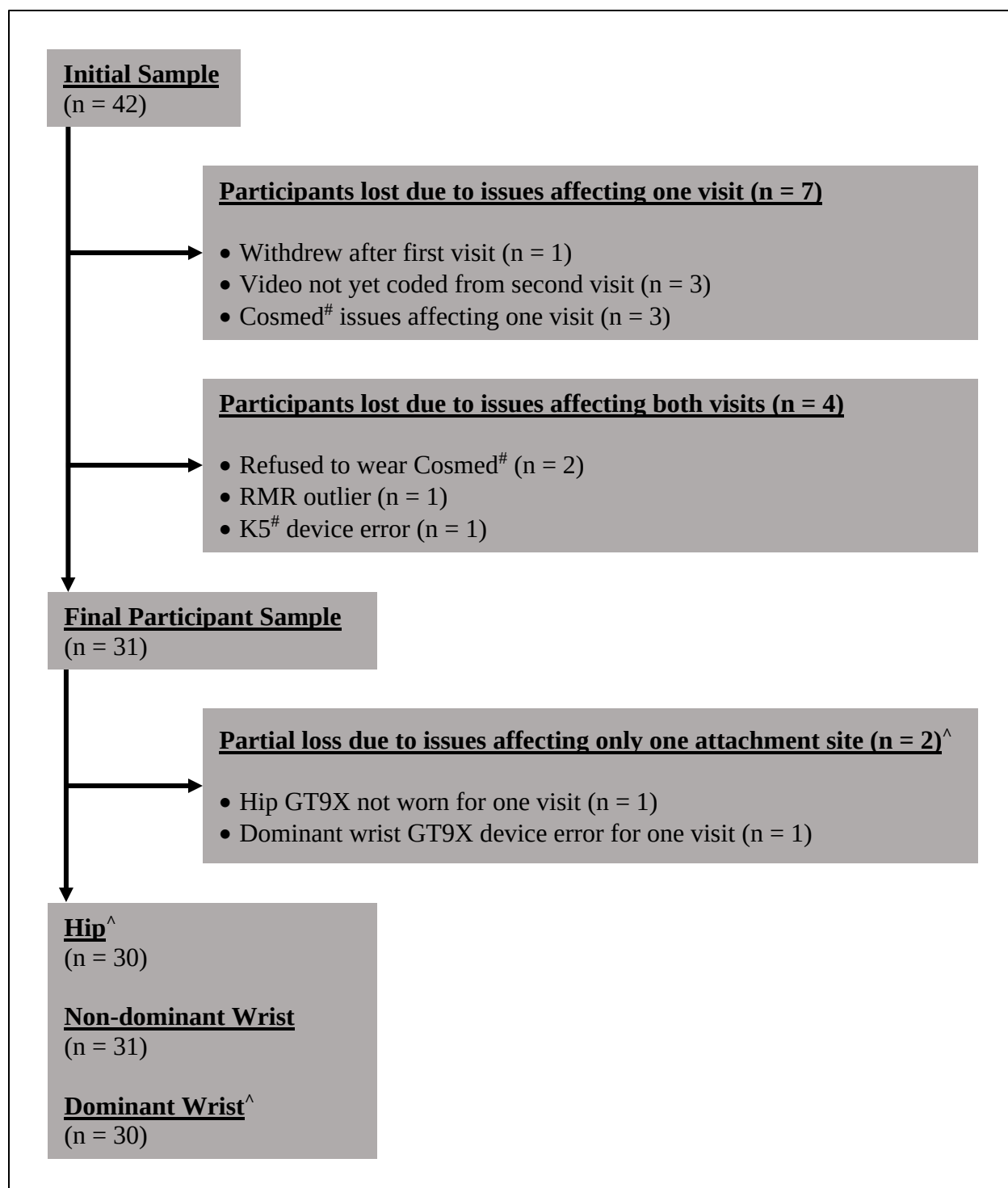


Figure 8. Summary of participant level data loss and retention for the free-living data set.

RMR- resting metabolic rate.

[#]The term “Cosmed” is used generically to refer to the K4b² or the K5, whereas the specific system is indicated where appropriate.

[^]Different participants were lost for the hip and dominant wrist.

samples differed for each attachment site, with 30 total participants for the hip (four of whom also participated in the previously-discussed laboratory study), 31 for the non-dominant wrist (five of whom also participated in the laboratory study), and 30 for the dominant wrist (five of whom also participated in the laboratory study). Within trials, there was sometimes partial loss of criterion data (e.g. when the participant was off camera or took the Cosmed mask off for a water break), but such loss was rare, accounting for mean $\pm$ SD 21.3 ± 24.7 min (or $9.5\% \pm 11.3\%$ of total measurement time) per participant. The affected data were not removed until analysis (see the following chapter), to allow sensor segmentation algorithms to run on continuously-collected GT9X data.

Summary of Final Data Set

Participants had mean $\pm$ SD 3.8 ± 0.4 h of data that were usable for at least one attachment site. The total volume of data amounted to more than four days (103.7 h for the hip, 107.0 h for the non-dominant wrist, and 103.0 h for the dominant wrist). Table 16 shows participant characteristics for the final sample. The data presented in this chapter will serve as a starting point for analyses in the next chapter, wherein the third research question will be addressed.

Table 16. Participant demographics for the free-living study. Values are n (%), except for age (mean $\pm$ SD).

	Female (n = 11)	Male (n = 20)	Total (N = 31)
Age (yr)	11.2 $\pm$ 3.5	12.6 $\pm$ 3.0	12.1 $\pm$ 3.2
Weight Status[#]			
Underweight (< 5 th percentile)	0 (0%)	2 (10%)	2 (6%)
Normal Weight (5.0-84.9 th percentile)	10 (91%)	13 (65%)	23 (74%)
Overweight (85.0-94.9 th percentile)	0 (0%)	2 (10%)	2 (6%)
Obese ($\geq$ 95 th percentile)	1 (9%)	3 (15%)	4 (13%)
Handedness			
Left	0 (0%)	1 (5%)	1 (3%)
Right	11 (100%)	19 (95%)	30 (97%)
Ethnicity			
Hispanic	0 (0%)	1 (5%)	1 (3%)
Non-Hispanic	11 (100%)	19 (95%)	30 (97%)
Race			
Asian	0 (0%)	2 (10%)	2 (6%)
Hawaiian/Pacific Islander	0 (0%)	1 (5%)	1 (3%)
White	11 (100%)	17 (85%)	28 (90%)

[#]Percentiles determined from the Centers for Disease Control and Prevention body-mass-index-for-age growth charts.

CHAPTER VII

RESEARCH QUESTION 3: METHODS AND RESULTS

Overview

In Chapter 5, versions of two gyroscope inclusive youth Sojourn models (i.e., Soj_{recall} and Soj_{precision}) were developed for the hip, non-dominant wrist, and dominant wrist. Laboratory data were used for the development, and also for making performance comparisons between Soj_{recall}, Soj_{precision}, and the original raw acceleration-based youth Sojourn models (34). Notably, there was no original model for dominant wrist data, and thus the original non-dominant wrist model was used for both wrists.

In this chapter, the free-living data set from Chapter 6 (pages 120-135) was used to perform an independent validation of Soj_{recall} and Soj_{precision}. The original raw acceleration-focused youth Sojourn models (34) were again included for comparison, using the non-dominant wrist model to obtain estimates for both wrists. Below, four analyses are presented. The analyses were designed to address the third research question, which was formulated as follows:

3. **Question:** In free-living settings, does the performance of youth Sojourn models improve when using sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification?

Hypothesis: When applied to free-living data, youth Sojourn models that use sensor fusion and change point detection for segmentation, and gyroscope-based neural networks for intensity classification will have improved performance, compared to existing youth Sojourn models.

Methods

Analysis of Model Runtime

Each data file contained pre-processed data that were obtained from a single participant, visit, and attachment site. The analysis was focused exclusively on algorithm runtime, meaning it

was not necessary to distinguish between participants, visits, or attachment sites. Thus, all files were treated as independent cases, and analysis was performed on a file-by-file basis. Each data file was read into R, at which point three youth Sojourn models were applied (i.e., the site-appropriate versions of $\text{Soj}_{\text{recall}}$, $\text{Soj}_{\text{precision}}$, and the original youth Sojourn model). The models were timed as they ran, and after processing all files, the model-wise runtimes were compared using repeated measures analysis of variance (ANOVA). Individual models were compared using pairwise paired t-tests were, and the false discovery rate correction (244) was used to account for multiple comparisons.

Analysis of Segmentation

The transition pairing method (TPM) of Hibbing et al. (35) was used to make site-specific comparisons of segmentation correctness for $\text{Soj}_{\text{recall}}$, $\text{Soj}_{\text{precision}}$, and the original youth Sojourn model. Comparisons were made for various settings of the spurious pairing threshold, namely 1-20 s in 1-s increments. Predicted and criterion transitions were first obtained for each participant and visit. The values were then pooled across visits. For each participant, performance metrics were then calculated, namely recall, precision, relative root mean squared error ($\text{RMSE}_{\%}$), and aggregated performance. Recall and precision are commonly used in many classification settings, but $\text{RMSE}_{\%}$ and aggregated performance are unique to the TPM. Ordinarily, root mean squared error (RMSE) is used on an absolute scale to represent timing error by a dynamic segmentation algorithm, and lower values reflect higher performance. Within the TPM, the maximum lag time is known (i.e., equal to the spurious pairing threshold value), as is the minimum lag time (i.e., 0 s). Therefore, it is possible to express RMSE on a relative scale (i.e., $\text{RMSE}_{\%}$) such that performance is reflected the same way as precision and recall, namely as a percentage between 0% and 100% where higher values reflect better performance. To calculate

RMSE%, the following formula can be used: $RMSE\% = 100\% - \left(\frac{RMSE}{SPT} * 100\% \right)$, where SPT is the spurious pairing threshold value. Aggregated performance is defined as the mean of recall, precision, and RMSE%. More information is available in the vignette for the R implementation of the TPM (see <https://cran.r-project.org/web/packages/PAutilities/vignettes/tpm.html>).

After calculating metrics for each participant and attachment site, the metrics were then compared across participants using the overlapping confidence interval method. Specifically, in each analysis 95% confidence intervals were calculated for each of the three youth Sojourn models, and non-overlapping confidence intervals were defined as being significantly different with $\alpha = 0.05$. For each analysis, missing data (e.g. due to a model predicting no true positives) were handled via listwise deletion to ensure the same cases were used for comparing each model.

Analysis of Activity Intensity Classification

The youth Sojourn models produced second-by-second predictions of activity intensity, namely sedentary behavior (SB), light physical activity (LPA), or moderate-to-vigorous physical activity (MVPA). At the participant level (pooling data across visits), 11 total metrics were used to assess classification agreement compared to the criterion measure. Two of the metrics reflected overall agreement (i.e., percent correct classification and kappa statistics), while the other nine reflected agreement for individual intensity categories. Specifically, for each of the three intensities (SB, LPA, and MVPA), agreement was assessed with sensitivity, specificity, and F1 (i.e., the harmonic mean of recall and precision).

After obtaining the metrics for each participant and model, the values were averaged across participants and compared statistically. Specifically, 33 tests were run (11 metrics times three attachment sites) using repeated measures analysis of variance (ANOVA). Pairwise post hoc tests were then used to compare Soj_{recall} , $Soj_{precision}$, and the original youth Sojourn model. The

false discovery rate correction (244) was applied to account for multiple comparisons. For each analysis, data were screened for missing values that occurred due to divide-by-zero errors (e.g. if F1 was undefined due to recall or precision of 0.0%). In such cases, listwise deletion was performed to ensure the same cases were used for comparing each model.

Analysis of Activity Time

For each participant (separately for each attachment site), data were pooled across visits to obtain total time spent in SB, LPA, and MVPA for the criterion method and the various youth Sojourn models. The values were then compared descriptively and statistically. For the descriptive analysis, methods were compared in terms of mean and standard deviation. Statistical analyses were carried out after converting SB, LPA, and MVPA estimates to percent of total time spent in each category. The latter transformation was applied to normalize the scale across participants prior to analysis. The predictions from the youth Sojourn models were then compared to the criterion values using a total of nine repeated measures ANOVAs (one for each intensity category and attachment site). Planned contrasts were used to perform post hoc tests ($\alpha = 0.05$). Specifically, paired t-tests were used to compare each model to the criterion measure, with multiple comparisons being addressed using the false discovery rate correction (244).

Results

Results for the four analyses are presented in separate subsections. Summary statistics are presented as mean $\pm$ standard deviation. Notably, the samples differed slightly for each attachment site (see Figure 8), and thus direct comparisons have been avoided across sites.

Analysis of Model Runtime

Across all attachment sites, there were 182 files processed, corresponding to 60 visits each for the hip and dominant wrist, and 62 visits for the non-dominant wrist. The individual files had

durations of 114.3 ± 18.3 min. Algorithm runtimes were 1.2 ± 0.3 s (original youth Sojourn), 16.3 ± 3.0 s (Soj_{recall}), and 5.2 ± 1.1 s (Soj_{precision}). There was a significant main effect of model choice on runtime ($F(2) = 5053.2$, $p < 0.001$), and all pairwise comparisons were significant with $p < 0.001$.

Analysis of Segmentation

MISSING PARTICIPANT DATA. Table 17 summarizes participant level data loss by attachment site and performance metric. There were no missing data for recall and precision. For RMSE_% and aggregated performance, missing data occurred whenever a model yielded no true positive predictions. The latter issue occurred only when the spurious pairing threshold was ≤ 4 s, resulting in loss of data from 1-5 participants (depending on the attachment site and the specific threshold setting).

REJECTION OF INDIVIDUAL TRANSITION PAIRINGS. For the original youth Sojourn algorithm, the TPM rejected a total of 14 non-sequential pairings (never more than one per participant) across all combinations of attachment site and spurious pairing threshold. For Soj_{recall}, there were 188 rejected pairings, 53.7% of which came from the hip (1-3 rejected pairings per participant, for up to 11 participants depending on the spurious pairing threshold). For the other 46.3% of cases (i.e., on the wrists) there were 1-2 rejected pairings per participant, for up to eight participants (depending on the specific wrist and spurious pairing threshold setting). For Soj_{precision}, there were 59 rejected pairings, 91.5% of which occurred for the hip (maximum of one rejected pairing per participant, for up to eight participants depending on the spurious pairing threshold). On the non-dominant wrist, there were no rejected pairings for Soj_{precision}, whereas on the dominant wrist there was one rejected pairing for one participant (spurious pairing threshold ≥ 16 s).

Table 17. Summary of data loss (participant level) for each analysis of segmentation using the free-living data set.

	Spurious Pairing Threshold (s)	Participants Removed (n)	Final Sample (n)
Hip (starting N = 30)			
Recall and precision	1-20	0	30
RMSE% and aggregated performance	1	3	27
	2-20	0	30
Non-Dominant Wrist (starting N = 31)			
Recall and precision	1-20	0	31
RMSE% and aggregated performance	1	5	26
	2	3	28
	3	1	30
	4-20	0	31
Dominant Wrist (starting N = 30)			
Recall and precision	1-20	0	30
RMSE% and aggregated performance	1	5	25
	2	3	27
	3	1	29
	4-20	0	30

RMSE%- relative root mean squared error.

Note: For a given metric and attachment site, participants were excluded whenever one or more models gave a missing value. All missing values occurred due to a model predicting no true positives.

OUTCOMES. Figure 9 shows results of the segmentation analysis. In terms of aggregated performance (Figure 9D, Figure 9H, and Figure 9L), the original youth Sojourn models performed the best for all attachment sites and spurious pairing threshold settings. Soj_{recall} had the next highest performance for all but spurious pairing threshold settings of 2 s (non-dominant wrist) and 1 s (dominant wrist), meaning Soj_{precision} generally had the lowest aggregated performance. For the original youth Sojourn algorithm, mean aggregated performance exceeded 50% at spurious pairing threshold values of 16 s (all attachment sites) and never exceeded 53.6% (dominant wrist, spurious pairing threshold of 20 s). Soj_{recall} and Soj_{precision} never differed from one another by more than 7.6%, and neither model achieved mean aggregated performance > 50%, with the maximum being 46.3% (Soj_{recall} on the hip, spurious pairing threshold of 20 s).

For recall (Figure 9A, Figure 9E, and Figure 9I), the original Sojourn algorithm had the highest mean for all wear locations (regardless of the spurious pairing threshold), followed by Soj_{recall} and Soj_{precision}. For the original youth Sojourn algorithm, mean recall exceeded 50% for threshold values of 7-8 s (depending on attachment site), and exceeded 75% for threshold values of 13-14 s (wrists only; maximum hip recall 72.4%, for spurious pairing threshold of 20 s). For Soj_{recall} and Soj_{precision}, mean recall differed by 4.0%-26.1%. For Soj_{recall}, mean recall exceeded 50% when the spurious pairing threshold was 17-19 s (hip and non-dominant wrist only; maximum of 48.6% for the dominant wrist, spurious pairing threshold of 20 s), and mean recall > 75% was never achieved (maximum of 53.3% for the hip, spurious pairing threshold of 20 s). For Soj_{precision}, mean recall did not exceed 50% for any setting of the spurious pairing threshold (maximum of 28.1% for the hip, spurious pairing threshold of 20 s).

For precision (Figure 9B, Figure 9F, and Figure 9J), the original youth Sojourn algorithm and Soj_{recall} performed similarly, never differing by more than 2.1% across attachment sites and

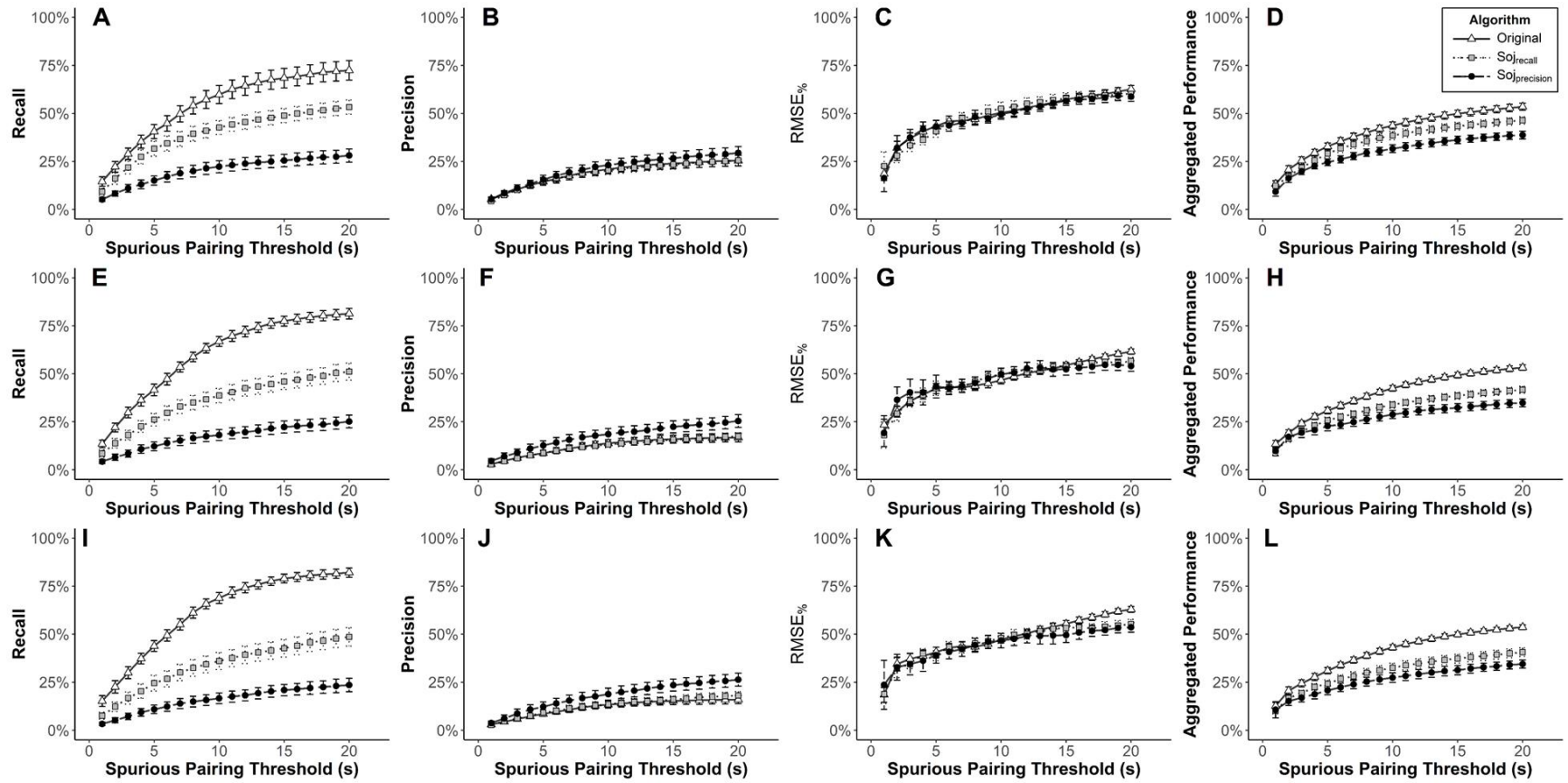


Figure 9. Performance of segmentation algorithms when applied to the free-living data set. Results are shown for recall, precision, RMSE%, and aggregated performance on the hip (A-D, respectively), non-dominant wrist (E-H, respectively), and dominant wrist (I-L, respectively). Values are means, and error bars are 95% confidence intervals. The spurious pairing threshold (x-axis) reflects the longest-allowable lag time for true positives.

spurious pairing threshold settings, and never exceeding 25.6% (Soj_{recall} for the hip, spurious pairing threshold of 20 s). Soj_{precision} had the highest mean precision for all attachment sites and spurious pairing threshold settings. However, the margins were small, ranging from 0.7%-4.0% (hip), 1.8%-8.6% (non-dominant wrist), and 1.0%-10.4% (dominant wrist). Mean precision for Soj_{precision} did not exceed 50% for any attachment site or threshold setting (maximum of 29.4% for the hip, spurious pairing threshold of 20 s).

For RMSE%, (Figure 9C, Figure 9G, and Figure 9K) the original youth Sojourn models had the highest mean when the spurious pairing threshold was ≥ 17 s (hip), ≥ 15 s (non-dominant wrist), and ≥ 12 s (dominant wrist). For lower threshold settings, no single model consistently had the highest or lowest mean RMSE%. No differences exceeded 9.3% (regardless of attachment site or spurious pairing threshold), and for spurious pairing threshold settings ≤ 17 s, differences were all $< 7.2\%$.

Analysis of Activity Intensity Classification

Table 18 summarizes participant level data loss by attachment site and performance metric. Apart from one missing SB F1 score (due to precision of 0.0% for the original youth Sojourn model), there was no data loss for analyses of overall agreement, SB, or LPA. For MVPA, up to eight participants (depending on attachment site) lacked true positive data for one or more models, resulting in undefined F1 scores. In three of those cases, F1 and sensitivity were both missing because the participants did not perform any MVPA at either visit.

A total of 33 ANOVAs were performed (11 metrics times three attachment sites), and main effects were significant for 29 of the 33 ($p < 0.03$). Two of the exceptions occurred for the non-dominant wrist (LPA specificity with $p = 0.08$, and MVPA F1 with $p = 0.29$), and the other two occurred for the dominant wrist (percent correct classification with $p = 0.05$, and MVPA F1 with

Table 18. Summary of data loss (participant level) for each analysis of classification agreement using the free-living data set.

	Overall Agreement		SB			LPA			MVPA		
	CC	Kappa	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
Hip (starting N = 30)											
Participants removed (n)	0	0	0	0	1 [#]	0	0	0	3 [^]	0	8 ^{^*}
Final sample (n)	30	30	30	30	29	30	30	30	27	0	22
Non-Dominant Wrist (starting N = 31)											
Participants removed (n)	0	0	0	0	0	0	0	0	3 [^]	0	8 ^{^*}
Final sample (n)	31	31	31	31	31	31	31	31	28	0	23
Dominant Wrist (starting N = 30)											
Participants removed (n)	0	0	0	0	0	0	0	0	3 [^]	0	6 ^{^*}
Final sample (n)	30	30	30	30	30	30	30	30	27	0	24

SB- sedentary behavior; LPA- light physical activity; MVPA- moderate-to-vigorous physical activity; CC- correct classification;

Se- sensitivity; sp- specificity.

[#]F1 undefined because original youth Sojourn model predicted no SB (n = 1)

[^]Sensitivity and F1 undefined because participants did not perform any MVPA (n = 3)

^{*}F1 undefined because one or more models predicted no true positive MVPA (n = 3-5)

$p = 0.96$). Figure 10 shows the results of the entire classification analysis. Below, results are presented sequentially for overall agreement, SB, LPA, and MVPA.

OVERALL AGREEMENT. For the hip, mean percent correct classification was 39.6%-42.9% higher when using Soj_{recall} ($68.0\% \pm 14.6\%$) or Soj_{precision} ($71.3\% \pm 14.4\%$) than when using the original model ($28.4\% \pm 16.1\%$; both $p < 0.001$). The differences were more modest for the non-dominant wrist (13.1%-15.0%; both $p < 0.001$), and the differences were non-significant for the dominant wrist (4.2%-4.8%; both $p = 0.16$). Kappa statistics followed similar trends to percent correct classification, with means ranging from 0.26-0.47 (for Soj_{recall} and Soj_{precision}) versus 0.08-0.23 (for the original models; $p = 0.00$ -0.25). When comparing Soj_{recall} and Soj_{precision} to one another across attachment sites, both models had similar means for percent correct classification (differences $\leq 3.3\%$) and kappa (differences ≤ 0.05), with mixed significance ($p = 0.00$ -0.61).

SEDENTARY BEHAVIOR. For the hip, mean sensitivity was more than three times as high for Soj_{recall} ($91.6\% \pm 8.9\%$) and Soj_{precision} ($89.0 \pm 11.0\%$) as for the original model ($24.2\% \pm 22.2\%$; both $p < 0.001$), whereas specificity was lower (means of 63.8%-70.1% for Soj_{recall} and Soj_{precision}, versus 92.1% for the original model; both $p < 0.001$). Mean F1 scores were 45.0%-47.0% higher for the new models than the original (both $p < 0.001$). For the wrists, mean sensitivity was 83.3%-95.1% for Soj_{recall} and Soj_{precision}, whereas it was significantly lower (63.6%-63.8%; all $p < 0.01$) for the original model. The trends were reversed for specificity, with means of 50.2%-65.7% for Soj_{recall} and Soj_{precision}, versus significantly higher means (76.2%-78.3%; all $p < 0.01$) for the original model. F1 means for the wrists were higher for the new models than the original by 9.3%-11.0% (all $p < 0.01$). When comparing Soj_{recall} and Soj_{precision} to one another across attachment sites, both models had similar mean values for sensitivity and F1

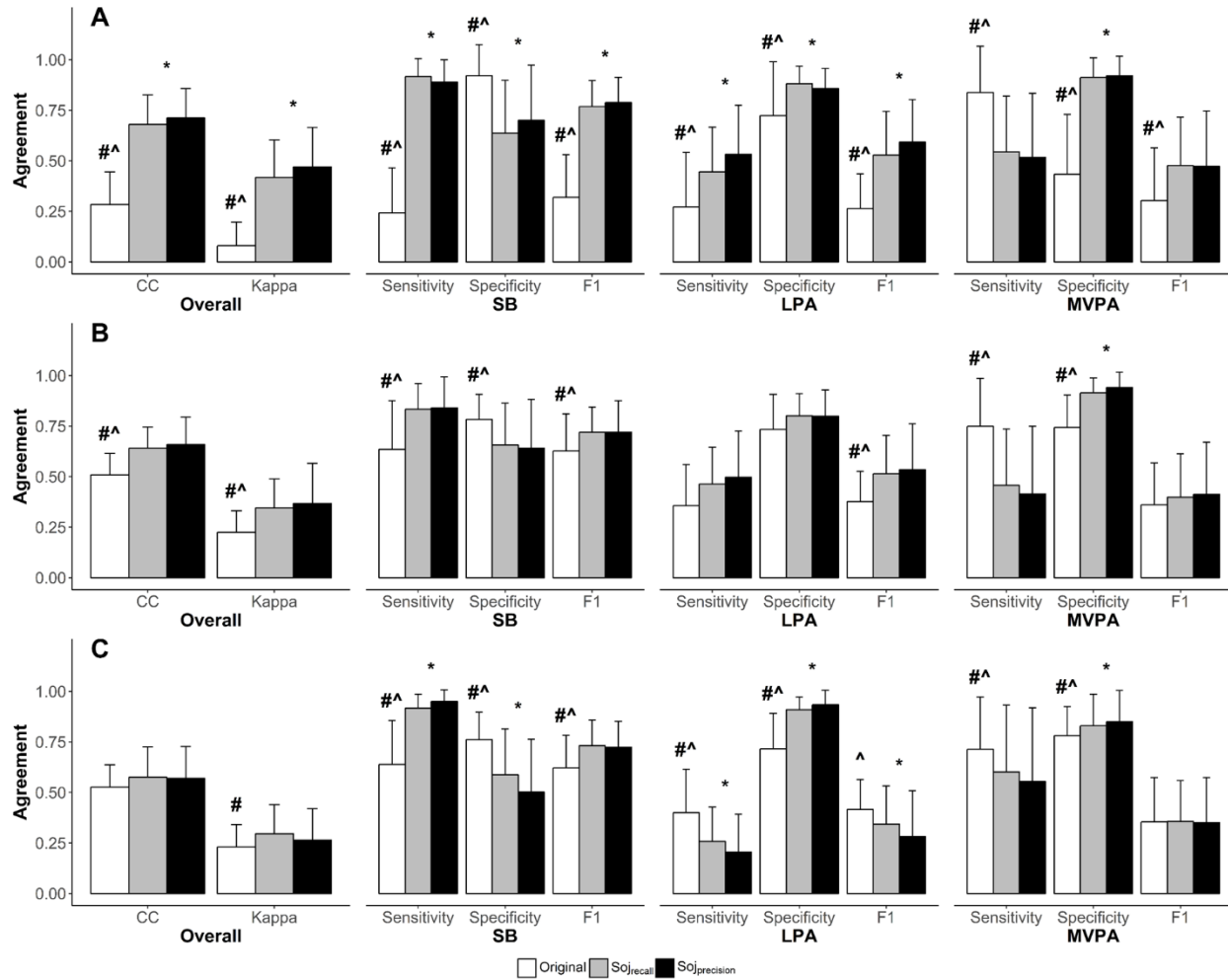


Figure 10. Classification agreement of the original youth Sojourn models and the newly developed, gyroscope inclusive Soj_{recall} and Soj_{precision} models when applied to the free-living data set. Assessments include overall agreement and intensity-specific agreement for sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA). Attachment sites are (A) hip, (B) non-dominant wrist, and (C) dominant wrist. Error bars are standard deviation. For the dominant wrist, the original youth Sojourn results were obtained using the model calibrated for the non-dominant wrist.

CC- correct classification

#Significant difference ($p < 0.05$) between original youth Sojourn and Soj_{recall}

^Significant difference ($p < 0.05$) between original youth Sojourn and Soj_{precision}

*Significant difference ($p < 0.05$) between Soj_{recall} and Soj_{precision}

(differences $\leq 3.4\%$; $p = 0.00-0.93$), whereas differences for specificity ranged from 1.5%-8.6% ($p = 0.00-0.20$).

LIGHT PHYSICAL ACTIVITY. For the hip, mean sensitivity was 17.5%-26.1% higher for Soj_{recall} (44.6% $\pm$ 22.1%) and Soj_{precision} (53.2% $\pm$ 24.2%) than the original model (27.1% $\pm$ 27.0%; both $p < 0.01$). Similarly, mean specificity was 13.3%-15.7% higher for Soj_{recall} (88.1% $\pm$ 8.8%) and Soj_{precision} (85.7% $\pm$ 10.0%) than the original model (72.4% $\pm$ 26.6%; both $p = 0.02$). F1 means were approximately twice as high for Soj_{recall} and Soj_{precision} (means of 52.8%-59.3%) than the original model (mean of 26.3%; both $p < 0.001$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the hip, mean sensitivity and F1 differed by 6.4%-8.7% (both $p < 0.001$), while mean specificity differed by 2.4% ($p = 0.02$).

On the non-dominant wrist, the patterns were largely the same as for the hip, with mean sensitivity of 46.4%-49.7% (for Soj_{recall} and Soj_{precision}) versus 35.6% (original model; both $p = 0.09$), and mean specificity of 73.4%-80.1% for all three models. F1 scores were 13.8%-15.8% higher for the new models than the original (both $p = 0.01$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the non-dominant wrist, mean sensitivity, specificity, and F1 all differed by $\leq 3.3\%$ ($p = 0.09-0.86$).

On the dominant wrist, sensitivity was lower for Soj_{recall} and Soj_{precision} (means of 20.5%-25.8%) than the original model (mean of 39.9%; both $p \leq 0.01$), whereas the opposite was true for specificity (means of 91.0%-93.4% for Soj_{recall} and Soj_{precision}, versus 71.5% for the original model; both $p < 0.001$). F1 means were 7.3%-13.3% lower for Soj_{recall} and Soj_{precision} than the original model ($p = 0.03-0.10$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the dominant wrist, mean sensitivity and F1 differed by 5.3%-6.1% ($p = 0.02-0.03$), while mean specificity differed by 2.5% ($p = 0.01$).

MODERATE-TO-VIGOROUS PHYSICAL ACTIVITY. For the hip, sensitivity was significantly lower for Soj_{recall} (54.4% $\pm$ 27.6%) and Soj_{precision} (51.7% $\pm$ 31.6%) than the original model (83.8% $\pm$ 23.0%; both $p < 0.001$), whereas the opposite was true for specificity (means of 91.2%-92.1% for Soj_{recall} and Soj_{precision}, versus 43.3% for the original model; both $p < 0.001$). Mean F1 was 17.0%-17.3% higher for the new models than the original (both $p < 0.001$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the hip, mean sensitivity, specificity, and F1 all differed by $\leq 2.7\%$ ($p = 0.00-0.89$).

On the non-dominant wrist, Soj_{recall} and Soj_{precision} had significantly lower mean sensitivity (41.5%-45.6%) than the original model (74.9%; both $p < 0.001$). In contrast, mean specificities were 91.5%-94.1% for Soj_{recall} and Soj_{precision}, versus 74.3% for the original models (both $p < 0.001$). Mean F1 was 3.7%-5.1% higher for the new models than the original (both $p = 0.38$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the non-dominant wrist, mean sensitivity, specificity, and F1 all differed by $\leq 4.2\%$ ($p = 0.00-0.58$).

On the dominant wrist, mean sensitivity was significantly lower for Soj_{recall} and Soj_{precision} (55.4%-60.2%) than the original model (71.3%; both $p < 0.01$). The three models all had similar mean values for specificity (78.1%-84.9%) and F1 (35.1%-35.6%), and significance was mixed ($p = 0.03-0.91$). In terms of how Soj_{recall} and Soj_{precision} compared to one another for the dominant wrist, mean sensitivity, specificity, and F1 all differed by $\leq 4.8\%$ ($p = 0.04-0.91$).

Analysis of Activity Time

Figure 11 summarizes the results of the activity time analyses. A total of nine ANOVAs were performed (three intensities times three attachment sites), and all main effects were significant with $p < 0.001$. Below, results are presented sequentially for time spent in SB, LPA, and MVPA. In each section, the descriptive emphasis is on raw minutes, whereas p values are reported from

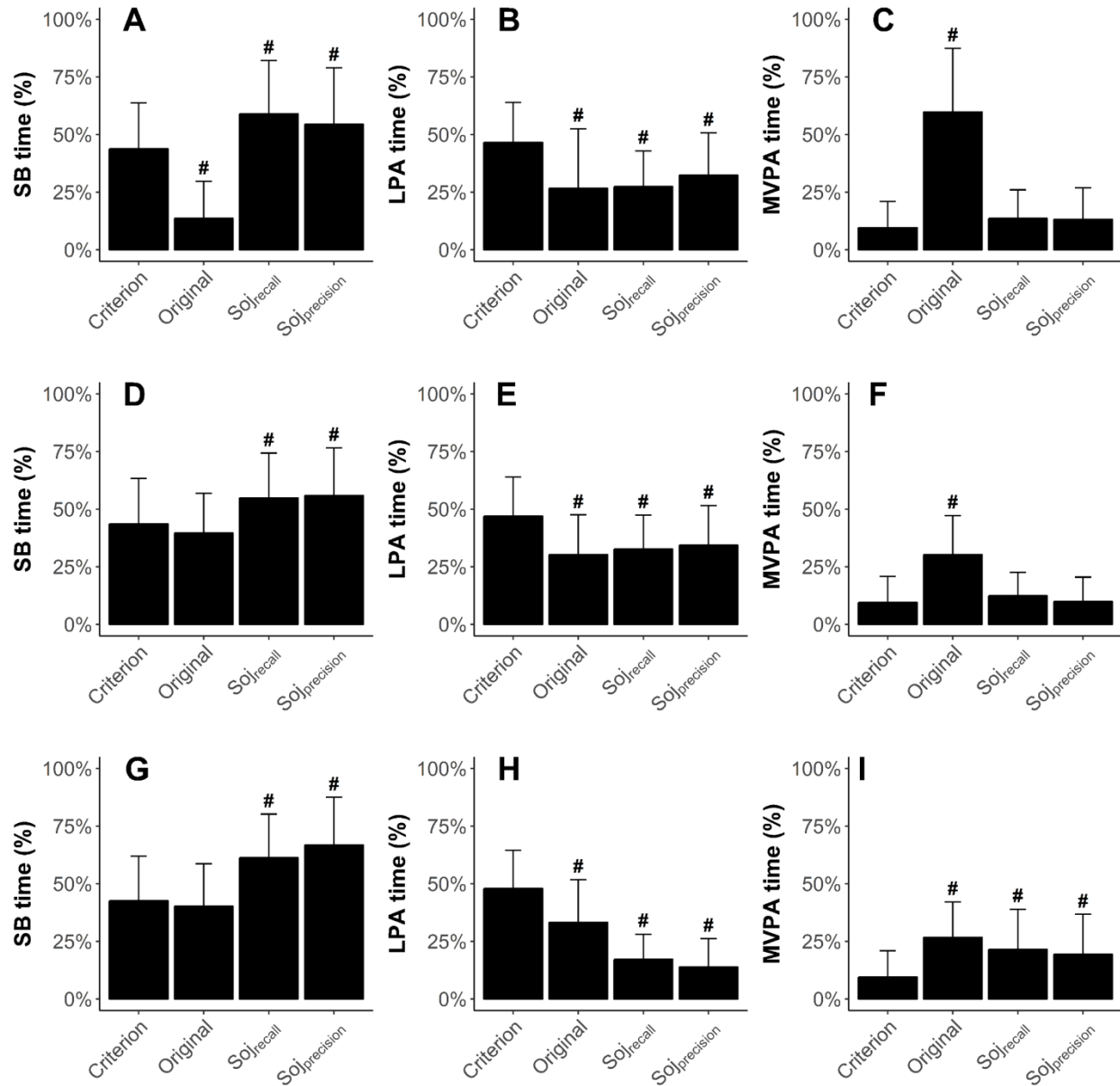


Figure 11. Measured and predicted percent of free-living time spent in sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA) for the hip (A-C, respectively), non-dominant wrist (D-F, respectively), and dominant wrist (G-I, respectively).

#Significant difference from the criterion measure ($p < 0.05$)

the percentage-based analyses (i.e., based on percent of total time spent in each category; see Figure 11). Summary values are given as mean $\pm$ SD.

SEDENTARY BEHAVIOR. At each attachment site, Soj_{recall} and $Soj_{precision}$ had higher mean SB than the criterion, whereas the original youth Sojourn model had lower mean SB than the criterion. On the hip (91.1 ± 41.7 min of criterion SB), the means for Soj_{recall} and $Soj_{precision}$ differed from the criterion by 23.2-32.9 min (both $p < 0.01$), while the mean for the original model differed by more (63.6 min; $p < 0.001$). On the non-dominant wrist (90.5 ± 41.1 min of criterion SB), the means for Soj_{recall} and $Soj_{precision}$ differed from the criterion by 24.6-26.1 min (both $p < 0.001$), while the mean for the original model differed by less (8.6 min; $p = 0.21$). Similarly on the dominant wrist (87.6 ± 38.7 min of criterion SB), the means for Soj_{recall} and $Soj_{precision}$ differed from the criterion by 39.4-51.3 min (both $p < 0.001$), while the mean for the original model differed by less (6.0 min; $p = 0.42$).

LIGHT PHYSICAL ACTIVITY. At each attachment site, all three models had lower mean LPA than the criterion. On the hip (97.6 ± 38.0 min of criterion LPA), all three models had similar mean differences from the criterion measure (30.3-41.7 min; all $p < 0.01$), and the same was true for the non-dominant wrist (98.1 ± 37.4 min of criterion LPA; mean differences of 27.1-34.7 min; all $p < 0.01$). On the dominant wrist (99.9 ± 36.7 min of criterion LPA), the means for Soj_{recall} and $Soj_{precision}$ differed from the criterion by 64.3-71.5 min (both $p < 0.001$), whereas the mean for the original model differed by considerably less (30.6 min; $p < 0.01$).

MODERATE-TO-VIGOROUS PHYSICAL ACTIVITY. At each attachment site, all three models had higher mean MVPA than the criterion. On the hip (18.7 ± 17.1 min of criterion MVPA), the means for Soj_{recall} and $Soj_{precision}$ differed from the criterion by 7.2-9.0 min (both $p = 0.05$), whereas the mean for the original model differed by more (103.8 min; $p < 0.001$). The patterns

were similar for the non-dominant wrist (18.5 ± 16.8 min of criterion MVPA), with means for Soj_{recall} and Soj_{precision} differing from the criterion by a smaller amount (1.0-6.4 min; $p = 0.10=0.82$) than the original model (43.3 min; $p < 0.001$). On the dominant wrist (18.4 ± 17.1 min of criterion MVPA), mean differences from the criterion were large for all three methods, though still smaller for Soj_{recall} and Soj_{precision} (20.3-25.0 min; both $p < 0.01$) than the original model (36.6 min; $p < 0.001$).

Summary and Transition

In this chapter, performance was assessed for Soj_{recall} and Soj_{precision} when applied to data collected from an independent sample assessed in free-living settings. In terms of runtime, Soj_{recall} and Soj_{precision} each ran more slowly than the original youth Sojourn models (34). In terms of segmentation, the change point detection-based approaches of Soj_{recall} and Soj_{precision} did not perform as well as the original models, particularly because the change point detection methods failed to retain the precision advantage that was shown in Chapter 5 (see Figure 6). In terms of second-by-second activity intensity classification, Soj_{recall} and Soj_{precision} generally performed better than the original models, while also performing similarly to one another. In terms of total time spent in SB, LPA, and MVPA, Soj_{recall} and Soj_{precision} tended to overpredict SB and MVPA time (which was compensated by underprediction of LPA time), whereas the original youth Sojourn models tended to underpredict SB and LPA time (which was compensated by overprediction of MVPA time). The following chapter will provide an integrated discussion in which the implications are discussed from Chapters 1-7.

CHAPTER VIII
DISCUSSION AND CONCLUSION

In this dissertation, two gyroscope inclusive youth Sojourn models were developed for objectively assessing activity intensity in youth. One model (Soj_{recall}) was designed to favor recall over precision by erring on the side of over-predicting transitions, and the other algorithm (Soj_{precision}) was designed to favor precision over recall by erring on the side of under-predicting transitions. Three research questions were investigated, which were related to the general merit of gyroscope-based modeling within the youth Sojourn framework (Chapter 4; see pages 81-101), and the performance of Soj_{recall} and Soj_{precision} in laboratory (Chapter 5; see pages 102-119) and free-living (Chapter 7; see pages 136-153) settings. In this chapter, the implications are discussed by integrating the evidence from all three research questions, and by contextualizing the evidence with respect to both past and future research. The next three sections (Summary of Findings, Implications of Model Design, and Implications of Study Design) are focused on presenting the key findings and addressing issues that affect how those findings are interpreted. The ensuing section (Core Implications) is focused on addressing at length the implications of this dissertation in key concept areas, namely sensor fusion and change point detection (separately and combined), as well as model generalizability. The final major section (Technical and Practical Implications) addresses the impact of Soj_{recall} and Soj_{precision} on continued practice in sensor-based physical activity (PA) assessment. Afterwards, the strengths and limitations of this dissertation are discussed, followed by a brief conclusion.

Summary of Findings

In terms of segmentation, the original youth Sojourn models (34) had high recall and low precision for both the laboratory and free-living data sets, indicating dramatic overprediction of transitions. In comparison, Soj_{recall} and Soj_{precision} each had a better balance of recall and precision

when applied to the laboratory data set. However, those advantages did not translate to the free-living data set, wherein Soj_{recall} and $Soj_{precision}$ each had low values for both recall and precision.

In terms of intensity classification, there was evidence in all three research questions to suggest that gyroscope data had unique value for detecting sedentary behavior (SB) and light physical activity (LPA), whereas for moderate-to-vigorous physical activity (MVPA) there were clear strengths and weaknesses for both accelerometer data (high sensitivity and low specificity) and gyroscope data (low sensitivity and high specificity). Overall classification agreement was generally higher for Soj_{recall} and $Soj_{precision}$ than the original youth Sojourn models, regardless of setting (laboratory or free-living) and attachment site (hip, non-dominant wrist, or dominant wrist). However, the distinctions were greater for the hip than the wrists. Compared to one another, Soj_{recall} and $Soj_{precision}$ had remarkably similar performance for most performance metrics (including those related to individual intensity categories).

In terms of free-living time spent in SB, LPA, and MVPA, a key finding was that Soj_{recall} and $Soj_{precision}$ yielded mean MVPA estimates that were much closer to the criterion than the estimates from the original models were. The latter finding was especially pronounced for the hip and non-dominant wrist. For SB and LPA, trends differed by attachment site, with Soj_{recall} and $Soj_{precision}$ outperforming the original models across the board on the hip, but not the wrists.

Altogether, Soj_{recall} and $Soj_{precision}$ demonstrated several distinct advantages compared to the original youth Sojourn models. However, there were also clear areas for improvement, as the next sections will discuss.

Implications of Model Design

The design characteristics of Soj_{recall} and $Soj_{precision}$ are important to consider when interpreting the present findings. Furthermore, the design process itself yielded insights that may

benefit future research (including continued refinement of Soj_{recall} and $Soj_{precision}$). Thus, the focus in this section will be on the implications of how Soj_{recall} and $Soj_{precision}$ were designed.

For both Soj_{recall} and $Soj_{precision}$, segmentation was accomplished partly using accelerometer data and partly using gyroscope data. Accelerometer data were used to apply a change point detection method called pruned exact linear time (PELT), which was developed by Killick et al. (196). Gyroscope data were used to perform the remaining segmentation, which was based on identification of prolonged periods of SB, as discussed on page 105 (for Soj_{recall}) and page 107 (for $Soj_{precision}$). After segmentation was complete, activity intensity was determined for each segment using gyroscope-based neural networks.

Since Soj_{recall} and $Soj_{precision}$ were among the first models of their kind, simplicity was emphasized when invoking PELT. Thus, the default settings were used, with the exception of the 5-s minimum segment length, which was implemented for consistency with the grid search results in Chapter 4 (see page 97). Notably, 5-s time spans were also used elsewhere in the segmentation algorithms for the same reason. Other window lengths were tested, but did not lead to improved performance.

Lagged differences (i.e., the time series derivative of the accelerometer data) were used as input to PELT, since that is how data were treated in the original demonstration of PELT (196). Notably, the latter demonstration was not focused on PA. For Soj_{recall} and $Soj_{precision}$, accelerometer data were used as input for PELT instead of gyroscope data, based on testing that showed better performance for the accelerometer. The latter finding is interesting to consider in light of the intensity classification analyses, wherein gyroscope data showed unique value for classifying low-movement activity (high SB sensitivity, but low MVPA sensitivity), while accelerometer data showed the opposite (low SB sensitivity, but high MVPA sensitivity).

Transitions generally involve movement rather than non-movement, which may explain why accelerometer data were more effective than gyroscope data for segmentation via the PELT method.

There were two purposes for identifying prolonged SB bouts as part of segmentation within Soj_{recall} and $Soj_{precision}$. First, the gyroscope-based neural networks showed excellent performance for classifying SB in Chapter 4 (see Table 12), making the bout search a logical way of identifying change points while preventing PELT from treating aberrant accelerometer values as change points. Second, the foundational approach in Sojourn models (32–34) has been to identify segments based on the start and end of low-movement sequences. The latter approach is intuitive, and also beneficial because the statistical properties of low-movement data are much different (and thus may warrant different treatment) than those of movement data. By targeting bouts of SB, Soj_{recall} and $Soj_{precision}$ mirrored the original rationale while adding a layer of sophistication.

Thus far, Soj_{recall} and $Soj_{precision}$ have been primarily discussed in tandem, with little emphasis on their distinctions. For Soj_{recall} , the defining features were limited rejection of PELT change points (only if they occurred within ± 5 s of the start or end of a bout of SB) and thorough determination of SB bouts using a 5-s rolling window. In contrast, $Soj_{precision}$ was designed to favor precision by rejecting all PELT change points that occurred during SB bouts. $Soj_{precision}$ was also designed to run faster than Soj_{recall} by determining bouts of SB via non-overlapping 5-s windows rather than rolling 5-s windows (thereby reducing computations by a factor of five). The differences between Soj_{recall} and $Soj_{precision}$ were specifically engineered to address potential associations between segmentation algorithm characteristics (especially tendencies to predict many versus few transitions) and eventual performance of intensity classification models. In

terms of classification performance, the findings from both the laboratory and free-living studies suggested that it makes little difference which model is used. Thus, it is appropriate to choose between them based on practical differences such as runtime. From that perspective, $Soj_{precision}$ is preferable, since its mean runtime was more than three times faster than the mean for Soj_{recall} .

Implications of Study Design

Powerful study designs were used to generate both the laboratory and free-living data sets, as discussed in Chapter 3 (pages 62-80) and Chapter 6 (pages 120-135), respectively. For the laboratory data set, an exceptionally large sample size was used, and diverse activities were performed with minimal structure. For the free-living data set, the sample size was smaller, but a large amount of data was available for each participant (3.8 h on average). Furthermore, the free-living study was conducted in a mostly independent sample. Overall, the data were far richer than the majority of previous studies (cf. Table 1). However, it is also important to consider the potential implications of collecting data from two separate visits, and the manner in which data were subsequently handled.

For the laboratory study, multiple visits did not pose a major concern. Under normal conditions, the protocol eventually led to all participants performing the same 16 activities (twice each), and missing data were rare, as shown in Figure 2 and Table 5. Data were pooled across visits, and all performance metrics were percentage-based, meaning the scale was identical regardless of whether data were available from one or both visits.

For the free-living data set, the small sample size and unstructured protocol led to greater potential for the findings to be influenced by whether or not participants had two days of data. It was still possible to calculate performance metrics on a standardized (i.e., percentage-based) scale, which was an important part of preventing analyses from being skewed by time scale

differences. However, an irresolvable issue was that those with two days of data had more opportunities than those with one day to engage in a variety of behavior, since the visits ideally took place in different contexts (e.g. one in the home and one at a park). Thus, performance metrics would not necessarily reflect the same thing for those with one versus two days of data. To address the latter issue in Chapter 7 (see page 136-153), the analyses were restricted to include only those with two days of data. However, there were 7-8 participants (depending on attachment site) who had only one day of data, and thus an important question was whether there were major performance changes when including versus excluding those participants from analyses.

Supplemental analyses were conducted to address the latter question. There were 38 participants included for the analyses at each attachment site. The intensity classification and activity time analyses from Chapter 7 (pages 139-140) were replicated, and performance metrics were compared between the original and supplemental analyses. In terms of overall classification agreement across attachment sites, percent correct classification means were always close to the original values (within $\pm 1.5\%$ for Soj_{recall} and $Soj_{precision}$, and within $\pm 4.8\%$ for the original youth Sojourn models). Similarly, kappa means were always within ± 0.04 of the original values for all three models. In terms of intensity-specific classification agreement (i.e., sensitivity, specificity, and F1), means for Soj_{recall} and $Soj_{precision}$ were always within $\pm 3.6\%$ of the original values. For the original youth Sojourn models, there were similar differences for SB (within $\pm 3.0\%$ of the original values), but larger differences for LPA and MVPA (within $\pm 15.7\%$ of the original values). In terms of mean percent of time spent in SB, LPA, and MVPA, criterion values were all within $\pm 1.6\%$ of their original values. For Soj_{recall} and $Soj_{precision}$, there were similar changes (within $\pm 2.0\%$ of the original values), whereas for the original models the changes were

similar for SB (within $\pm 2.5\%$ of the original values) but not LPA or MVPA (within $\pm 14.9\%$ of the original values).

As a whole, the supplemental analyses showed that there were marginal differences between analyses that included versus excluded those with only one day of data. Specifically, criterion data changed very little (on a percentage basis) at the group level, and the same was true for the performance of $\text{Soj}_{\text{recall}}$ and $\text{Soj}_{\text{precision}}$. There were notable changes in performance for the original youth Sojourn models, but those changes most likely reflected inconsistent performance of the models from person to person rather than issues with study design. The latter findings underscore an important advantage of $\text{Soj}_{\text{recall}}$ and $\text{Soj}_{\text{precision}}$ compared to the original models, namely more consistent performance from person to person.

Core Implications

Sensor fusion and change point detection were core concepts for this dissertation. Previous research has examined the two concepts separately, with studies focusing on either sensor fusion or change point detection. However, almost no studies have explored using them simultaneously. This dissertation was specifically geared toward addressing the latter gap, as the segmentation algorithms of $\text{Soj}_{\text{recall}}$ and $\text{Soj}_{\text{precision}}$ used accelerometer and gyroscope data (i.e., cooperative fusion), as well as the PELT method for change point detection. Furthermore, although sensor fusion was not directly used for the intensity classification neural networks, the comparisons of accelerometer and gyroscope neural networks yielded insights that may have relevance for future applications of sensor fusion. Thus, it is crucial to interpret the findings in light of sensor fusion, change point detection, and their combined usage. It is also important to consider the implications from a generalizability perspective, which was addressed in this dissertation through

the independent free-living validation. The preceding factors are discussed below in separate subsections.

Sensor Fusion

Sensor fusion has been used in many previous studies (20, 22–25, 41, 147, 151, 156–172), with accelerometers and gyroscopes being the most frequently used sensors. Almost all studies have focused on activity recognition (20, 22, 23, 41, 147, 156–168), providing consistent evidence that accelerometer and gyroscope data have a complementary relationship. However, there is considerably less evidence to suggest how accelerometer and gyroscope data may complement one another for other tasks besides activity recognition.

A few studies have examined energy expenditure (EE) or oxygen consumption (VO_2) when using accelerometer data, gyroscope data, or both combined. Pärkkä et al. (170) compared the performance of linear models that were developed to predict EE using either accelerometer or gyroscope data obtained from hip-worn, non-dominant wrist-worn, or ankle-worn (same side as wrist) devices. They showed that accelerometer-based models tended to have better performance than gyroscope-based models, although there were certain activities (notably stair walking for all three attachment sites) for which the reverse was true. Vathsangam et al. (24) combined accelerometer and gyroscope data to predict VO_2 during walking. Their findings showed that prediction error decreased by up to 58.8% when combining accelerometer and gyroscope data, compared to using either alone. Taken together, the studies of Pärkkä et al. (170) and Vathsangam et al. (24) provided evidence that accelerometer and gyroscope data can complement one another for assessing EE and VO_2 . However, neither study protocol involved any SB, which was shown by Hibbing et al. (25) to be a key area of strength for gyroscope data. Specifically, in their study receiver operating characteristic curve analyses were used to develop

SB classifiers based on gyroscope or accelerometer data collected from one of five attachment sites (hip, either wrist, or either ankle). All gyroscope classifiers achieved 100% correct classification of SB, whereas the accelerometer classifiers achieved 76.7%-96.7% correct classification (depending on the attachment site). Hibbing et al. (25) also showed that combining accelerometer and gyroscope data led to mean absolute percent error reductions up to 6.0%, when predicting EE.

The present study is among the first to examine accelerometer and gyroscope data within models that predicted activity intensity. For Soj_{recall} and $Soj_{precision}$, segmentation algorithms were developed that used cooperative fusion to detect transitions from both accelerometer and gyroscope data. As discussed previously, the laboratory findings (Chapter 5) showed that the fusion-based segmentation algorithms had a stronger balance of recall and precision than the existing accelerometer-based algorithms. However, that advantage was not maintained in free living (Chapter 7). Thus, sensor fusion showed limited utility for enhancing segmentation.

Apart from segmentation, sensor fusion was not directly addressed in this study. Specifically, intensity classification was examined for accelerometer and gyroscope data separately, but not both combined. Nevertheless, the findings are still informative for future work that may involve calibration of fusion-based classifiers. In Chapter 4, it was shown that neural networks (hip and non-dominant wrist) had higher sensitivity for SB and LPA prediction when using gyroscope data than accelerometer data, by margins of 19.3%-74.2%. The latter finding accords with what was shown for the SB classifiers in the study by Hibbing et al. (25). For MVPA, sensitivity was higher when using accelerometer data than gyroscope data (by margins of 18.7%-24.2%), which may indicate that accelerometer data are particularly useful for MVPA classification. In Chapter 5 and Chapter 7, mean percent correct classification was higher for Soj_{recall} and $Soj_{precision}$

(gyroscope-based classification) than the original youth Sojourn models (raw acceleration-based classification), by margins of 2.3%-20.7% (Chapter 5) and 4.2%-42.9% (Chapter 7). Taken together, the present findings demonstrate the value of gyroscope data within the youth Sojourn framework, while also suggesting that there may be key areas in which fusion of accelerometer and gyroscope data could enhance classification by leveraging the different strengths of each sensor.

Only one other study (158) has examined accelerometer and gyroscope data in the context of activity intensity classification. The study was conducted in a small sample of adults ($N = 9$) who performed up to 18 structured activities. The authors used accelerometer and gyroscope data to classify LPA, moderate PA (MPA), or vigorous PA (VPA), with the models being trained against labels obtained from the compendium of physical activities (180). Using both accelerometer and gyroscope data, a k nearest neighbors model was able to achieve 96.7% correct classification, providing evidence that accelerometer and gyroscope data have complementary value for activity intensity classification. Although Soj_{recall} and $Soj_{precision}$ did not achieve such a high percent correct classification in the present study (maximum of $77.6\% \pm 6.9\%$, for $Soj_{precision}$ on the hip in Chapter 5), there are notable differences between the present study and the study by Reiss and Stricker (158). First, the populations differed, with Reiss and Stricker (158) assessing adults while youth were assessed for the present study. Youth can be more challenging to assess than adults, due to their erratic behavior patterns (231, 246), which may partially explain the lower performance seen in this study. Additionally, the models of Reiss and Stricker (158) predicted different intensities than the models in the present study. Specifically, their models classified LPA, MPA, and VPA, whereas the present models classified SB, LPA, and MVPA. Differences between the categories may have influenced the findings, but

the biggest difference was between the criterion measures in both studies. Reiss and Stricker (158) used the compendium as a criterion measure, whereas the present study used a combination of indirect calorimetry (for EE) and direct observation (for posture). The compendium values assume no individual differences in EE for a given activity, which simplifies the classification task compared to the approach used for the present study. Specifically, indirect calorimetry and direct observation account for the fact that the same activity may have differing intensity for two individuals (e.g. LPA for one and MVPA for the other), whereas the compendium contains only one intensity per activity. The latter distinction makes modeling much easier when using the compendium than when using indirect calorimetry and direct observation, especially considering the unique challenges of assessing EE in youth (232, 247). Despite the differences between the present study and the study by Reiss and Stricker (158), both studies provided strong evidence that accelerometer and gyroscope data have value for activity intensity classification.

It was previously noted that gyroscope data showed unique value for classifying SB. That utility has favorable implications for several modeling approaches. Naturally, the Sojourn method is a key example due to its emphasis on detecting low-movement periods as part of segmentation. However, the benefits of gyroscope-based SB detection are not limited to Sojourn modeling. “Divide-and-conquer” modeling (220, 221) could also benefit from incorporating gyroscope data for SB recognition. In divide-and-conquer models, an initial classification step is used to identify the type of activity being performed, and that information is then used to select which of several EE models should be used. By using gyroscope data to identify SB, divide-and-conquer models would likely show improved EE estimations for both SB and non-SB, since

gyroscope data showed strong performance in terms of both sensitivity and specificity (89.1%-98.9% in Chapter 4) when classifying SB.

Apart from SB classification, a key finding in the present study was that neither accelerometer data nor gyroscope data showed strong performance for MVPA classification. As discussed previously, the two sensors had contrasting classification characteristics, with the accelerometer showing high sensitivity and low specificity, and the gyroscope showing the opposite. The latter finding may indicate that the accelerometer and gyroscope are especially complementary for MVPA detection. If so, future studies should aim to harmonize the two in ways that leverage their individual strengths. An obvious approach would be to develop models that use data from both sensors, which has shown promise in several studies discussed previously (24, 25). Alternatively, there may be value in taking an ensemble approach, in which a metamodel is trained to make a final intensity classification based on initial classifications made separately from accelerometer and gyroscope data.

Change Point Detection

Change point detection differs from other unique segmentation approaches (including the original youth Sojourn segmentation algorithms) in its focus on data distributions. That is, rather than identifying changes in signal magnitude (36) or shape (38), change point detection methods identify changes in statistical distribution. Several studies have explored change point detection for PA applications, and the present study is the first to do so using the PELT method.

Feuz et al. (42) used a method called relative unconstrained least squares importance fitting (RuLSIF) (195) to identify activity transitions using smart home data from both structured (N = 15 adults, nine females) and unstructured (N = 8 older adults, six females) activity protocols, with the former involving a total of 12 activities. For both data sets, RuLSIF had true positive

rate > 80% and false positive rate < 15%, both of which were better than baseline segmentation methods. For the present study, RuLSIF was examined as a potential approach during development, but it showed neither stronger performance nor reasonable runtime compared to PELT. There are several reasons RuLSIF may have been effective in the study by Feuz et al. (42), but not in the present study. A key reason is that the smart home data were much more robust (30 total sensors) than data from the devices worn in this study. However, a more subtle reason has to do with analysis. In the present study, the transition pairing method (TPM) of Hibbing et al. (35) was used to evaluate segmentation correctness, whereas Feuz et al. (42) used the common approach of simply looking within ± 20 s of each criterion transition to see if a prediction occurred. If multiple predictions occurred, the best-timed prediction was considered a true positive, and the others were ignored. The choice to ignore certain transitions may have caused spurious increases in the performance metrics that would not occur when using the TPM. Thus, the different findings for this study versus the study by Feuz et al. (42) may have been explained in part by differences in the analytical approach taken.

The issue of analytical methodology also affects comparisons between this study and others that have used change point detection. Ni et al. (44) used the multivariate online change detection algorithm (198, 199) in a laboratory-based activity recognition framework. Participants (N = 10) wore an accelerometer on the wrist while performing 12 activities, including three stationary activities (e.g. sitting), three dynamic activities (e.g. walking), and six transitional activities (e.g. sit-to-stand). After applying the segmentation algorithm, various classification models were applied, of which random forest performed the best (95.5% correct classification). Agreement decreased when using standard segmentation methods, namely sliding windows with 50% overlap (94.7% correct classification) and no overlap (93.6% correct classification). Thus,

the study by Ni et al. (44) provided some evidence in favor of change point detection. However, the key performance metric was a classification-focused metric (i.e., percent correct classification), rather than a segmentation-focused metric such as recall or root mean squared error. Hibbing et al. (35) demonstrated that classification-focused metrics can take on high values even if segmentation-focused metrics take on low values. Thus, it is possible that the change point detection method presented by Ni et al. (44) did not actually perform as well as the classification-focused analysis made it seem. The latter issue requires especially careful consideration in light of the fact that performance increases from change point detection were only 0.8%-1.9% compared to sliding windows, which is likely not meaningful in practice. Altogether, it is difficult to make segmentation-specific comparisons between the present study and the study by Ni et al. (44), due to the different analytical approaches taken in each study.

For the present study, the PELT method was used, in part due to its speed-related scalability for future applications. That is, many change point detection methods run slowly when applied to high volumes of data, which are typical in sensor-based PA assessment. Thus, the speed of PELT played a considerable role in its use. Compared to previous studies, the PELT method showed weak performance, but as discussed above, much of that difference may be attributable to the nature of the TPM versus other analytical methods.

In the laboratory study, PELT showed stronger performance than the youth Sojourn segmentation algorithms, which was reflected in a stronger balance of recall and precision. However, PELT performed poorly in the free-living cross-validation. Furthermore, classification agreement was similar for Soj_{recall} and $Soj_{precision}$ in both the laboratory and free-living studies, which could potentially call into question whether segmentation plays a large role in intensity classification to begin with. The latter notion was explored in a supplemental analysis by

applying the gyroscope neural networks with features calculated in 5-s rolling windows (i.e., without using a specialized segmentation algorithm).

In the supplemental analysis, data from 38 participants were used for all attachment sites. Participants were included in the analysis whether or not they had data from both visits. When using the rolling windows, mean percent correct classification decreased (compared to both Soj_{recall} and $Soj_{precision}$) by 3.4%-5.9% for the hip and non-dominant wrist, whereas there were smaller decreases (0.7%-1.5%) on the dominant wrist. Similarly, mean kappa decreased by 0.05-0.10 for the hip and non-dominant wrist, whereas the changes were -0.02-0.02 for the dominant wrist. In terms of intensity-specific agreement (i.e., sensitivity, specificity, and F1), most means were lower for the rolling windows than Soj_{recall} and $Soj_{precision}$, by margins of 0.6%-18.9%. When that was not the case, the differences were typically small (0.3%-3.6%), with the exceptions of SB specificity on the wrists (differences of 6.5%-21.5% for the wrists) and LPA sensitivity and F1 on the dominant wrist (differences of 11.0%-19.7%). In terms of mean percent of time spent in SB, LPA, and MVPA, the findings differed by attachment site. For the hip, rolling windows yielded mean SB and LPA estimates that were further from the criterion than the means for Soj_{recall} and $Soj_{precision}$, by 0.7%-7.6%. For MVPA, all three methods yielded means within 1.1% of one another. For the non-dominant wrist, SB estimates were closer to the criterion using rolling windows than Soj_{recall} or $Soj_{precision}$, by margins of 6.7%-7.4%. The opposite was true for MVPA, with Soj_{recall} and $Soj_{precision}$ being closer to the criterion by 5.2%-7.8%. For LPA, all three methods yielded means within 3.3% of one another. For the dominant wrist, means for SB and LPA were closer to the criterion when using rolling windows than Soj_{recall} or $Soj_{precision}$, by 13.7%-18.8%. For MVPA, all three methods yielded means within 2.2% of one another.

As a whole, the supplemental analysis yielded equivocal evidence. That is, rolling windows outperformed Soj_{recall} and $Soj_{precision}$ in some areas, but not in others. While it seems unclear whether change point detection ultimately enhances the prediction of activity intensity, it is difficult to judge given the poor performance of the PELT-based segmentation algorithms to begin with. A future study could address the issue more directly by comparing rolling window predictions to those generated under a condition of perfect segmentation (i.e., by using criterion data to assign the segments), similar to the approach taken by Albinali et al. (72). Such a study would provide a litmus test to assess whether continued innovation in change point detection is worthwhile. That is, if rolling windows were to perform as well or better than predictions under ideal segmentation, there would be evidence to suggest that change point detection is not an area worthy of continued innovation. Given the ostensibly high volume of work it would take to make change point detection effective, such a study would potentially spare a great deal of needless effort and wasted resources in the field.

Combining Sensor Fusion and Change Point Detection

The present study was one of the first to approach segmentation by combining sensor fusion and change point detection. The findings suggested that sensor fusion may be more promising than change point detection, in terms of developing improved models for predicting activity intensity. However, ongoing research may shed new light on ways the two can be effectively combined. Recently, Twaites et al. (40) presented an adaptation of online Bayesian change point detection (OBCPD) (200), which they used for identifying postural transitions. Although postural transitions are not highly relevant to the focus in this study (i.e., activity intensity classification), it is nevertheless notable that OBCPD has the ability to process multivariate data. The latter capability may allow OBCPD to leverage the advantages of accelerometer and

gyroscope data simultaneously when detecting activity transitions. Furthermore, OBCPD runs quickly, meaning it is scalable for PA research. Altogether, OBCPD seems to be a promising avenue for combining sensor fusion and change point detection in future research. However, there is currently not a robust infrastructure for applying OBCPD, meaning further development is needed before it becomes fully viable for further calibration, validation, and deployment.

Only one other study (38) has combined sensor fusion with a specialized segmentation algorithm. Although the segmentation algorithm was not truly a change point detection method, the study still warrants mention here. Gharghabi et al. (38) presented a multi-purpose algorithm called fast low-cost online semantic segmentation (FLOSS), which identified transition points when the shape of the time series changed. Several demonstrations of FLOSS were given, including one focused on detecting activity transitions. In that analysis, an informal graphical approach was taken, wherein visualizations showed strong alignment between the occurrence of actual transitions (annotated using direct observation) and predictions by FLOSS, which were based on accelerometer and gyroscope data from a hip-worn device. Similar to OBCPD, FLOSS holds promise for future studies, but currently lacks a suitable infrastructure for widespread use.

Model Generalizability

A notable finding in this dissertation was that some aspects of model performance were similar in laboratory and free-living settings (with independent samples), while other aspects were much lower in the latter than the former. Reduced performance in free living is a well-documented problem (248, 249), and thus it is crucial to examine any factors that can contribute to improved model generalizability. In that regard, gyroscope data showed some promise, which warrants further comment.

As noted previously, Soj_{recall} and $Soj_{precision}$ each exhibited poor segmentation performance in the free-living study, compared to what was shown in the laboratory study. Similarly, mean percent correct classification decreased by 6.3%-14.2% across attachment sites. Thus, generalizability was not a major strength of the models overall. However, for LPA classification specifically, generalizability was outstanding, with half of the metrics increasing (by 0.8%-6.0%), and most others decreasing by small amounts ($\leq 3.3\%$). For SB classification, all metrics decreased, but the decreases for sensitivity and F1 were 5.0%-10.4% for all attachment sites. Given the unique value of gyroscope data for detecting SB and LPA, it is likely that gyroscope data played a driving role in the promising level of generalizability that was seen for those intensity categories.

Although it is well-known that free-living studies are essential, such studies have rarely been conducted, especially on the scale used for this dissertation. With a free-living data set like the one presented in Chapter 6 (see pages 120-135), it becomes possible to explore the novel concept of using free-living data to train models, which has been demonstrated by Ermes et al. (23). The latter approach could potentially increase model generalizability even further, and thus it warrants attention in future research.

Technical and Practical Implications

To date, sensor fusion and change point detection models have had limited impact on general practice in sensor-based PA assessments. The few existing models have generally been developed for devices that are not commercially available, which is a major limitation addressed by Soj_{recall} and $Soj_{precision}$ (through development for ActiGraph GT9X devices). Beyond the commercial availability of devices, another limitation in recent PA assessment research has been poor distribution of new models (216). Soj_{recall} and $Soj_{precision}$ also offer improvements in the

latter area, due to the existing infrastructure for handling ActiGraph data through open source channels (234, 241). Specifically, the models can be readily added to the Sojourn R package (241) for easy dissemination and implementation.

Although Soj_{recall} and $Soj_{precision}$ represent considerable advances in sensor-based PA assessment, battery life currently poses a challenge for free-living applications. That is, gyroscopes use inordinate amounts of battery (161), which makes extended assessments challenging. For the GT9X, battery life is < 48 h when the gyroscope is enabled, which is not suitable for many applications. Although limited battery life is a concern, the issue can likely be resolved through ongoing hardware innovation, as has already been seen with accelerometer technology (10). Furthermore, there is already precedent (from studies involving global positioning systems) for asking participants to charge devices throughout extended wear protocols (250, 251). Thus, solutions exist in both the short and long term.

When considering applications for Soj_{recall} and $Soj_{precision}$, attachment site is an important factor. Performance was reasonably similar between the hip and wrists, and thus Soj_{recall} and $Soj_{precision}$ are compatible with the increasingly widespread use of wrist-worn monitors for free-living assessments (see page 12-13). However, an interesting outcome was that Soj_{recall} and $Soj_{precision}$ performed better on the non-dominant wrist than the dominant wrist, while the opposite was true for the original models (see Figure 11). The latter finding was especially interesting in light of the fact that the original models were calibrated for the non-dominant wrist, not the dominant wrist. Several previous studies have shown that performance differs very little when using the non-dominant versus dominant wrist (34, 252), and it has also been noted in the past that there is not definitive evidence of one wrist being better (10, 76). However, the existing evidence is based on accelerometer data. Given the wrist differences that were seen in this

dissertation when using gyroscope data, it may be possible that gyroscope data are more influenced than accelerometer data by the issue of dominance. In particular, non-PA-related motions of the dominant limb (e.g. writing and eating) may have contributed to the lower performance on the dominant side. The latter concerns have been commonly posed with respect to accelerometer data (253, 254), but new investigations may be warranted for gyroscope data.

Notably, the laboratory and free-living studies both included limited numbers of left-handed participants, which may affect the evidence presented in this dissertation. Left-handed individuals are also rare in the population (255), which is important to consider. A compounding issue is that switching a monitor from one wrist to the other results in inversion of one axis (256). While the latter issue is largely resolvable through the use of vector magnitudes, there may be cause for advocating wrist attachments by side rather than by dominance. Since the left wrist is non-dominant for the majority of individuals, that may be the preferred attachment site (97). However, more research is needed, particularly as sensor fusion becomes more commonplace.

Strengths & Limitations

This dissertation had notable strengths and limitations. The sample size of the laboratory study was a major strength compared to previous sensor fusion studies that have often used samples of 10-20 (see Table 1). Furthermore, most previous studies have not included free-living data. Thus, the independent validation in this dissertation was a major strength. Although the free-living sample size was small, it also had strengths including a primarily independent sample (all but 4-5 of the 30-31 participants, depending on the attachment site) and a high volume of data (mean of 3.8 h per participant). In both the laboratory and free-living studies, the participant ages ranged from 6-18 years old, which was both a strength and a limitation. Specifically, the

age range increased the diversity of the sample, but the findings may also have been influenced by differences (e.g. in metabolism, movement mechanics, or free-living activity selection) among participants of different ages.

The use of both energy expenditure and posture to classify intensity was a strength of this study, as many studies have only used energy expenditure or posture (via direct observation) despite consensus that SB requires both low energy expenditure and seated or lying posture (235). Despite the strengths of intensity classification in this dissertation, the use of indirect calorimetry may have also posed a limitation for the continuous time course data that were examined in both the laboratory and free-living data sets. Specifically, aerobic metabolism responds slowly to changes in activity, and that lag may have adversely influenced the classification analyses.

Conclusion

Sensor-based PA assessment is an evolving field driven by innovation in many disciplines, such as engineering, computer science, statistics, and exercise physiology. In this dissertation, key innovations were explored within the youth Sojourn model framework (34). Specifically, sensor fusion, change point detection, and gyroscope-based neural networks were examined as potential pathways for improving objective estimates of activity intensity. Two models (i.e., Soj_{recall} and Soj_{precision}) were developed and tested in both laboratory and free-living settings. Compared to existing models, Soj_{recall} and Soj_{precision} had better overall agreement (percent correct classification and kappa) and greater generalizability from the laboratory study to the independent free-living validation. However, there were also areas in which Soj_{recall} and Soj_{precision} could benefit from continued refinement. Altogether, the work presented in this

dissertation represents a step forward in sensor-based PA assessment, and future work can build on several of the findings that have been presented.

REFERENCES

1. US Department of Health and Human Services. *Physical activity and health: a report of the Surgeon General*. Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 1996. 278 p.
2. Tipton CM. The history of “Exercise Is Medicine” in ancient civilizations. *Adv Physiol Educ*. 2014;38(2):109–17.
3. Warburton DER. Health benefits of physical activity: The evidence. *Can Med Assoc J*. 2006;174(6):801–9.
4. Morris JN, Heady JA, Raffle PAB, Roberts CG, Parks JW. Coronary heart-disease and physical activity of work. *Lancet*. 1953;262(6796):1111–20.
5. Penedo FJ, Dahn JR. Exercise and well-being: A review of mental and physical health benefits associated with physical activity. *Curr Opin Psychiatry*. 2005;18(2):189–93.
6. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. *Public Health Rep*. 1985;100(2):126–31.
7. Katch VL. Physical conditioning of children. *J Adolesc Health Care*. 1983;3(4):241–6.
8. Booth M. Assessment of physical activity: An international perspective. *Res Q Exerc Sport*. 2000;71(sup2):114–20.
9. Welk G. *Physical Activity Assessments for Health-related Research*. Champaign (IL): Human Kinetics; 2002. 288 p.
10. 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.

11. Warren JM, Ekelund U, Besson H, Mezzani A, Geladas ND, Vanhees L. Assessment of physical activity - a review of methodologies with reference to epidemiological research: A report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. *Eur J Cardiovasc Prev Rehabil*. 2010;17(2):127–39.
12. Troiano RP. Large-Scale applications of accelerometers: New frontiers and new questions. *Med Sci Sports Exerc*. 2007;39(9):1501.
13. John D, Freedson P. ActiGraph and Actical physical activity monitors: A peek under the hood. *Med Sci Sports Exerc*. 2012;44(1 Suppl):S86–9.
14. Staudenmayer J, Zhu W, Catellier DJ. Statistical considerations in the analysis of accelerometry-based activity monitor data. *Med Sci Sports Exerc*. 2012;44(1 Suppl):S61–7.
15. Liu S, Gao R, Freedson P. Computational methods for estimating energy expenditure in human physical activities. *Med Sci Sports Exerc*. 2012;44(11):2138–46.
16. Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J Appl Physiol*. 2015;119(4):396–403.
17. Crouter SE, Clowers KG, Bassett DR. A novel method for using accelerometer data to predict energy expenditure. *J Appl Physiol*. 2006;100(4):1324–31.
18. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. *Med Sci Sports Exerc*. 2006;38(9):1626–34.

19. Luo RC, Kay MG. Multisensor integration and fusion in intelligent systems. *IEEE Trans Syst Man Cybern.* 1989;19(5):901–31.
20. Zhang Y, Markovic S, Sapir I, Wagenaar RC, Little TDC. Continuous functional activity monitoring based on wearable tri-axial accelerometer and gyroscope. In: *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*; 2011 May 23-26: Dublin (Ireland); 2011. p. 370–3.
21. Chen KY, Bassett DR. The technology of accelerometry-based activity monitors: Current and future. *Med Sci Sports Exerc.* 2005;37(11 Suppl):S490-500.
22. Pärkkä J, Ermes M, Korpipää P, Mäntylä J, Peltola J, Korhonen I. Activity classification using realistic data from wearable sensors. *IEEE Tran Inf Technol Biomed.* 2006;10(1):119–28.
23. Ermes M, Pärkkä J, Mäntylä J, Korhonen I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Trans Inf Technol Biomed.* 2008;12(1):20–6.
24. Vathsangam H, Emken B, Schroeder E, Spruijt-Metz D, Sukhatme GS. Energy estimation of treadmill walking using on-body accelerometers and gyroscopes. In: *Conf Proc IEEE Eng Med Biol Soc.* 2010 Aug 31 - Sep 4: Buenos Aires (Argentina). 2010; p. 6497–501.
25. Hibbing PR, Lamunion SR, Kaplan AS, Crouter SE. Estimating energy expenditure with ActiGraph GT9X inertial measurement unit. *Med Sci Sports Exerc.* 2018;50(5):1093–102.

26. Marcotte RT, Bassett DR, Weinhandl JT, Crouter SE. Application of the ActiGraph GT9X IMU for the assessment of turning during walking and running. *Biomed Phys Eng Express* [Internet]. 2018;4(6). doi: 10.1088/2057-1976/aad0d0.
27. Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors—a review of classification techniques. *Physiol Meas*. 2009;30(4):R1–33.
28. Ayabe M, Kumahara H, Morimura K, Tanaka H. Epoch length and the physical activity bout analysis: An accelerometry research issue. *BMC Res Notes*. 2013;6(1):20.
29. Freedson P, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. *Med Sci Sports Exerc*. 1998;30(5):777–81.
30. Rowlands A, Powell SM, Humphries R, Eston RG. The effect of accelerometer epoch on physical activity output measures. *J Exerc Sci Fit*. 2006;4(1):52–8.
31. Crouter SE, Kuffel E, Haas JD, Frongillo EA, Bassett DR. Refined two-regression model for the ActiGraph accelerometer. *Med Sci Sports Exerc*. 2010;42(5):1029–37.
32. 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.
33. 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.
34. Hibbing PR, Ellingson LD, Dixon PM, Welk GJ. Adapted Sojourn models to estimate activity intensity in youth: A suite of tools. *Med Sci Sports Exerc*. 2018;50(4):846–54.

35. Hibbing PR, LaMunion SR, Hilafu H, Crouter SE. Evaluating the performance of sensor-based bout detection algorithms: The transition pairing method. *J Mea Phys Behav.* in Press.
36. Kozina S, Luštrek M, Gams M. Dynamic signal segmentation for activity recognition. In: *Workshop Proceedings on Space, Time, and Ambient Intelligence, International Joint Conference on Artificial Intelligence*; 2011 Jul 16-22: Barcelona (Spain); 2011. p. 93–8.
37. Noor MHM, Salcic Z, Wang KI-K. Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer. *Pervasive Mob Comput* [Internet]. 2017 [cited 2020 April 13]; 38. Available from: <http://www.sciencedirect.com/science/article/pii/S1574119216302280>. doi:10.1016/j.pmcj.2016.09.009.
38. Gharghabi S, Yeh C-CM, Ding Y, et al. Domain agnostic online semantic segmentation for multi-dimensional time series. *Data Min Knowl Discov.* 2019;33(1):96–130.
39. Aminikhanghahi S, Cook DJ. A survey of methods for time series change point detection. *Knowl Inf Syst.* 2017;51(2):339–67.
40. Twaites J, Everson R, Langford J, Hillsdon M. Transition detection for automatic segmentation of wrist-worn acceleration data: A comparison of new and existing methods. *J Meas Phys Behav.* 2019;3(1):19–28.
41. Shoaib M, Scholten H, Havinga PJM. Towards physical activity recognition using smartphone sensors. In: *2013 IEEE 10th international conference on ubiquitous intelligence and computing and 2013 IEEE 10th international conference on autonomic and trusted computing*; 2013 Dec 18-21: Sorrento Peninsula (Italy); 2013. p. 80–7.

42. Feuz KD, Cook DJ, Rosasco C, Robertson K, Schmitter-Edgecombe M. Automated detection of activity transitions for prompting. *IEEE Trans Hum Mach Syst.* 2015;45(5):575–85.
43. Sprint G, Cook DJ, Schmitter-Edgecombe M. Unsupervised detection and analysis of changes in everyday physical activity data. *J Biomed Inform.* 2016;63:54–65.
44. Ni Q, Patterson T, Cleland I, Nugent C. Dynamic detection of window starting positions and its implementation within an activity recognition framework. *J Biomed Inform.* 2016;62:171–80.
45. Mathie MJ, Coster ACF, Lovell NH, Celler BG. Accelerometry: Providing an integrated, practical method for long-term, ambulatory monitoring of human movement. *Physiol Meas.* 2004;25(2):R1–20.
46. Aminian K, Najafi B. Capturing human motion using body-fixed sensors: Outdoor measurement and clinical applications. *Comp Anim Virtual Worlds.* 2004;15(2):79–94.
47. Corder K, Ekelund U, Steele RM, Wareham NJ, Brage S. Assessment of physical activity in youth. *J Appl Physiol.* 2008;105(3):977–87.
48. Chen KY, Janz KF, Zhu W, Brychta RJ. Redefining the roles of sensors in objective physical activity monitoring. *Med Sci Sports Exerc.* 2012;44(1 Suppl):S13–23.
49. Physical Activity Guidelines Advisory Committee. *Physical Activity Guidelines Advisory Committee Report.* Washington, D.C.: U.S. Department of Health and Human Services. ; 2008. 683 p. Available from:
<https://health.gov/paguidelines/report/pdf/committeereport.pdf>.
50. World Health Organization. *Global recommendations on physical activity for health.* World Health Organization; 2010. 58 p.

51. de Almeida Mendes M, da Silva ICM, Ramires VV, Reichert FF, Martins RC, Tomasi E. Calibration of raw accelerometer data to measure physical activity: A systematic review. *Gait Posture*. 2018;61:98–110.
52. Bouten CVC, Koekkoek KTM, Verduin M, Kodde R, Janssen JD. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Trans Biomed Eng*. 1997;44(3):136–47.
53. Redmond DP, Hegge FW. Observations on the design and specification of a wrist-worn human activity monitoring system. *Behav Res Methods Instrum Comput*. 1985;17(6):659–69.
54. Esliger DW, Copeland JL, Barnes JD, Tremblay MS. Standardizing and optimizing the use of accelerometer data for free-living physical activity monitoring. *J Phys Act Health*. 2005;2(3):366–83.
55. Cleland I, Kikhia B, Nugent C, et al. Optimal placement of accelerometers for the detection of everyday activities. *Sensors*. 2013;13(7):9183–200.
56. Cavagna G, Saibene F, Margaria R. A three-directional accelerometer for analyzing body movements. *J Appl Physiol*. 1961;16(1):191–191.
57. Wong TC, Webster JG, Montoye HJ, Washburn R. Portable accelerometer device for measuring human energy expenditure. *IEEE Trans Biomed Eng*. 1981; 28(6):467–71.
58. Montoye HJ, Washburn R, Servais S, Ertl A, Webster JG, Nagle FJ. Estimation of energy expenditure by a portable accelerometer. *Med Sci Sports Exerc*. 1983;15(5):403–7.
59. Iannacci J. *RF-MEMS technology for high-performance passives the challenge of 5G mobile applications*. Bristol (UK): IOP Publishing; 2017. 164 p.

60. Roylance LM, Angell JB. A batch-fabricated silicon accelerometer. *IEEE Trans Electron Devices*. 1979;26(12):1911–7.
61. Chen P, Muller RS, Shiosaki T, White RM. WP-B6 silicon cantilever beam accelerometer utilizing a PI-FET capacitive transducer. *IEEE Trans Electron Devices*. 1979;26(11):1857–1857.
62. Intille SS, Lester J, Sallis JF, Duncan G. New horizons in sensor development. *Med Sci Sports Exerc*. 2012;44(1 Suppl):S24–31.
63. Troiano RP. Translating accelerometer counts into energy expenditure: Advancing the quest. *J Appl Physiol*. 2005;100(4):1107–8.
64. Godfrey A, Conway R, Meagher D, ÓLaighin G. Direct measurement of human movement by accelerometry. *Med Eng Phys*. 2008;30(10):1364–86.
65. Hendelman D, Miller E, Baggett C, Debold E, Freedson P. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med Sci Sports Exerc*. 2000;32(9 Suppl):S442-9.
66. Bouten CVC, Sauren AAHJ, Verduin M, Janssen JD. Effects of placement and orientation of body-fixed accelerometers on the assessment of energy expenditure during walking. *Med Biol Eng Comput*. 1997;35(1):50–6.
67. Mackintosh KA, Montoye AHK, 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.
68. Trost SG, Mciver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exerc*. 2005;37(11 Suppl):S531–43.

69. Heil DP, Brage S, Rothney MP. Modeling Physical Activity Outcomes from Wearable Monitors: *Med Sci Sports Exerc.* 2012;44(1 Suppl):S50–60.
70. Bassett DR. Validity of four motion sensors in measuring moderate intensity physical activity. *Med Sci Sports Exerc.* 2000;32(9 Suppl):S471–S480.
71. Strath SJ, Bassett Jr DR, Swartz AM. Comparison of MTI accelerometer cut-points for predicting time spent in physical activity. *Int J Sports Med.* 2003;24(04):298–303.
72. Albinali F, Intille S, Haskell W, Rosenberger M. Using wearable activity type detection to improve physical activity energy expenditure estimation. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*; 2010 Sep 26-29: Copenhagen (Denmark); 2010. p. 311–20.
73. Pate RR, O’neill JR, Lobelo F. The evolving definition of "sedentary". *Exerc Sport Sci Rev.* 2008;36(4):173–178.
74. 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):402–407.
75. Welk GJ. Principles of design and analyses for the calibration of accelerometry-based activity monitors. *Med Sci Sports Exerc.* 2005;37(11 Suppl):S501–11.
76. Crouter SE, Flynn JI, Bassett DR. Estimating physical activity in youth using a wrist accelerometer. *Med Sci Sports Exerc.* 2015;47(5):944–51.
77. Schaefer CA, Nigg CR, Hill JO, Brink LA, Browning RC. Establishing and evaluating wrist cutpoints for the GENEActiv accelerometer in youth. *Med Sci Sports Exerc.* 2014;46(4):826–33.

78. Schaefer CA, Nace H, Browning R. Establishing wrist-based cutpoints for the Actical accelerometer in elementary school-aged children. *J Phys Act Health*. 2014;11(3):604–13.
79. Janz KF. Validation of the CSA accelerometer for assessing children’s physical activity. *Med Sci Sports Exerc*. 1994;26(3):369–75.
80. Arvidsson D, Fridolfsson J, Börjesson M, et al. Re-examination of accelerometer data processing and calibration for the assessment of physical activity intensity. *Scand J Med Sci Sports*. 2019;29(10):1442-52.
81. 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.
82. Zhang S, Rowlands AV, Murray P, Hurst TL. Physical activity classification using the GENEa wrist-worn accelerometer: *Med Sci Sports Exerc*. 2012;44(4):742–8.
83. Migueles JH, Cadenas-Sanchez C, Ekelund U, et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Med*. 2017;47(9):1821-45.
84. Granat MH. Event-based analysis of free-living behaviour. *Physiol Meas*. 2012;33(11):1785–800.
85. 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(1 Suppl):S1–4.
86. Strath SJ, Pfeiffer KA, Whitt-Glover MC. Accelerometer use with children, older adults, and adults with functional limitations. *Med Sci Sports Exerc*. 2012;44(1 Suppl):S77–85.

87. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y. Physical human activity recognition using wearable sensors. *Sensors*. 2015;15(12):31314–38.
88. Najafi B, Aminian K, Loew F, Blanc Y, Robert PA. Measurement of stand-sit and sit-stand transitions using a miniature gyroscope and its application in fall risk evaluation in the elderly. *IEEE Trans Biomed Eng*. 2002;49(8):843–51.
89. Bonomi AG, Goris AHC, Yin B, Westerterp KR. Detection of type, duration, and intensity of physical activity using an accelerometer. *Med Sci Sports Exerc*. 2009;41(9):1770–7.
90. Ward DS, Evenson KR, Vaughn A, Rodgers AB, Troiano RP. Accelerometer use in physical activity: Best practices and research recommendations. *Med Sci Sports Exerc*. 2005;37(11 Suppl):S582–8.
91. Brooks RR, Iyengar SS. *Multi-sensor fusion: fundamentals and applications with software*. Upper Saddle River (NJ): Prentice-Hall, Inc.; 1998. 488 p.
92. 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):S450–6.
93. Heil DP. Predicting activity energy expenditure using the Actical® activity monitor. *Res Q Exerc Sport*. 2006;77(1):64–80.
94. Yngve A, Nilsson A, Sjöström M, Ekelund U. Effect of monitor placement and of activity setting on the MTI accelerometer output. *Med Sci Sports Exerc*. 2003;35(2):320–6.

95. Mannini A, Rosenberger M, Haskell WL, Sabatini AM, Intille SS. Activity recognition in youth using single accelerometer placed at wrist or ankle. *Med Sci Sports Exerc.* 2017;49(4):801–12.
96. Crouter SE, Oody JF, Bassett DR. Estimating physical activity in youth using an ankle accelerometer. *J Sports Sci.* 2018;36(19):2265–71.
97. Montoye AHK, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA. Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior. *AIMS Public Health.* 2016;3(2):298-312.
98. Berendsen BAJ, Hendriks MRC, Willems P, Meijer K, Schaper NC, Hans H C M Savelberg. A 20 min window is optimal in a non-wear algorithm for tri-axial thigh-worn accelerometry in overweight people. *Physiol Meas.* 2014;35(11):2205.
99. 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.
100. Dowd KP, Harrington DM, Bourke AK, Nelson J, Donnelly AE. The measurement of sedentary patterns and behaviors using the activPAL™ Professional physical activity monitor. *Physiol Meas.* 2012;33(11):1887–99.
101. 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.
102. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: Hip versus wrist data. *Physiol Meas.* 2014;35(11):2183–9.

103. Doherty A, Jackson D, Hammerla N, et al. Large scale population assessment of Physical activity using wrist worn accelerometers: The UK Biobank study. *PLoS One* [Internet]. 2017 [cited 2020 April 13];12(2). Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169649>. doi: 10.1371/journal.pone.0169649.
104. Morgenthaler T, Alessi C, Friedman L, et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: An update for 2007. *Sleep*. 2007;30(4):519–29.
105. Slater JA, Botsis T, Walsh J, King S, Straker LM, Eastwood PR. Assessing sleep using hip and wrist actigraphy. *Sleep Biol Rhythms*. 2015;13(2):172–80.
106. Bergmann JHM, Graham S, Howard N, McGregor A. Comparison of median frequency between traditional and functional sensor placements during activity monitoring. *Measurement*. 2013;46(7):2193–200.
107. Troiano RP, Berrigan D, Dodd KW, MâSse LC, Tilert T, Mcdowell M. Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc*. 2008;40(1):181–8.
108. Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc*. 2011;43(2):357–64.
109. Ehrlich S, Casteel A, Crouter S, et al. Alternative wear-time estimation methods compared to traditional diary logs for wrist-worn ActiGraph accelerometers in pregnant women [published online ahead of print, Feb 18, 2020]. *J Meas Phys Behav*. doi: 10.1123/jmpb.2019-0049.

110. Plasqui G. Smart approaches for assessing free-living energy expenditure following identification of types of physical activity: Assessing free-living energy expenditure. *Obes Rev.* 2017;18:50–5.
111. Plasqui G, Bonomi AG, Westerterp KR. Daily physical activity assessment with accelerometers: New insights and validation studies: Accelerometer validity. *Obes Rev.* 2013;14(6):451–62.
112. 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.
113. Rowlands AV, Stiles VH. Accelerometer counts and raw acceleration output in relation to mechanical loading. *J Biomech.* 2012;45(3):448–54.
114. Rowlands AV, Fraysse F, Catt M, et al. Comparability of measured acceleration from accelerometry-based activity monitors. *Med Sci Sports Exerc.* 2015;47(1):201–10.
115. Kiani K, Snijders CJ, Gelsema ES. Recognition of daily life motor activity classes using an artificial neural network. *Arch Phys Med Rehabil.* 1998;79(2):147–54.
116. Rothney MP, Neumann M, Béziat A, Chen KY. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J Appl Physiol.* 2007;103(4):1419–27.
117. 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.

118. Vähä-Ypyä H, Vasankari T, Husu P, et al. Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (MAD). *PLoS One* [Internet]. 2015 [cited 2020 April 13];10(8). Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134813>. doi: 10.1371/journal.pone.0134813.
119. van Hees VT, Gorzelniak L, León ECD, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One* [Internet]. 2013 [cited 2020 April 13];8(4). Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061691>. doi: journal.pone.0061691.
120. Migueles JH, Rowlands AV, Huber F, Sabia S, van Hees VT. GGIR: A research community-driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *J Meas Phys Behav*. 2019;2(3):188-196.
121. van Hees VT, Fang Z, Langford J, et al. Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents. *J Appl Physiol*. 2014;117(7):738–44.
122. Bai J, He B, Shou H, Zipunnikov V, Glass TA, Crainiceanu CM. Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics*. 2014;15(1):102–16.
123. van Hees VT, Thaler-Kall K, Wolf K-H, et al. Challenges and opportunities for harmonizing research methodology: Raw accelerometry. *Methods Inf Med*. 2016;55(6):525–32.

124. Welk GJ, McClain J, Ainsworth BE. Protocols for evaluating equivalency of accelerometry-based activity monitors. *Med Sci Sports Exerc.* 2012;44(1 Suppl):S39–49.
125. O’Driscoll R, Turicchi J, Beaulieu K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. *Br J Sports Med.* 2020;54(6):332-40.
126. Zhang K, Werner P, Sun M, Pi-Sunyer FX, Boozer CN. Measurement of human daily physical activity. *Obes Res.* 2003;11(1):33–40.
127. Ryan CG. The validity and reliability of a novel activity monitor as a measure of walking. *Br J Sports Med.* 2006;40(9):779–84.
128. Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Estimation of free-living energy expenditure using a novel activity monitor designed to minimize obtrusiveness. *Obesity.* 2010;18(9):1845–51.
129. Lee J-M, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. *Med Sci Sports Exerc.* 2014;46(9):1840–8.
130. Wright SP, Hall Brown TS, Collier SR, Sandberg K. How consumer physical activity monitors could transform human physiology research. *Am J Physiol Regul Integr Comp Physiol.* 2017;312(3):R358–67.
131. Wijndaele K, Westgate K, Stephens SK, et al. Utilization and harmonization of adult accelerometry data: Review and expert consensus. *Med Sci Sports Exerc.* 2015;47(10):2129–39.
132. Cain KL, Conway TL, Adams MA, Husak LE, Sallis JF. Comparison of older and newer generations of ActiGraph accelerometers with the normal filter and the low frequency extension. *Int J Behav Nutr Phys Act.* 2013;10(1):51.

133. Robusto KM, Trost SG. Comparison of three generations of ActiGraph™ activity monitors in children and adolescents. *J Sports Sci.* 2012;30(13):1429–35.
134. Rothney MP, Apker GA, Song Y, Chen KY. Comparing the performance of three generations of ActiGraph accelerometers. *J Appl Physiol.* 2008;105(4):1091–7.
135. Jones DM, Hart HF, Crossley KM, Ackerman IN, Kemp JL. What is the agreement between two generations of commercial accelerometer in a free-living environment for young to middle-aged adults? *J Meas Phys Behav.* 2019;2(2):49–57.
136. Puyau MR, Adolph AL, Vohra FA, Butte NF. Validation and calibration of physical activity monitors in children. *Obes Res.* 2002;10(3):150–157.
137. Butte NF, Ekelund U, Westerterp KR. Assessing physical activity using wearable monitors: Measures of physical activity. *Med Sci Sports Exerc.* 2012;44(1 Suppl):S5–12.
138. John D, Sasaki J, Staudenmayer J, Mavilia M, Freedson P. Comparison of raw acceleration from the GENEa and ActiGraph™ GT3X+ activity monitors. *Sensors.* 2013;13(11):14754–63.
139. Rowlands AV, Yates T, Davies M, Khunti K, Edwardson CL. Raw accelerometer data analysis with GGIR R-package: Does accelerometer brand matter? *Med Sci Sports Exerc.* 2016;48(10):1935–41.
140. Rowlands AV, Mirkes EM, Yates T, et al. Accelerometer-assessed physical activity in epidemiology: Are monitors equivalent? *Med Sci Sports Exerc.* 2018;50(2):257–65.
141. Dorminy CA, Choi L, Akohoue SA, Chen KY, Buchowski MS. Validity of a multisensor armband in estimating 24-h energy expenditure in children. *Med Sci Sports Exerc.* 2008;40(4):699–706.

142. Welk GJ, McClain JJ, Eisenmann JC, Wickel EE. Field validation of the MTI ActiGraph and BodyMedia Armband monitor using the IDEEA monitor. *Obesity*. 2007;15(4):918–28.
143. Bai Y, Hibbing P, Mantis C, Welk GJ. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. *J Sports Sci*. 2017;1–8.
144. Montoye AHK, Conger SA, Connolly CP, et al. Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings. *Meas Phys Educ Exerc Sci*. 2017;21(4):223–34.
145. Park S, Toth LP, Hibbing PR, et al. Dominant vs. non-dominant wrist placement of activity monitors: Impact on steps per day. *J Meas Phys Behav*. 2019;2(2):118–23.
146. Rhudy MB, Dreisbach SB, Moran MD, Ruggiero MJ, Veerabhadrapa P. Cut points of the ActiGraph GT9X for moderate and vigorous intensity physical activity at four different wear locations. *J Sports Sci*. 2020;38(5):503–10.
147. Nam Y, Park JW. Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor. *IEEE J Biomed Health Inform*. 2013;17(2):420–6.
148. Zhang K, Pi-Sunyer FX, Boozer CN. Improving energy expenditure estimation for physical activity. *Med Sci Sports Exerc*. 2004;36(5):883–9.
149. Strath SJ, Bassett DR, Thompson DL, Swartz AM. Validity of the simultaneous heart rate-motion sensor technique for measuring energy expenditure: *Med Sci Sports Exerc*. 2002;34(5):888–94.

150. Tapia EM, Intille SS, Haskell W, et al. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: *Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers*; 2007 Oct 11-13: Boston (MA); 2007. p. 1–4.
151. Reiss A, Stricker D. Introducing a modular activity monitoring system. In: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*; 2011 Aug 30 - Sep 3: Boston (MA); 2011. p. 5621–5624.
152. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Med Sci Sports Exerc.* 2010;42(11):2134–40.
153. Arvidsson D, Slinde F, Hulthén L. Free-living energy expenditure in children using multi-sensor activity monitors. *Clin Nutr.* 2009;28(3):305–12.
154. LaMunion SR, Blythe AL, Hibbing PR, Kaplan AS, Clendenin BJ, Crouter SE. Use of consumer monitors for estimating energy expenditure in youth. *Appl Physiol Nutr Metab.* 2020;45(2):161-8.
155. Hedegaard M, Anvari-Moghaddam A, Jensen BK, Jensen CB, Pedersen MK, Samani A. Prediction of energy expenditure during activities of daily living by a wearable set of inertial sensors. *Med Eng Phys.* 2020;75:13–22.
156. Lau H-Y, Tong K-Y, Zhu H. Support vector machine for classification of walking conditions using miniature kinematic sensors. *Med Biol Eng Comput.* 2008;46(6):563–73.

157. Boughorbel S, Breebaart J, Bruekers F, Flinsenberg I, ten Kate W. Child-activity recognition from multi-sensor data. In: *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*; 2010 Aug 24-27: Eindhoven (The Netherlands); 2010. p. 1–3.
158. Reiss A, Stricker D. Introducing a new benchmarked dataset for activity monitoring. In: *Proceedings of the 2012 16th International Symposium on Wearable Computers*; 2012 June 18-22: Newcastle (United Kingdom); 2012. p. 108–9.
159. Baños O, Damas M, Pomares H, Rojas I, Tóth MA, Amft O. A benchmark dataset to evaluate sensor displacement in activity recognition. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*; 2012 Sep 5-8: Pittsburgh (PA); 2012. p. 1026-35.
160. Liu S, Gao RX, Mo L, Freedson PS. Wearable sensing for physical activity measurement: Design and performance evaluation. *IFAC Proceedings Volumes*. 2013;46(5):53–60.
161. Moncada-Torres A, Leuenberger K, Gonzenbach R, Luft A, Gassert R. Activity classification based on inertial and barometric pressure sensors at different anatomical locations. *Physiol Meas*. 2014;35(7):1245–63.
162. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJM. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*. 2014;14(6):10146-76.
163. Reyes-Ortiz J-L, Oneto L, Samà A, Parra X, Anguita D. Transition-aware human activity recognition using smartphones. *Neurocomputing*. 2016;171:754–67.
164. Zebin T, Scully PJ, Ozanyan KB. Human activity recognition with inertial sensors using a deep learning approach. In: *Proceedings of IEEE Sensors 2016*; 2016 Oct 30 - Nov 2: Orlando (FL); 2016. p. 1–3.

165. Zebin T, Scully PJ, Ozanyan KB. Inertial sensor based modelling of human activity classes: Feature extraction and multi-sensor data fusion using machine learning algorithms. In: Giokas K, Bokor L, Hopfgartner F, editors. *eHealth 360°*. Cham: Springer International Publishing; 2017. p. 306–14.
166. Zebin T, Scully PJ, Ozanyan KB. Evaluation of supervised classification algorithms for human activity recognition with inertial sensors. In: *Proceedings of IEEE SENSORS 2017*; 2017 Oct 29 - Nov 1: Glasgow (Scotland). Scottish Event Campus; 2017. p. 1–3.
167. Zebin T, Sperrin M, Peek N, Casson AJ. Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2018 Jul 17-21: Honolulu (HI); 2018. p. 1–4.
168. D'souza W, R K. Human activity recognition using accelerometer and gyroscope sensors. *Int J Eng Technol*. 2017;9(2):1171–9.
169. Najafi B, Aminian K, Paraschiv-Ionescu A, Loew F, Bula CJ, Robert P. Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Trans Biomed Eng*. 2003;50(6):711–23.
170. Pärkkä J, Ermes M, Antila K, Gils M van, Mänttari A, Nieminen H. Estimating intensity of physical activity: A comparison of wearable accelerometer and gyro sensors and 3 sensor locations. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2007 Aug 23-26: Lyon (France); 2007. p. 1511–4.
171. John D, Liu S, Sasaki JE, et al. Calibrating a novel multi-sensor physical activity measurement system. *Physiol Meas*. 2011;32(9):1473–89.

172. Liu S, Gao RX, John D, Staudenmayer JW, Freedson PS. Multisensor data fusion for physical activity assessment. *IEEE Trans Biomed Eng.* 2012;59(3):687–96.
173. Leuenberger K, Gassert R. Low-power sensor module for long-term activity monitoring. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2011 Aug 30 - Sep 3: Boston (MA); 2011. p. 2237–41.
174. Wu W, Dasgupta S, Ramirez EE, Peterson C, Norman GJ. Classification accuracies of physical activities using smartphone motion sensors. *J Med Internet Res* [Internet]. 2012 [cited 2020 April 13];14(5):e130. Available from: <https://www.jmir.org/2012/5/e130/>. doi: 10.2196/jmir.2208.
175. Anjum A, Ilyas MU. Activity recognition using smartphone sensors. In: *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*; 2013 Jan 11-14: Las Vegas (NV); 2013. p. 914–9.
176. Incel OD, Kose M, Ersoy C. A review and taxonomy of activity recognition on mobile phones. *Bionanoscience.* 2013;3(2):145–71.
177. Martín H, Bernardos AM, Iglesias J, Casar JR. Activity logging using lightweight classification techniques in mobile devices. *Pers Ubiquit Comput.* 2013;17(4):675–95.
178. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A public domain dataset for human activity recognition using smartphones. In: *ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*; 2013 Apr 24-26: Bruges (Belgium). 2013. Available from <http://www.i6doc.com/en/livre/?GCOI=28001100131010>.

179. Baños O, Garcia R, Holgado-Terriza JA, et al. mHealthDroid: A novel framework for agile development of mobile health applications. In: Pecchia L, Chen LL, Nugent C, Bravo J, editors. *Ambient Assisted Living and Daily Activities*. Cham: Springer International Publishing; 2014. p. 91–8.
180. 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–S504.
181. Farrahi V, Niemelä M, Kangas M, Korpelainen R, Jämsä T. Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches. *Gait Posture*. 2019;68:285–99.
182. Nyan MN, Tay FEH, Seah KHW, Sitoh YY. Classification of gait patterns in the time–frequency domain. *J Biomech*. 2006;39(14):2647–56.
183. Sekine M, Tamura T, Togawa T, Fukui Y. Classification of waist-acceleration signals in a continuous walking record. *Med Eng Phys*. 2000;22(4):285–91.
184. Karas M, Strączkiewicz M, Fadel W, Harezlak J, Crainiceanu CM, Urbanek JK. Adaptive empirical pattern transformation (ADEPT) with application to walking stride segmentation [published online ahead of print, September 23, 2019]. *Biostatistics*. doi:10.1093/biostatistics/kxz033.
185. Chua S-L, Marsland S, Guesjen HW. Spatio-temporal and context reasoning in smart homes. In: *Proceeding of COSIT Workshop on Spatial and Temporal Reasoning for Ambient Intelligence System*; 2009 Sep 21-25. Wrac'h (France); 1992. p. 9-20.

186. Hong X, Nugent CD. Partitioning time series sensor data for activity recognition. In: *2009 9th International Conference on Information Technology and Applications in Biomedicine*; 2009 Nov 5-7: Larnaca (Cyprus); 2009. p. 1–4.
187. Ortiz J, Olaya AG, Borrajo D. A dynamic sliding window approach for activity recognition. In: Konstan JA, Conejo R, Marzo JL, Oliver N, editors. *User Modeling, Adaption and Personalization*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 219–30.
188. Okeyo G, Chen L, Wang H, Sterritt R. Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. *Pervasive Mob Comput*. 2014;10:155–72.
189. Krishnan NC, Cook DJ. Activity recognition on streaming sensor data. *Pervasive Mob Comput*. 2014;10:138–54.
190. Shahi A, Woodford BJ, Deng JD. Event classification using adaptive cluster-based ensemble learning of streaming sensor data. In: Pfahringer B, Renz J, editors. *AI 2015: Advances in Artificial Intelligence*. Cham: Springer International Publishing; 2015. p. 505–16.
191. Shahi A, Woodford BJ, Lin H. Dynamic real-time segmentation and recognition of activities using a multi-feature windowing approach. In: Kang U, Lim E-P, Yu JX, Moon Y-S, editors. *Trends and Applications in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing; 2017. p. 26–38.
192. Bifet A, Gavaldà R. Learning from time-changing data with adaptive windowing. In: *Proceedings of the 2007 SIAM International Conference on Data Mining*; 2007 April 26-28: Minneapolis (MN); 2007. p. 443–8.

193. Núñez M, Fidalgo R, Morales R. Learning in environments with unknown dynamics: Towards more robust concept learners. *J Mach Learn Res.* 2007;8:2595–628.
194. Kawahara Y, Sugiyama M. Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*; 2009 Jul 6-10: Denver (United States). Colorado Convention Center; 2009. p. 389-400.
195. Liu S, Yamada M, Collier N, Sugiyama M. Change-point detection in time-series data by relative density-ratio estimation. *Neural Netw.* 2013;43:72–83.
196. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. *J Am Stat Assoc.* 2012;107(500):1590–8.
197. Diaz KM, Krupka DJ, Chang MJ, et al. Fitbit®: An accurate and reliable device for wireless physical activity tracking. *Int J Cardiol.* 2015;185:138–40.
198. Patterson T, McClean S, Nugent C, Zhang S, Galway L, Cleland I. Online change detection for timely solicitation of user interaction. In: Hervás R, Lee S, Nugent C, Bravo J, editors. *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services*. Cham: Springer International Publishing; 2014. p. 116–23.
199. Patterson T, Khan N, McClean S, et al. Sensor-based change detection for timely solicitation of user engagement. *IEEE Trans Mob Comput.* 2017;16(10):2889–900.
200. Adams RP, MacKay DJC. Bayesian online changepoint detection [Internet]. 2007 [cited 2020 Mar 9]; available from: <http://arxiv.org/abs/0710.3742>.
201. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Protein Struct.* 1975;405(2):442–51.

202. Mannini A, Sabatini AM. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*. 2010;10(2):1154–75.
203. Gao L, Bourke AK, Nelson J. Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. *Med Eng Phys*. 2014;36(6):779–85.
204. Zhang M, Sawchuk A. A feature selection-based framework for human activity recognition using wearable multimodal sensors. In: *Proceedings of the 6th International ICST Conference on Body Area Networks*; 2011 Nov 7-8: Beijing (China); 2011. p. 92-8.
205. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507–17.
206. Preece SJ, Goulermas JY, Kenney LPJ, Howard D. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans Biomed Eng*. 2009;56(3):871–9.
207. Kate RJ, Swartz AM, Welch WA, Strath SJ. Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data. *Physiol Meas*. 2016;37(3):360–79.
208. Kheirkhahan M, Tudor-Locke C, Axtell R, et al. Actigraphy features for predicting mobility disability in older adults. *Physiol Meas*. 2016;37(10):1813–33.
209. Koch P, Konen W, Hein K. Gesture recognition on few training data using slow feature analysis and parametric bootstrap. In: *The 2010 International Joint Conference on Neural Networks*; 2010 July 18-23: Barcelona (Spain); 2010. p. 1–8.
210. Kheirkhahan M, Chakraborty A, Wanigatunga AA, Corbett DB, Manini TM, Ranka S. Wrist accelerometer shape feature derivation methods for assessing activities of daily living. *BMC Med Inform Decis Mak*. 2018;18(4):124.

211. 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*. 2009;107(4):1300–7.
212. Freedson PS, Lyden K, Kozey-Keadle S, Staudenmayer J. Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample. *J Appl Physiol*. 2011;111(6):1804–12.
213. 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.
214. Ellingson LD, Hibbing PR, Kim Y, Frey-Law LA, Saint-Maurice PF, Welk GJ. Lab-based validation of different data processing methods for wrist-worn ActiGraph accelerometers in young adults. *Physiol Meas*. 2017;38(6):1045–60.
215. Ellingson LD, Hibbing PR, Welk GJ, et al. Choice of processing method for wrist-worn accelerometers influences interpretation of free-living physical activity data in a clinical sample. *J Meas Phys Behav*. 2019;1–9.
216. Hibbing PR, Bassett DR, Crouter SE. Modifying accelerometer cut-points affects criterion validity in simulated free-living for adolescents and adults [published online ahead of print, February 5, 2020]. *Res Q Exerc Sport*. 2020. doi: 10.1080/02701367.2019.1688227.
217. Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG. Ensemble methods for classification of physical activities from wrist accelerometry. *Med Sci Sports Exerc*. 2017;49(9):1965–73.

218. Trost SG, Wong W-K, Pfeiffer KA, Zheng Y. Artificial neural networks to predict activity type and energy expenditure in youth. *Med Sci Sports Exerc.* 2012;44(9):1801–9.
219. Trost SG, Cliff DP, Ahmadi MN, Tuc NV, Hagenbuchner M. Sensor-enabled activity class recognition in preschoolers: Hip versus wrist data. *Med Sci Sports Exerc.* 2018;50(3):634–41.
220. Bonomi AG, Plasqui G. “Divide and conquer”: Assessing energy expenditure following physical activity type classification. *J Appl Physiol.* 2012;112(5):932–932.
221. Freedson P, Lyden K, Kozey-Keadle S, Staudenmayer J. Reply to Bonomi and Plasqui. *J Appl Physiol.* 2012;112(5):933–933.
222. Crouter SE, Bassett DR. A new 2-regression model for the Actical accelerometer. *Br J Sports Med.* 2008;42(3):217–24.
223. Crouter SE, DellaValle DM, Horton M, Haas JD, Frongillo EA, Bassett DR. Validity of the Actical for estimating free-living physical activity. *Eur J Appl Physiol.* 2011;111(7):1381–9.
224. Crouter SE, DellaValle DM, Haas JD, Frongillo EA, Bassett DR. Validity of ActiGraph 2-regression model, Matthews cut-points, and NHANES cut-points for assessing free-living physical activity. *J Phys Act Health.* 2013;10(4):504–14.
225. Crouter SE, Horton M, Bassett DR. Use of a two-regression model for estimating energy expenditure in children. *Med Sci Sports Exerc.* 2012;44(6):1177–85.
226. Crouter SE, Horton M, Bassett DR. Validity of ActiGraph child-specific equations during various physical activities. *Med Sci Sports Exerc.* 2013;45(7):1403–9.

227. Dongwoo K, Kim HC. Activity energy expenditure assessment system based on activity classification using multi-site triaxial accelerometers. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2007 Aug 23-26: Lyon (France); 2007. p. 2285–7.
228. Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. *J Appl Physiol*. 2009;107(3):655–61.
229. Ruch N, Joss F, Jimmy G, Melzer K, Hanggi J, Mader U. Neural network versus activity-specific prediction equations for energy expenditure estimation in children. *J Appl Physiol*. 2013;115(9):1229–36.
230. Mu Y, Lo HZ, Ding W, Amaral K, Crouter SE. Bipart: Learning block structure for activity detection. *IEEE Trans Knowl Data Eng*. 2014;26(10):2397–409.
231. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Med Sci Sports Exerc*. 2005;37(11 Suppl):S523–30.
232. Hibbing PR, Bassett DR, Coe DP, LaMunion SR, Crouter SE. Youth Metabolic Equivalents Differ Depending on Operational Definitions [published online ahead of print, February 18, 2020]. *Med Sci Sports Exerc*. doi: 10.1249/MSS.0000000000002299.
233. McLaughlin JE, King GA, Howley ET, Bassett Jr DR, Ainsworth BE. Validation of the COSMED K4 b² portable metabolic system. *Int J Sports Med*. 2001;22(04):280–4.
234. Hibbing PR. *AGread: Read data files from ActiGraph monitors*. 2018. [cited 2018 Jul 31] Available from: <https://CRAN.R-project.org/package=AGread>.

235. Tremblay MS, Aubert S, Barnes JD, et al. Sedentary behavior research network (SBRN) – terminology consensus project process and outcome. *Int J Behav Nutr Phys Act.* 2017;14:75.
236. Brønd JC, Arvidsson D. Sampling frequency affects the processing of Actigraph raw acceleration data to activity counts. *J Appl Physiol.* 2016;120(3):362–9.
237. Clevenger KA, Pfeiffer KA, Mackintosh KA, et al. Effect of sampling rate on acceleration and counts of hip- and wrist-worn ActiGraph accelerometers in children. *Physiol Meas.* 2019;40(9):095008.
238. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. *Med Sci Sports Exerc.* 2011;43(8):1561–7.
239. Crouter SE, Hibbing PR, LaMunion SR. Use of objective measures to estimate sedentary time in youth. *J Meas Phys Behav.* 2018;1(3):136–42.
240. Gale D, Shapley LS. College admissions and the stability of marriage. *Am Math Mon.* 1962;69(1):9–15.
241. Hibbing PR, Lyden K, Schwabacher IJ. *Sojourn: Apply Sojourn methods for processing ActiGraph accelerometer data.* 2019. [cited 2019 Jul 10] Available from: <https://CRAN.R-project.org/package=Sojourn>.
242. Jackson B, Scargle JD, Barnes D, et al. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process Lett.* 2005;12(2):105–8.
243. Killick R, Haynes K, Eckley I, Fearnhead P, Lee J. *Changepoint: Methods for changepoint detection.* 2016. [cited 2020 Jan 2] Available from: <https://CRAN.R-project.org/package=changepoint>.

244. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57(1):289–300.
245. Crouter SE, LaMunion SR, Hibbing PR, Kaplan AS, Jr DRB. Accuracy of the Cosmed K5 portable calorimeter. *PLoS One* [Internet]. 2019 [cited 2020 April 14];14(12). Available from:
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226290>. doi: 10.1371/journal.pone.0226290.
246. Nilsson A, Ekelund U, Yngve A, Söström M. Assessing physical activity among children with accelerometers using different time sampling intervals and placements. *Pediatr Exerc Sci.* 2002;14(1):87–96.
247. Herrmann SD, McMurray RG, Kim Y, Willis EA, Kang M, McCurdy T. The influence of physical characteristics on the resting energy expenditure of youth: A meta-analysis. *Am J Hum Biol* [Internet]. 2017 [cited 2020 April 14];29(3):e22944. Available from:
https://onlinelibrary.wiley.com/doi/full/10.1002/ajhb.22944?casa_token=gmtFaq-pU3UAAAAA%3Alktg964TRANZN359HQDiBuygN0jBZG5lCDalGiwSqxOtuCtgCrLO_5bLlyW0G05o6C3lSivS0AVrRw. doi: 10.1002/ajhb.22944.
248. 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):2656–2663.
249. Bastian T, Maire A, Dugas J, et al. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: Laboratory-based calibrations are not enough. *J Appl Physiol.* 2015;118(6):716–22.

250. Demant Klinker C, Schipperijn J, Toftager M, Kerr J, Troelsen J. When cities move children: Development of a new methodology to assess context-specific physical activity behaviour among children and adolescents using accelerometers and GPS. *Health Place*. 2015;31:90–9.
251. Stewart T, Duncan S, Schipperijn J. Adolescents who engage in active school transport are also more active in other contexts: A space-time investigation. *Health Place*. 2017;43:25–32.
252. Montoye AHK, 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.
253. Phillips LRS, Parfitt G, Rowlands AV. Calibration of the GENEa accelerometer for assessment of physical activity intensity in children. *J Sci Med Sport*. 2013;16(2):124–8.
254. 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* [Internet]. 2016 [cited 2018 May 16];16(1). Available from: <http://www.biomedcentral.com/1471-2458/16/223>. doi:10.1186/s12889-016-2901-8. doi: 10.1186/s12889-016-2901-8.
255. Hardyck C, Petrinoich LF. Left-handedness. *Psychol Bull*. 1977;84(3):385–404.
256. Montoye AHK, 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.

APPENDICES

Recruitment Flyer

Seeking Research Participants

Study Details

- » Perform common daily activities, including household chores and sports, while wearing activity monitors
- » Two visits (~2.5 hours total)
- » Compensation up to \$75
- » Excellent opportunity for research and technology exposure!



Who can Participate?

- » Boys and girls ages 6-18 years old
- » Active or non-active youth eligible
- » All testing will take place at the University of Tennessee campus



**Are you Interested?
Contact us below**

Email: APLStudy@utk.edu

Phone: (865) 974-5651

Health History Questionnaire

HEALTH HISTORY QUESTIONNAIRE

Today's Date: _____

Name of parent/legal guardian completing the form: _____

ABOUT THE PARTICIPANT

Please answer the following questions about your child.

Name: _____

Address: _____

City: _____ **Zip Code:** _____

Parent's Phone: _____ **Date of Birth (month/day/year):** _____

Age: _____ **Gender:** ____ M ____ F

Current School (if summer, School entering in Fall): _____

Current Grade (if summer, Grade entering in the Fall): _____

How do you identify your child?

- ____ Asian, Non-Hispanic
- ____ Asian, Hispanic
- ____ Black/African American, Non-Hispanic
- ____ Black/African American, Hispanic
- ____ Native Hawaiian/Pacific Islander, Non-Hispanic
- ____ Native Hawaiian/Pacific Islander, Hispanic
- ____ Native American/Alaskan, Non-Hispanic
- ____ Native American/Alaskan, Hispanic
- ____ White, Non-Hispanic
- ____ White, Hispanic

Is this child Multiracial? _____ Y _____ N

Who resides in the household with your child?

____ Both parents ____ Single Mother ____ Single Father ____ Grandparent(s) ____ Other

Please Turn Over

Health History

Has your child ever been diagnosed with any of the follow conditions? If yes, please explain.

	NO	YES	Current Condition	Explanation
Heart Disease				
Diabetes/High Blood Sugar				
High Blood Pressure				
Seizure				
Asthma or Other Lung Condition				
Cancer				
Musculoskeletal Condition or Injury				
Other Serious Illness				
Allergy				

Please describe any additional medical conditions that may affect your child's participation in physical activity.

Is your child taking any medication (including prescription and non-prescription? If yes, please state below.

YES _____ NO _____

Name of Medication

Reason for Taking

For How Long?

Emergency Contact

Name: _____

Relationship: _____ Phone: Work: _____

Home: _____

Parent Informed Consent Form

Informed Consent Statement for Parent or Guardian

Title of Research: Novel Approaches for Predicting Unstructured Short Periods of Physical Activities in Youth: Study 1 – Development of New Techniques

Principal Investigator: Dr. Scott Crouter

Location: Applied Physiology Laboratory, 1914 Andy Holt Ave., University of Tennessee, Knoxville, TN 37996

PURPOSE

You are being asked to give permission for your child to participate in a research project that will examine new methods to measure the amount of physical activity youth obtain using small devices called accelerometers. These devices are attached to a belt which is worn around the waist, wrist, or ankle similar to a pedometer. Your child will be eligible if he or she is: 1) 6 to 18 years of age, and 2) you and your child can read and speak English. Your child will not be eligible if: 1) he or she has a serious health problem or injury that would prevent them from participating safely in regular physical activity, 2) he or she has a Pacemaker or other internal medical device.

The study will take place at the Applied Physiology Laboratory in the Health, Physical Education, and Recreation (HPER) building at The University of Tennessee. The lead researcher is Scott Crouter, PhD, Assistant Professor in the Department of Kinesiology, Recreation, and Sport Studies. Please read this form and feel free to ask questions. If you have further questions, Dr. Crouter will discuss them with you. You can contact him by telephone (617-287-7509) or email (scrouter@utk.edu).

PARTICIPANT'S INVOLVEMENT IN THIS STUDY

Here is what the study involves:

- 1) You will be asked to fill out a health history questionnaire pertaining to your child's health status to determine eligibility for the study.
- 2) Your child will be asked to visit the Applied Physiology Laboratory in the HPER building on 2 separate days for approximately 100 minutes on visit 1 and 60 minutes on visit 2, for a total time commitment of 160 minutes (2 hrs 40 minutes).
 - a. During visit 1 your child will have his/her height, weight, percent body fat measured, resting metabolic rate measured, and will perform 8 activities randomly selected from a list of 16 activities.
 - b. During visit 2 your child will perform the remaining 8 activities, from the list of 16 that they did not perform during visit 1. The 2 testing days will occur within 2 weeks of each other.
- 3) The percent body fat will be measured using bio-electrical impedance (BIA). BIA is a non-invasive technique that involves standing on a scale while also holding on to a handlebar. During this time an undetectable and harmless electrical current is passed through the body.
- 4) The resting metabolic rate test will take 35 minutes. This test will involve having your child rest in a lying position for 25 minutes and it will take 10 minutes to connect the equipment and prepare for the test. During this time he/she will have 3 accelerometers placed on their waist, 3 on each wrist, 2 on each ankle, and 1 on each shoe using an elastic belt. Another device will also be attached to each thigh using a sticky pad. Your child will also be attached to a Cosmed portable system, which will measure the air that they breathe. This device is light weight and straps onto the body. It also requires the use of a facemask, over the nose and mouth. While wearing the facemask your child will be able to breathe normally and talk, and the mask does not block or restrict airflow.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/16/2018
IRB EXPIRATION DATE: 09/13/2019

_____ **Participant's Initials**

- 5) Your child will perform 16 different activities. During each visit 8 activities will be performed which include:

1) self-paced slow walk, 2) self-paced brisk walk, 3) self-paced running, 4) self-paced stair climbing, 5) shooting a basketball, 6) playing catch with a ball, 7) playing soccer, 8) sitting in a reclined position, 9) sitting upright while reading, 10) sitting upright while searching the internet, 11) sitting upright while playing computer games, 12) lying, 13) cycling at a moderate pace, 14) performing jumping jacks, 15) sweeping, and 16) dusting

- a. While performing the activities your child will be wearing the elastic belts with the accelerometers and the Cosmed metabolic cart as described above. Each visit your child will receive a list of 8 activities from which to choose from and they will be instructed to perform each activity for two times; once for between 60-90 seconds and once for between 4-5 minutes. A trained research assistant will work with your child to ensure they perform the activities correctly.
- b. While performing the activities a trained research assistant will also video record your child so that exact transition times between activities can be obtained. This is an essential part of this project for development of better tracking methods in real-world situations.

RISKS

There are minimal risks associated with participation in the research study. Since the study involves children engaging in physical activities, there is the possibility of minor injuries such as cuts, bruises, and muscle soreness, and a slight risk of sprains. The risk is similar to that involved in a physical education class or a community sport program. There is also a risk of skin irritation from wearing the sticky pad on the thigh, which will be minimized by cleaning the area after it is removed. If irritation does occur during testing it will be removed immediately. Additionally, there is also a potential for social and psychological risks associated with wearing the Cosmed K4b² and being followed during the activities. Lastly, there is a risk of the identity of your child being known through the use of the video recording.

BENEFITS

The benefits of the proposed research include improved methods for measuring free-living physical activity. This has implications for intervention researchers and those looking at the health of children and adolescents. Participants will learn about their general health characteristics (e.g. BMI) and strategies to become more active.

CONFIDENTIALITY

Information and records included in this study will be kept confidential. Data will be stored in a secure location and will only be made available to the people conducting the study, unless you specifically give permission in writing to do otherwise. The video recordings will be kept on a secure password protected server and only investigators involved with this study will have access to these files. The results of the study will be published, but no reference will be made in oral or written reports that could link participants to the study.

The data files will also be sent to study investigators at the University of Massachusetts Boston and the University of California Riverside. Only non-identifiable data will be sent to these investigators and they will not receive the videos of your child, thus they would not be able to identify who your child is. The Office of Human Research Protections (OHRP) may also review the study records at any time.

COMPENSATION

For participation in the study, your child will receive a \$30 gift card for completion of visit 1 and a \$45 gift card for completion of visit 2, for a total of \$75 in gift cards if both days are completed.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/16/2018
IRB EXPIRATION DATE: 09/13/2019

_____ Participant's Initials

EMERGENCY MEDICAL TREATMENT

The University of Tennessee does not "automatically" reimburse subjects for medical claims or other compensation. If physical injury is suffered in the course of research, or for more information, please notify the investigator in charge, Dr. Scott Crouter, at (865) 974-1272.

CONTACT INFORMATION

If you have questions at any time about the study or the procedures (or you experience adverse effects as a result of participating in this study), you should immediately contact the principal investigator, Dr. Scott Crouter, 334 HPER Building, The University of Tennessee, Knoxville, TN 37996, (865) 974-1272. If you have questions concerning your rights as a participant, contact the Office of Research Compliance Officer at (865) 974-7697.

PARTICIPATION

Your participation in this study is voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

STATEMENT OF CONSENT

I have read the above information, and I have received a copy of this form. I consent for my child to participate in this study. I also certify that I am 18 years of age or older.

Printed name of your child _____

Parent or Legal Guardian signature _____ Date _____

Investigator's signature _____ Date _____

STATEMENT OF VIDEO CONSENT

I consent for my child to be videotaped while completing the activity described above. Not signing, does not exclude your child from the study and they can still participate, however they will not be videotaped.

Parent or Legal Guardian signature _____ Date _____

I consent for my child's videotape to be used for training purposes in the future. This might include use for demonstration in an undergraduate or graduate kinesiology course or training for future studies where investigators are using the same procedures. By not signing below, your child can still be videotaped for use in the current study, but their file will not be used for future training or educational purposes.

Parent or Legal Guardian signature _____ Date _____

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/16/2018
IRB EXPIRATION DATE: 09/13/2019

Youth Assent Form

Children's Assent form

Title of Research: Novel Approaches for Predicting Unstructured Short Periods of Physical Activities in Youth: Study 1 – Development of New Techniques

Principal Investigator: Dr. Scott Crouter

Location: Applied Physiology Laboratory, 1914 Andy Holt Ave., University of Tennessee, Knoxville, TN 37996

PURPOSE

You are being asked to participate in a research project that will use small devices called accelerometers to measure how much you move. These devices are attached to a belt and can be worn around the waist, hip, or ankle. The study will take place at the Applied Physiology Laboratory in the Health, Physical Education, and Recreation (HPER) building at The University of Tennessee. The lead researcher is Scott Crouter, PhD, Assistant Professor in the Department of Kinesiology, Recreation, and Sport Studies. Please read this form and feel free to ask questions. If you have further questions, Dr. Crouter or a research assistant will discuss them with you. You can contact him by telephone (617-287-7509) or email (scrouter@utk.edu).

PARTICIPANT'S INVOLVEMENT IN THIS STUDY

If you decide that you want to be in this study, this is what you will be asked to do:

- 1) Visit the Applied Physiology Laboratory in the HPER building on 2 different days for approximately 1.5 hours on visit 1 and 1 hour on visit 2.
- 2) Have your height, weight, and percent body fat measured, which requires you to stand barefoot on a scale.
 - a. During your 2 visits you will be asked to do the following tests:
 - b. Resting metabolic rate (visit 1): During this test you will be asked to rest quietly while lying on a table. During the test you will have 3 small devices placed on your waist, 3 on each wrist, 2 on each ankle, and 1 on each shoe using an elastic belt. Another device will also be attached to each thigh using a sticky pad. You will also have devices placed on your shoes and You will also be asked to wear a mask over your nose and mouth which will be connected to a device that is worn on your chest. The mask over your face will not affect how you breathe and you will be able to talk. This test will take 35 minutes to complete.
- 3) Perform 16 different activities (visits 1 and 2): During each visit 8 activities from the list below will be performed:
 - 1) walk slowly, 2) walk quickly, 3) run as fast or as slow as you want, 4) walk up and down stairs, 5) play basketball, 6) play catch with a ball, 7) playing soccer, 8) sit in a reclined position, 9) sit in a chair and read, 10) sit in a chair and play on the internet, 11) sit in a chair and play computer games, 12) lay on your back, 13) indoor cycling, 14) jumping jacks, 15) sweeping a floor, and 16) dusting a table
 - a. While performing the activities you will be asked to wear the elastic belts facemask as you will do for the resting metabolic measurement test. You will do each activity two for either 60-90 seconds or 4-5 minutes. A research assistant will work with you to perform each activity.
 - b. While doing the activities a research assistant will also video record what you are doing.

CAN ANYTHING BAD HAPPEN TO ME?

Although there are very few risks to taking part in this study, you could get hurt just like with any type of physical activity such as what you do in PE class. Dr. Crouter and the other research team members will help you to avoid being hurt or upset. There is also the risk that your skin will become itchy while wearing the sticky pad on your thigh.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/16/2018
IRB EXPIRATION DATE: 09/13/2019

_____ **Participant's Initials**

CAN ANYTHING GOOD HAPPEN TO ME?

By participating in this study, you will learn about other physical activities that will be good for your health while also having fun.

DO I HAVE OTHER CHOICES?

You can choose not to be in this study if you don't want to.

WILL ANYONE KNOW I AM IN THE STUDY?

We won't tell anyone you took part in this study. Other people who are in the Applied Physiology Laboratory will know that you are being tested but we will not tell them any of your personal information. When we are done with the study, we will write a report about what we found out. We won't use your name in the report.

WHAT WILL YOU RECEIVE?

You will receive a \$30 gift card for completing visit 1 and a \$45 gift card for completing visit 2. If you complete both days of testing you will get a total of \$75 in gift cards.

WHAT IF I DO NOT WANT TO DO THIS?

You don't have to be in this study. It's up to you. If you say yes now, but you change your mind later, that's okay too. All you have to do is tell us.

Before you tell us your decision about being a part of this study; be sure to ask Dr. Crouter or one of his research assistants to tell you more about anything you don't understand.

STATEMENT OF CONSENT

I have read the above information, and I have received a copy of this form. By signing this form I give assent to participate in this study.

Participant's signature _____ Date _____

Investigator's signature _____ Date _____

STATEMENT OF VIDEO CONSENT

I give assent to be videotaped while completing the activity described above. Not signing, does not exclude you from the study and you can still participate, however you will not be videotaped.

Participant's signature _____ Date _____

I give assent for my videotape to be used for training purposes in the future. This might include use for demonstration in an undergraduate or graduate kinesiology course or training for future studies where investigators are using the same procedures. By not signing below, you can still be videotaped for use in the current study, but there file will not be used for future training or educational purposes.

Participant's signature _____ Date _____

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/16/2018
IRB EXPIRATION DATE: 09/13/2019

Recruitment Flyer

Seeking Research Participants

Study Details

- » Perform common daily activities, including household chores and sports, while wearing activity monitors
- » Three visits (~6 hours total)
- » Compensation up to \$75
- » Excellent opportunity for research and technology exposure!



Who can Participate?

- » Boys and girls ages 6-18 years old
- » Active or non-active youth eligible
- » All testing will take place at participants home and after school programs



**Are you Interested?
Contact us below**

Email: APLStudy@utk.edu

Phone: (865) 974-5651

Parent Informed Consent Form

Informed Consent Statement for Parent or Guardian

Title of Research: Novel Approaches for Predicting Unstructured Short Periods of Physical Activities in Youth: Study 2 – Validation of New Techniques

Principal Investigator: Dr. Scott Crouter

Location: Applied Physiology Laboratory, 1914 Andy Holt Ave., University of Tennessee, Knoxville, TN 37996

PURPOSE

You are being asked to give permission for your child to participate in a research project that will examine new methods to measure the amount of physical activity youth obtain using small devices called accelerometers. These devices are attached to a belt which is worn around the waist, wrist, or ankle similar to a pedometer. Your child will be eligible if he or she is: 1) 6 to 18 years of age, and 2) you and your child can read and speak English. Your child will not be eligible if: 1) he or she has a serious health problem or injury that would prevent them from participating safely in regular physical activity, 2) he or she has a Pacemaker or other internal medical device.

The study will take place at the Applied Physiology Laboratory in the Health, Physical Education, and Recreation (HPER) building at The University of Tennessee. The lead researcher is Scott Crouter, PhD, Assistant Professor in the Department of Kinesiology, Recreation, and Sport Studies. Please read this form and feel free to ask questions. If you have further questions, Dr. Crouter will discuss them with you. You can contact him by telephone (617-287-7509) or email (scrouter@utk.edu).

PARTICIPANT'S INVOLVEMENT IN THIS STUDY

Here is what the study involves:

- 1) You will be asked to fill out a health history questionnaire pertaining to your child's health status to determine eligibility for the study.
- 2) Your child will be asked to be tested on 3 different days. Day 1 will take place at the Applied Physiology Laboratory in the HPER building (approximately 60 minute visit) and days 2 and 3 will take place afterschool at a local facility (e.g. YMCA, the school they attend), public park, or in their home. Days 2 and 3 will take approximately 2 hrs and 10 minutes to complete on each day. The three testing days will be completed over a 2 week period. If your child is enrolled in an afterschool program where we are testing, day 1 will be completed at the program facility. In addition, we may combine days 1 and 2 into a single testing day, if time allows.
 - a. **Day 1:** Your child will have his/her height, weight, percent body fat and resting metabolic rate measured.
 - i. The percent body fat will be measured using bio-electrical impedance (BIA). BIA is a non-invasive technique that involves standing on a scale while also holding on to a handlebar. During this time an undetectable and harmless electrical current is passed through the body.
 - ii. The resting metabolic rate test will take 35 minutes. This test will involve having your child rest in a lying position for 25 minutes and it will take 10 minutes to connect the equipment and prepare for the test. During this time he/she will have 3 accelerometers placed on their waist, 3 on each wrist, 2 on each ankle, and 1 on each shoe using an elastic belt. Another device will also be attached to each thigh using a sticky pad. Your child will also be attached to a Cosmed portable system, which will measure the air that they breathe. This device is light weight and straps onto the body. It also requires the use of a facemask, over the nose and mouth. While wearing the facemask your child will be able to breathe normally and talk, and the mask does not block or restrict airflow.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/26/2019
IRB EXPIRATION DATE: 09/12/2020

- b. **Days 2 and 3:** Your child will be asked to wear the Cosmed portable system and belt with the accelerometers for a 2-hour period. On one of the days this testing will take place at an afterschool facility and on the other day it will take place in their home. If testing in the home or an afterschool facility is not possible, alternative testing sites may be found (e.g. public park). The home visit will be scheduled in advance so that a parent or guardian can be there. During this time he/she will be free to choose the activities they perform using what is available at the testing location. During all times a trained research assistant will be with your child. While performing the activities a trained research assistant will also video record your child so that exact transition times between activities can be obtained. This is an essential part of this project for development of better tracking methods in real-world situations.

RISKS

There are minimal risks associated with participation in the research study. Since the study involves children engaging in physical activities, there is the possibility of minor injuries such as cuts, bruises, and muscle soreness, and a slight risk of sprains. The risk is similar to that involved in a physical education class or a community sport program. There is also a risk of skin irritation from wearing the sticky pad on the thigh, which will be minimized by cleaning the area after it is removed. If irritation does occur during testing it will be removed immediately. Additionally, there is also a potential for social and psychological risks associated with wearing the Cosmed K4b2 and being followed during the activities. Lastly, there is a risk of the identity of your child being known through the use of the video recording.

BENEFITS

The benefits of the proposed research include improved methods for measuring free-living physical activity. This has implications for intervention researchers and those looking at the health of children and adolescents. Participants will learn about their general health characteristics (e.g. BMI) and strategies to become more active.

CONFIDENTIALITY

Information and records included in this study will be kept confidential. Data will be stored in a secure location and will only be made available to the people conducting the study, unless you specifically give permission in writing to do otherwise. The video recordings will be kept on a secure password protected server and only investigators involved with this study will have access to these files. The results of the study will be published, but no reference will be made in oral or written reports that could link participants to the study.

The data files will also be sent to study investigators at the University of Massachusetts Boston and the University of California Riverside. Only non-identifiable data will be sent to these investigators and they will not receive the videos of your child, thus they would not be able to identify who your child is. The Office of Human Research Protections (OHRP) may also review the study records at any time.

WHAT WILL HAPPEN TO INFORMATION AFTER THE STUDY IS OVER?

We will keep your child's information to use for future research ideas and training for future studies. Your child's name and other information that can directly identify them will be kept secure and stored separately from their research data collected as part of the study

We may share your child's research data with other researchers without asking for your permission again, but it will not contain information that could directly identify your child. This sharing of data may include being deposited into a data repository as required by publishers for publication or public repository.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/26/2019
IRB EXPIRATION DATE: 09/12/2020

COMPENSATION

For participation in the study, your child will receive a \$25 gift card for completing visit one, a \$25 gift card for completing visit 2, and a \$25 gift card for completing visit 3; for a total of \$75 in gift cards if all 3 days are completed. If visit 1 and 2 testing is completed on the same day then your child will receive \$50 for the visit and \$25 for the second visit.

EMERGENCY MEDICAL TREATMENT

The University of Tennessee does not "automatically" reimburse subjects for medical claims or other compensation. If physical injury is suffered in the course of research, or for more information, please notify the investigator in charge, Dr. Scott Crouter, at (865) 974-1272.

CONTACT INFORMATION

If you have questions at any time about the study or the procedures (or you experience adverse effects as a result of participating in this study), you should immediately contact the principal investigator, Dr. Scott Crouter, 334 HPER Building, The University of Tennessee, Knoxville, TN 37996, (865) 974-1272. If you have questions concerning your rights as a participant, contact the Office of Research Compliance Officer at (865) 974-7697.

PARTICIPATION

Your participation in this study is voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

STATEMENT OF CONSENT

I have read the above information, and I have received a copy of this form. I consent for my child to participate in this study. I also certify that I am 18 years of age or older.

Printed name of your child _____

Parent or Legal Guardian signature _____ Date _____

Investigator's signature _____ Date _____

STATEMENT OF VIDEO CONSENT

I consent for my child to be videotaped while completing the activity described above. Not signing, does not exclude your child from the study and they can still participate, however they will not be videotaped.

Parent or Legal Guardian signature _____ Date _____

I consent for my child's videotape to be used for training purposes in the future. This might include use for demonstration in an undergraduate or graduate kinesiology course or training for future studies where investigators are using the same procedures. By not signing below, your child can still be videotaped for use in the current study, but their file will not be used for future training or educational purposes.

Parent or Legal Guardian signature _____ Date _____

IRB NUMBER: UTK-IRB-15-02487-FB
IRB APPROVAL DATE: 08/26/2019
IRB EXPIRATION DATE: 09/12/2020

Youth Assent Form

Children's Assent form

Title of Research: Novel Approaches for Predicting Unstructured Short Periods of Physical Activities in Youth: Study 2 – Development of New Techniques

Principal Investigator: Dr. Scott Crouter

Location: Applied Physiology Laboratory, 1914 Andy Holt Ave., University of Tennessee, Knoxville, TN 37996

PURPOSE

You are being asked to participate in a research project that will use small devices called accelerometers to measure how much you move. These devices are attached to a belt and can be worn around the waist, hip, or ankle. The study will take place at the Applied Physiology Laboratory in the Health, Physical Education, and Recreation (HPER) building at The University of Tennessee. The lead researcher is Scott Crouter, PhD, Assistant Professor in the Department of Kinesiology, Recreation, and Sport Studies. Please read this form and feel free to ask questions. If you have further questions, Dr. Crouter or a research assistant will discuss them with you. You can contact him by telephone (617-287-7509) or email (scrouter@utk.edu).

PARTICIPANT'S INVOLVEMENT IN THIS STUDY

If you decide that you want to be in this study, this is what you will be asked to do:

- 1) Complete testing on 3 different days. Day 1 will take place at the Applied Physiology Laboratory in the HPER building (approximately 1 hour visit) and days 2 and 3 will take place afterschool at a local facility (e.g. YMCA, the school they attend), public park, or your home. Days 2 and 3 will take approximately 2 hrs to complete. If you are enrolled in an afterschool program that we are testing at we will complete day 1 at the program facility and may combine days 1 and 2 into one day of testing.
 - a. **Day 1:** You will have your height, weight, percent body fat and resting metabolic rate measured.
 - i. The percent body fat measurement will require you to stand barefoot on a scale.
 - ii. For the resting metabolic rate test you will be asked to rest quietly while lying on a table. During the test you will have 3 small devices placed on your waist, 3 on each wrist, 2 on each ankle, and 1 on each shoe using an elastic belt. Another device will also be attached to each thigh using a sticky pad. There will also be a device attached to a pair of eyeglasses that you will wear. You will also be asked to wear a mask over your nose and mouth which will be connected to a device that is worn on your chest. The mask over your face will not affect how you breathe and you will be able to talk. This test will take 35 minutes to complete.
 - b. **Days 2 and 3:** You will be asked to wear the facemask and elastic belt as described above for 2 hours on each day. During this time you will be able to freely choose what activities you do based on what is available. On one of the days this testing will take place at an afterschool facility and on the other day it will take place in their home. If this is not possible, alternative testing sites might be found (e.g. public park). You will have a research assistant with you at all times to ensure your safety. While performing the activities a research assistant will also video record what you are doing.

CAN ANYTHING BAD HAPPEN TO ME?

Although there are very few risks to taking part in this study, you could get hurt just like with any type of physical activity such as what you do in PE class. Dr. Crouter and the other research team members will help you to avoid being hurt or upset. There is also the risk that your skin will become itchy while wearing the sticky pad on your thigh.

CAN ANYTHING GOOD HAPPEN TO ME?

By participating in this study, you will learn about other physical activities that will be good for your health while also having fun.

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 10/25/2019
IRB EXPIRATION DATE: 09/12/2020

_____ **Participant's Initials**

DO I HAVE OTHER CHOICES?

You can choose not to be in this study if you don't want to.

WILL ANYONE KNOW I AM IN THE STUDY?

We won't tell anyone you took part in this study. Other people who are in the Applied Physiology Laboratory will know that you are being tested but we will not tell them any of your personal information. When we are done with the study, we will write a report about what we found out. We won't use your name in the report.

WHAT WILL HAPPEN TO MY INFORMATION AFTER THE STUDY IS OVER?

We will keep your information to use for future research ideas and training. Your name and other information that can directly identify you will be kept in a safe location separate from your research data.

We may share your research data with other researchers without asking for your permission again, but we will not give them your name or any information that would allow them to identify who you are. Your research data might also be placed into a data repository when we publish the report.

WHAT WILL YOU RECEIVE?

You will receive a \$25 gift card for completing visit 1, a \$25 gift card for completing visit 2, and a \$25 gift card for completing visit 3. If you complete all 3 days of testing you will get a total of \$75 in gift cards. If you do day 1 and 2 testing on the same day then you will receive \$50 for the visit and \$25 for the second visit.

WHAT IF I DO NOT WANT TO DO THIS?

You don't have to be in this study. It's up to you. If you say yes now, but you change your mind later, that's okay too. All you have to do is tell us.

Before you tell us your decision about being a part of this study; be sure to ask Dr. Crouter or one of his research assistants to tell you more about anything you don't understand.

STATEMENT OF CONSENT

I have read the above information, and I have received a copy of this form. By signing this form I give assent to participate in this study.

Participant's signature _____ Date _____

Investigator's signature _____ Date _____

STATEMENT OF VIDEO CONSENT

I give assent to be videotaped while completing the activity described above. Not signing, does not exclude you from the study and you can still participate, however you will not be videotaped.

Participant's signature _____ Date _____

I give assent for my videotape to be used for training purposes in the future. This might include use for demonstration in an undergraduate or graduate kinesiology course or training for future studies where investigators are using the same procedures. By not signing below, you can still be videotaped for use in the current study, but there file will not be used for future training or educational purposes.

Participant's signature _____ Date _____

IRB NUMBER: UTK IRB-15-02487-FB
IRB APPROVAL DATE: 08/26/2019
IRB EXPIRATION DATE: 09/12/2020

VITA

Paul Hibbing was born in Ames, IA, where he lived on the property of Iowa State University's animal science teaching farm until just before his graduation from Ames High School. After moving across town and graduating, he enrolled at Iowa State with no intentions to speak of. Halfway through a bachelor's degree in Kinesiology and Health, he was advised that he would be a good fit for graduate education, which sparked his interest in pursuing an academic career. After completing his bachelor's degree, Paul stayed at Iowa State to complete a master's degree in Kinesiology. Along the way, he developed an interest in physical activity measurement research, which ultimately connected him with Dr. Scott Crouter, who eventually became his Ph.D. advisor at the University of Tennessee, Knoxville. In rapid succession, Paul married Sarah Corfman, finished his master's degree, and moved to Knoxville, where he and Sarah had two children during the four years it took to complete the degree. Paul accepted a postdoctoral fellowship from Children's Mercy Hospital in Kansas City, Missouri, where he conducted research in the Center for Children's Healthy Lifestyles and Nutrition.