ANALYSIS OF CHANGES IN SENSOR DATA
FOR MOBILITY ASSESSMENT

By
GINA LEE SPRINT

A dissertation submitted in partial fulfillment of
the requirements for the degree of
DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science
MAY 2016

© Copyright by GINA LEE SPRINT, 2016
All Rights Reserved
To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of GINA LEE SPRINT find it satisfactory and recommend that it be accepted.

________________________
Diane J. Cook, Ph.D., Chair

________________________
Hassan Ghasemzadeh, Ph.D.

________________________
Maureen Schmitter-Edgecombe, Ph.D.

________________________
Douglas Weeks, Ph.D.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Diane J. Cook, for her support, guidance, and encouragement. I would not be where I am today without her. I would like to thank my committee members, Dr. Hassan Ghasemzadeh, Dr. Maureen Schmitter-Edgecombe, and Dr. Douglas Weeks for their valuable feedback. I would like to thank my collaborators Vladimir Borisov, Dr. Douglas Weeks, and Prafulla Dawadi for their expertise and contributions to this work. In addition, I would like to thank our physical therapy collaborators at St. Luke’s Rehabilitation Institute for their clinical insights and suggestions. Next, I would like to thank Jordana Dahmen, Shirin Shahsavand, Catherine Sumida, and Thao Vo for their help in data collection. I would like to thank the members of the CASAS group for their research support and friendship. I would like to thank Michael Turi, Daniel Iparraguire, Steve Wang, and Nick Zhang for their camaraderie in my early years at WSU. Finally, I would like to recognize those who encouraged me to take breaks and allocate time for friends, family, and fun: Jaclyn Sprint, Iryna Malova, Prafulla Dawadi, Jess Dahmen, Anthony La, Bryan Minor, Elizabeth Edwards, Jason Fairey, Chris Cain, and Yuchao Ma.
ANALYSIS OF CHANGES IN SENSOR DATA
FOR MOBILITY ASSESSMENT

Abstract

by Gina Lee Sprint, Ph.D.
Washington State University
May 2016

Chair: Diane J. Cook

Age, injury, or disease-related impairments can severely diminish one’s ability to be mobile and perform everyday tasks. To detect changes in abilities, healthcare professionals administer standardized assessments. Sensor technology can be utilized to complement clinical assessments to gain a more objective and detailed view of functionality. Specifically for rehabilitation and everyday living situations, sensor technology is able to provide more information about mobility and reduce subjectivity in outcome measures. We hypothesize that data from sensors can be analyzed using machine learning techniques to provide insights on mobility changes related to rehabilitation and daily behavior. We validate this hypothesis by analyzing data from three diverse settings: wearable sensor data collected during rehabilitation, wearable sensor data collected during everyday living, and ambient sensor data collected from everyday living in smart home environments.

To investigate mobility assessment for rehabilitation, we analyze wearable sensor data collected from rehabilitation patients as they perform a sequence of ambulatory tasks that closely resemble everyday activities. We present algorithms to process sensor signals, compute metrics that describe ambulation performance, and quantify changes in mobility
over one week of rehabilitation. Furthermore, we train machine learning algorithms with
sensor-derived features and clinical information as part of our Hybrid Clinical Sensor
Prediction (HCSP) approach. We are able to achieve higher prediction accuracy with
HCSP by including our wearable sensor-derived features.

For analyzing changes in everyday mobility data, we formalize the problem of un-
supervised physical activity change detection and address the problem with our Physical
Activity Change Detection (PACD) approach. PACD is a framework that detects changes
between time periods, determines significance of the detected changes, and analyzes the
nature of the changes. We illustrate and evaluate PACD with physical activity data col-
lected from older adults who participated in a health intervention study and data collected
from ambient sensors embedded in smart home environments. Results indicate PACD de-
tects several mobility changes in the datasets. The proposed algorithms and analysis
methods are useful data mining techniques for unsupervised change detection with poten-
tial to track physical activity, detect behavior changes due to health events, and motivate
progress toward health goals.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Mobility Assessment for Rehabilitation</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Mobility Assessment for Everyday Living</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Dissertation Contributions</td>
<td>7</td>
</tr>
<tr>
<td>2. BACKGROUND MATERIAL</td>
<td>10</td>
</tr>
<tr>
<td>2.1 Inertial Measurement Units</td>
<td>11</td>
</tr>
<tr>
<td>2.2 The Timed Up and Go</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Activity-based Change Detection</td>
<td>22</td>
</tr>
<tr>
<td>3. WEARABLE SENSOR-BASED MOBILITY ANALYSIS</td>
<td>25</td>
</tr>
<tr>
<td>3.1 Functional Independence Measure</td>
<td>26</td>
</tr>
<tr>
<td>3.2 Participants</td>
<td>28</td>
</tr>
<tr>
<td>3.3 The Ambulation Circuit</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Sensor Data Processing</td>
<td>34</td>
</tr>
<tr>
<td>3.5 Computed Metrics</td>
<td>41</td>
</tr>
<tr>
<td>3.6 Assistive Device Analysis</td>
<td>45</td>
</tr>
<tr>
<td>3.7 Study Limitations</td>
<td>52</td>
</tr>
<tr>
<td>4. MEASURING GAIT CHANGES</td>
<td>54</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>4.1</td>
<td>Prior Work on Sensor-based Gait Assessment</td>
</tr>
<tr>
<td>4.2</td>
<td>Quantifying Group Changes</td>
</tr>
<tr>
<td>4.3</td>
<td>Individual-level Comparisons</td>
</tr>
<tr>
<td>4.4</td>
<td>Results</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
</tr>
<tr>
<td>5.1</td>
<td>Prior Work on Clinical Outcome Prediction</td>
</tr>
<tr>
<td>5.2</td>
<td>Additional Medical Record Data</td>
</tr>
<tr>
<td>5.3</td>
<td>Predictor Variables</td>
</tr>
<tr>
<td>5.4</td>
<td>Hybrid Clinical Sensor Prediction</td>
</tr>
<tr>
<td>5.5</td>
<td>Joint Patient Prediction</td>
</tr>
<tr>
<td>5.6</td>
<td>Results</td>
</tr>
<tr>
<td>5.7</td>
<td>Discussion</td>
</tr>
<tr>
<td>6.1</td>
<td>Prior Work on Collecting Physical Therapist Feedback</td>
</tr>
<tr>
<td>6.2</td>
<td>Interview Design</td>
</tr>
<tr>
<td>6.3</td>
<td>Results</td>
</tr>
<tr>
<td>6.4</td>
<td>Discussion</td>
</tr>
<tr>
<td>7.1</td>
<td>Prior Work on Sensor-based Physical Activity Monitoring</td>
</tr>
<tr>
<td>7.2</td>
<td>Window-based Comparisons</td>
</tr>
<tr>
<td>7.3</td>
<td>Physical Activity Datasets</td>
</tr>
<tr>
<td>7.4</td>
<td>Results</td>
</tr>
<tr>
<td>7.5</td>
<td>Discussion</td>
</tr>
<tr>
<td>8.1</td>
<td>Prior Work on Activity Analysis for Health Assessment</td>
</tr>
<tr>
<td>8.2</td>
<td>Smart Home Environments</td>
</tr>
<tr>
<td>8.3</td>
<td>Smart Home Resident Health Events</td>
</tr>
<tr>
<td>8.4</td>
<td>Results</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8.5 Discussion</td>
<td>190</td>
</tr>
<tr>
<td>9. CONCLUSIONS</td>
<td>201</td>
</tr>
<tr>
<td>9.1 Summary and Conclusions</td>
<td>201</td>
</tr>
<tr>
<td>9.2 Future Work</td>
<td>203</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>204</td>
</tr>
<tr>
<td>A Mini-Cog Examination</td>
<td>204</td>
</tr>
<tr>
<td>B Ambulation Study Questionnaire</td>
<td>205</td>
</tr>
<tr>
<td>C Acronyms</td>
<td>207</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>212</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Wearable technology benefits and limitations.</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>FIM tasks.</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>FIM scoring.</td>
<td>28</td>
</tr>
<tr>
<td>3.3</td>
<td>Participant characteristics.</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td>Clinical assessments of progress metrics.</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Whole body movement metrics.</td>
<td>46</td>
</tr>
<tr>
<td>3.6</td>
<td>Gait features.</td>
<td>47</td>
</tr>
<tr>
<td>4.1</td>
<td>CAP metric results.</td>
<td>63</td>
</tr>
<tr>
<td>4.2</td>
<td>WBM metric results.</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>GF metric results.</td>
<td>65</td>
</tr>
<tr>
<td>4.4</td>
<td>Individual limb GF metric results.</td>
<td>65</td>
</tr>
<tr>
<td>4.5</td>
<td>Cane gait parameter results.</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>Additional medical record data.</td>
<td>80</td>
</tr>
<tr>
<td>5.2</td>
<td>Admission features.</td>
<td>83</td>
</tr>
<tr>
<td>5.3</td>
<td>Predictor and FIM correlations.</td>
<td>102</td>
</tr>
<tr>
<td>5.4</td>
<td>Top ranked features.</td>
<td>103</td>
</tr>
<tr>
<td>5.5</td>
<td>FIM motor prediction results.</td>
<td>105</td>
</tr>
<tr>
<td>5.6</td>
<td>SMOTE FIM motor prediction results.</td>
<td>106</td>
</tr>
<tr>
<td>5.7</td>
<td>JPP FIM motor prediction results.</td>
<td>110</td>
</tr>
<tr>
<td>5.8</td>
<td>FIM cognitive prediction results.</td>
<td>112</td>
</tr>
<tr>
<td>6.1</td>
<td>Therapist-ratings for metric usefulness.</td>
<td>135</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Instrumenting the Timed Up and Go</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Ambulation circuit study timeline</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>The ambulation circuit</td>
<td>31</td>
</tr>
<tr>
<td>3.3</td>
<td>The St. Luke’s Rehabilitation Institute Community</td>
<td>32</td>
</tr>
<tr>
<td>3.4</td>
<td>Sensor placement and axes orientation</td>
<td>35</td>
</tr>
<tr>
<td>3.5</td>
<td>Sensor data processing</td>
<td>36</td>
</tr>
<tr>
<td>3.6</td>
<td>Example AC sensor signals</td>
<td>41</td>
</tr>
<tr>
<td>3.7</td>
<td>Gait cycles</td>
<td>44</td>
</tr>
<tr>
<td>3.8</td>
<td>Stance and swing phase plot for #015 S2</td>
<td>51</td>
</tr>
<tr>
<td>3.9</td>
<td>Stance and swing phase plot for #005 S2</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>Example reliable change index plot</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td>Metric box plots</td>
<td>61</td>
</tr>
<tr>
<td>4.3</td>
<td>Metric reliable change index plots</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>Stance and swing phase plot for #005 S1</td>
<td>63</td>
</tr>
<tr>
<td>5.1</td>
<td>Participant FIM scores</td>
<td>82</td>
</tr>
<tr>
<td>5.2</td>
<td>Cumulative model construction</td>
<td>88</td>
</tr>
<tr>
<td>5.3</td>
<td>Separate model construction</td>
<td>88</td>
</tr>
<tr>
<td>5.4</td>
<td>Example SMOTE relevance function</td>
<td>90</td>
</tr>
<tr>
<td>5.5</td>
<td>Patient similarity</td>
<td>95</td>
</tr>
<tr>
<td>5.6</td>
<td>FIM motor prediction results</td>
<td>107</td>
</tr>
<tr>
<td>5.7</td>
<td>FIM motor SMOTE histograms</td>
<td>108</td>
</tr>
</tbody>
</table>
5.8 FIM task prediction correlations ........................................ 113
5.9 Individual participant FIM prediction results ....................... 119
6.1 Task duration bar plot. ............................................. 129
6.2 Gait cycle bar plots. ................................................. 131
6.3 Effect size forest plot. ............................................. 133
7.1 PACD overview. ................................................... 148
7.2 Grey level co-occurrence matrix example. ......................... 152
7.3 Texture comparison between two windows. ......................... 153
7.4 Sliding window change scores. .................................. 158
7.5 HS0 activity density map. ......................................... 166
7.6 HS1 activity density map. ......................................... 166
7.7 HS2 activity density map. ......................................... 167
7.8 HS3 activity density map. ......................................... 167
7.9 HS4 activity density map. ......................................... 168
7.10 Hybrid-synthetic normalized change scores. .................... 172
7.11 Hybrid-synthetic decision trees for HS1-4. ...................... 173
7.12 Health intervention study decision trees ......................... 175
8.1 Floorplans and sensor layouts. .................................. 186
8.2 SH1 change results. ............................................... 191
8.3 SH2 change results. ............................................... 192
8.4 SH3 change results. ............................................... 193
8.5 Smart home decision trees. ..................................... 194
8.6 SH1 activity density maps. ...................................... 195
8.7 SH2 activity density maps. ...................................... 196
8.8 SH3 activity density maps. .................................................. 197
Dedication

This work is dedicated to my parents, Cathy and Gary, and to my sister, Jaclyn.
In recent years, there has been an increase in life expectancy which has resulted in a global aging of the population. In 2050 there will be an estimated 88.5 million individuals aged 65 and older in the United States alone, a 120% increase over the elderly population in 2010 [1]. This growing older population is placing a heavy burden on our healthcare systems. According to the Association of American Medical Colleges, the increasing demand for healthcare will cause a shortage of 124,400 physicians by 2025 [2]. The future of healthcare availability and quality of services is uncertain. To meet these demands, healthcare should utilize technology in order to scale to the needs of a larger population with health concerns.

To address this problem, proposed technological solutions are being researched and developed for health applications. For example, to aid individuals in living safely and independently in their homes, researchers are designing ambient intelligent and smart environment technologies. Therapy and rehabilitation in the home is becoming more prevalent with inexpensive teleconferencing systems and networked gaming platforms like the Nintendo Wii U. Mobile applications for smartphones and tablets are being developed to assist individuals with cognitive impairments as they navigate their activities of daily living (ADLs). Wearable sensors are ubiquitously collecting health status data such as vital signs and physical activity levels in order to perform continuous health assessments.

The aforementioned technology examples provide evidence that sensors have become
pervasive in both the healthcare domain and in everyday living situations. Sensors are ambient in the environment, embedded in smartphones, and worn on the body. Data collected from sensors form a time series, where each sample of data is paired with an associated timestamp. This sensor-based time series data is valuable when monitoring human behavior to detect and analyze changes. Such analysis can be used to detect functional deficits, seasonal physical activity variations, and health events. Analyzing sensor-based time series data can also be used to monitor changes in mobility as a person recovers from an injury or makes progress toward a fitness goal. Recovering from an injury or making a significant lifestyle change often takes weeks or months of establishing new behavior patterns [3], which can be challenging to sustain. Automatically detecting and tracking mobility changes from sensor data can provide a valuable motivating and monitoring tool. Sensor data in particular is proving to be useful for mobility assessment because of the detailed physical activity information that sensors provide. Sensor-based mobility assessment has applications in several health-related areas, particularly in the following two application areas: inpatient rehabilitation and everyday living situations.

We hypothesize that sensor data can be analyzed using machine learning techniques to provide insights on mobility changes related to rehabilitation and daily behavior. We validate this hypothesis by analyzing data from three diverse settings: wearable sensor data collected during inpatient rehabilitation, wearable sensor data collected during everyday living, and ambient sensor data collected from everyday living in smart home environments.
1.1 Mobility Assessment for Rehabilitation

Interdisciplinary researchers are instrumenting inpatient rehabilitation with technology to gain a more detailed view of patient functionality. In the rehabilitation settings, technology is able to provide more details and insight about patient performance. For example, the extensively-researched Timed Up and Go (TUG) test has been widely used in the clinic and in the home to assess functionality for over 20 years [4]. The TUG test has recently been instrumented with technology in several studies [5], yielding promising results toward the future of technology for quantifying clinical assessments and assessing patient movements. A few benefits of the TUG technology implementations include additional performance parameters, generated reports, and the potential for self-administration. These benefits also represent necessary steps to address the healthcare crisis and achieve widespread adoption by clinicians and patients alike.

Typically, therapists provide clinical observations that are used to assess patient movement status and progress. More precise quantitative measurements of patient performance can be collected via pervasive technology, such as wearable inertial measurement units (IMUs). Wearable IMUs contain inertial sensors, such as accelerometers and gyroscopes, which can be used in addition to clinical observations to collect fine-grained movement data from patients as they undergo therapy. IMUs are an ideal technology for tracking patient movement because they are unobtrusive, relatively inexpensive, and can easily be attached to the body. Computing algorithms operating on inertial measurements can identify subtle performance changes during rehabilitation that are difficult to
observe. Furthermore, computations based on data collected from wearable IMU sensors can provide therapists with measures that are not subject to the inter-observer bias that is possible with subjective clinical judgments. By analyzing wearable sensor data we can provide insights into patient mobility, gait, and rehabilitation. These insights are valuable for validating physical therapy regimens, quantifying patient progress, and determining appropriate patient discharge status. Therapy-collected inertial sensor data offers several benefits in comparison to human observation alone:

- More accurate tracking of the time course of recovery.
- Reduction in the subjectivity of therapist observations between admission and discharge.
- More ecologically valid assessment of patient ability by avoiding the “test” situation which is often not representative of everyday functioning [6].

In this dissertation we present algorithmic approaches to analyze data collected from rehabilitation patients for changes in mobility. To do this, we collect IMU data from patients receiving inpatient rehabilitation services for impairments such as stroke and brain injury. We present algorithms to objectively characterize mobility performance based upon the collected inertial sensor data. Computational components of our system include algorithms to extract mobility metrics, quantify performance changes, train and test clinical assessment prediction models, and generate metric visualizations. To produce clinically-meaningful metrics, we introduce a standardized ambulation performance task, titled the ambulation circuit, which involves a range of gait and transfer tasks. While inpatient reha-
habilitation patients perform the ambulation circuit, they wear three inertial sensors for data collection. We fix the interval of time over which repeated measurements of ambulation circuit performance is assessed (7 days) in order to quantify changes in movement parameters over one week of rehabilitation. Our algorithms statistically analyze the sensor-based metrics to identify clinically significant changes in the repeated measures data.

In addition to being clinically relevant on their own, we hypothesize that wearable sensor-derived metrics also provide power for predicting future patient performance on clinical rating scales. While data mining techniques have been applied to medical records to predict clinical performance, augmenting medical records with sensor data (such as IMU data) represents a new direction of research. To validate our hypothesis, we propose the Hybrid Clinical Sensor Prediction (HCSP) approach to utilize patient medical record information, sensor data measured at mid-stay during inpatient rehabilitation, and a combination of the two data sources to predict Functional Independence Measure (FIM) motor scores at discharge [7]. The FIM is a clinical assessment measuring functional status for various ADLs. Our HCSP machine learning-based approach provides insight into individual patient progress between admission and discharge measured on the FIM scoring scale by using movement data collected during therapy tasks, without the need to re-administer the entire FIM assessment.

Highly useful quantitative information and visual presentations of wearable sensor data are critical in gaining therapist acceptance of the technology and improving the therapy experience for patients. To bridge the gap between design of mobility monitoring technology and actual use of the technology, we report responses from interviews conducted
with physical therapy providers at an inpatient rehabilitation facility. Data collected from the ambulation circuit wearable sensor study were presented to physical therapy providers to collect their perceived clinical utility of wearable sensor data for mobility assessment. We present and discuss the responses received from the therapists and suggest future computing research directions to potentially enhance therapy services and increase the adoption of wearable sensor systems for mobility assessment.

1.2 Mobility Assessment for Everyday Living

Recently, wearable fitness trackers have increased in popularity as people aspire to be more conscientious of their physical health. Many consumers purchase a pedometer or wearable fitness device in order to track their physical activity, often in pursuit of goals such as increasing cardiovascular strength, losing weight, or improving overall health. To track different types of changes in physical activity data, two or more time periods, or windows, of physical activity data can be quantitatively and objectively compared. If the two time windows contain significantly different sensor data then this may indicate a significant behavior change. Existing off-the-shelf change point detection methods are available to detect change in time series data, but the methods do not provide context or explanation regarding the detected change. For physical activity data, algorithmic approaches to change detection require additional information about what type of change is detected and its magnitude to potentially report progress to users for motivation and encouragement purposes. Furthermore, existing approaches often do not provide a method
for determining if a detected change is *significant*, meaning the magnitude of change is high enough to suspect it likely resulted from a lifestyle alteration. A personalized, data-driven approach to significance testing for fitness tracker users is a necessary feature of physical activity change detection.

We present algorithmic approaches to analyze changes in everyday physical activity data. In this dissertation we formalize the problem of unsupervised physical activity change detection and address the problem with our Physical Activity Change Detection (PACD) approach. PACD is a framework that 1) segments time series data into time periods, 2) detects changes between time periods, 3) determines significance of the detected changes, and 4) analyzes the nature of the significant changes. We evaluate the ability of four different change detection approaches to capture pattern changes in synthetic physical activity data. We also illustrate how the change approaches are used to monitor, quantify, and explain behavior differences in Fitbit data collected from older adults who participated in a health behavior intervention and ambient motion sensor data collected from smart home residents.

### 1.3 Dissertation Contributions

Our work focuses on analyzing mobility changes in two application areas: 1) rehabilitation and 2) everyday living. The major contributions of this dissertation include:

1. Design of a novel ambulation circuit in an indoor simulated community. The circuit consists of standardized gait tasks to measure patient performance in real-world
settings.

2. Introduction of a wearable sensor data processing and feature extraction architecture for mobility analysis. We present algorithms to map raw acceleration and angular velocity data into meaningful movement and gait parameters and visualizations.

3. Creation of a methodology for computing changes in gait and ambulation circuit performance for individual patients and for groups of patients.

4. Design of our HCSP computational approach to predict clinical outcomes by combining clinical information with wearable sensor-derived features and associated statistics quantifying changes. This includes the introduction of machine learning techniques to improve prediction accuracy and adjust the algorithms to be effective for low sample sizes.

5. Evaluation of the clinical utility of wearable sensor metrics, associated data visualizations, and sensor-based clinical outcome prediction. We interviewed seven physical therapy providers to gather their perceived usefulness of wearable technology for inpatient rehabilitation.

6. Introduction of a new physical activity change detection approach to perform unsupervised physical activity change detection. PACD is a framework that detects changes between time periods, determines significance of the detected changes, and analyzes the nature of the changes. We apply the PACD approach to synthetic and real-world acceleration-based datasets to detect mobility changes. We also use
PACD to detect physical activity and behavior changes indicative of health events from smart home sensor data.

In the remainder of this dissertation we describe the current state of change analysis for mobility assessment and our contribution to the field. We begin with a review of related work (Chapter 2). We then describe the ambulation circuit and introduce our approach to processing wearable data collected from the circuit (Chapter 3). Next, we introduce statistical methods to quantify patient performance changes (Chapter 4). We then describe our HCSP approach that utilizes wearable sensor metrics and change statistics to predict patient functional outcomes at discharge (Chapter 5). We culminate the ambulation circuit study with an evaluation of the clinical utility of wearable sensor data and visualizations of patient performance change for physical therapy providers (Chapter 6). To highlight our contributions to physical activity change detection, we address the unsupervised change detection problem with our proposed PACD approach and validate PACD in the context of data collected from wearables (Chapter 7). We also validate PACD in the context of smart homes by analyzing unobtrusively collected smart home sensor data to detect behavior changes indicative of major health events (Chapter 8). Finally, we discuss conclusions of the work and directions for future research investigating changes in sensor data for mobility assessment (Chapter 9).
In the past decade, technology has become more advanced and inexpensive than ever before. Recent research has focused on designing custom technology and adapting off-the-shelf solutions to healthcare applications. With the advent of wireless technologies such as Bluetooth, WiFi, and Zigbee, sensors have become quite popular for activity logging and healthcare applications. There is a large body of previous research related to the application of sensors for gait analysis, rehabilitation, and physical activity. These works include studies that measure device accuracy, that utilize different types of sensor devices, that validate methodologies for different patient populations, and that investigate solutions for the clinic, at home, or on the go (mobile) [8]. In this chapter, we first describe wearable sensor technology and its suitability for use in rehabilitation settings. We then take a closer look at how wearable technology has been applied to the Timed Up and Go test to gather more objective movement and gait measurements for various injuries and illnesses. This body of work instrumenting the TUG test with technology is the foundation for our experimental setup for analyzing mobility changes during inpatient rehabilitation (see Chapter 3 for details regarding the ambulation circuit study design). Finally, we discuss previous work related to detecting changes in physical activity data collected from wearable sensors and ambient sensors installed in the home.
2.1 Inertial Measurement Units

Inertial measurement units (IMUs) are movement tracking devices that contain accelerometers and gyroscopes. An accelerometer measures acceleration in meters/second$^2$ and a gyroscope measures angular velocity in degrees/second. Because of their low cost, portability, and reliability, IMUs have been utilized extensively in healthcare applications, [9–12]. IMUs are an ideal technology for tracking changes in movement because they can easily be attached to the body. Furthermore, IMUs operate as a self-contained wireless network which can enable testing outside the lab, for any sequence of tasks, without restricting the wearer’s movement. Table 2.1 provides a summary of the benefits and limitations of wearable IMU technology.

Several studies have used IMUs for analyzing gait and movement as an inexpensive and unobtrusive substitute to other technologies [9, 12, 13]. Performance on common clinical assessments, such as the Timed Up and Go test, have been characterized with IMUs (see Section 2.2). Similar to studies pairing wearable technology and the TUG test, we utilize wearable IMUs for mobility assessment. Consequently, research related to instrumenting the TUG test is highly relevant to the studies we present in this dissertation.

2.2 The Timed Up and Go

The Timed Up and Go test is a widely used method of evaluating basic mobility maneuvers [14]. It is based on the Get Up and Go (GUG) test that was originally
Table 2.1: Wearable inertial sensor technology benefits and limitations.

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small form factor.</td>
<td>Need to be routinely charged.</td>
</tr>
<tr>
<td>Comfortable attachment minimizes user aware-</td>
<td>Difficult to mount sensors on one’s own body.</td>
</tr>
<tr>
<td>ness.</td>
<td></td>
</tr>
<tr>
<td>Units contain several sensors (e.g. accelerometers, gyroscopes, etc.).</td>
<td>Need to be well-positioned and oriented.</td>
</tr>
<tr>
<td>Do not require skin surface contact.</td>
<td>Sensors are noisy and suffer from drift.</td>
</tr>
<tr>
<td>Inexpensive.</td>
<td>May require calibration.</td>
</tr>
<tr>
<td>Attachable anywhere on the body.</td>
<td>May be easily noticeable.</td>
</tr>
<tr>
<td>Portable; the testing space is not constrained.</td>
<td>May be uncomfortable or interfere with natural movement.</td>
</tr>
<tr>
<td>Wireless.</td>
<td>Low on-board computational processing power.</td>
</tr>
</tbody>
</table>

proposed by Mathias et al. [15] in 1986. The GUG test begins with the subject seated in an armchair. The subject rises from the chair, walks 3 meters in a linear path, performs a 180° turn, walks back to the chair, and sits down (see Figure 2.1). Typical instructions given to the subject are: “When I say ‘go’, I want you to stand up and walk to the line, turn, and then walk back to the chair and sit down again. Walk at your normal pace” [4]. GUG performance is subjectively evaluated by the observer on a five-point ordinal scale: normal, very slightly abnormal, mildly abnormal, moderately abnormal, and severely abnormal [15]. The TUG is a timed version of the GUG that attempts to address the subjectivity of the ordinal scale with the introduction of an objective measure, the total time to complete the task [4]. For the TUG, an examiner records the number of seconds it takes for the subject to perform the task using a stopwatch. Several clinical trials and research endeavors have discovered that this duration measure is representative of an individual’s ambulatory abilities, balance, and possibly risk of falling [14].

The TUG test has become one of the most popular functional assessments for several
Figure 2.1: Experimental setup for instrumenting the Timed Up and Go with inertial sensors and cameras (left) [16]. The red cross on the floor denotes the turnaround point. Shimmer inertial sensor (5.4 cm × 1.9 cm × 3.2 cm) with coordinate axes (right). ©2010 IEEE.
reasons. First off, the TUG tests several different mobility skills. These include sit-to-stand and stand-to-sit chair transitions, turning, straight-ahead gait, balance control, and the ability to sequence tasks [17–19]. The TUG requires minimal materials and setup. All that is required is a chair, 3 meters of walking space, and tape for marking the turnaround point. Furthermore, the TUG is simple to score, requiring minimal training and no expertise in mobility analysis. In the seminal TUG paper, Podsiadlo et al. [4] found the TUG to have good test-retest reliability, inter-rater reliability, and concurrent validity. More recently, Hafsteinsdottir et al. [20] and Rydwik et al. [21] reviewed TUG studies for analysis of test reliability, validity, and responsiveness.

It is evident that the TUG test is an important standardized test with several benefits, but the TUG is not without limitations:

- The duration measure is not always sensitive to falls risk in healthy older populations [22–26].
- With three highly different subtasks (chair transition, straight-ahead gait, and 180° turn) there is opportunity for various movement strategies. For example, the 180° turn introduces variability as people with different gait and balance impairments compensate differently when performing turns. The subject may turn on the spot or in a curve, as is often the case with the use of an assistive device such as a walker [27].
- Movement deficiencies exhibited on the complex subtasks are ignored. The effects of a new medication or therapy could go unnoticed when only analyzing the course-grained measurement of duration [11]. Three meters is not long enough to produce
high reliability and discriminate amongst healthy and Parkinson’s Disease (PD) populations [28].

- The TUG is fairly sensitive to subject and environmental conditions. For example, test-retest reliability is low when subjects wear different footwear [29]. A similar conclusion has been formed regarding the usage of assistive devices during the TUG test [30].

- The choice of chair can introduce variability. For example, if the chair has arms, then the arms can be used for assistance rising from or lowering to the chair [11]. For this reason, several studies opt to explicitly use armless chairs [4,31,32].

There are several limitations that are not specific to the TUG and are common amongst other clinical assessments, including variability in the provided instructions, subjectivity of examiners, and documentation differences [33]. It is known that performance in a lab setting does not fully represent the abilities of an individual [27], as it does not replicate ecological conditions [34]. People are more aware of the “test” situation in a laboratory or clinical setting and thus are more conscientious of their performance, often resulting in better performance [35].

Several variations of the TUG have been proposed to address the limitations of the standard TUG and to perform additional assessments. A second task has been added to the TUG, producing Timed Up and Go-Dual Tasks (TUG-DT) [36]. For several studies, the second task involves a cognitive component. Beauchet et al. [37] reported that gait parameters are affected by which cognitive task is chosen as the secondary task. The
countdown task requires the use of working memory [38], which consequently has the highest perturbation on gait parameters [37]. To address the limitations of the standard TUG and to perform additional assessments, researchers have proposed using IMUs to instrument the TUG test. In the following sections, technology-infused TUG assessment studies are grouped together based on the subject populations the technology was used to investigate.

2.2.1 Parkinson’s Disease

The largest body of research published on instrumenting the TUG test comes from Salarian and colleagues working on iTUG, the Instrumented Timed Up and Go [11,39–44]. The iTUG extended the standard 3 meter TUG length to 7 meters to allow for more gait cycles during the walking phase. To observe total body movement, several inertial sensors were mounted on the body: one bi-axial (2-dimensions) gyroscope on each forearm, one uni-axial (1 dimension) gyroscope on each shank, one uni-axial gyroscope on each thigh, and one bi-axial gyroscope and tri-axial (3 dimensions) accelerometer on the sternum.

The iTUG was broken down into four sections: sit-to-stand, steady-state gait, turning, and turn-to-sit. Each component was automatically detected in the sensor signals and has a set of parameters computed for each body part involved. For example, the $180^\circ$ turning phase consists of duration, trunk peak angular velocity, average step time, maximum step time, last step time before turn, and number of steps. The system was utilized to study 12 subjects with idiopathic, early-to-moderate stage Parkinson’s Disease.
and 12 healthy age-matched controls. Gait, turns, and turn-to-sit sections of the iTUG demonstrated significant differences between the two populations. Cadence was found to be the most reliable metric ($\rho = 0.94$) and other measures of gait exhibited high reliability as well. With the additional sensors on the arms that are not included in most technological TUG studies, the authors discovered range and amplitude of arm-swing to be sensitive to the early stages of PD, whereas the standard TUG total duration was not sensitive enough to pick up on the early PD changes. The research and success surrounding iTUG led to a commercial sensing company, APDM \cite{45, 46}. APDM has an extensive customer list, greatly contributing to the use of iTUG in the clinic and in research.

2.2.2 Falls Risk

The TUG test has widely been used to determine falls risk and to classify fallers and non-fallers. Narayanan and colleagues \cite{47–49} were one of the first to instrument the TUG test with inertial sensors for falls risk detection. They proposed a battery of common clinical assessments that could be self-administered and performed daily in the home. They called this set the “directed routine” and it consisted of the following five clinical assessments: 3 meter TUG test, Near-tandem Standing Balance, Alternative Step Test, Five Times Sit-to-stand Chair Transfer, and Simple Reaction Time. While performing these assessments, subjects wore a PreventaFall Ambulatory Monitor which contained a single tri-axial accelerometer on their waist. To start the routine, the subject pressed a button on the device and audio cues began guiding them. The collected acceleration
signals were uploaded each evening and processed on a remote server. Physicians could then access the data and monitor the status of subjects via a web interface.

The total TUG duration was computed by analyzing the medial-lateral acceleration signal. The acceleration signal was divided into each TUG component: time to stand, time to reach the 3 meter turnaround point, time to turn around, time to reach the chair, and time to sit down in the chair. The system was evaluated with 36 elderly participants. An estimation of falls risk with a linear least squares model yielded a root mean squared error of 0.69 (\( \rho = 0.58, p < 0.0002 \)) [48].

In another seminal paper, Greene et al. [50] proposed the qTUG, the Quantitative Timed Up and Go test, to compute falls risk. Two tri-axial IMUs were placed on the front of the midpoint of each shank (see Figure 2.1). Temporal gait parameters including cadence, number of gait cycles, stride time, swing time, stance time, step time, double support percent, and single support percent were computed from the inertial signals. Additional temporal parameters from the TUG components were calculated: TUG time, walk time, turn time, return time, walk-turn time ratio. Forty-four parameters from the sensors were computed in total, 29 of which were able to differentiate between fallers and non-fallers with statistical significance (\( p < 0.05 \)). Additionally, using the computed IMU metrics the authors were able to achieve a mean test accuracy of 76.8\% for retrospectively estimating falls risk. This method is reported as more accurate than the total TUG duration and Berg Balance Scale prediction. An additional paper discussing the reliability of qTUG for falls risk assessment was published by McGrath et al. [51].
2.2.3 Hemiplegia

For investigating Hemiplegic subjects, Higashi et al. [31] attached one IMU on the lower back and another IMU on the upper thigh of the leg that takes the first step when initiating gait. In total, 30 participants, 10 healthy and 20 hemiplegic, performed the TUG while wearing the sensors. Of the 20 hemiplegic participants, 10 had gait levels classified as independent and the other 10 were classified as supervised. The TUG tests were video recorded for scoring by a therapist at a later time. In addition to total TUG duration, therapists were instructed to record the following component times: standing-up, walking, turning, and sitting-down. These times were measured by analysis of the sensor signals and compared to the therapist times. A strong correlation ($r = 0.998$) was found between the two scoring mechanisms. In addition, metrics such as cadence, acceleration root mean square (RMS), and acceleration coefficient of variation (CV) were calculated for the gait component of the test. Using the RMS and CV values, the hemiplegic participants with independent gait were able to be distinguished from the supervised hemiplegic participants with statistical significance ($p < 0.01$).

2.2.4 Disability Levels

SankarPandi et al. [52] recently undertook a study to investigate the predictive utility of a single wrist-worn accelerometer for disability levels. The disability level was computed by summing 17 responses to questions regarding ADLs. If the participant and/or caregiver
stated they could perform the task in question, such as walking at least 400 yards, a one was scored, zero otherwise. Forty features extracted from the wrist acceleration signals were used to classify disability levels with a mean accuracy of 62.16%, which was higher than the 39.10% accuracy achieved by the total TUG duration alone. Sixteen features were representative enough of the population to discriminate all disability levels and different features were found to be more significant depending on the subject’s gender.

### 2.2.5 Cognitive Impairment

Gillain et al. [53] investigated an IMU-based TUG on populations of healthy controls, participants with mild cognitive impairment (MCI), and participants with Alzheimer’s Disease (AD). A single Locometrix tri-axial accelerometer was attached on the lower back at the level of the L3 vertebra. The gait parameters of speed, stride frequency, stride length, stride regularity, and stride symmetry were reported for the TUG test and for TUG-DT (counting down sequentially from 50, i.e. “50, 49, 48, …”). Several gait parameters were observed to be useful for differentiating amongst the three subject groups. For example, TUG-DT gait speed was found to differentiate the three participant groups. AD participants had lower stride length and gait regularity than MCI and healthy groups. MCI participants exhibited lower stride frequency than the healthy participants.

Furthermore in the area of cognitive impairment research, work by Greene et al. [50] served as a foundation for a longitudinal study by Greene and Kenny [16]. In this work, the qTUG was utilized to predict cognitive decline as measured by the Mini Mental State
Examination.

2.2.6 TUG Summary

Of the TUG studies, eight used only accelerometers (44.44% of the TUG studies) [47,52–58], seven used IMUs (38.89%) [11,31,34,50,59–61], one used inertial sensors plus magnetometers (5.56%) [62], one used an accelerometer and magnetometer (5.55%) [63], and one used surface electromyography (5.55%) [64]. Eleven of the studies utilized only a single sensing unit (61.11%) [34,47,52–59,63], two studies utilized two sensors (11.11%) [31,50], and five studies utilized three or more sensors (27.78%) [11,60–62,64]. The most common location to mount a sensor was the lower back (55.56%), close to the center of mass. The next most common choice was the lower limbs, a choice which yields a high number of gait parameters. Five studies [11,52,58,60,62] investigated accelerometers or gyroscopes on the upper limbs and only one study [54] researched movements of the head.

Several of the studies reported high numbers of computed metrics, with the most common parameters being TUG subtask durations, number of steps, cadence, stride length, and peak angular velocity. For our own wearable sensor studies presented in this dissertation, we compute these metrics and introduce additional metrics for mobility assessment. We also adopt several of the aforementioned sensor placement, processing, and suggestions put forth by previous IMU researchers. We aim to compute clinically-meaningful metrics and analyze changes in mobility to extend several areas of research, including IMU data processing, gait analysis, and rehabilitation research.
2.3 Activity-based Change Detection

In the literature, a few studies have aimed to detect change specifically in human behavior patterns. These approaches have quantified change statistically [65, 66], graphically [66–68], and algorithmically [67, 69–71]. Recently, Merilahti et al. [65] extracted features derived from actigraphy data collected for at least one year. Each feature was individually correlated with a component of the Resident Assessment Instrument [72] for insights into how longitudinal changes in actigraphy and functioning are associated. While this approach provides insight into the relationship between wearable sensor data and clinical assessment scores, this study does not directly quantify sensor-based change.

Wang et al. [67] introduced another activity-based change detection approach in which passive infrared motion sensors were installed in apartments and utilized to estimate physical activity in the home as well as time away from home. The data were converted into co-occurrence matrices for computation of image-based texture features. Their case studies suggested the proposed texture method can detect lifestyle changes, such as knee replacement surgery and recovery. Though the approach did not provide explanation of the detected changes over time, visual inspection of the data was suggested with activity density maps. More recently, Tan et al. [68] applied the texture method to data from Fitbit Flex sensors for tracking changes in daily activity patterns for elderly participants. Another approach for activity monitoring includes the Permutation-based Change Detection in Activity Routine (PCAR) algorithm [69]. PCAR researchers modeled activity distributions for time windows spanning three months of data. Changes between windows
were quantified with probabilities of change acquired via hypothesis testing.

The change detection algorithms described previously are intended for monitoring human activity behavior. There are several additional approaches that are not specific to activity data, but instead represent generic statistical approaches to detecting changes in time series data. Change point detection, the problem of identifying abrupt changes in time series data [73], constitutes an extensive body of research as there are many applications requiring efficient, effective algorithms for reliably detecting variation. There are many families of change detection algorithms that are suitable for different applications [74]. Algorithms appropriately handling unlabeled data are most relevant to the current study due to their data-driven change score computation and no need for ground truth information. Unsupervised change detection approaches include subspace models and likelihood ratio methods [70]. One particular subgroup of likelihood ratio methods, direct density ratio estimator methods, is commonly used because the methods have a lower complexity than other likelihood ratio methods [75, 76]. Relative Unconstrained Least-Squares Importance Fitting (RuLSIF) [70] is one such direct density estimator approach that is used to measure the difference between two samples of data surrounding a candidate change point. Other recent change point detection research includes work on multidimensional [77, 78] and streaming time series data [74].

The above approaches are effective methods for detecting change between two samples of data; however, they are not explanatory methods as they only identify if two samples are different and do not provide information on how the samples are different. Once a change is detected and determined significant, additional analyses are required to explain
the change that occurred. Hido et al. [71] formalized this problem as change analysis, a method of examination beyond change detection to explain the nature of discrepancy. Hido’s solution to change analysis utilizes supervised machine learning algorithms, specifically virtual binary classifiers, to identify and describe changes in unsupervised data. Research by Ng and Dash [79] and Yamada et al. [73] have also explored methods for detecting and explaining change in time series data.

Our PACD method makes use of some of these change detection methods and also introduces new enhancements that are particularly well suited for analyzing wearable and smart home sensor data. We begin with a study of mobility analysis using sensor data in rehabilitation environments, introduced in the next chapter.
CHAPTER 3. WEARABLE SENSOR-BASED MOBILITY

ANALYSIS

We postulate that wearable sensor data can be analyzed to provide insights on mobility recovery. To validate our hypothesis, we introduce methods to provide quantitative insights on mobility progress in a rehabilitation context. This chapter describes the first component of this study, a mobility assessment protocol we designed in collaboration with physical therapists at St. Luke’s Rehabilitation Institute, an inpatient rehabilitation hospital. For this assessment, patients wore IMUs on two occasions separated by 7 days while performing functional gait tasks in a simulated community setting. Adding ecological context in a simulated community environment has been shown to better represent an individual’s functionality than a controlled laboratory environment [6]. This is because patients adapt their movements to accomplish challenging tasks, such as transitioning from sitting to standing, crossing different flooring surfaces, and transferring into and out of motor vehicles.

The study followed a single-arm prospective cohort design with repeated measures of participant performance on standardized gait tasks on two different testing sessions separated by 7 days (see Figure 3.1). The first test session (S1) occurred shortly after the participant became physically able to walk the distance required of the gait task. The second test session (S2) occurred within the final week of care. During each test session, participant performance on the gait tasks was recorded two times, producing two separate
3.1 Functional Independence Measure

The FIM is a well-validated clinical assessment used to measure patient functioning at inpatient rehabilitation hospitals [7]. The FIM is measured at two distinct points in time: admission (FIM_A) and discharge (FIM_D). The FIM measures the level of assistance required to perform 18 ADL tasks (see Table 3.1) [80]. The tasks are categorized as either motor (13 tasks) or cognitive (5 tasks). Each task is scored on a 7-point ordinal scale to measure independence as determined by the amount of assistance required to perform each ADL task (see Table 3.2). A score of 7 denotes that a helper is not required for the patient to perform the task and a score of 1 denotes total assistance from a helper is
Table 3.1: Functional Independence Measure (FIM) tasks and associated categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>Task Type</th>
<th>#</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motor</td>
<td></td>
<td>1</td>
<td>Eating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Grooming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Bathing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Upper body dressing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Lower body dressing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Toileting</td>
</tr>
<tr>
<td></td>
<td>Sphincter control</td>
<td>7</td>
<td>Bladder management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Bowel management</td>
</tr>
<tr>
<td></td>
<td>Transfers</td>
<td>9</td>
<td>Bed to chair transfer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Toilet transfer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Tub/shower transfer</td>
</tr>
<tr>
<td>Locomotion</td>
<td></td>
<td>12</td>
<td>Walk/wheelchair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Stairs</td>
</tr>
<tr>
<td>Cognitive</td>
<td>Communication</td>
<td>14</td>
<td>Comprehension</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Expression</td>
</tr>
<tr>
<td></td>
<td>Social cognition</td>
<td>16</td>
<td>Social interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17</td>
<td>Problem solving</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18</td>
<td>Memory</td>
</tr>
</tbody>
</table>

required for the patient to perform the task [80].

The FIM motor aggregate score ($FIM_{motor}$) is the sum of all 13 individual motor task scores. The cognitive aggregate score ($FIM_{cog}$) is the sum of all five individual cognitive task scores. Finally, the total FIM score is the sum of all individual task scores. The change in FIM from admission to discharge is important in the clinical setting, representing the improvement or regression exhibited by the patient during their stay at the rehabilitation hospital. The change in FIM is represented by:

$$\Delta FIM = FIM_D - FIM_A$$ \hspace{1cm} (3.1)

Furthermore, the rehabilitation efficiency ratio (RER), also known as FIM efficiency, de-
Table 3.2: Functional Independence Measure (FIM) scoring scale.

<table>
<thead>
<tr>
<th>Independence Level</th>
<th>Score</th>
<th>Score Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>No helper required</td>
<td>7</td>
<td>Complete independence</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Modified independence (patient requires use of a device, but no physical assistance)</td>
</tr>
<tr>
<td>Helper (modified dependence)</td>
<td>5</td>
<td>Supervision or setup</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Minimal contact assistance (patient can perform ≥ 75% of the task)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Moderate assistance (patient can perform 50% to 74% of task)</td>
</tr>
<tr>
<td>Helper (complete dependence)</td>
<td>2</td>
<td>Maximal assistance (patient can perform 25% to 49% of task)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Total assistance (patient can perform &lt; 25% of task or requires more than one person to assist)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>Activity does not occur</td>
</tr>
</tbody>
</table>

termits the average rate of FIM change per day:

\[
RER = \frac{\Delta FIM}{LOS}
\]  

(3.2)

where LOS is the length of stay at the rehabilitation facility, measured in number of days.

### 3.2 Participants

For our sensor-based mobility assessment study, participants were recruited from the patient population at a large inpatient rehabilitation facility. All participants met the following eligibility criteria: any rehabilitation diagnosis (e.g., stroke, spinal cord injury, debility), ≥ 18 years of age, no more than minimally cognitively impaired as signified by a score of one or greater on the Mini-Cog examination [81,82] (see Appendix A), able and willing to perform the walking task as signified by a score of four or better (minimal stand-by assistance) on the locomotion independence item of the FIM (using an ambulatory aid
such as a cane or walker was acceptable). Patients were ineligible for the study if one of the following conditions applied: a significant cognitive or comprehension impairment as signified by a score of zero on the Mini-Cog examination, uncorrected vision or hearing problems, signs of delirium, insufficient awareness to respond to verbal or visual stimuli, a diagnosis of a significant preexisting or active psychiatric disorder or dementia severe enough to prevent them from completing the assessment protocol, inability to walk (e.g., wheelchair-bound), or inability to perform the walking task as signified by a score of three or less (more than minimal stand-by assistance required to walk) on the locomotion independence item of the FIM. The study was approved by a regional hospital institutional review board and all participants gave written informed consent.

Twenty participants (Male = 14, Female = 6), between the ages of 52 and 88 years old (71.55 ± 10.62 years), participated in both testing sessions of the study (see Table 3.3 for participant characteristics). The majority (70%) of participants required a wheeled walker during both testing sessions. Three (15%) participants used a cane during both testing sessions. One participant transitioned from a walker to a cane between the sessions. Two (10%) participants were able to complete both sessions without an assistive device. All participants required stand-by assistance either with or without contact guarding from a physical therapist during performance of the trials. Medical record review revealed an average group FIM score at the time of admission of 58.65 ± 12.52. Rehabilitation diagnoses were varied, with fourteen (70%) participants undergoing post-stroke rehabilitation. Hemiparesis was present in eleven (55%) post-stroke participants. Hemiparesis involved the dominant side for three individuals.
Table 3.3: Characteristics and Functional Independence Measure (FIM) scores for participants in the wearable sensor study.

<table>
<thead>
<tr>
<th>PID</th>
<th>RIC</th>
<th>Involved</th>
<th>Gender</th>
<th>Age</th>
<th>CMs</th>
<th>LOS</th>
<th>#Days A→S1</th>
<th>#Days S2→D</th>
<th>FIM_A-c</th>
<th>FIM_D-c</th>
<th>FIM_A-m</th>
<th>FIM_D-m</th>
<th>Total RER</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>73</td>
<td>N</td>
<td>31</td>
<td>16</td>
<td>8</td>
<td>23</td>
<td>34</td>
<td>25</td>
<td>70</td>
<td>1.81</td>
</tr>
<tr>
<td>002</td>
<td>Cardiac</td>
<td>N/A</td>
<td>M</td>
<td>84</td>
<td>Y</td>
<td>14</td>
<td>4</td>
<td>3</td>
<td>26</td>
<td>30</td>
<td>33</td>
<td>62</td>
<td>2.36</td>
</tr>
<tr>
<td>003</td>
<td>Misc</td>
<td>N/A</td>
<td>M</td>
<td>68</td>
<td>Y</td>
<td>19</td>
<td>8</td>
<td>4</td>
<td>32</td>
<td>33</td>
<td>37</td>
<td>65</td>
<td>1.53</td>
</tr>
<tr>
<td>004</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>75</td>
<td>Y</td>
<td>21</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>26</td>
<td>24</td>
<td>52</td>
<td>1.86</td>
</tr>
<tr>
<td>005</td>
<td>Stroke</td>
<td>NoP</td>
<td>M</td>
<td>63</td>
<td>Y</td>
<td>23</td>
<td>15</td>
<td>1</td>
<td>25</td>
<td>33</td>
<td>44</td>
<td>74</td>
<td>1.65</td>
</tr>
<tr>
<td>006</td>
<td>Stroke</td>
<td>L</td>
<td>F</td>
<td>82</td>
<td>Y</td>
<td>29</td>
<td>18</td>
<td>5</td>
<td>21</td>
<td>27</td>
<td>25</td>
<td>59</td>
<td>1.38</td>
</tr>
<tr>
<td>007</td>
<td>NTBI</td>
<td>N/A</td>
<td>M</td>
<td>52</td>
<td>Y</td>
<td>22</td>
<td>8</td>
<td>7</td>
<td>21</td>
<td>31</td>
<td>43</td>
<td>62</td>
<td>1.32</td>
</tr>
<tr>
<td>009</td>
<td>Stroke</td>
<td>NoP</td>
<td>F</td>
<td>85</td>
<td>Y</td>
<td>20</td>
<td>8</td>
<td>5</td>
<td>22</td>
<td>30</td>
<td>46</td>
<td>58</td>
<td>1.00</td>
</tr>
<tr>
<td>010</td>
<td>NTBI</td>
<td>N/A</td>
<td>F</td>
<td>67</td>
<td>N</td>
<td>24</td>
<td>17</td>
<td>0</td>
<td>16</td>
<td>28</td>
<td>25</td>
<td>40</td>
<td>1.13</td>
</tr>
<tr>
<td>011</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>74</td>
<td>N</td>
<td>20</td>
<td>11</td>
<td>2</td>
<td>16</td>
<td>23</td>
<td>28</td>
<td>49</td>
<td>1.40</td>
</tr>
<tr>
<td>013</td>
<td>NTSCI</td>
<td>N/A</td>
<td>M</td>
<td>76</td>
<td>N</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>25</td>
<td>30</td>
<td>26</td>
<td>58</td>
<td>2.47</td>
</tr>
<tr>
<td>014</td>
<td>Stroke</td>
<td>NoP</td>
<td>M</td>
<td>55</td>
<td>Y</td>
<td>17</td>
<td>9</td>
<td>1</td>
<td>16</td>
<td>25</td>
<td>45</td>
<td>78</td>
<td>2.47</td>
</tr>
<tr>
<td>015</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>85</td>
<td>N</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>32</td>
<td>33</td>
<td>55</td>
<td>80</td>
<td>2.00</td>
</tr>
<tr>
<td>016</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>54</td>
<td>N</td>
<td>21</td>
<td>13</td>
<td>1</td>
<td>23</td>
<td>29</td>
<td>37</td>
<td>69</td>
<td>1.81</td>
</tr>
<tr>
<td>018</td>
<td>Stroke</td>
<td>L</td>
<td>M</td>
<td>88</td>
<td>N</td>
<td>29</td>
<td>19</td>
<td>3</td>
<td>26</td>
<td>32</td>
<td>27</td>
<td>60</td>
<td>1.34</td>
</tr>
<tr>
<td>019</td>
<td>Stroke</td>
<td>R</td>
<td>M</td>
<td>65</td>
<td>N</td>
<td>14</td>
<td>5</td>
<td>2</td>
<td>30</td>
<td>31</td>
<td>49</td>
<td>72</td>
<td>1.71</td>
</tr>
<tr>
<td>020</td>
<td>Misc</td>
<td>N/A</td>
<td>M</td>
<td>74</td>
<td>Y</td>
<td>28</td>
<td>17</td>
<td>4</td>
<td>21</td>
<td>30</td>
<td>31</td>
<td>61</td>
<td>1.39</td>
</tr>
<tr>
<td>021</td>
<td>Stroke</td>
<td>R</td>
<td>F</td>
<td>74</td>
<td>N</td>
<td>16</td>
<td>9</td>
<td>0</td>
<td>24</td>
<td>33</td>
<td>40</td>
<td>80</td>
<td>3.06</td>
</tr>
<tr>
<td>024</td>
<td>Stroke</td>
<td>L</td>
<td>F</td>
<td>63</td>
<td>N</td>
<td>18</td>
<td>9</td>
<td>2</td>
<td>29</td>
<td>31</td>
<td>38</td>
<td>73</td>
<td>2.06</td>
</tr>
<tr>
<td>025</td>
<td>Stroke</td>
<td>R</td>
<td>F</td>
<td>74</td>
<td>N</td>
<td>21</td>
<td>12</td>
<td>2</td>
<td>19</td>
<td>22</td>
<td>37</td>
<td>61</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Mean - - - 71.55 - 20.75 11.15 2.65 23.10 29.55 35.75 64.15 1.752
SD - - - 10.62 - 5.35 4.75 2.25 5.20 3.43 9.27 10.51 0.53

A = admission, CM = comorbidities, c = cognitive, D = discharge, F = female, FIM = Functional Independence Measure, L = left, LOS = length of stay, m = motor, M = male, N = no, NoP = no paresis, NTBI = non-traumatic brain injury, NTSCI = non-traumatic spinal cord injury, PID = patient identification, R = right, RER = rehabilitation efficiency ratio, RIC = rehabilitation impairment category, SD = standard deviation, Y = yes.
Figure 3.2: The ambulation circuit. The solid line represents the circuit path and the dashed line represents the mirrored return portion. Key circuit tasks are labeled with distances in meters.

### 3.3 The Ambulation Circuit

We designed a standardized ambulation circuit (AC) to assess mobility of the participants during the test sessions. The AC is a continuous sequence of activities performed in a simulated community environment at the rehabilitation facility consisting of several indoor and outdoor modules. The ecological context provided by a simulated environment has been shown to produce a more representative assessment of an individual’s functionality than a controlled laboratory setting [6]. In ecological environments, patients adapt their movements to accomplish challenging tasks, such as transitioning from sitting to standing, navigating different flooring surfaces, and transferring into and out of motor vehicles.
Figure 3.3: The St. Luke’s Rehabilitation Institute Community utilized for the ambulation circuit.
Figure 3.2 illustrates the AC. The AC begins in a simulated hotel lobby area with the participant seated in a chair on a rectangular shag rug (see Figure 3.3). The chair faces a linear path that leads to an outdoor area with several motor vehicles. On beginning the circuit, the participant rises from the sitting position, performing a sit-to-stand transition. Once standing, the participant walks across the remaining length of the shag rug. When the edge of the rug is reached, the participant performs a surface transition from the shag rug to smooth wood flooring. After walking for 3.66 meters (12 feet) on the smooth floor, a researcher asks the participant “When is your birthday?” during the first trial and “What is the date today?” during the second trial, to assess whether the participant stops walking or slows down. This is a variant of the Stops Walking When Talking Test, a simple method to determine risk for falling when simultaneously engaged in motor and cognitive tasks [83,84]. Participants are not informed that the question will be asked.

Next, the participant approaches the front of a sport utility vehicle and begins a curvilinear path around the vehicle to approach an open passenger side door. The curvilinear path length is approximately 6.7 meters (22 feet). The curvilinear path contains a simulated sewer drain lid (manhole cover) over which the participant has to maneuver. As the participant approaches the vehicle passenger seat, the participant performs a transfer into and then out of the vehicle front passenger seat. After transferring out of the vehicle, the participant walks the AC route in reverse, returning to the chair in the simulated hotel lobby and sits down, ending the AC. The Stops Walking When Talking Test is not performed on the way back in order to provide 6.4 meters (21 feet) of uninterrupted smooth linear walking for gait analysis. Time taken to complete the AC officially stops once the
participant’s back is fully rested against the back of the chair.

In summary, the AC is an extension of the common clinical assessment, the TUG (see Section 2.2), including a greater range of functional tasks (e.g., car transfers) and situational challenges (e.g., different flooring surfaces; a curvilinear pathway) than is found in more common assessments. This greater range of cognitive and motor challenges enhances the potential usefulness of the sensor data as a means to show change across time. The majority of the metrics we report can be computed from any assessment in any environment involving a chair transfer and walking (5 Times Sit-to-Stand, TUG, etc.).

3.4 Sensor Data Processing

Using four Shimmer3 [85] wireless IMUs, we recorded participant motion as they ambulated through the AC. The Shimmer3 platform contains a tri-axial accelerometer and a tri-axial gyroscope. The accelerometers and gyroscopes of all three sensor platforms were calibrated using the software provided by the manufacturer. One IMU was placed centrally on the lumbar spine at the level of the third vertebrae, near the individual’s center of mass (COM) [10]. Additionally, one sensor was placed on each shank, above the ankle and in line with the tibia. Positioning the sensor along the tibia reduced mounting error as the sensors were always positioned at approximately the same angle relative to the sagittal plane [86]. The flatness of the tibia bone also prevented the sensor from moving during the activities. The sensor modules were securely attached to the body with elastic straps. Figure 3.4 illustrates the shank mounting locations and axes of the sensors. If the
Figure 3.4: Sensor placement and axes orientation. Sensor units are mounted on the center of mass (COM), left shank (LS), and right shank (RS).
Figure 3.5: Wearable inertial sensor data are processed before computing mobility metrics. Processing steps include timestamp alignment, orientation, and filtering.

If a patient used a cane or a walker as an assistive device, a fourth sensor was placed on the device. The accelerometer range was set to ± 2g for the COM sensor [87] and ± 4g for the shanks (and the assistive device sensor if the participant used a cane). The gyroscope ranges for the shank and COM sensors (and the assistive device sensor if the participant used a cane) were set at 500 °/s and 250 °/s, respectively. The data were collected at a sampling frequency of 51.2 Hz for all sensor platforms.

To process the AC wearable sensor data, we present several computational tools. Figure 3.5 provides an overview of the sensor data processing steps. First, we align timestamps recorded by each of the sensor platforms. This step is required in order to compare movements of the COM with individual limbs and cane assistive devices. The timestamp
alignment approach presented in this dissertation is based on information provided by the sensor manufacturer [88]. The method requires one Shimmer sensor platform be designated as the “master” and the remaining sensor platforms be designated as “slaves.” Each Shimmer platform contains an on-board 32 kHz clock that begins counting at zero. When each sensor platform begins logging data to the local SD card, the platform saves the current clock time as a 32-bit timestamp for use in alignment post-processing. Subsequent timestamps are saved at 16-bit resolution. In this configuration, while collecting data, the master sensor broadcasts its on-board clock time (at 32-bit resolution) at regular one minute intervals wirelessly over Bluetooth. The slave sensors receive the master’s clock time and each slave computes the difference between its local clock time and the master’s clock time. This difference consists of 1 byte representing the sign of the offset (0 if local time \(\geq\) master time and 1 otherwise) and 4 bytes for an unsigned integer representing the offset magnitude. The sign and magnitude values are stored to a binary file on the SD card at the start of every write operation; however, data blocks are written much more frequently than master timestamps are received. When it is time to perform a write operation and a new master timestamp has not been received, an invalid value is written for the offset magnitude. Once data collection is complete, post-processing of the stored binary file is required to produce new slave timestamps that are aligned with the master timestamps. Algorithm 1, TimestampAlignment, provides details regarding the slave timestamp alignment process. A high level overview of the algorithm is as follows:

1. Identify the initial 32-bit timestamp at which data logging started. This information is stored in the header of the binary file stored on the SD card.
2. Identify the initial 16-bit timestamp. This is the first 16-bit timestamp in the data.

3. Convert the 16-bit timestamps to monotonically increasing integer timestamps (see Algorithm 2).

4. Identify the timestamps at which master timestamps were received. Compute the corresponding offset values.

5. Train a linear regression model to map timestamps onto offset values (see Algorithm 3).

6. Utilize the trained model to interpolate the offsets for timestamps between the received master offset values.

7. Subtract each interpolated offset from each corresponding timestamp. Convert the resulting timestamp to milliseconds.

After timestamp alignment, we transform the sensor local coordinate system \((X, Y, Z)\) to the body coordinate system \((X_B, Y_B, Z_B)\) to correct for the orientation of the shank sensors along the tibia \([89]\); a right handed system with the X-axis along the anterior-posterior body axis, the Y-axis along the vertical body axis, and the Z-axis along the medial-lateral body axis. To minimize error due to misaligned sensor mounting, we utilize the mounting calibration technique proposed by Chen \([90]\). For this method, we measure two acceleration vectors, \(Y_B = [a_x, a_y, a_z]^T\) and \(g' = [a'_{x}, a'_{y}, a'_{z}]^T\), for each shank sensor while the sensor is mounted on the participant. \(Y_B\) is measured by orienting the shank such that the leg is vertical with respect to the floor. \(g'\) is measured by orienting
Algorithm 1 TimestampAlignment($X$)

1: Input: $X$ = binary sensor data recorded by “slave” sensor platform
2: Output: $T_{aligned}$ = aligned timestamps in milliseconds
3: Initialize: $ts_{prev} = 0, ts_{16b_{prev}} = 0, offsets = dictionary, T_{continuous} = list$
4: $ts_{32b_{initial}}$ = initial 32-bit timestamp when logging started (retrieved from data header)
5: $ts_{16b_{initial}}$ = initial 16-bit timestamp when logging started
6: for each binary block $B$ in $X$:
7: \hspace{1em} sign = unsigned integer($B[0]$)
8: \hspace{1em} magnitude = unsigned integer($B[1 : 4]$)
9: \hspace{1em} $ts_{16b_{curr}}$ = unsigned integer($B[5 : 6]$)
10: \hspace{1em} $ts_{curr} = \text{ContinuousTimestamp}(ts_{16b_{curr}}, ts_{16b_{prev}}, ts_{prev})$ (see Algorithm 2)
11: \hspace{1em} Append $ts_{curr}$ to $T_{continuous}$
12: \hspace{1em} $ts_{16b_{prev}} = ts_{16b_{curr}}$
13: \hspace{1em} $ts_{prev} = ts_{curr}$
14: \hspace{1em} if magnitude is valid:
15: \hspace{2em} offset = $(1 - 2 \times \text{sign}) \times \text{magnitude}$
16: \hspace{2em} offsets[$ts_{curr}$] = offset
17: \hspace{1em} for each data sample $S$ in $B$:
18: \hspace{2em} $ts_{16b_{curr}}$ = unsigned integer($S[1 : 2]$)
19: \hspace{2em} $ts_{curr} = \text{ContinuousTimestamp}(ts_{16b_{curr}}, ts_{16b_{prev}}, ts_{prev})$ (see Algorithm 2)
20: \hspace{2em} Append $ts_{curr}$ to $T_{continuous}$
21: \hspace{2em} $ts_{16b_{prev}} = ts_{16b_{curr}}$
22: \hspace{2em} $ts_{prev} = ts_{curr}$
23: \hspace{1em} end for
24: $T_{aligned} = \text{InterpolateTimestamps}(T_{continuous}, offsets, ts_{32b_{initial}}, ts_{16b_{initial}})$ (see Algorithm 3)
25: return $T_{aligned}$

Algorithm 2 ContinuousTimestamp($ts_{16b_{curr}}, ts_{16b_{prev}}, ts_{prev}$)

1: Input: $ts_{16b_{curr}}$ = current 16-bit timestamp
2: Input: $ts_{16b_{prev}}$ = previous 16-bit timestamp
3: Input: $ts_{prev}$ = previous continuous timestamp
4: Output: $ts_{curr}$ = current continuous timestamp
5: Initialize: elapsed = 0
6: if $ts_{16b_{curr}} < ts_{16b_{prev}}$:
7: \hspace{1em} elapsed = $ts_{16b_{curr}} + (2^{16} - ts_{16b_{prev}})$
8: \hspace{1em} else:
9: \hspace{2em} elapsed = $ts_{16b_{curr}} - ts_{16b_{prev}}$
10: \hspace{1em} return $ts_{curr} + ts_{prev}$
Algorithm 3 InterpolateTimestamps($T_{continuous}$, offsets, $ts32b_{initial}$, $ts16b_{initial}$)

1: Input: $T_{continuous}$ = list of continuous timestamps
2: Input: offsets = pairs of timestamps to slave offset values
3: Output: $T_{aligned}$ = list of aligned timestamps in milliseconds
4: Initialize: $T_{aligned}$ = list
5: regressor = train linear regression to map offsets keys (timestamps) to offsets values (offsets)
6: for each $ts$ in $T_{continuous}$:
7:   offset = regressor($ts$)
8:   $ts = ts32b_{initial} + (ts - ts16b_{initial}) - offset$
9:   Append $ts$ to $T_{aligned}$
end for
10: return $T_{aligned}$

the shank such that the leg is horizontal with respect to the floor. To measure $g'$ from AC participants, the participant sits in a chair and the experimenter cautiously lifts the participant’s leg and rests it on a chair of equal height as the chair the participant is sitting in. To compute $Z_B$, the cross product between $Y_B$ and $g'$ is computed and normalized: $Z_B = Y_B \times g'$. Similarly, $X_B$ is computed as $X_B = Y_B \times Z_B$. Next, a rotation matrix $R = [X_B, Y_B, Z_B]$ is constructed and multiplied by the acceleration ($a = [a_x, a_y, a_z]^T$) and gyroscope data ($\omega = [\omega_x, \omega_y, \omega_z]^T$) to transform the sensor local coordinates to body coordinates: $a_B = Ra$ and $\omega_B = R\omega$.

After performing orientation correction to minimize mounting error, we filter acceleration data with a $4^{th}$ order zero-phase band pass Butterworth filter using cutoff frequencies of 0.1 Hz and 3 Hz for the COM accelerometer [91] and 0.1 Hz and 10 Hz for the shanks (and the assistive device sensor if the participant used a cane) [92]. Similarly, we low pass filter the gyroscope signals for all sensors platforms at 4 Hz [93].

We compute metrics from the processed data representing participant performance on the AC. Experimenters recorded AC task durations using a stopwatch. The times are
Figure 3.6: Example sensor signals recorded from participant #006 as she performed the ambulation circuit. The center of mass (top figure: accelerometer) sensor signals and the shank (bottom figure: gyroscope) sensor signals are shown.

As can be seen by the signal waveforms, the activities involving the chair, walking, and the car transfer are quite distinct from each other.

### 3.5 Computed Metrics

For a unique analysis of sensor-based gait information in a rehabilitation setting, we compute metrics from three main components of the AC. The first component consists of the chair sit-to-stand and stand-to-sit movements at the beginning and ending of the AC.
The second component consists of the vehicle transfer, and the third component consists of the ambulation occurring between the chair and the vehicle. This ambulation section includes the linear path on the smooth floor that is used to compute the majority of the gait cycle metrics. In order to facilitate comparison across different participants, we normalize a few of the metrics. For example, it has previously been found that an individual’s stride length is directly proportional to their body height [94]. For this reason, metrics based on steps and strides are normalized by stature.

For the ambulation section, we present an algorithm, GaitCycleEventDetection (see Algorithm 4), to detect the gait cycle events of initial contact (IC), terminal contact (TC), and mid-swing (MS) [95]. We define initial contact as the moment the heel strikes the ground and terminal contact is the moment the toes leave contact with the ground. By locating these key gait events, we can define the gait cycle as the time interval between two successive initial contacts of the same leg and we can introduce several metrics related to walking. The GaitCycleEventDetection algorithm operates on the left and right shank medial-lateral (Z-axis) gyroscope data. If the participant uses a cane, we perform additional gait cycle analysis (see Section 3.6). The algorithm utilizes peak detection and thresholding techniques that were implemented with high accuracy by previous studies [11,96,97]. Figure 3.7 shows Z-axis gyroscope data and the associated gait events for participant #015’s left shank, right shank, and cane. The outline of the algorithm is as follows:

1. Detect MS events. MS events correspond to the highest peaks in the Z-axis gyroscope signal (square points in Figure 3.7).
2. Detect IC events. IC events correspond to the local minimum after a MS event (diamonds in Figure 3.7).

3. Detect TC events. TC events correspond to the local minimum before a MS point (circles in Figure 3.7).

4. Identify MS, IC, and TC events corresponding to individual gait cycles for the lead leg. The lead leg is defined as the leg corresponding to the first detected IC event of the linear walking portion of the AC.

5. Identify MS, IC, and TC events for the non-lead leg and cane with respect to the cycles defined by the lead leg.

**Algorithm 4 GaitCycleEventDetection**

1: Input: $lead_Z, other_Z, cane_Z$ Gyroscope Z-axis signals
2: Output: IC, TC, MS event timestamp vectors for $lead_Z, other_Z, cane_Z$
3: Detect and store $lead_Z$ IC, TC, MS events
4: Identify gait cycles based on $lead_{IC}$ events
5: for each gait cycle:
6: Detect and store $other_Z$ and $cane_Z$ IC, TC, MS events
end for
7: Perform validity checks on the identified gait cycle events
8: return IC, TC, MS event timestamp vectors for $lead_Z, other_Z, cane_Z$

We group the computed metrics into three categories. The categories include clinical assessments of progress (CAP), whole body movement metrics (WBM), and gait features (GF):

1. Clinical assessments of progress. CAP metrics are commonly used approaches for assessing mobility in a clinical setting by recording the duration of a standardized
Figure 3.7: Participant #015’s left shank, right shank, and cane sensor Z-axis gyroscope signals recorded during the linear walking portion of the AC. The individual gait cycles are identified as alternating shaded regions.
activity, such as walking a fixed distance, rising from a chair, or the TUG assessment [4].

2. Whole body movement. WBM metrics are computed primarily from data collected from the COM sensor. An example WBM metric is COM peak angular velocity.

3. Gait features. GF metrics are computed from data collected from the shank sensors. For each trial \( t \) of the AC, there are \( N_t \) gait cycles detected during the ambulation section. For each gait cycle \( i \), several gait features are computed. Also, the mean, standard deviation, and coefficient of variation are computed for each gait feature for all \( N_t \) gait cycles of each trial \( t \). Examples include parameters of gait, such as cadence and shank range of motion, which are based on the aforementioned gait cycle event detection algorithm (see Algorithm 4).

Tables 3.4, 3.5, and 3.6 summarize the metrics computed for each category. Furthermore, we compute additional metrics based on the IMU data collected from sensors attached to cane assistive devices. The following section (Section 3.6) provides a description of the analysis we perform for cane assistive devices.

### 3.6 Assistive Device Analysis

Age, injury, or disease-related mobility impairments can severely diminish one’s functional independence. Particularly for walking, moving independently is often necessary to maintain strength, mobility, and a reasonable quality of life. Many individuals with walking impairments use assistive devices (e.g. canes, walkers, and wheelchairs) during ambulation.
Table 3.4: Clinical assessments of progress metrics computed from wearable inertial sensor data.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>s</td>
<td>Total time to complete the ambulation circuit or an individual task of the ambulation circuit.</td>
</tr>
<tr>
<td>Floor surface speed ratio</td>
<td></td>
<td>Measures the effect of walking speed on two different floor surfaces.</td>
</tr>
<tr>
<td>Walking speed</td>
<td>m/s</td>
<td>Walking speed as determined by distance divided by time (normalized by leg length).</td>
</tr>
</tbody>
</table>

m = meters, s = seconds.

Table 3.5: Whole body movement metrics computed from wearable inertial sensor data.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COM peak angular velocity</td>
<td>°/s</td>
<td>Maximum rotational velocity of the COM around the Z-axis.</td>
</tr>
<tr>
<td>Root mean square (RMS)</td>
<td>m/s²/s</td>
<td>Square root of the mean of the squares of each COM acceleration signal (normalized by time). Represents acceleration magnitude. Synonymous with movement intensity (MI).</td>
</tr>
<tr>
<td>Smoothness index (harmonic ratio)</td>
<td></td>
<td>Ratio of even to odd harmonics of the vertical Y-axis COM acceleration signal. A higher harmonic ratio represents a smoother walking pattern.</td>
</tr>
<tr>
<td>Smoothness of RMS</td>
<td>m/s³/s</td>
<td>Square root of the mean of the squares of each COM acceleration signal derivative (normalized by time). Synonymous with RMS of jerk and smoothness of MI.</td>
</tr>
</tbody>
</table>

COM = center of mass, m = meters, MI = movement intensity, RMS = root mean square, s = seconds, ° = degrees.
Table 3.6: Gait features computed from wearable inertial sensor data.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Units</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cadence</td>
<td>steps/min</td>
<td>The average number of steps taken per minute.</td>
<td></td>
</tr>
<tr>
<td>Double support percent</td>
<td>%</td>
<td>Percentage of the gait cycle that both feet are on the ground. Computed as the sum of the initial double support time and the terminal double support time.</td>
<td>[11]</td>
</tr>
<tr>
<td>Gait cycle time</td>
<td>s</td>
<td>Duration to complete one stride (time between two consecutive initial contacts of the same foot).</td>
<td>[11]</td>
</tr>
<tr>
<td>Number of gait cycles</td>
<td></td>
<td>Total number of complete gait cycles (strides) that occurred.</td>
<td></td>
</tr>
<tr>
<td>Shank peak angular velocity</td>
<td>°/s</td>
<td>Maximum rotational velocity of the shank around the Z-axis during the gait cycle. This occurs during the swing phase.</td>
<td></td>
</tr>
<tr>
<td>Shank range of motion (ROM)</td>
<td>°</td>
<td>Range of integrated Z-axis angular velocity for each gait cycle. Provides an estimate of the degrees of shank movement.</td>
<td>[11]</td>
</tr>
<tr>
<td>Step length</td>
<td>m</td>
<td>Distance between initial contacts of opposite feet (normalized by leg length).</td>
<td>[95]</td>
</tr>
<tr>
<td>Step regularity</td>
<td>%</td>
<td>Regularity of the acceleration of sequential steps. Computed using the autocorrelation of the vertical Y-axis of the COM acceleration.</td>
<td>[10]</td>
</tr>
<tr>
<td>Stride regularity</td>
<td>%</td>
<td>Regularity of the acceleration of sequential strides (see step regularity).</td>
<td>[10]</td>
</tr>
<tr>
<td>Step symmetry</td>
<td>%</td>
<td>Ratio of step regularity to stride regularity.</td>
<td>[10]</td>
</tr>
</tbody>
</table>

COM = center of mass, m = meters, min = minute, RMS = root mean square, ROM = range of motion, s = seconds, ° = degrees.
to increase their base of support. A study from 2011 revealed 8.5 million adults in the United States aged 65 and older reported using a mobility device within the last month [98]. The study found a cane to be the most commonly used mobility device, with 16.4% of adults 65 years and older using one.

Although utilizing a walking aid is intended to help the user, research has revealed this is not always the case. Assistive devices often are used without consultation of a medical professional [99], are set for inappropriate height, or are faulty [100], and may cause an increase in falls risk [101]. Assistive devices are also often used incorrectly. An estimated 28% of cane users incorrectly hold the cane on their weak side, 11% occasionally swing the cane with the ipsilateral leg, and 14% occasionally hold the cane in the air for multiple steps [99]. Furthermore, using a walking aid alters spatiotemporal parameters of gait [101]. Objective assessments describing cane use while walking can offer useful information to users, clinicians, and caregivers. Wearable inertial measurement units are an ideal technology for providing these measures because they can be worn by the user and placed on the walking aid. To extend research on assistive devices and gait analysis, our work utilizes objective IMU-derived measurements of lower limb and cane movement to quantitatively assess assistive device usage in populations with functional walking deficits. More specifically, we investigate the relationship between swing and stance timing and variability in movement to provide insight into assistive device usage and potentially offer feedback to users, clinicians, and caregivers about proper cane movement.

Several studies have applied wearable IMUs for monitoring human movement and gait analysis [102]; however, fewer IMU studies have accounted for gait impairments
that require the use of an assistive device. These studies can be grouped into those investigating lower limb movement (sensors attached to the user and not to the device) \[103\], device movement (sensors attached to the device and not to the user) \[104–106\], and the relationship between user and device movement (sensors attached both to the device and user) \[107–109\]. The studies investigating the movement and gait of both the user and device are most relevant to the current study. Hester \textit{et al.} \[107\] computed statistical features from an ankle-mounted accelerometer and a cane equipped with two accelerometers and a load cell to train a neural network for activity recognition. Hassan \textit{et al.} \[108\] utilized IMUs mounted on a cane and affected limb of hemiplegic participants to estimate the movement intention of the user to control an exoskeleton. Recently, Lancini \textit{et al.} \[109\] added several sensors, including an accelerometer, to crutches used by powered exoskeleton users.

While the aforementioned studies have investigated human movement and device usage with several types of technologies, IMU-based gait analysis with respect to cane usage is a new direction of research. Our research in this area is unique because we introduce algorithms designed for IMUs attached to both lower limbs and a cane device. The current study extends several areas of research, including IMU data processing, gait analysis, and quantitative assessment of assistive device movement.
3.6.1 Cane and Gait Features

Of the included AC participants, two participants who used canes at both S1 and S2, #005 and #015, are included in the cane assistive device analysis (see Table 3.3 for participant characteristics). Both participants were undergoing post-stroke rehabilitation, with #015 having left side hemiparesis, and #005 having no paresis but profound imbalance secondary to stroke.

For each trial \( t \) of the AC, there are \( N_t \) gait cycles detected during the ambulation section. For each gait cycle \( i \), the following gait features related to cane movement are computed:

- Stance %. Leg and cane stance periods as a percentage of the total gait cycle.
- Mid-swing \( ^\circ /s \). Leg and cane MS angular velocity.
- Cane stance ratio. Ratio of contralateral limb stance percent to cane stance percent:
  \[
  \frac{\text{ContralateralStance}_i}{\text{CaneStance}_i}
  \]
  Represents the similarity in amount of time the cane and the contralateral limb are supporting the body.
- Cane swing offset. Cane MS temporal offset from the contralateral leg MS:
  \[|\text{CaneMS}_i - \text{ContralateralMS}_i|\]
  Represents how closely the cane swings with the contralateral limb.
- Double support %. The overlap in stance phases for the left and right legs as a percentage of the total gait cycle.
Figure 3.8: Stance and swing phase plot for participant #015 recorded from trial 3 at session 2 testing.

- Triple support %. The overlap in stance phases for the left leg, right leg, and cane as a percentage of the total gait cycle.

To visualize the variability of timing between both legs and a cane, we introduce stance and swing phase plots. Figure 3.8 and Figure 3.9 show stance and swing phase plots for participant #015 and #005 at S2 testing. For each sensor location, stance (gray) and swing (blue) phases are represented with horizontal bars. The sensor location stance and swing phases are grouped together for each gait cycle (the Y-axis). The X-axis shows an estimate of the percentage of the gait cycle, based on the average gait cycle duration. Overlaid on top of the stance and swing bars are periods of double (star hatch) and triple (cross hatch) support periods.
Figure 3.9: Stance and swing phase plot for participant #005 recorded from trial 3 at session 2 testing.

3.7 Study Limitations

A limitation of this mobility assessment protocol is that the metric computations have not been laboratory validated; however, the algorithms are derived from previously-published and validated sources. In particular, the gait cycle event detection algorithm is based on one such study utilizing IMUs from the same manufacturer (Shimmer Sensing). The study reported mean true errors for gyroscope-based gait cycle event detection of $-5.5 \pm 7.3$ ms for detecting IC points and $40.6 \pm 19.2$ ms for detecting TC points [96, 110]; therefore, the timings of gyroscope-detected gait events should be interpreted as estimates of the true durations. Other limitations include:
• Low sample size. Including more participants exhibiting various impairments would constitute a more representative sample.

• Non-uniform days between admission and S1 AC testing. Participants were tested when their physical therapists determined they were able to safely perform the AC.

• The use of human-operated stopwatch times to segment the AC into its subtasks. The times recorded by the researchers could impose non-systematic error.

• Only three sensor platforms were used to measure body movements. Additional sensor placement on the thigh would offer more detailed data about the upper and lower legs, as well as insight into the range of motion of the knee joint. We did not include these sensors in order to be as unobtrusive as possible.

• Recording only two trials within a session. Having a broader range of testing times within a single day could offer a more thorough quantification of capabilities and recovery.

In this chapter we introduced our mobility assessment protocol that utilizes wearable sensor data for patients undergoing inpatient rehabilitation. We also defined metrics for analyzing mobility and gait based on the collected data. The following three chapters present our approaches and results for the following AC inertial sensor data analyses. Chapter 4 describes our proposed method to measure gait changes, Chapter 5 introduces our computational approach to predicting clinical outcomes, and Chapter 6 summarizes feedback from physical therapy providers based on visualizations of our defined gait and assessment metrics.
CHAPTER 4. MEASURING GAIT CHANGES

Restoration of functional independence in gait and vehicle transfer ability is a common goal of inpatient rehabilitation. Currently, ambulation changes tend to be subjectively assessed. To support our hypothesis that changes in mobility assessment of patient progress can be computed from wearable sensor data, we quantitatively assess gait and transfer performances using wearable inertial sensors for 20 patients receiving inpatient rehabilitation services. Statistics are calculated to identify clinically significant changes in the repeated measures data [111, 112]. The statistical analyses we apply to the wearable sensor data at the group and individual levels are summarized in Sections 4.2 and 4.3, respectively.

4.1 Prior Work on Sensor-based Gait Assessment

Wearable IMUs have been utilized extensively in healthcare applications [113], particularly for gait analysis [11] and rehabilitation [9]. In addition, performance on common clinical assessments, such as the TUG test, have been characterized with IMUs [5, 11, 114]. Section 2.2 discusses prior research related to gait analysis and the TUG test in more detail.

While research in applications of IMUs for gait analysis and instrumenting clinical assessments is prevalent in the literature, to date only a few studies have focused on uti-
lizing IMUs to quantify changes in mobility and gait parameters of impaired populations. These studies have investigated improvement in gait following surgery, such as hip arthroplasty surgery [115]; changes in gait after treatment for a specific injury or illness, such as Parkinson’s Disease [114,116]; the relationship between changes in longitudinally collected gait parameters and changes in falls risk [117]; and changes in daily walking time over the course of rehabilitation for stroke inpatients [118]. However, research quantifying fine-grained gait and transfer ability changes exhibited during rehabilitation with wearable inertial sensors represents a new direction to investigate. We analyze changes in wearable sensor-derived metrics to gain insight into the recovery process and provide clinicians with more objective measurements of patient progress in clinical and natural environments.

4.2 Quantifying Group Changes

We utilize a standardized mean difference (SMD) effect size (ES) based on Cohen’s $d$ for the repeated measures (RM) data to quantify the strength of changes in the computed metrics [119,120]:

$$d_{RM} = \frac{\overline{X}_{S2} - \overline{X}_{S1}}{S_D}$$

(4.1)

Where $\overline{X}_{S1}$ is the mean group score from data collected at S1, $\overline{X}_{S2}$ is the mean group score from data collected at S2, and $S_D$ represents the standard error of difference between S1 and S2 scores [121]. In cases where the S1 and S2 scores have equal variance, $S_D$ is calculated using the formula from Morris and DeShon [122]:
In Equation 4.2, \( s_{S1} \) represents the SD for the S1 participant pool and \( r \) is the test-retest reliability of the measurement between trial 3 and 4 at S2 testing. However, if S1 and S2 variances are not equal, as determined by the Levene test of equal variances, we apply an adjustment to the estimate of \( S_D \) [120]:

\[
S_D^\# = s_{S1} \sqrt{2(1 - r)}
\]  

(4.2)

Additionally, an unbiased estimation of population test-retest reliability is derived using \( r \) [123]:

\[
\hat{\rho} = r + \frac{r(1 - r^2)}{2(n - 4)}
\]  

(4.3)

We use the effect size measurement of change, \( d_{RM} \) (Equation 4.1), to evaluate group changes in gait parameters over the course of one week of inpatient rehabilitation. Responsiveness is categorized as no effect, small, moderate, or large based on corresponding \( d_{RM} \) values of 0.0, 0.2, 0.5, and 0.8, respectively [120]. Confidence intervals are derived for each ES describing the variability in metrics between the two sessions. The confidence intervals are computed using a small sample size approximation with \( \alpha = 95\% \) [124]:

\[
d_{RM} \pm q \times \hat{\sigma}^2_d(L1)
\]  

(4.5)

Where \( q \) is the \( 100 \times (1 - \alpha / 2)^{th} \) quantile of the standard normal distribution and \( \hat{\sigma}^2_d(L1) \) is defined in the following equation [124]:
\[
\hat{\sigma}_d^2(L1) = \sqrt{\frac{2(1 - \hat{\rho})}{n} + \frac{d_{RM}^2}{2(n - 1)}} \tag{4.6}
\]

4.3 Individual-level Comparisons

At the individual level, we characterize changes in gait metrics one week apart with the reliable change index (RCI) [112]:

\[
\text{RCI} = \frac{x_{S2} - x_{S1}}{S_D} \tag{4.7}
\]

Where \(x_{S1}\) is an individual participant’s score from data collected at S1, \(x_{S2}\) is the same participant’s score from data collected at S2, and \(S_D\) is computed from Equations 4.2 and 4.3. Classifications of reliable change are based on the criteria suggested by Wise [111]: substantial changes are represented by RCI values that exceed the 95% CI for the difference among sessions. This indicates changes exceeding that which could be considered due to measurement error. RCI values exceeding the 90% CI or 80% CI, are classified as mild or minimal clinically reliable responses. RCIs that do not exceed the 80% CI are classified as indeterminate or unchanged.

In addition to numeric RCI statistics, comparison of individuals to the group for change between S1 and S2 are accomplished graphically with RCI plots. Figure 4.1 shows an example RCI plot of the walking smoothness index metric. The values measured for the walking smoothness index at S1 (X-axis) are plotted against S2 (Y-axis). The red diagonal line intersecting the figure represents an absence of change from S1 to S2.
The shaded gray diagonal areas represent confidence intervals based on standard error of measurement and the previously mentioned criteria suggested by Wise [111]. The green bands represent the mean value for S1 plus one and two standard deviations, respectively. As Figure 4.1 shows, the majority of participants demonstrate improvement between S1 and S2 for walking smoothness. Four smoothness indices fall below the diagonal line (indicating lack of response) and ten smoothness indices appear above the 95% confidence interval, indicating substantial response.

4.4 Results

Tables 4.1, 4.2, 4.3, and 4.4 contain results for CAP, WBM, GF, and individual limb GF metrics, respectively. We report statistics for each metric, including the mean ($\mu_{S1}$, $\mu_{S2}$) and standard deviation ($SD_{S1}$, $SD_{S2}$) for S1 (the average of the two trials at S1) and S2 (the average of the two trials at S2). Furthermore, results include the change in the value expressed as a difference and as a percent ($\Delta$ and $\Delta\%$), the standardized mean difference effect size, and the associated effect size confidence interval. We also report cane gait parameter results for participants #005 and #015 in Table 4.5. We include the mean ($\mu_{S1}$, $\mu_{S2}$) and coefficient of variation ($CV_{S1}$, $CV_{S2}$) for each cane metric.

We display several metrics of interest from Tables 4.1, 4.2, 4.3, 4.4, and 4.5 visually. At the group level, box plots in Figure 4.2 display average participant performance across the group for two trials within a session. The performance distributions for 9 selected metrics demonstrate the wide range of abilities between participants. Figures 4.2a, 4.2b,
Figure 4.1: The smoothness index metric as an example reliable change index plot. Participant session 1 (S1) scores are plotted against session 2 (S2) scores. Also plotted are confidence intervals (CI) and session 1 standard deviations (SD). Select individuals are labeled with their participant identification number (see Table 3.3 for participant characteristics).
and 4.2c show group changes for the selected CAP metrics of total AC duration (see Figure 4.2a), sit-to-stand duration (see Figure 4.2b), and vehicle transfer duration (see Figure 4.2c). The remaining box plots show performance on the selected WBM and GF metrics of cadence (see Figure 4.2d), step symmetry (see Figure 4.2e), walking speed ratio (see Figure 4.2f), involved side shank peak angular velocity (see Figure 4.2g), uninvolved side shank peak angular velocity (see Figure 4.2h), and double support percent (see Figure 4.2i).

To facilitate analysis and insights at the individual patient level, we display smoothness index (see Figure 4.1), gait speed (see Figure 4.3a), step regularity (see Figure 4.3b), involved side shank range of motion (see Figure 4.3c), and uninvolved side shank range of motion (see Figure 4.3d) as RCI plots. Furthermore, stance and swing phase plots are included for participant #015 in Figure 3.8 and for participant #005 in Figure 3.9 and Figure 4.4.

4.5 Discussion

We are investigating the insights that sensor-based quantifiable measures can supply in addition to observations by clinicians. While analyzing changes at the group level provides insight about the effects of therapy from a research perspective, the effects of rehabilitation on an individual basis can be established with wearable sensors and applied directly to patient care [125]. Consequently, we discuss the results of the wearable sensor data first at the group level and then as applied to changes exhibited for individual patients.
(a) Total duration.  
(b) Sit-to-stand duration.  
(c) Vehicle challenge duration.  
(d) Cadence.  
(e) Step symmetry.  
(f) Walking speed ratio (shag carpet to smooth flooring).  
(g) Involved side peak angular velocity.  
(h) Uninvolved side peak angular velocity.  
(i) Double support percent.

Figure 4.2: Metric box plots. Box plots provide a visual summary of the group’s distribution. The boxes encompass the interquartile range of the population. The horizontal line indicates the median of the distribution. The mean difference between session 1 (S1) and session 2 (S2) are provided: a) -40.49 s, b) -2.04 s, c) -13.06 s, d) 5.67 steps/min, e) 7.23%, f) 0.05, g) 18.69 °/s, h) 23.82 °/s, and i) -1.61%.
(a) Gait speed.

(b) Step regularity.

(c) Involved side shank range of motion (ROM).

(d) Uninvolved side shank range of motion (ROM).

Figure 4.3: Metric reliable change index plots. Individuals are labeled with their participant identification number (see Table 3.3 for participant characteristics). RCI plots from left to right and top to bottom: a) gait speed, b) step regularity, c) involved side shank range of motion, and d) uninvolved side shank range of motion.
Table 4.1: Clinical assessments of progress (CAP) metric results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\mu_{S1}$</th>
<th>$SD_{S1}$</th>
<th>$\mu_{S2}$</th>
<th>$SD_{S2}$</th>
<th>$\Delta$</th>
<th>$\Delta%$</th>
<th>$ES_{cat}$</th>
<th>ES</th>
<th>$CI_L$</th>
<th>$CI_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curvilinear walking duration</td>
<td>24.43</td>
<td>18.14</td>
<td>19.48</td>
<td>12.23</td>
<td>-4.95</td>
<td>-20.26</td>
<td>Large</td>
<td>-1.03*</td>
<td>-1.38</td>
<td>-0.68</td>
</tr>
<tr>
<td>Floor surface speed ratio</td>
<td>177.85</td>
<td>129.53</td>
<td>137.36</td>
<td>88.56</td>
<td>-40.49</td>
<td>-22.77</td>
<td>Large</td>
<td>-1.08*</td>
<td>-1.45</td>
<td>-0.72</td>
</tr>
<tr>
<td>Sit-to-stand duration</td>
<td>0.75</td>
<td>0.13</td>
<td>0.80</td>
<td>0.14</td>
<td>0.05</td>
<td>6.93</td>
<td>Small</td>
<td>0.48*</td>
<td>0.08</td>
<td>0.88</td>
</tr>
<tr>
<td>Stand-to-sit duration</td>
<td>6.84</td>
<td>6.10</td>
<td>4.79</td>
<td>3.03</td>
<td>-2.04</td>
<td>-29.89</td>
<td>Small</td>
<td>-0.49*</td>
<td>-0.83</td>
<td>-0.15</td>
</tr>
<tr>
<td>Vehicle challenge duration</td>
<td>12.94</td>
<td>12.05</td>
<td>11.13</td>
<td>7.13</td>
<td>-1.81</td>
<td>-13.98</td>
<td>Small</td>
<td>-0.34*</td>
<td>-0.56</td>
<td>-0.12</td>
</tr>
<tr>
<td>Walking speed</td>
<td>47.81</td>
<td>36.36</td>
<td>34.75</td>
<td>28.10</td>
<td>-13.06</td>
<td>-27.32</td>
<td>Moderate</td>
<td>-0.55*</td>
<td>-0.89</td>
<td>-0.22</td>
</tr>
<tr>
<td>NWS</td>
<td>0.47</td>
<td>0.22</td>
<td>0.57</td>
<td>0.28</td>
<td>0.10</td>
<td>21.10</td>
<td>Large</td>
<td>1.58*</td>
<td>1.06</td>
<td>2.09</td>
</tr>
<tr>
<td>NWS</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>20.39</td>
<td>Large</td>
<td>1.46*</td>
<td>0.98</td>
<td>1.95</td>
</tr>
</tbody>
</table>

CI = confidence interval, ES = effect size, $ES_{cat}$ = ES category, L = lower, NWS = normalized walking speed, $S1 = session 1$, $S2 = session 2$, $SD = standard deviation$, $U = upper$, $\mu = mean$, $\Delta = change in S1 and S2 means$, $\Delta\% = percent change in means$, * = significant at the 95% confidence level.

Figure 4.4: Stance and swing phase plot for participant #005 recorded from the first four gait cycles detected during trial 1 at session 1 testing.
Table 4.2: Whole body movement (WBM) metric results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\mu_{S1}$</th>
<th>$SD_{S1}$</th>
<th>$\mu_{S2}$</th>
<th>$SD_{S2}$</th>
<th>$\Delta$</th>
<th>$\Delta%$</th>
<th>$ES_{cat}$</th>
<th>ES</th>
<th>$CI_L$</th>
<th>$CI_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>COM RMS</td>
<td>0.08</td>
<td>0.06</td>
<td>0.11</td>
<td>0.10</td>
<td>0.03</td>
<td>37.88</td>
<td>Large</td>
<td>1.87*</td>
<td>1.26</td>
<td>2.48</td>
</tr>
<tr>
<td>COM Smoothness of RMS</td>
<td>0.75</td>
<td>0.59</td>
<td>1.10</td>
<td>1.04</td>
<td>0.35</td>
<td>46.73</td>
<td>Large</td>
<td>1.90*</td>
<td>1.28</td>
<td>2.51</td>
</tr>
<tr>
<td>Sit-to-stand RMS</td>
<td>0.65</td>
<td>0.64</td>
<td>0.85</td>
<td>0.67</td>
<td>0.20</td>
<td>31.43</td>
<td>Moderate</td>
<td>0.71*</td>
<td>0.41</td>
<td>1.01</td>
</tr>
<tr>
<td>Sit-to-stand peak angular velocity</td>
<td>84.02</td>
<td>37.96</td>
<td>72.46</td>
<td>44.47</td>
<td>-11.57</td>
<td>-13.76</td>
<td>Small</td>
<td>-0.37</td>
<td>-0.75</td>
<td>0.02</td>
</tr>
<tr>
<td>Smoothness index</td>
<td>1.60</td>
<td>0.65</td>
<td>2.01</td>
<td>0.97</td>
<td>0.41</td>
<td>25.76</td>
<td>Large</td>
<td>1.82*</td>
<td>1.22</td>
<td>2.42</td>
</tr>
<tr>
<td>Stand-to-sit RMS</td>
<td>0.38</td>
<td>0.45</td>
<td>0.36</td>
<td>0.30</td>
<td>-0.02</td>
<td>-4.13</td>
<td>None</td>
<td>-0.07</td>
<td>-0.30</td>
<td>0.16</td>
</tr>
<tr>
<td>Stand-to-sit peak angular velocity</td>
<td>126.87</td>
<td>36.75</td>
<td>118.45</td>
<td>42.08</td>
<td>-8.42</td>
<td>-6.64</td>
<td>Small</td>
<td>-0.26</td>
<td>-0.65</td>
<td>0.13</td>
</tr>
<tr>
<td>Vehicle load RMS</td>
<td>0.10</td>
<td>0.11</td>
<td>0.16</td>
<td>0.16</td>
<td>0.06</td>
<td>59.24</td>
<td>Large</td>
<td>1.37*</td>
<td>0.90</td>
<td>1.84</td>
</tr>
<tr>
<td>Vehicle load peak angular velocity</td>
<td>81.94</td>
<td>36.62</td>
<td>78.27</td>
<td>26.32</td>
<td>-3.67</td>
<td>-4.48</td>
<td>None</td>
<td>-0.10</td>
<td>-0.59</td>
<td>0.39</td>
</tr>
<tr>
<td>Vehicle unload RMS</td>
<td>0.16</td>
<td>0.12</td>
<td>0.29</td>
<td>0.37</td>
<td>0.13</td>
<td>82.25</td>
<td>Large</td>
<td>2.71*</td>
<td>1.83</td>
<td>3.59</td>
</tr>
<tr>
<td>Vehicle unload peak angular velocity</td>
<td>74.45</td>
<td>47.85</td>
<td>68.45</td>
<td>35.89</td>
<td>-5.99</td>
<td>-8.05</td>
<td>None</td>
<td>-0.15</td>
<td>-0.52</td>
<td>0.23</td>
</tr>
</tbody>
</table>

COM = center of mass, CI = confidence interval, ES = effect size, $ES_{cat}$ = ES category, L = lower, RMS = root mean square, S1 = session 1, S2 = session 2, SD = standard deviation, U = upper, $\mu$ = mean, $\Delta$ = change in S1 and S2 means, $\Delta\%$ = percent change in means, * = significant at the 95% confidence level.
Table 4.3: Gait features (GF) metric results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\mu_{S1}$</th>
<th>$SD_{S1}$</th>
<th>$\mu_{S2}$</th>
<th>$SD_{S2}$</th>
<th>$\Delta$</th>
<th>$\Delta%$</th>
<th>$ES_{cat}$</th>
<th>ES</th>
<th>$CI_L$</th>
<th>$CI_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cadence</td>
<td>64.88</td>
<td>17.67</td>
<td>70.54</td>
<td>20.32</td>
<td>5.67</td>
<td>8.73</td>
<td>Large</td>
<td>1.38*</td>
<td>0.93</td>
<td>1.83</td>
</tr>
<tr>
<td>Double support percent</td>
<td>33.79</td>
<td>11.97</td>
<td>32.19</td>
<td>13.72</td>
<td>-1.61</td>
<td>-4.76</td>
<td>Small</td>
<td>-0.49*</td>
<td>-0.68</td>
<td>-0.29</td>
</tr>
<tr>
<td>Gait cycle time</td>
<td>1.96</td>
<td>0.66</td>
<td>1.87</td>
<td>0.68</td>
<td>-0.09</td>
<td>-4.76</td>
<td>Moderate</td>
<td>-0.64*</td>
<td>-0.86</td>
<td>-0.41</td>
</tr>
<tr>
<td>Number of gait cycles</td>
<td>18.95</td>
<td>9.32</td>
<td>16.38</td>
<td>5.59</td>
<td>-2.58</td>
<td>-13.59</td>
<td>Large</td>
<td>-0.90*</td>
<td>-1.22</td>
<td>-0.59</td>
</tr>
<tr>
<td>Step length</td>
<td>0.21</td>
<td>0.07</td>
<td>0.23</td>
<td>0.06</td>
<td>0.02</td>
<td>10.62</td>
<td>Moderate</td>
<td>0.64*</td>
<td>0.34</td>
<td>0.94</td>
</tr>
<tr>
<td>Step length normalized</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>9.93</td>
<td>Moderate</td>
<td>0.54*</td>
<td>0.26</td>
<td>0.82</td>
</tr>
<tr>
<td>Step regularity</td>
<td>37.23</td>
<td>22.36</td>
<td>46.91</td>
<td>28.11</td>
<td>9.62</td>
<td>25.81</td>
<td>Large</td>
<td>1.31*</td>
<td>0.87</td>
<td>1.75</td>
</tr>
<tr>
<td>Stride regularity</td>
<td>40.88</td>
<td>22.73</td>
<td>51.53</td>
<td>24.45</td>
<td>10.65</td>
<td>26.04</td>
<td>Moderate</td>
<td>0.55*</td>
<td>0.13</td>
<td>0.96</td>
</tr>
<tr>
<td>Step symmetry</td>
<td>63.57</td>
<td>27.50</td>
<td>70.80</td>
<td>26.31</td>
<td>7.23</td>
<td>11.37</td>
<td>Small</td>
<td>0.35</td>
<td>-0.00</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*CI = confidence interval, ES = effect size, $ES_{cat}$ = ES category, L = lower, S1 = session 1, S2 = session 2, SD = standard deviation, U = upper, $\mu$ = mean, $\Delta$ = change in S1 and S2 means, $\Delta\%$ = percent change in means, * = significant at the 95% confidence level.

Table 4.4: Individual limb gait features (GF) metric results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\mu_{S1}$</th>
<th>$SD_{S1}$</th>
<th>$\mu_{S2}$</th>
<th>$SD_{S2}$</th>
<th>$\Delta$</th>
<th>$\Delta%$</th>
<th>$ES_{cat}$</th>
<th>ES</th>
<th>$CI_L$</th>
<th>$CI_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involved side peak angular velocity</td>
<td>190.09</td>
<td>67.88</td>
<td>208.79</td>
<td>72.50</td>
<td>18.69</td>
<td>9.83</td>
<td>Large</td>
<td>1.52*</td>
<td>0.85</td>
<td>2.20</td>
</tr>
<tr>
<td>Involved side shank range of motion</td>
<td>47.34</td>
<td>13.22</td>
<td>50.02</td>
<td>11.44</td>
<td>2.68</td>
<td>5.66</td>
<td>Large</td>
<td>1.20*</td>
<td>0.66</td>
<td>1.73</td>
</tr>
<tr>
<td>Left side peak angular velocity</td>
<td>195.98</td>
<td>61.01</td>
<td>213.09</td>
<td>68.84</td>
<td>17.12</td>
<td>8.73</td>
<td>Large</td>
<td>1.30*</td>
<td>0.88</td>
<td>1.73</td>
</tr>
<tr>
<td>Left side shank range of motion</td>
<td>47.24</td>
<td>12.66</td>
<td>50.93</td>
<td>12.59</td>
<td>3.69</td>
<td>7.80</td>
<td>Large</td>
<td>1.73*</td>
<td>1.18</td>
<td>2.29</td>
</tr>
<tr>
<td>Right side peak angular velocity</td>
<td>217.65</td>
<td>43.80</td>
<td>244.53</td>
<td>51.02</td>
<td>26.88</td>
<td>12.35</td>
<td>Large</td>
<td>2.02*</td>
<td>1.37</td>
<td>2.68</td>
</tr>
<tr>
<td>Right side shank range of motion</td>
<td>50.41</td>
<td>10.91</td>
<td>54.70</td>
<td>9.43</td>
<td>4.29</td>
<td>8.50</td>
<td>Large</td>
<td>1.45*</td>
<td>0.98</td>
<td>1.93</td>
</tr>
<tr>
<td>Uninvolved side peak angular velocity</td>
<td>231.38</td>
<td>39.33</td>
<td>255.20</td>
<td>40.62</td>
<td>23.82</td>
<td>10.30</td>
<td>Large</td>
<td>1.91*</td>
<td>1.05</td>
<td>2.77</td>
</tr>
<tr>
<td>Uninvolved side shank range of motion</td>
<td>51.91</td>
<td>11.96</td>
<td>55.69</td>
<td>8.40</td>
<td>3.78</td>
<td>7.28</td>
<td>Large</td>
<td>1.56*</td>
<td>0.87</td>
<td>2.25</td>
</tr>
</tbody>
</table>

*CI = confidence interval, ES = effect size, $ES_{cat}$ = ES category, L = lower, S1 = session 1, S2 = session 2, SD = standard deviation, U = upper, $\mu$ = mean, $\Delta$ = change in S1 and S2 means, $\Delta\%$ = percent change in means, * = significant at the 95% confidence level.
Table 4.5: Cane gait parameter results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>#015 $\mu_{S1}$ (#015 CV$_{S1}$)</th>
<th>#015 $\mu_{S2}$ (#015 CV$_{S2}$)</th>
<th>#005 $\mu_{S1}$ (#005 CV$_{S1}$)</th>
<th>#005 $\mu_{S2}$ (#005 CV$_{S2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cycles per trial</td>
<td>8, 7</td>
<td>8, 7</td>
<td>14, 12</td>
<td>9, 9</td>
</tr>
<tr>
<td>Cycle duration in milliseconds</td>
<td>1377.79</td>
<td>1221.33</td>
<td>1398.95</td>
<td>1308.21</td>
</tr>
<tr>
<td>Left stance %</td>
<td>-4.04%</td>
<td>-5.18%</td>
<td>-5.29%</td>
<td>-7.68%</td>
</tr>
<tr>
<td>Right stance %</td>
<td>-4.45%</td>
<td>-3.73%</td>
<td>-7.37%</td>
<td>-4.53%</td>
</tr>
<tr>
<td>Cane stance %</td>
<td>58.82%</td>
<td>58.18%</td>
<td>58.82%</td>
<td>59.26%</td>
</tr>
<tr>
<td>Cane stance ratio</td>
<td>-6.11%</td>
<td>-8.48%</td>
<td>-23.04%</td>
<td>-13.23%</td>
</tr>
<tr>
<td>Left mid-swing $^\circ$/s</td>
<td>282.86</td>
<td>321.70</td>
<td>249.76</td>
<td>292.47</td>
</tr>
<tr>
<td>Right mid-swing $^\circ$/s</td>
<td>286.91</td>
<td>320.81</td>
<td>240.43</td>
<td>276.11</td>
</tr>
<tr>
<td>Cane mid-swing $^\circ$/s</td>
<td>-9.79%</td>
<td>-14.71%</td>
<td>-18.74%</td>
<td>-9.93%</td>
</tr>
<tr>
<td>Cane swing offset</td>
<td>140.88</td>
<td>153.24</td>
<td>58.45</td>
<td>66.61</td>
</tr>
<tr>
<td>Double support %</td>
<td>1.24</td>
<td>1.30</td>
<td>1.33</td>
<td>1.49</td>
</tr>
<tr>
<td>Triple support %</td>
<td>-7.26%</td>
<td>-9.17%</td>
<td>-22.88%</td>
<td>-13.20%</td>
</tr>
<tr>
<td>CV = coefficient of variation, $S1 = session 1$, $S2 = session 2$, and $\mu = mean$.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.5.1 Group Responsiveness to Therapy

Clinical Assessments of Progress

As a group, AC metrics related to CAP demonstrate moderate improvements based on the magnitude of effect sizes (see Table 4.1), particularly for total AC duration and walking speed. For total AC duration (see Figure 4.2a), there is a 22.77% decrease in the total time required to complete the AC. The increased ambulatory capabilities, specifically completing the curvilinear section of the AC 20.26% faster, indicate improved functionality. Responsiveness on the vehicle load/unload challenge is moderate with a 27.32% decrease in duration observed on average (see Figure 4.2c). These changes in simple, easy-to-understand metrics corroborate for clinicians that patients as a group are functionally improving [126–128]. Task durations provide an efficient performance summary because their execution incorporates temporal-spatial variables of locomotion, physical capabilities, movement strategies, and navigation skills [125]. For example, the vehicle transfer requires the simultaneous involvement of both physical and cognitive capabilities, presenting a test of awareness, upper and lower limb coordination, and strength.

Walking speed, the ratio of distance and duration, and stature-normalized walking speed (NWS), are typical metrics for comparison among populations and for indication of ambulatory improvements. Group average performance during both sessions (walking speed S1 0.47 ± 0.22 m/s; walking speed S2 0.57 ± 0.28 m/s) as well as the large responsiveness across time (see Figure 4.3a for an RCI plot) are in agreement with previous published studies analyzing walking speed in post-stroke populations [126,129]. During
the course of recovery, increased walking speed can be the result of numerous changes: larger step or stride lengths, higher cadences, shorter double or single support or swing times, higher joint range of motion, and increased ground reaction forces during push-off [130,131]; however, a faster walking speed does not necessarily imply that the underlying gait has become more stable or that movement coordination has improved [130]. Performance should ideally be measured with metrics that take into consideration the profile of motion displayed by the musculoskeletal system to gain insight into recovery, such as whole body movement metrics and gait features.

**Whole Body Movements**

A rudimentary quantification of movement patterns can be achieved by reducing the musculoskeletal system into a point mass moving through space. This simplified approach allows for summary quantifications of whole body movement patterns (see Table 4.2). For example, the time-normalized root mean square of COM acceleration provides a quantitative description of movement intensity [12]. COM acceleration RMS is also used as a standard gait analysis metric to explore relationships between stability, walking speed, and limb strength [9, 132, 133]. For the AC, we observe large levels of responsiveness for RMS during linear path gait. While there is a demonstrated relationship between COM RMS and walking speed [9], the large responsiveness across time is due in part to the participant-dictated speed of movement. Interpretation is also contingent on the high patient reliance (90%) on assistive devices to support locomotion. During the vehicle transfer, the ES for COM RMS during the load and unload tasks suggest substantial progress from one week of rehabilitation therapy. The increased RMS observed during the
vehicle transfer reveals deliberate and pronounced trunk movements, indicating gains in strength and mobility over time. Similar changes in COM RMS are also present on the chair task.

Another WBM metric, the smoothness index of walking (see Figure 4.1), characterizes gait harmonics to quantify cyclical movements independent of speed [9]. The index is computed as the ratio of even to odd harmonics of trunk motion during gait and represents coordination [134]. The calculated ES for change in smoothness index emphasizes the influence rehabilitation has on developing a more stable pattern of locomotion. As a group, the participants demonstrate a 25.76% improvement in the smoothness index. Changes in the cyclical behavior of global movement patterns indicate that recovery improves the movement itself, regardless of how fast participants are moving [9,134].

Gait Features

Changes in metrics describing gait quality in terms of symmetry, regularity, and consistency were observed during the straight path portion of the AC (see Tables 4.3 and 4.4). During one week of rehabilitation, the increased walking speed is accompanied by an average increase of 8.73% in cadence. A faster cadence suggests that gait cycle duration should decrease, but the 4.76% decrease of cycle duration is less pronounced than the cadence increase. Another important outcome is the 4.76% decrease in the amount of double support in the gait cycle. These changes indicate the therapeutic efforts over the course of one week produced positive global changes in participant gait features. In addition, improvement is observed in gait consistency, measured with stride and step regularity (see Figure 4.3b for an RCI plot). These metrics indicate that patients produce
more consistent walking patterns over one week, increasing the load carried by the affected limb.

We also observe changes in individual leg movements. Large levels of responsiveness exist in peak angular velocity, measured at each shank. Along with faster leg movements during the swing phase, there is a strong indication of increased limb range of motion during gait. To perform sub-group analyses of patients with hemiparesis, each limb is considered involved or uninvolved, instead of left and right. The re-categorization produces a slightly different ES for shank peak angular velocity (see Figures 4.2g and 4.2h) and range of motion (see Figures 4.3c and 4.3d). Tracking changes in the affected side of the body offers additional insight for clinicians treating stroke patients and injuries affecting one side of the body more than the other side. These range of motion improvements indicate that patients are increasing hemiplegic limb control by re-integrating knee flexion, as opposed to swinging the leg in a motion facilitated by the hip (hip hiking) [135].

4.5.2 Individual Responsiveness to Therapy

Our analyses of RCI plots suggest that recovery is not consistent for all patients over one week of inpatient rehabilitation (see Figures 4.1 and 4.3). For example, participant #014 demonstrates a substantial amount of recovery compared to the rest of the participants as assessed by RCI plots. This finding is corroborated by the conventional method of using the FIM to characterize functioning at admission (#014, admission FIM = 61) and discharge (#014, discharge FIM = 103). By contrast, participant #015 does not
demonstrate significant change in smoothness of walking or step regularity. At admission to the inpatient facility the functional capabilities for this participant were close to independent (#015, admission FIM = 87), rendering a small window for improvement across time (#015, discharge FIM = 113). The RCI visualization of performance at the individual level can track progress by assessing performance on multiple metrics. For example, a few participants with moderate responsiveness for walking speed (#007, #015, and #020) do not show change in the smoothness index metric and vice versa (#004). Therefore, analysis of multiple metrics, such as smoothness index along with walking speed, highlights the differences in individual recovery.

We also utilize wearable IMUs to perform fine-grained, quantitative assessment of gait and cane usage for two patients undergoing stroke rehabilitation. The metrics we compute reveal several differences between the two patients (see Table 4.5). For example, participant #015 has shorter gait cycles, spends less time in double and triple support, and swings both his legs and cane with higher angular velocity. Participant #015 also generally demonstrates lower variability in gait and cane usage, with the exception of left leg mid-swing angular velocity and the percentage of time spent in double support. Considering that participant #015 has a left side involvement (see Table 3.3), this anomaly is understandable.

Stance and swing plots allow several observations of the timing and variability of individual gait cycles. For example, inspecting Figure 3.8 reveals several important characteristics of participant #015’s gait and cane usage. First, the cane correctly swings in phase with the affected (left) leg. Second, the initial double support period appears
more consistent in occurrence and duration than the second double support period. Figure 3.9 reveals participant #005 has more variable gait. In particular, he swings his cane such that there is only one period of triple support each cycle, indicating asymmetrical cane support between the first and second double support periods. This is different cane movement behavior than participant #005 exhibited one week earlier at S1 testing (see Figure 4.4). Additional analysis of changes after one week of inpatient rehabilitation reveal each participant had a reduction in total cycle duration, 11.4% for participant #015 and 6.5% for participant #005. Similar improvements are seen in stance and double/triple support percentages. The observed differences between participants #015 and #005 are corroborated by participant #005’s lower admission and discharge FIM scores (see Table 3.3).

Wearable IMU data can also be used to detect incorrect cane use. Figure 4.4 shows #005 exhibiting incorrect cane use [99] by swinging the cane in phase with the ipsilateral leg for the first detected gait cycle (GC #0) of the first trial. Algorithms operating on IMU data can also automatically detect when cane users hold the cane in the air for multiple steps [99] by detecting a missing cane IC within a gait cycle.

In summary, inpatient rehabilitation contains a wide spectrum of challenges that are tackled uniquely by different patients, depending on their pre-morbid state, injury, drive to improve, and compensatory strategies. Wearable IMUs provide a viable platform for gaining insight into these complex recovery processes by measuring changes in movement profiles over the course of one week of therapy. We utilize an ecological environment to enable data collection to capture performance of real-world challenges with the ambulation
circuit. Several movement features exhibit statistically significant differences in value from session one to session two, which indicates that these may be practical metrics for clinicians to use in addition to observation to quantify gait improvement. Specifically, metrics showing the greatest change are walking speed, stride regularity, acceleration root mean square, walking smoothness, shank peak angular velocity, and shank range of motion. Of these metrics, only walking speed does not make use of wearable inertial sensor data, indicating that wearable sensors can capture details about changes in movement patterns that cannot be acquired from standardized subjective clinical assessments. Furthermore, with the large number of assistive device users in the United States, technology to monitor device usage for gait analysis and safety is of increasing importance. Consequently, we investigate the application of wearable inertial sensors for fine-grained, quantitative assessment of cane and lower limb movement during straight-path gait for two patients undergoing inpatient rehabilitation for stroke recovery. The gait parameters computed and sensor data visualization introduced provide a foundation for future IMU monitoring of gait and assistive device movements.

The sensor-based mobility metrics and associated change statistics presented thus far are useful measures on their own; however, these mobility quantifications have additional utility as indicators of patient functioning at discharge. In the next chapter, we present a computational framework and results for utilizing sensor-based metrics as features to machine learning algorithms to predict discharge clinical outcomes.
In addition to computing changes in AC wearable sensor metrics, we also design computational methods to predict patient clinical outcomes at discharge based on the collected sensor data. In this chapter, we introduce our Hybrid Clinical Sensor Prediction algorithmic approach that fuses sensor data and clinical information to perform clinical outcome prediction. To improve prediction results, we also introduce our Joint Patient Prediction algorithmic approach in this chapter. The aim of clinical outcome prediction is to provide evidence in support of our hypothesis that sensor data and machine learning techniques can provide insights on rehabilitation progress.

5.1 Prior Work on Clinical Outcome Prediction

Several studies have built models to predict clinical outcomes, such as assessment scores and hospital length of stay, for patients at inpatient rehabilitation facilities. These studies can be grouped into two categories based on the features that are used as predictive variables: non-technology-based clinical metrics and technology-based metrics.

5.1.1 Clinical Predictors

Sonoda et al. [136] used linear regression to predict discharge FIM motor scores for 131 first-stroke patients at an inpatient rehabilitation hospital. Features used for prediction
included age, days since onset of the stroke to admission, admission FIM cognitive and motor scores, and the reciprocal of the admission FIM motor score. The regression models yielded correlations of $r = 0.89$ for a training group and $r = 0.93$ for a validation group. Similar studies predicting FIM scores using only clinical predictors include Matsugi et al. [137], Jeremic et al. [138], Fujiwara et al. [139], Tsuji et al. [140], and Jeong et al. [141]. Jeong and colleagues predicted discharge FIM to investigate the differences between two stroke groups: 4,311 patients admitted to acute hospitals ($R^2 = 0.78$) and 1,941 patients admitted to convalescent hospitals ($R^2 = 0.66$).

Sakurai et al. [142] investigated the predictive ability of admission FIM scores of patients with stroke ($N = 286$) to determine functional independence. Independence was classified as either completely dependent/requiring maximal assistance, moderately dependent/requiring minimal assistance, or completely independent/requiring supervision. The study concluded the motor and cognitive scores of the FIM are valid predictors of functional independence, whereas the individual FIM tasks alone are not useful predictors.

In addition to predicting clinical assessments scores such as the FIM, a fair amount of research has been performed to predict individual patient length of stay [143–146]. Tan et al. [143] considered motor function on admission and the effects of patients’ socioeconomic status and family structure on LOS for patients with stroke. Franchignoni et al. [144] found individual FIM task scores on admission to be strong predictors of patients’ LOS, with the tasks related to transfers having the highest predictive ability. Dvorak et al. [147] used regression modeling and Wilcoxon’s rank sum test to assess the association between independent variables and the outcome of Asia Spinal Cord Injury (ASIA) motor scores. The
researchers found level of education, spasticity, age, and surgical treatment to be strong predictors of motor recovery in traumatic central cord syndrome participants. Similarly, Brosseau et al. [145] used successive multiple linear regressions to predict length of stay for inpatient rehabilitation. In this paper they also looked at the strength of the predictors and discovered that age, functional status at one week after admission, perceptual status, and balance status accounted for 43.6% of the total variance in the rehabilitation LOS for stroke patients. Furthermore, functional status at admission, rehabilitation program, motor status, communication problems, and medical complications were indirect predictors of LOS.

5.1.2 Technology-based Predictors

In addition to utilizing clinical metrics, several studies have investigated mapping technology-based measurements into clinical assessment scores. Zariffa et al. [148] considered the relationship between robot-collected kinematic data and the Graded Redefined Assessment of Strength, Sensibility, and Prehension, Action Research Arm Test (ARAT), and Spinal Cord Independence Measure. Olesh et al. [149] collected data from a Kinect sensor and mapped it to the Fugl-Meyer Assessment (FMA) and ARAT. Similarly, Wang et al. [150] mapped accelerometer data from upper arm movements to the FMA for shoulder-elbow. Finally, Simila et al. [151] analyzed lower-back accelerometer data to estimate Berg Balance Scale scores for identifying subjects with high or low risk of falling.

The aforementioned studies have primarily examined the relationships between technology-
based metrics and associated clinical rating scales. These studies do not utilize collected data to project into the future and predict discharge assessment scores. On the other hand, Mostafavi et al. [152] predicted several clinical scores using metrics collected from the Kinensiological Instrument for Normal and Altered Reaching Movements (KINARM) rehabilitation robotic device. Data were collected from two tasks: an upper limb reaching task and a positioning task. Using linear regression, robot-based measurements from these tasks for 126 stroke patients were mapped to predictions of FIM total score, FIM motor score, LOS, the Purdue pegboard test, and the modified Ashworth score with statistically significant accuracy (see [152], Table 1). This work differs from our study in the choice of technology (robotic device vs. wearable inertial sensors), the involved parts of the body (upper limb vs. whole body), and the timeline of participant data collection.

The clinical outcome prediction methods we introduce aim to further move the field forward in several aspects. First, we combine clinical metrics with movement data collected using IMU technology that is relatively inexpensive and unobtrusive. Furthermore, while wearing the inertial sensors, participants in our study perform a sequence of ambulatory tasks that are representative of the patient’s ecological environment. Potentially, our sensor platform will be able to collect movement profiles from various therapy tasks and map this information into clinical assessment scores. Finally, previous studies have focused on single, homogeneous populations, often only considering patients who do not have other medical complications [136, 142] or have a LOS greater than a certain duration [142]. While these restrictions are useful to narrow the scope of findings, we are investigating several rehabilitation populations with varying LOS and comorbidities. We do not distinguish
between medical conditions because our proposed wearable sensor platform, algorithms, and machine learning methodologies are applicable to all individuals undergoing inpatient rehabilitation. In summary, we aim to lay the foundation for a monitoring system that fuses sensor data collected during therapy with medical record information to predict discharge clinical scores at any point during rehabilitation.

To perform clinical outcome prediction, our approach consists of three steps:

1. Collect patient data from two sources:
   
   (a) Patient wearable inertial sensor data as they ambulated throughout an ecological environment (see Chapter 3).

   (b) Medical records both from patients who participated in the wearable sensor study and from patients who were not involved in the wearable sensor study. The latter are collected to provide additional training instances for comparison to a baseline FIM prediction model using only clinical features available upon admission (see Section 5.2).

2. Compute sensor-based features and analyze each feature’s predictive utility (see Sections 5.3 and 5.6.1).

3. Train and test machine learning models to predict discharge FIM scores (see Section 5.6).

The following sections provide details on each of these steps.
5.2 Additional Medical Record Data

For training baseline admission models, we collected additional data from the inpatient rehabilitation hospital during October 2010-December 2013. The dataset contains data for 4,936 patients of various rehabilitation impairment categories (RICs). These patients did not participate in the AC wearable sensor study. Consequently, this dataset is henceforth referred to as the non-AC (NAC) dataset. Table 5.1 provides patient demographics and FIM performance summaries for the NAC patient data. As can be seen in Table 5.1, NAC patients are primarily in the RICs of stroke and lower extremity joint replacement.

The information in this dataset represents traditional medical record data. Other projects have focused on mining medical records and predicting patient health from this information alone. Our goal is to show that prediction of rehabilitation outcomes can be enhanced by including sensor data in the predictive model alongside medical record data. As we will see, including both sources of information can present a challenge because the medical record data is data rich (in number of patient records) while the sensor data is fairly sparse.
<table>
<thead>
<tr>
<th>RIC</th>
<th>N (%)</th>
<th>Males (SD)</th>
<th>Age (%)</th>
<th>No comorbidities (SD)</th>
<th>LOS (SD)</th>
<th>FIM&lt;sub&gt;admission&lt;/sub&gt; (SD)</th>
<th>FIM&lt;sub&gt;discharge&lt;/sub&gt; (SD)</th>
<th>FIM&lt;sub&gt;admission&lt;/sub&gt; (SD)</th>
<th>FIM&lt;sub&gt;discharge&lt;/sub&gt; (SD)</th>
<th>FIM&lt;sub&gt;motor admission&lt;/sub&gt; (SD)</th>
<th>FIM&lt;sub&gt;motor discharge&lt;/sub&gt; (SD)</th>
<th>Total RER (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amputation of lower extremity</td>
<td>96 (1.94%)</td>
<td>61 (63.54%)</td>
<td>69.91</td>
<td>32 (33.35%)</td>
<td>12.69</td>
<td>25.56 (6.18)</td>
<td>30.19 (4.78)</td>
<td>37.95 (11.93)</td>
<td>56.62 (14.60)</td>
<td>2.23 (1.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiac disorders</td>
<td>344 (6.97%)</td>
<td>202 (58.72%)</td>
<td>78.12</td>
<td>179 (52.03%)</td>
<td>11.04</td>
<td>23.74 (6.05)</td>
<td>29.67 (4.68)</td>
<td>44.29 (9.67)</td>
<td>61.24 (13.07)</td>
<td>2.61 (1.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fracture of lower extremity</td>
<td>338 (6.85%)</td>
<td>117 (34.62%)</td>
<td>80.27</td>
<td>240 (71.01%)</td>
<td>12.37</td>
<td>23.64 (4.49)</td>
<td>29.29 (5.48)</td>
<td>35.60 (9.63)</td>
<td>56.35 (13.71)</td>
<td>2.37 (1.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guillain Barre syndrome</td>
<td>30 (0.61%)</td>
<td>17 (56.67%)</td>
<td>48.43</td>
<td>15 (11.74)</td>
<td>14.80</td>
<td>26.30 (10.48)</td>
<td>32.10 (6.39)</td>
<td>35.67 (16.39)</td>
<td>61.73 (21.21)</td>
<td>2.63 (1.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMT-BSCI</td>
<td>86 (1.74%)</td>
<td>51 (59.30%)</td>
<td>45.84</td>
<td>54 (63.54%)</td>
<td>18.91</td>
<td>21.40 (7.53)</td>
<td>29.91 (5.25)</td>
<td>33.48 (15.12)</td>
<td>56.09 (19.17)</td>
<td>2.60 (1.81)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMT-NBSCI</td>
<td>155 (3.14%)</td>
<td>79 (19.81%)</td>
<td>64.36</td>
<td>110 (70.97%)</td>
<td>11.20</td>
<td>25.88 (5.34)</td>
<td>31.77 (4.13)</td>
<td>37.92 (10.77)</td>
<td>61.70 (13.25)</td>
<td>3.02 (1.66)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>386 (7.82%)</td>
<td>203 (17.19)</td>
<td>71.80</td>
<td>166 (43.01%)</td>
<td>11.44</td>
<td>24.36 (6.14)</td>
<td>29.92 (5.97)</td>
<td>40.74 (11.43)</td>
<td>60.41 (15.16)</td>
<td>2.72 (2.26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-traumatic brain injury</td>
<td>295 (5.98%)</td>
<td>161 (17.59)</td>
<td>62.82</td>
<td>129 (43.73%)</td>
<td>13.16</td>
<td>17.17 (7.00)</td>
<td>24.50 (6.75)</td>
<td>41.86 (13.72)</td>
<td>62.74 (17.27)</td>
<td>2.68 (2.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-traumatic spinal cord injury</td>
<td>176 (3.57%)</td>
<td>94 (15.29)</td>
<td>66.16</td>
<td>129 (73.30%)</td>
<td>15.49</td>
<td>26.38 (9.62)</td>
<td>31.23 (6.09)</td>
<td>35.32 (12.92)</td>
<td>56.43 (18.06)</td>
<td>2.90 (1.85)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurological conditions</td>
<td>279 (5.65%)</td>
<td>128 (14.78)</td>
<td>68.93</td>
<td>194 (69.53%)</td>
<td>11.37</td>
<td>25.32 (5.92)</td>
<td>30.37 (6.62)</td>
<td>38.36 (11.18)</td>
<td>59.37 (16.58)</td>
<td>2.83 (1.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other orthopaedic</td>
<td>281 (5.69%)</td>
<td>100 (74.38%)</td>
<td>68.49</td>
<td>72 (47.98%)</td>
<td>18.10</td>
<td>25.71 (6.54)</td>
<td>32.73 (4.99)</td>
<td>43.42 (14.58)</td>
<td>67.57 (2.07)</td>
<td>2.88 (2.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>1 (0.02%)</td>
<td>0 (0.00%)</td>
<td>88 (N/A)</td>
<td>1 (N/A)</td>
<td>15 (N/A)</td>
<td>19 (N/A)</td>
<td>25 (N/A)</td>
<td>30 (N/A)</td>
<td>47 (N/A)</td>
<td>1.53 (N/A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain syndromes</td>
<td>10 (0.20%)</td>
<td>2 (20.00%)</td>
<td>73.60</td>
<td>8 (80.00%)</td>
<td>10 (N/A)</td>
<td>23.40 (6.99)</td>
<td>30.50 (7.74)</td>
<td>40.20 (5.19)</td>
<td>55.20 (16.61)</td>
<td>2.72 (3.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulmonary disorders</td>
<td>59 (1.20%)</td>
<td>31 (52.54%)</td>
<td>74.32</td>
<td>34 (57.63%)</td>
<td>10.06</td>
<td>25.80 (4.85)</td>
<td>29.71 (8.61)</td>
<td>41.56 (4.30)</td>
<td>59.76 (12.16)</td>
<td>2.82 (2.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower joint replacement</td>
<td>259 (5.07%)</td>
<td>154 (73.77%)</td>
<td>73.56</td>
<td>54 (72.72%)</td>
<td>9.49</td>
<td>27.20 (5.30)</td>
<td>32.23 (5.29)</td>
<td>43.42 (3.17)</td>
<td>67.57 (10.10)</td>
<td>2.88 (1.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>1 (0.02%)</td>
<td>1 (100.00%)</td>
<td>17 (N/A)</td>
<td>0 (N/A)</td>
<td>7 (N/A)</td>
<td>28 (N/A)</td>
<td>35 (N/A)</td>
<td>50 (N/A)</td>
<td>43 (N/A)</td>
<td>0 (N/A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>1270 (25.73%)</td>
<td>659 (51.89%)</td>
<td>71.55</td>
<td>943 (74.25%)</td>
<td>16.71</td>
<td>17.40 (9.17)</td>
<td>25.24 (6.56)</td>
<td>36.25 (6.54)</td>
<td>55.84 (14.47)</td>
<td>2.14 (1.70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traumatic brain injury</td>
<td>272 (5.51%)</td>
<td>188 (69.12%)</td>
<td>57.95</td>
<td>140 (51.47%)</td>
<td>17.52</td>
<td>13.65 (14.06)</td>
<td>22.28 (6.70)</td>
<td>38.71 (7.79)</td>
<td>61.19 (18.17)</td>
<td>2.55 (23.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traumatic spinal cord injury</td>
<td>113 (2.29%)</td>
<td>91 (80.53%)</td>
<td>50.71</td>
<td>64 (56.64%)</td>
<td>28.61</td>
<td>27.64 (20.79)</td>
<td>32.60 (6.37)</td>
<td>32.60 (3.42)</td>
<td>48.12 (13.47)</td>
<td>1.59 (21.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4936 (24.44%)</td>
<td>2144 (49.51%)</td>
<td>70.24</td>
<td>3188 (64.59%)</td>
<td>13.63</td>
<td>22.12 (8.99)</td>
<td>28.46 (7.67)</td>
<td>38.58 (6.27)</td>
<td>59.53 (12.98)</td>
<td>2.60 (1.80)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A = admission, BSCI = brain injury or spinal cord injury, cog = cognitive, D = discharge, FIM = Functional Independence Measure, LOS = length of stay, MMT = major multiple trauma, N = count, NBSCI = no brain injury or spinal cord injury, RER = rehabilitation efficiency ratio, RIC = rehabilitation impairment category, SD = standard deviation.
Useful clinical outcomes to predict include discharge FIM motor and cognitive scores, which represent patient functioning at the end of inpatient rehabilitation (see Section 3.1 for a description of the FIM assessment). To see the distribution and changes of these scores between admission and discharge, Figure 5.1 shows box-and-whisker plots for AC and NAC participant motor and cognitive FIM scores. Since AC performance metrics primarily measure motor functioning, our models focus on predicting the FIM motor score. To further explore the predictive abilities of the AC, we also train models to predict FIM cognitive and individual FIM item scores.

5.3 Predictor Variables

As AC participants undergo rehabilitation, data becomes available at four points displaced in time: admission, AC S1, AC S2, and discharge. We utilize metrics from data collected at admission, AC S1, and AC S2 as inputs to machine learning models, which are trained to predict FIM scores at discharge.

5.3.1 Admission Predictors

Data available at admission for both AC and NAC patients include patient characteristics such as age, gender, and RIC, as well as FIM task scores (see Table 5.2 for all admission features). In addition, we include the reciprocal of the FIM motor score as suggested by Sonoda and colleagues [136]. Although additional data from medical records are available, we only include features that apply to all populations. For example, the number
Figure 5.1: Admission and discharge Functional Independence Measure (FIM) scores for motor and cognitive FIM. AC represents the participants in the ambulation circuit study and NAC represents the non-AC patients used for training baseline admission models.
Table 5.2: Features extracted from medical records available at admission.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient characteristics</td>
<td>Age</td>
<td>Age in years</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>Male or female</td>
</tr>
<tr>
<td></td>
<td>RIC</td>
<td>Rehabilitation impairment category. Table 5.1 lists RICs in the NAC dataset</td>
</tr>
<tr>
<td></td>
<td>Comorbidity tier</td>
<td>No relevant comorbidities, tier 1 (most severe/expensive), tier 2 (medium severe/expensive), tier 3 (least severe/expensive)</td>
</tr>
<tr>
<td></td>
<td>Case mix group (CMG) relative weight</td>
<td>LOS modifier determined by presence of comorbidities and complications</td>
</tr>
<tr>
<td>Aggregated FIM</td>
<td>FIM&lt;sub&gt;A&lt;/sub&gt; motor score</td>
<td>Sum of the 13 FIM motor task scores</td>
</tr>
<tr>
<td></td>
<td>FIM&lt;sub&gt;A&lt;/sub&gt; cognitive score</td>
<td>Sum of the 5 FIM cognitive task scores</td>
</tr>
<tr>
<td></td>
<td>Reciprocal FIM&lt;sub&gt;A&lt;/sub&gt; motor score</td>
<td>Reciprocal of admission FIM score [137]</td>
</tr>
<tr>
<td>Individual tasks</td>
<td>FIM</td>
<td>17 FIM&lt;sub&gt;A&lt;/sub&gt; task scores 17 total scores, one score for each FIM task</td>
</tr>
</tbody>
</table>

A = admission, D = discharge, FIM = Functional Independence Measure, NAC = non-ambulation circuit, RIC = rehabilitation impairment category, SD = standard deviation.

of days since stroke onset is only applicable to stroke populations and is not included as a predictor.

### 5.3.2 AC Predictors

Sensor-based metrics of AC performance are grouped into three categories: clinical assessments of progress, whole body movements, and gait features (see Section 3.5). CAP metrics refer to commonly-used approaches for assessing mobility in a clinical setting, such as the duration of a task (see Table 3.4). WBM metrics are computed from the sensor placed on the COM (see Table 3.5). Finally, gait features refer to quantifications of steps and strides while walking (see Table 3.6). Gait features are primarily computed from gait cycles derived from the gait cycle event detection algorithm applied to the shank sensor.
angular velocity signals [50]. Tables 3.4, 3.5, and 3.6 in Chapter 3 summarize the CAP, WBM, and GF metrics, respectively.

Often, motor skills on only one side of the body are affected, called the involved or paretic side. Since several metrics computed in Table 3.6 are based on the left or right shank (e.g., shank peak angular velocity), we recast the left and right metrics as greater or lesser in value. For example, participant #001 exhibits average left shank peak angular velocity of 208.61 °/s, while right shank is 222.07 °/s at S1 testing. For #001, the left side is cast as the lesser peak angular velocity limb, and the right side as the greater. This classification aligns with the medical record data, which reports participant #001 experienced a stroke with the left side of the body as the involved side (see Table 3.3).

5.3.3 AC Change Predictors

At AC S2, we compute additional metrics to quantify the changes exhibited over one week of therapy from S1 to S2. For example, the percentage change for any given metric $x$ is computed as the difference between the S1 and S2 metric scores for $x$ and is normalized by the S1 metric score:

$$\Delta \% = \frac{x_{S2} - x_{S1}}{x_{S1}}$$  \hspace{1cm} (5.1)

Another metric we use to quantify the changes between S1 and S2 is the standardized mean difference effect size for repeated measures [120] (see Section 4.2). The SMD ES is applied to gait cycle metrics. For example, gait cycle duration is the amount of time for the completion of one gait cycle (stride). For this study, a gait cycle corresponds to the
time interval between one initial contact (heel strike) and the next initial contact of the same leg. If an individual took 15 strides at S1, then we derive 15 gait cycle durations, from which we compute the average gait cycle duration at S1 ($\bar{X}_{S1}$).

5.4 Hybrid Clinical Sensor Prediction

We express discharge FIM score prediction as a supervised learning task that maps the admission and AC features to predicted discharge FIM scores. Our approach, called Hybrid Clinical Sensor Prediction, is outlined in Algorithm 5. For HCSP, we extract features from clinical information available upon admission and from sensor data collected from therapy sessions mid-stay of rehabilitation. These clinical and sensor-based features are combined to form training data for a machine learning model constructed to predict discharge clinical assessment scores. HCSP makes use of an epsilon support vector machine ($\epsilon$-SVM) for its machine learning model. SVMs utilize a subset of the training data, called support vectors, to identify boundaries of maximal distance from the support vectors. In the case of regression, the SVM learns a function $F(x) \rightarrow w \cdot x - b$ to approximate a target variable $y_i$ under $\epsilon$ precision for each feature vector $x_i$. The vector $w$ is the learned weights, or coefficients, representing the relative importance of each feature for the SVM. We compare the prediction results of the SVM with linear regression and random forest with 100 regression trees. The 100 trees criterion is chosen because of its success in a previous inertial sensor and clinical assessment study [153]. These three machine learning algorithms are chosen because of demonstrated accuracy found in previous technology-
based clinical assessment studies [152–154].

5.4.1 Model Construction

For AC participants, we collect data from three different points in time and build the corresponding HCSP models. \( M_1 \) is a model trained with data available upon admission, \( M_2 \) is a model trained with data available at AC S1, and \( M_3 \) is a model trained with data available at AC S2. Figure 3.1 depicts this timeline and the associated models. For NAC participants, only admission data are available for training \( M_1 \). Each model \( M_1 \), \( M_2 \), and \( M_3 \) produces a prediction (\( P_1 \), \( P_2 \), and \( P_3 \)) for the same clinical outcome. These predictions represent the change in model prediction accuracy over time as new data are collected. Next, an ensemble learner (\( M_E \)) takes \( P_1 \), \( P_2 \), and \( P_3 \) as inputs and produces a fourth prediction (\( P_E \)). An ensemble learner combines the predictions of multiple learning algorithms to produce a final prediction. Ensemble learners are usually applied as an effort to achieve higher performance than the individual algorithms achieve alone. For comparison with the ensemble result, predictions \( P_1 \), \( P_2 \), and \( P_3 \) are also averaged to produce a fifth prediction (\( P_{avg} \)).

We construct predictive models using two different approaches, separate and cumulative, in order to explore the performance of \( M_2 \) and \( M_3 \) with different training features. Each approach builds three models (\( M_1 \), \( M_2 \), \( M_3 \)) representing the three different points during rehabilitation (admission, AC S1, AC S2); however, depending on the approach, different input features are utilized at \( M_2 \) and \( M_3 \). In the cumulative model construction
Algorithm 5 HCSP(\(A_{\text{train}}, S_{\text{train}}, D_{\text{train}}, A_{\text{test}}, S_{\text{test}}\))

1: Input: \(A_{\text{train}}\) = admission clinical training data
2: Input: \(S_{\text{train}}\) = mid-stay sensor training data
3: Input: \(D_{\text{train}}\) = discharge clinical assessment training scores
4: Input: \(A_{\text{test}}\) = admission clinical testing data
5: Input: \(S_{\text{test}}\) = mid-stay sensor testing data
6: Output: \(D_{\text{pred}}\) = discharge clinical predictions
7: Initialize: \(M\) = prediction model
8: Extract features from \(A_{\text{train}}, S_{\text{train}}, A_{\text{test}},\) and \(S_{\text{test}}\)
9: Train \(M\) with \(A_{\text{train}}\) features combined with \(S_{\text{train}}\) features to predict \(D_{\text{train}}\) scores
10: for each patient \(p\) in \(A_{\text{test}}\):
    11:     Test \(M\) on \(p\)
    12:     Append prediction to \(D_{\text{pred}}\)
  end for
13: return predictions \(D_{\text{pred}}\)

approach, we train \(M_2\) and \(M_3\) on all previous and current data available up to and including the corresponding point in time (see Figure 5.2). In the separate model approach, we do not train \(M_2\) and \(M_3\) with previously collected data but only with the data collected at that point in time (see Figure 5.3). Our approach then combines the results of \(M_1\), \(M_2\), and \(M_3\) with an ensemble learner to produce the final prediction. We include the separate model construction approach to examine the predictive power of \(M_2\) and \(M_3\) without utilizing any admission features (or in the case of \(M_3\), without utilizing admission features or AC S1 features).

5.4.2 SMOTE Oversampling

As mentioned earlier, the relatively few AC values pose a machine learning challenge, particularly in comparison with the large number of NAC records that are available. With such a small AC population, the prediction algorithm is at risk of overfitting the training
Figure 5.2: *Cumulative* model construction approach.

Figure 5.3: *Separate* model construction approach.
data. One method to compensate for a small number of samples is to oversample the data by replicating data points or adding synthetic data [155]. Resampled or synthetic data are only used during the training process and are designed to more thoroughly represent the space of possible data points.

To accommodate our low sample size, we employ the oversampling technique of synthetic minority oversampling technique (SMOTE) to the AC dataset [156]. SMOTE is an alternative to oversampling with replacement that creates synthetic data examples from the available training data. Synthetic examples are created by randomly interpolating features along the line segments joining any or all of the $k$ nearest neighbors for each existing data point. The algorithm is typically applied to correct a class imbalance problem. Since our clinical outcomes are continuous target variables, we employ a version of SMOTE for regression (SMOTE-R) [157]. In SMOTE-R, the minority class is considered to be rare, extreme values in the target variable. By applying SMOTE to these instances, the distribution of the target variable becomes more uniform. SMOTE has shown success in a variety of applications that share characteristics with this data and is therefore used to boost the size and diversity of the training data for this prediction problem.

To apply SMOTE-R, a function $\phi(y_k)$ maps each value of the target variable $y_k$ (discharge FIM scores in our study) to a notion of relevance (in this case, rarity and extremeness) in the range [0,1]. For example, if the target variable is assumed normal, the relevance function can be approximated by the complement of the variable's probability density function. In this case, highly relevant values will be near the tails of the distributions.
Figure 5.4: Example relevance function \( \phi \) for discharge Functional Independence Measure (FIM) motor score. The three scatter points denote the control points used to fit a cubic polynomial.
To build a relevance function for rare extreme values, our HCSP algorithm uses three control points to fit a cubic interpolating piecewise polynomial representing target variable relevance. We estimate the three control points \( (CP) \) using quartiles \( (Q) \): 

\[
CP_H = Q_3 + C \times IQR
\]

with relevance 1.0, 

\[
CP_L = Q_1 - C \times IQR
\]

with relevance 1.0, and the median, 

\[
CP_Y
\]

with relevance 0.0, where \( IQR = Q_3 - Q_1 \) (the interquartile range). \( C \) is a parameter reflecting the extent to which a sample is considered an outlier, where lower values of \( C \) imply more samples will be considered outliers. We set \( C = 1.2 \) due to our small sample size. All data points above \( CP_H \) or below \( CP_L \) are tagged as outliers and comprise the rare high extreme outliers and the rare low extreme outliers, respectively. An example relevance function \( \phi \) for discharge total motor score is shown in Figure 5.4. The three points in the Figure 5.4 correspond to the control points \( CP_L = 41.7 \), \( CP_Y = 62.0 \), and \( CP_H = 89.3 \). In this example, values less than 41.7 or greater than 89.3 are considered to be outliers with relevance 1.0. Finally, samples with relevance greater than a threshold \( t_E \) are considered relevant sample points and used to generate new synthetic samples. For our dataset we use \( t_E = 0.5 \) to obtain more relevant sample points on either tail of the discharge FIM score distributions. As the method suggests, the motivation behind this approach is to sample points around the outliers in order to give those points more representation in the training data.

We introduce a sampling variation that is a combination of existing methods. For nominal variables, we use a majority vote of \( k \) nearest neighbors [156]. Discharge FIM scores are assigned using the SMOTE-R approach. For this method, a weighted average of the two seed samples’ \( x \) and its nearest neighbor \( nn \) discharge FIM score is computed.
where $d_1$ and $d_2$ are the distances of the generated point to each of the two seed examples. This weights the sample with the smaller distance to the new synthetic data point higher. Usually SMOTE-based oversampling is coupled with undersampling of the majority class, or in the case of regression problems, undersampling of the more frequent samples. Since we want to retain all of our real AC data points, we do not undersample our dataset. Instead we apply SMOTE-R to the entire training set and perform additional sampling of the relevant samples. On each leave-one-out cross validation (LOOCV) fold, $N - 1$ training points are sampled to create $N - 1$ new synthetic data points that are added to the training set. SMOTE-R is applied in two configurations. The first configuration, *extremes*, only applies SMOTE to the outliers on the high and low tails of the training distribution [157]. The second configuration, *all data*, applies SMOTE to the entire training set.

5.4.3 Feature Selection

In order to understand the predictive utility of each feature and remove noisy or redundant features, we utilize feature selection techniques [158]. First, we correlate individual features with the target variable to investigate their individual predictive ability. Features with a Pearson correlation coefficient $r < 0.1$ are determined to be noisy and are not considered useful. Next, we apply a wrapper-based recursive feature elimination algorithm with cross validation (RFECV) to identify the optimal set of features [159]. The
model used in RFECV is a linear SVM trained with 10-fold cross validation with mean squared error scoring. Starting with all features, an SVM learns a vector of weights, or coefficients, representing the relative importance of the feature in learning the separating hyper plane. The feature with the smallest SVM weight is then removed from the feature set and the model is retrained on the remaining features. This process is repeated until the set with the lowest mean squared error is identified. Additionally, the size of this set denotes the optimal number of features. The top ranked features are then selected as the inputs to the prediction models.

5.5 Joint Patient Prediction

To achieve higher prediction accuracy with HCSP, we propose enhancements of HCSP based on patient similarity (PS) and joint prediction (JP). We hypothesize a combination of PS and JP techniques, called Joint Patient Prediction, will improve accuracy of discharge FIM motor score prediction using HCSP for the AC participant group. While the techniques we propose here are presented in the context of inpatient rehabilitation, our patient similarity and joint prediction approaches are general and suitable for other application domains.

5.5.1 Patient Similarity

Many applications require people to be grouped together into clusters based on similarity. For example, to improve human activity inference, researchers grouped similar
smartphone users into “communities” based on their behavior [160,161]. With the advent of electronic health records, a notion of similarity between people for healthcare applications has emerged, called “patient similarity”. Researchers have utilized patient similarity to propose novel electronic health record data mining techniques [162–166]. Based on the success of these previous studies using patient similarity, we propose to utilize patient similarity to improve our HCSP approach.

In machine learning, k-means clustering is an unsupervised learning approach to group N samples into k clusters where each cluster contains samples of similar nature. To form the clusters in k-means clustering, a distance function computes a dissimilarity measure between two samples. Once cluster centers are defined, individual data points are assigned to the closest cluster based on distance to the corresponding cluster center. Another type of machine learning algorithm that utilizes this notion of distance is k nearest neighbors (NN), or k-NN [167]. This supervised learning algorithm identifies the k labeled training data points that are closest to an unlabeled data point. The unlabeled point is then labeled using the label that appears most frequently among the k closest training data points (nearest neighbors).

In the case of patient similarity, a distance function, like those used in k-NN, computes a dissimilarity measure between two patients. For the distance function in HCSP, we use the heterogeneous Euclidean overlap metric (HEOM) [168] that was utilized in previous patient similarity work by Hielscher et al. [164]. We compute a pairwise distance matrix, dist, relating each AC participant to each other using HEOM. Let the current test patient held out during each fold of LOOCV be denoted as \( p_i \), \( 1 \leq i \leq N \). For our proposed
1. Identify $\text{NN}_i$, the $kPS$ NN of $p_i$

3. Test $M_i$ on $p_i$

2. Train $M_i$ w/$\text{NN}_i$

Figure 5.5: Using patient similarity to select training samples for a leave-out patient $p_i$ in leave-one-out cross validation. NN stands for nearest neighbors.

PS algorithm, $\text{dist}$ is queried to identify the $kPS$ ($1 \leq kPS < N$) nearest neighbors of the test participant $p_i$. The set of $kPS$ nearest neighbors, $\text{NN}_i$, forms the training set for the current fold. This approach trains a model using only patients that are most similar to the current test patient. The diagram presented in Figure 5.5 outlines the patient similarity algorithm.

There are several options for features to utilize when comparing two patients during generation of the distance matrix ($\text{dist}$). Lane et al. [161] proposed different types of similarity networks for building communities of smartphone users for activity inference. The researchers utilized physical, lifestyle, and sensor-data similarity networks. We adapt
the concept of multiple similarity networks proposed by Lane and colleagues by exploring
the following patient similarity features used for computing $dist$:

1. $dist_{all}$: Using all admission features (see Section 5.3.1 for a description of the ad-
mission features).

2. $dist_{demo}$: Using only admission demographic features.

3. $dist_{FIM-A}$: Using only admission FIM scores.

4. $dist_{M_1}$, $dist_{M_2}$, $dist_{M_3}$: Using only the top ranked features as determined by RFECV
(see Section 5.4.3 for details about feature selection) for each model $M_1$, $M_2$, and
$M_3$ individually.

In the next section, we discuss another approach, joint prediction, that also makes use of
pairwise patient distance matrices.

5.5.2 Joint Prediction

Joint prediction represents a recent direction of machine learning research. To ex-
plain joint prediction, consider a model $M$ trained to learn a mapping from a feature space
$X$ to a target variable $Y$, $M : X \rightarrow Y$. For joint prediction, each sample $x_j \in X$ is aug-
mented with additional “joint” features. The joint features depend on the machine learning
problem formulation, but are typically features of the same nature as the target variable
$Y$. For example, work by Minor et al. [169] applied joint prediction to a multi-output
regression learner trained to forecast activity occurrences in smart home environments.
For their forecasting approach, sensor-based features ($X_{sensor}$) were mapped into predictions of the time until each activity $a \in A$ would occur (the target variable $Y$), where $A = \{a_1, a_2, ..., a_T\}$ is a set of $T$ activity labels. The researchers applied joint prediction to the activity forecasting problem by augmenting the sensor-based features ($X_{sensor}$) with previous occurrence predictions for each activity ($\hat{Y}$) as joint features: $X = X_{sensor} \oplus \hat{Y}$.

Minor and colleagues reported a decrease of 85.11% in prediction error when utilizing joint features over the baseline model (no joint prediction). Furthermore, the researchers also explored utilizing ground truth time values ($Y$) for the joint features in lieu of $\hat{Y}$ predictions. In this case, the associated predictor is called the oracle predictor.

To apply the concept of joint prediction to HCSP, we include discharge FIM motor score predictions ($\hat{Y}$) from other participants in the training set as joint features. The predictions from other patients can be achieved by one of the following methods, where $p_i$ is the leave-out patient in LOOCV:

- **Oracle**: Utilize the ground truth values ($Y$) from N-1 other patients in the training set as features to predict the ground truth value for the current patient $p_i$.

- **Single pass**: Train a model with N-1 patients (holding out $p_i$). Then, using the same model that was just trained, make N-1 predictions ($\hat{Y}$) for the same N-1 patients used to train the model. Use these N-1 predictions as joint features to predict the ground truth value for the current patient $p_i$.

- **Double pass**: Using a second layer of LOOCV, generate predictions for N-1 patients ($\hat{Y}$). Use these N-1 predictions as joint features to predict the ground truth value
for the current patient $p_i$.

The differences between the above joint prediction methods lies in which “predictions” are included as joint features. The choice for which of the aforementioned approaches to use depends on the application context. If ground truth information for all training patients is available, then the oracle approach is appropriate. However, if ground truth information is not available for all patients, then predictions from a previously trained model need to be substituted for ground truth information. Both of the non-oracle predictors (single and double pass) will produce predictions representative of “true” predictions; however, the double pass predictor approach should produce less accurate predictions than the single pass predictor since the model has not previously been trained on the test sample. In this dissertation, we consider oracle joint features. Future work involves investigating single and double pass joint prediction for larger datasets.

In our joint prediction approach, we borrow concepts from patient similarity that were presented in Section 5.5.1. Specifically, we use the pairwise distance matrices to identify similar participants to include their discharge FIM motor scores as joint features. We also use patient similarity to pare down the size of the training set each fold as described previously. We name the combined patient similarity and joint prediction approach Joint Patient Prediction (JPP). For JPP, the $k_{JP}$ ($1 \leq k_{JP} < N$) nearest neighbors of each patient $p_j$ are identified as the set $NN_j$. $NN_j$ can be generated by querying a distance matrix $dist$ (see Section 5.5.1 for details regarding the computation of $dist$). Later, we will discuss an alternative approach for building $NN_j$. The target $Y$ values of the samples in $NN_j$ (discharge FIM scores for HCSP) are extracted as the set of joint features of
$p_j$, $NNJP_j$. $NNJP_j$ is the set of $kJP$ oracle joint features. This approach to generate $NNJP_j$ produces $kJP$ joint features that are ranked based on each neighbor participant’s proximity to $p_j$. Finally, the feature vector of $p_j$, $x_j$, is augmented with the joint features: $x_j = x_j \oplus NNJP_j$. Algorithm 6, Joint Patient Prediction, presents a general version of the JPP algorithm with a single distance matrix $dist$ and an example LOOCV model $M_i$.

Finally, we propose an alternative generation of $NN_j$ to avoid features that are ranked by position. Instead, we fix $kJP$ to 3 joint features for $M_1$, 4 joint features for $M_2$, and 5 joint features for $M_3$. We compute these joint features by querying the nearest neighbor of $p_j$ in each of the distance matrices $dist_{demo}$, $dist_{FIM-A}$, $dist_{M1}$, $dist_{M2}$, and $dist_{M3}$. Consequently, the models $M_1$, $M_2$, and $M_3$ will each have 3, 4, and 5 joint features, respectively. For example, $NN_j$ for $M_3$ is generated as $NN_j = \{NN_{demo}, NN_{FIM-A}, NN_{M1}, NN_{M2}, NN_{M3}\}$. Consequently, $NNJP_j$ is the set of $Y$ values for each of the patients included in $NN_j$.

### 5.5.3 Evaluation Methods

To evaluate the quality of the predicted clinical outcomes, we use several evaluation metrics. The regression models are evaluated using the mean absolute error (MAE), root mean squared error (RMSE) and normalized RMSE (NRMSE). RMSE and NRMSE are defined in Equations 5.3 and 5.4:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{actual,i} - Y_{predicted,i})^2}$$  \hspace{1cm} (5.3)

$$NRMSE = \frac{RMSE}{Y_{actual,max} - Y_{actual,min}}$$  \hspace{1cm} (5.4)
Algorithm 6 JointPatientPrediction($X$, $Y$, $kPS$, $kJP$

1: Input: $X = N \times M$ matrix of $N$ patients’ feature vectors $x$, $|x| = M$
2: Input: $Y =$ vector of ground truth values for each patient in $X$
3: Input: $kPS =$ the number of nearest neighbors to include in the training set
4: Input: $kJP =$ the number of joint features to include
5: Output: $\hat{Y} =$ a vector of $N$ patients’ predictions
6: Compute $dist$, a $N \times N$ matrix of pairwise distances between patients in $X$
7: for each patient feature vector $x_j$ in $X$:
   8: $NN_j = dist[j][: kJP] \quad // kJP nearest neighbors of $x_j$
   9: $NNJP_j = Y[NN_j] \quad // the joint features$
   10: $x_j = x_j \oplus NNJP_j \quad // adding the joint features$
end for
11: Perform feature selection
12: for each patient feature vector $x_i$ in $X$:
13: Initialize: prediction model $M_i$
14: $NN_i = dist[i][: kPS] \quad // kPS nearest neighbors of $p_i$
15: Train $M_i$ with $NN_i$
16: $\hat{Y}_i = Test M_i$ on $x_i$
17: Store $\hat{Y}_i$ at $\hat{Y}[i]$
end for
18: return prediction vector $\hat{Y}$

where $Y$ is the predicted clinical outcome and $n$ is the number of predicted samples.

Pearson correlation coefficients, $r$, and associated $p$-values are also reported, as defined in Equation 5.5:

$$r = \frac{\sum_{i=1}^{n} (Y_{actual,i} - \bar{Y}_{actual}) \cdot (Y_{predicted,i} - \bar{Y}_{predicted})}{\sqrt{\sum_{i=1}^{n} (Y_{actual,i} - \bar{Y}_{actual})^2 \cdot \sum_{i=1}^{n} (Y_{predicted,i} - \bar{Y}_{predicted})^2}} \quad (5.5)$$

5.6 Results

All data are processed with the Sci-kit Learn machine learning library. Prior to training, admission and AC data are standardized by subtracting the mean and scaling
to unit variance. Unless otherwise stated, an SVM with a linear kernel is trained and evaluated using leave-one-out cross validation.

5.6.1 Feature Selection: FIM Motor Score

Features are correlated with the FIM motor score at discharge and noisy features are removed. Table 5.3 lists the 10 most highly correlated predictors and their correlation coefficients, grouped by the time points of admission, AC S1, and AC S2. Wrapper-based feature elimination results for discharge FIM motor score are shown in Table 5.4. This table contains the top 10 ranked features for each model, where $M_1$ is trained with AC participant admission data only (not including NAC patient admission data).
Table 5.3: Correlations ($r$) between individual predictors and discharge Functional Independence Measure (FIM) motor score.

<table>
<thead>
<tr>
<th>Admission</th>
<th>$r$</th>
<th>AC S1</th>
<th>$r$</th>
<th>AC S2</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reciprocal admission total</td>
<td>-0.68**</td>
<td>COM acceleration stand-to-sit Z peak angular velocity</td>
<td>0.62**</td>
<td>S2 peak angular velocity average (lesser side)</td>
<td>0.65**</td>
</tr>
<tr>
<td>motor score</td>
<td></td>
<td>Shank range of motion average (greater side)</td>
<td>0.62**</td>
<td>S2 vehicle challenge duration</td>
<td>-0.60**</td>
</tr>
<tr>
<td>Admission bladder</td>
<td>0.62**</td>
<td>Step length average</td>
<td>0.59**</td>
<td>S2 range of motion average (lesser side)</td>
<td>0.59**</td>
</tr>
<tr>
<td>Admission upper body dressing</td>
<td>0.61**</td>
<td>Vehicle challenge duration</td>
<td>-0.56**</td>
<td>S2 number of gait cycles</td>
<td>-0.56**</td>
</tr>
<tr>
<td>CMG relative weight</td>
<td>-0.60**</td>
<td>Shank range of motion average (lesser side)</td>
<td>0.55*</td>
<td>Cadence percent change</td>
<td>0.51*</td>
</tr>
<tr>
<td>Admission grooming</td>
<td>0.60**</td>
<td>Number of gait cycles</td>
<td>-0.51*</td>
<td>S2 swing percent CV</td>
<td>-0.50*</td>
</tr>
<tr>
<td>Admission problem solving</td>
<td>0.59**</td>
<td>Shank peak angular velocity average (greater side)</td>
<td>0.48*</td>
<td>Peak angular velocity SMD (greater side)</td>
<td>0.50*</td>
</tr>
<tr>
<td>Admission memory</td>
<td>0.56*</td>
<td>COM acceleration vehicle unload RMS</td>
<td>0.47*</td>
<td>S2 duration</td>
<td>-0.49*</td>
</tr>
<tr>
<td>Admission bed-to-chair transfer</td>
<td>0.53*</td>
<td>COM acceleration vehicle unload RMS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission toilet transfer</td>
<td>0.50*</td>
<td>COM acceleration stand-to-sit RMS</td>
<td>0.46*</td>
<td>S2 COM acceleration RMS</td>
<td>0.48*</td>
</tr>
<tr>
<td>Admission comprehension</td>
<td>0.46*</td>
<td>Walking speed</td>
<td>0.44</td>
<td>S2 COM acceleration RMS</td>
<td>0.46*</td>
</tr>
</tbody>
</table>

AC = ambulation circuit, CMG = case mix group, COM = center of mass, CV = coefficient of variation, $r$ = Pearson correlation coefficient, RMS = root mean square, $S1 =$ session 1, $S2 =$ session 2, SMD = standardized mean difference, * = $p < 0.05$, ** = $p < 0.01$. 
Table 5.4: Cumulative recursive feature elimination with cross validation (RFECV) top ranked features for discharge Functional Independence Measure (FIM) motor score.

<table>
<thead>
<tr>
<th></th>
<th>Rank</th>
<th></th>
<th>Rank</th>
<th></th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission upper body dressing*</td>
<td>1</td>
<td>COM acceleration stand-to-sit Z peak angular velocity*</td>
<td>1</td>
<td>COM acceleration stand-to-sit Z peak angular velocity*</td>
<td>1</td>
</tr>
<tr>
<td>Admission memory*</td>
<td>1</td>
<td>Admission memory*</td>
<td>1</td>
<td>Admission memory*</td>
<td>1</td>
</tr>
<tr>
<td>RIC</td>
<td>1</td>
<td>COM acceleration stand-to-sit RMS*</td>
<td>1</td>
<td>Range of motion SMD (lesser side)</td>
<td>1</td>
</tr>
<tr>
<td>Admission bladder*</td>
<td>1</td>
<td>Shank range of motion average (greater side)*</td>
<td>1</td>
<td>Step length average*</td>
<td>1</td>
</tr>
<tr>
<td>Admission grooming*</td>
<td>1</td>
<td>Admission grooming*</td>
<td>1</td>
<td>Admission grooming*</td>
<td>1</td>
</tr>
<tr>
<td>Admission problem solving*</td>
<td>1</td>
<td>Double support percent CV</td>
<td>1</td>
<td>COM vehicle unload Z peak angular velocity percent change</td>
<td>1</td>
</tr>
<tr>
<td>Admission tub/shower transfer</td>
<td>1</td>
<td>Admission upper body dressing*</td>
<td>1</td>
<td>Admission upper body dressing*</td>
<td>1</td>
</tr>
<tr>
<td>Admission lower body dressing</td>
<td>1</td>
<td>Admission bed-to-chair transfer*</td>
<td>2</td>
<td>S2 peak angular velocity average (lesser side)*</td>
<td>1</td>
</tr>
<tr>
<td>Reciprocal admission total motor score*</td>
<td>2</td>
<td>Swing percent CV</td>
<td>3</td>
<td>Cycle duration CV percent change</td>
<td>1</td>
</tr>
<tr>
<td>CMG relative weight*</td>
<td>3</td>
<td>Limp average</td>
<td>4</td>
<td>S2 double support percent CV</td>
<td>1</td>
</tr>
</tbody>
</table>

CMG = case mix group, COM = center of mass, CV = coefficient of variation, M = model, RIC = rehabilitation impairment category, RMS = root mean square, S2 = session 2, SMD = standardized mean difference, * = listed in Table 5.3.
5.6.2 Prediction Results

FIM Motor Score

Table 5.5 shows results for predicting discharge total motor score with LOOCV. Two admission models ($M_1$) are trained, one with AC participant data only and a second model including NAC patient admission data. To visualize the FIM motor score predictions, each participant’s actual discharge score is plotted together with the predictions generated by $M_1$, $M_2$, and $M_3$ (see Figure 5.6).
Table 5.5: Discharge Functional Independence Measure (FIM) motor score prediction results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear SVM</th>
<th>Linear Regression</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>NRMSE</td>
<td>r</td>
</tr>
<tr>
<td>M(_1) (w/o NAC)</td>
<td>4.66</td>
<td>11.65%</td>
<td>0.89**</td>
</tr>
<tr>
<td>M(_1)</td>
<td>7.36</td>
<td>18.41%</td>
<td>0.82**</td>
</tr>
<tr>
<td>Separate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M(_2)</td>
<td>8.55</td>
<td>21.38%</td>
<td>0.60*</td>
</tr>
<tr>
<td>M(_3)</td>
<td>5.54</td>
<td>13.86%</td>
<td>0.85**</td>
</tr>
<tr>
<td>M(_{avg})</td>
<td>5.54</td>
<td>13.86%</td>
<td>0.87**</td>
</tr>
<tr>
<td>M(_E)</td>
<td>5.50</td>
<td>13.74%</td>
<td>0.84**</td>
</tr>
<tr>
<td>Cumulative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M(_2)</td>
<td>5.43</td>
<td>13.57%</td>
<td>0.86**</td>
</tr>
<tr>
<td>M(_3)</td>
<td>3.40</td>
<td>8.50%</td>
<td>0.95**†</td>
</tr>
<tr>
<td>M(_{avg})</td>
<td>3.66</td>
<td>9.15%</td>
<td>0.96**†</td>
</tr>
<tr>
<td>M(_E)</td>
<td>3.24</td>
<td>8.11%</td>
<td>0.95**†</td>
</tr>
</tbody>
</table>

avg = average, E = ensemble, M = model, NAC = non-ambulation circuit, NRMSE = normalized root mean square error, r = Pearson correlation coefficient, RMSE = root mean square error, SVM = support vector machine, * = \(p < 0.05\), ** = \(p < 0.01\), † = significantly \((p < 0.05)\) improved results from M\(_1\).
Table 5.6: Discharge Functional Independence Measure (FIM) motor score with synthetic training data prediction results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Extremes SMOTE-R</th>
<th>All data SMOTE-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>NRMSE</td>
</tr>
<tr>
<td>M₁ (w/o NAC)</td>
<td>4.86</td>
<td>12.14%</td>
</tr>
<tr>
<td>M₁</td>
<td>7.34</td>
<td>18.36%</td>
</tr>
<tr>
<td>M₂</td>
<td>8.39</td>
<td>20.97%</td>
</tr>
<tr>
<td>M₃</td>
<td>5.40</td>
<td>13.49%</td>
</tr>
<tr>
<td>M₄avg</td>
<td>5.44</td>
<td>13.61%</td>
</tr>
<tr>
<td>M₅E</td>
<td>5.35</td>
<td>13.38%</td>
</tr>
<tr>
<td>M₆avg</td>
<td>6.15</td>
<td>15.36%</td>
</tr>
<tr>
<td>M₇E</td>
<td>6.35</td>
<td>15.38%</td>
</tr>
<tr>
<td>M₉avg</td>
<td>3.95</td>
<td>9.87%</td>
</tr>
<tr>
<td>M₉E</td>
<td>4.32</td>
<td>10.80%</td>
</tr>
</tbody>
</table>

avg = average, E = ensemble, M = model, NAC = non-ambulation circuit, NRMSE = normalized root mean square error, r = Pearson correlation coefficient, RMSE = root mean square error, * = p < 0.05, ** = p < 0.01.

**FIM Motor with SMOTE**

We apply SMOTE-R (see Section  5.4.2) to the discharge FIM motor distribution to generate additional training data. Table 5.6 shows the results of training a linear SVM with synthetic data in both configurations. To illustrate the resulting synthetic data, Figure 5.7 shows an example of how SMOTE-R in both configurations affects the distribution of the discharge FIM motor score in an example fold of LOOCV. Figure 5.7a shows the histogram of the original training discharge FIM motor score, while Figures 5.7b and 5.7c show the target variable distribution of the all data and extremes approach, respectively. For the example fold, the original data has a discharge FIM motor score of 63.32 ± 9.83. SMOTE-R for all data generates 19 new sample points which changes the values to 63.87 ± 7.44. SMOTE-R for extreme data points yields a mean and standard deviation of 62.06 ± 10.36.
Figure 5.6: Discharge Functional Independence Measure (FIM) motor score prediction results. For all figures, $M_1$ is trained with non-ambulation circuit admission data.
Figure 5.7: Histograms for discharge Functional Independence Measure (FIM) motor score for an example fold of cross validation before and after applying SMOTE.
FIM Motor with Joint Patient Prediction

To evaluate HCSP with PS and JP, we compare our JPP results for each model (M₁, M₂, M₃, M_avg, and M_E) to the cumulative prediction model results previously presented in Table 5.5. For further explanation, we compare M₁ without JPP to M₁ with JPP, M₂ without JPP to M₂ with JPP, and so on. We test JPP with multiple parameter values for patient similarity, specifically we run experiments for \( k_{PS} = \{1, 2, ..., N - 1\} \).

For M₁ and M₃, utilizing the entire training set (\( k_{PS} = 19 \)) yields the highest prediction accuracy. The joint features for these models are not selected by RFECV, indicating JPP does not result in an improvement in prediction accuracy for M₁ and M₃. However, 15 M₂ models with patient similarity-selected training sets and joint features yield an improvement in prediction accuracy. The JPP results from M₂ experiments that are improvements over the baseline M₂ model are shown in Table 5.7. Furthermore, slight improvements are exhibited by the average learner M_avg and the ensemble learner M_E. JPP results for the average learner M_avg indicate \( k_{PS} = 19 \) with joint prediction offers a prediction improvement of \( r = 0.96 \) (RMSE = 3.44) over the baseline M_avg (\( r = 0.96, \) RMSE = 3.66). JPP results for the ensemble learner M_E indicate \( k_{PS} = 19 \) with joint prediction demonstrates a prediction improvement of \( r = 0.95 \) (RMSE = 3.22) over the baseline M_E (\( r = 0.95, \) RMSE = 3.24).

FIM Cognitive Score

To explore the possible relationship between cognitive functioning and performance on the AC, we train additional models to predict the FIM cognitive score at discharge.
Table 5.7: Cumulative M2 joint patient prediction results for discharge Functional Independence Measure (FIM) motor score. Only results demonstrating improvement over the baseline M2 predictor are shown. Results are sorted in ascending root mean square error (RMSE) order.

<table>
<thead>
<tr>
<th>kPS</th>
<th>Joint Prediction</th>
<th>RMSE</th>
<th>NRMSE</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Yes</td>
<td>3.14</td>
<td>7.84%</td>
<td>0.96** †</td>
</tr>
<tr>
<td>19</td>
<td>Yes</td>
<td>3.15</td>
<td>7.87%</td>
<td>0.96** †</td>
</tr>
<tr>
<td>17</td>
<td>Yes</td>
<td>3.22</td>
<td>8.05%</td>
<td>0.96** †</td>
</tr>
<tr>
<td>16</td>
<td>Yes</td>
<td>3.94</td>
<td>9.84%</td>
<td>0.94**</td>
</tr>
<tr>
<td>15</td>
<td>Yes</td>
<td>4.02</td>
<td>10.04%</td>
<td>0.93**</td>
</tr>
<tr>
<td>13</td>
<td>Yes</td>
<td>4.02</td>
<td>10.05%</td>
<td>0.94**</td>
</tr>
<tr>
<td>14</td>
<td>Yes</td>
<td>4.08</td>
<td>10.20%</td>
<td>0.93**</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>4.20</td>
<td>10.49%</td>
<td>0.93**</td>
</tr>
<tr>
<td>11</td>
<td>Yes</td>
<td>4.76</td>
<td>11.90%</td>
<td>0.91**</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>4.84</td>
<td>12.09%</td>
<td>0.92**</td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>5.05</td>
<td>12.63%</td>
<td>0.90**</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>5.09</td>
<td>12.72%</td>
<td>0.90**</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>5.35</td>
<td>13.38%</td>
<td>0.89**</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>5.41</td>
<td>13.52%</td>
<td>0.89**</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>5.42</td>
<td>13.55%</td>
<td>0.88**</td>
</tr>
<tr>
<td>19 (baseline)</td>
<td>No</td>
<td>5.43</td>
<td>13.57%</td>
<td>0.86**</td>
</tr>
</tbody>
</table>

NRMSE = normalized root mean square error, r = Pearson correlation coefficient, RMSE = root mean square error, ** = p < 0.01, † = significantly (p < 0.05) improved results from baseline M2.
Table 5.8 shows discharge FIM cognitive prediction results for a SVM with a linear kernel, linear regression, and a random forest with 100 regression trees.
Table 5.8: Discharge Functional Independence Measure (FIM) cognitive score prediction results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear SVM</th>
<th>Linear Regression</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>NRMSE</td>
<td>r</td>
</tr>
<tr>
<td>M1 (w/o NAC)</td>
<td>2.42</td>
<td>20.19%</td>
<td>0.70**</td>
</tr>
<tr>
<td>M1</td>
<td>2.34</td>
<td>19.49%</td>
<td>0.73**</td>
</tr>
<tr>
<td>M2</td>
<td>3.10</td>
<td>25.82%</td>
<td>0.51*</td>
</tr>
<tr>
<td>M3</td>
<td>3.74</td>
<td>31.17%</td>
<td>-0.34</td>
</tr>
<tr>
<td>Mavg</td>
<td>2.56</td>
<td>21.36%</td>
<td>0.68**</td>
</tr>
<tr>
<td>ME</td>
<td>2.66</td>
<td>22.14%</td>
<td>0.64**</td>
</tr>
<tr>
<td>M2</td>
<td>3.44</td>
<td>28.69%</td>
<td>0.19</td>
</tr>
<tr>
<td>M3</td>
<td>2.42</td>
<td>20.19%</td>
<td>0.70**</td>
</tr>
<tr>
<td>Mavg</td>
<td>2.40</td>
<td>20.01%</td>
<td>0.73**</td>
</tr>
<tr>
<td>ME</td>
<td>2.71</td>
<td>22.59%</td>
<td>0.59*</td>
</tr>
</tbody>
</table>

avg = average, E = ensemble, M = model, NAC = non-ambulation circuit, NRMSE = normalized root mean square error, r = Pearson correlation coefficient, RMSE = root mean square error, SVM = support vector machine, * = p < 0.05, ** = p < 0.01, † = significantly (p < 0.05) improved results from M1.
Individual FIM Tasks

Each discharge FIM task, in turn, is predicted using models $M_1$, $M_2$, and $M_3$. The correlation results for each task prediction are plotted in Figure 5.8. The plot shows which tasks are more closely represented by the admission, AC S1, and AC S2 metrics and the corresponding models $M_1$, $M_2$, and $M_3$.

5.7 Discussion

We introduce a new method, called Hybrid Clinical Sensor Prediction (or HCSP), to predict discharge FIM scores using data collected from wearable inertial sensors. We trained regression models to predict the discharge FIM motor score, cognitive score, and
individual FIM task scores for participants who performed two testing sessions of the
ambulation circuit.

5.7.1 Predictor Strength

Of the predictors available at admission, the reciprocal of admission FIM motor score is the most highly correlated with discharge FIM motor score ($r = -0.68$). Sonoda et al. [136] also found this feature highly correlated ($r = -0.896$) with discharge FIM motor score for first stroke survivors without the presence of other diseases. Other strong predictors at admission include admission bladder item scores, admission upper body dressing scores, and case mix group (CMG) relative weight. Highly uncorrelated predictors include admission walk/wheelchair scores and admission eating scores. Admission eating in particular is not an informative feature as it exhibits zero variance for our AC participants. For the AC metrics, vehicle challenge duration, shank peak angular velocity, shank range of motion, number of gait cycles, and COM RMS are all highly correlated with discharge FIM motor scores, whereas sit-to-stand duration, step symmetry, and walking speed percentage change are not strong predictors on their own. This finding is particularly interesting for walking speed percent change because walking speed is a common clinical measurement of gait functioning [170].

Often when considered on their own, strong individual predictors are outperformed by a linear combination of weaker predictors [158]. The prediction results using RFECV is such an example. Discharge FIM motor score correlations for the separate model con-
struction approach without RFECV are $r = 0.70$, 0.40, and 0.48 compared to $r = 0.89$, 0.60, and 0.85 with RFECV for $M_1$, $M_2$, and $M_3$, respectively. For the cumulative model construction approach, discharge FIM motor score correlations without RFECV are $r = 0.55$ and 0.55 compared to $r = 0.86$ and 0.95 with RFECV for $M_2$ and $M_3$, respectively. For $M_1$, admission tub/shower transfer and admission lower body dressing are highly ranked; however, these features are not as highly correlated with FIM motor scores when considered alone. Furthermore, reciprocal admission motor FIM and CMG relative weight are not top ranked by RFECV and therefore not used to train $M_1$. Similar trends are seen in $M_2$ and $M_3$, which had 7 and 10 top ranked features, respectively.

5.7.2 Predictions

FIM Motor Score

When considering $M_1$ with AC data only, $M_1$ with NAC data, and $M_2$ and $M_3$ independently (the separate model construction approach), the correlations associated with discharge FIM motor score prediction are $r = 0.89$, 0.82, 0.60, and 0.85, respectively (see Table 5.5). Much higher accuracy is achieved when utilizing all features from previous points in time as well, as is the case with the cumulative approach with $M_2$ and $M_3$ that yields correlations $r = 0.86$ and 0.95. As we expect, the correlations increase as additional AC features representing participants’ performance are included. Consequently, the strongest correlations are achieved with the cumulative $M_3$ model ($RMSE = 3.40$ and $r = 0.95$). Even though $M_1$ alone is already a high performing model, cumulative $M_3$ predictions produce a significantly higher correlation ($p < 0.01$) than $M_1$ with NAC data.
(RMSE = 7.36, r = 0.82). These results indicate that wearable sensors enhance prediction of clinical rehabilitation outcomes over medical records alone.

Our results for M₁ are consistent with earlier literature for discharge FIM motor score prediction. Using only clinical information (no technology-based predictors) for stroke patients, Jeong et al. [141] reported a prediction correlation of $r = 0.88$ ($R^2 = 0.77$), Fujiwara et al. [139] reported $R^2$ between 0.66 and 0.75, and Tsuji et al. [140] reported $R^2 = 0.68$. These studies utilized additional clinical rating scale scores as predictors that are unavailable in our study. Despite this difference, our results for predictions based only upon admission features are consistent with previously reported results. Using technology-based predictors, Mostafavi et al. [152] utilized metrics from robot-assistive KINARM devices during upper limb reaching and positioning tasks to predict FIM at discharge with accuracy of RMSE = 11.8 (NRMSE = 17.3%). There are differences between this study and ours, including the technology (robot device vs. wearable sensors) and monitored body parts (upper limb vs. whole body). These differences make comparisons between the results of Mostafavi and colleagues and our results difficult; however, our lower error (RMSE = 3.40, NRMSE = 8.50%) does suggest the viability of wearable inertial sensors for work involving clinical outcome prediction of ambulatory tasks.

It is also worth noting that M₁ with AC participant data only outperforms M₁ with NAC participant data used for training ($r = 0.89, 0.82$ and RMSE = 4.66, 7.34 respectively). This difference suggests that our small sample size of 20 AC participants might result in overfit models. As this may be the case, there are also several differences between the two datasets to discuss. AC participants are patients who are able to physically per-
form the AC and have higher cognitive awareness on average than NAC patients (FIM$_{A\text{-cog}}$ = 23.10, 22.12 and FIM$_{D\text{-cog}}$ = 29.55, 28.46, respectively). This difference in cognition can be attributed to the requirement of passing the Mini-Cog exam in order to be a participant in the AC study. Furthermore, the variance in the NAC training data is higher, as there are additional RICs not represented by the AC participants. Several NAC patients also show FIM regression, which is not evident in the AC participant dataset. The inclusion of the NAC patients suggests our results might be an overestimate of the accuracy that could be obtained for M$_2$ and M$_3$ when considering AC data from any patient, ignoring the study recruitment strategies.

When comparing the linear SVM results to other machine learning techniques, linear regression performs similarly to the SVM, whereas the random forest does not perform as well (see Table 5.5 for results). The random forest still yields statistically significant correlations with $r = 0.73$ and 0.59 ($p < 0.05$) for M$_1$ and cumulative M$_2$ respectively. The regression trees would most likely perform better with additional training examples, as is the case for M$_1$ AC only compared with M$_1$ with NAC ($r = 0.61$, 0.73 respectively). Additional synthetic data are generated using SMOTE for regression and used for training a linear SVM (see Table 5.6 for results). Cumulative M$_3$ performance is not as high as without the synthetic data (original RMSE = 3.40, $r = 0.95$; extremes SMOTE RMSE = 3.50, $r = 0.94$; all data SMOTE RMSE = 3.85, $r = 0.93$), suggesting additional AC participant data might attenuate the prediction accuracy.

In Section 5.5 we hypothesized that patient similarity and joint prediction would improve HCSP accuracy for predicting discharge FIM motor scores. While our JPP al-
algorithm does not improve prediction accuracy for $M_1$ and $M_3$, it does improve prediction accuracy for $M_2$, $M_{\text{avg}}$, and $M_E$. Particularly for $M_2$, three configurations of PS and JP yield results that are significantly ($p < 0.05$) more accurate than the baseline HCSP $M_2$ model (see Table 5.7). Furthermore, $M_2$ with $kPS = 18$ and joint features is as accurate as the best HCSP predictor with JPP (cumulative $M_3$). This result suggests that with JPP only one AC session is sufficient to produce highly accurate FIM motor score predictions from wearable sensor data collected mid-stay of inpatient rehabilitation. Future work aims to investigate additional JPP techniques to improve $M_1$ and $M_3$. However, we do suspect improving FIM motor score predictions for $M_3$ may pose a challenge because the results for $M_3$ are already quite accurate.

**FIM Cognitive Score**

The models we generate based only on admission data are able to predict discharge FIM cognitive score fairly accurately (see Table 5.8). $M_1$ with AC data only achieves RMSE = 2.42, $r = 0.70$ and $M_1$ with NAC data demonstrating slightly stronger results (RMSE = 2.34, $r = 0.73$). Highly ranked predictors by RFECV include admission memory task score, walk/wheelchair task score, and tub/shower transfer score. When adding features from AC data, the results are not as strong as $M_1$ for separate $M_2$, separate $M_3$, and cumulative $M_2$. Cumulative $M_3$ performs as well as $M_1$ without NAC data, while averaging cumulative $M_1$, $M_2$, and $M_3$ predictions barely outperform $M_1$ without NAC data (RMSE = 2.40, $r = 0.73$). Considering the AC is primarily a motor task of physical functioning, it is not surprising that it is difficult to improve an initial prediction at $M_1$ with NAC data. On a related note, future work includes programmatically determining if the participant
(a) Participant #010 prediction results (cumulative method).

(b) Participant #015 prediction results (cumulative method).

Figure 5.9: AC participant mean absolute error (MAE) for each Functional Independence Measure (FIM) task. M_1 is trained with non-ambulation circuit data.

slowed down or stopped for the Stops Walking When Talking Test performed during the AC (see Section 3.3 for details about the ambulation circuit). This feature and others like it could provide cognitive information for the machine learning algorithms and improve discharge FIM cognitive score predictions.

**Individual FIM Tasks**

Analyzing the relevance of the AC features for predicting individual FIM task scores at discharge provides interesting insights about what tasks the AC most closely represents. Figure 5.8 shows correlations for predictions of all 18 tasks for separate and cumulative construction approaches. As reported for FIM motor score predictions, cumulative features together (see Figure 5.8b) offer more predictive power than considering each time point
separately (see Figure 5.8a). $M_3$ outperforms the other models on several tasks, most notably bathing, bladder management, all the transfers, and stairs. $M_2$ does particularly well on grooming, toileting, and expression, but particularly poorly on comprehension. $M_1$ performs consistently near $r = 0.50$ for all tasks, and better than $M_2$ and $M_3$ for most cognitive tasks. This confirms the cognitive score results for predicting total discharge FIM cognitive scores (see Table 5.8). Finally, the improvement of $M_3$ over $M_1$ for bladder and bowel control is especially interesting. Perhaps motor functioning on the AC exhibits a relationship between underlying mechanisms of sphincter control. Another possibility includes the existence of a third variable of which both the AC and sphincter control are related. Investigating this phenomenon is a direction for future work.

The FIM motor, cognitive, and individual task prediction results offer insight into the recovery process at the group level. The data can also be used to examine predictors and predictions made for individual patients. For example, participant #010 has the lowest FIM motor score at discharge ($FIM_{D-motor} = 40$, see Table 3.3). Predicting individual FIM tasks for participant #010 reveals the models do not generalize well to the profile of participant #010 (see Figure 5.9a). The models perform particularly poorly for the FIM tasks of upper body dressing, bladder sphincter control, and problem solving. On the contrary, participant #015 has the highest FIM score of the group ($FIM_{D-motor} = 80$, tied with participant #021). With the exception of the bladder control task ($M_2$) and stairs ($M_1$), scores for participant #015 are well predicted by the models (see Figure 5.9b). The juxtaposition between these two participants is interesting from a research perspective, but its usefulness for clinicians remains to be seen. We address this gap in the literature
with our work in Chapter 6, which includes results measuring the clinical utility of FIM predictions.

### 5.7.3 Limitations

The low number of available AC data points is an important limitation of this study; however, for the current investigation we apply SMOTE-R to increase the sample size with synthetic data to overcome this drawback. Another limitation is all data are collected from the same inpatient hospital. A wide variety of patients attending other rehabilitation facilities would be more representative of the population and potential clinical utility of the models. Finally, the AC participant population is primarily recovering from a stroke (70%). A wider variety of patient impairments would be more representative of all types of patients admitted to inpatient rehabilitation.

In summary, we present our Hybrid Clinical Sensor Prediction approach to investigate the combined predictive abilities of clinical information available at admission and features derived from wearable inertial sensor data. We apply HCSP to predict discharge clinical outcomes of the Functional Independence Measure for 20 patients at an inpatient rehabilitation hospital. While models trained only on admission data performed well, we are able to achieve an even higher level of prediction accuracy with HCSP by incorporating inertial sensor-based features. Correlations as high as $r = 0.95$ (RMSE = 3.40, NRMSE = 8.50%) for LOOCV predicting discharge FIM motor score are obtained. This research adds unique findings over previous studies by mapping longitudinal sensor data collected
during physical therapy into the prediction of clinical outcomes. Often, technology-based features do not correspond to what standard clinical assessments are evaluating [148], thus it is reassuring that the results presented in Section 5.6.2 demonstrated strong correspondence and can be leveraged to improve upon the predictive power of standard clinical assessments at admission.

There are several opportunities and directions for future work, including recruiting additional participants of various impairments. A large enough sample size could allow for training RIC-specific models that take into account the differences and similarities exhibited for individual RIC populations. We also plan to design and implement a mobile application to automate data collection, processing, and prediction to generate predictions mid-stay of inpatient rehabilitation. Finally, further exploring patient similarity and joint prediction techniques to improve prediction performance, as well as exploring other machine learning techniques and optimization methods, represents another direction of future work.

This research lays the foundation for a sensor-based system that collects data from ambulatory tasks of physical rehabilitation. Models based on our HCSP approach can map sensor data from heterogeneous sources into an appropriate clinical assessment to provide updates about patient progress in a more universal domain; however, it is not the intention of such a system to replace the expertise of a trained clinician, but instead to provide the therapist with additional, meaningful information. Viewed in this regard, a wearable sensor system and its associated algorithms are potentially a tool to inform therapists and help them better provide services to their patients during recovery. To research and
design wearable technology and software as a tool for rehabilitation, we interviewed physical therapy providers to gather their feedback regarding the sensor-based metrics, FIM predictions, and data visualizations. The next chapter discusses this informative feedback we received from therapy providers.
To evaluate and improve the wearable sensor experience for therapists and patients, we interviewed physical therapy providers regarding the utility of wearable technology and our algorithms for helping provide therapy services for patients. We presented data collected from the AC wearable sensor study to physical therapists (N = 5) and physical therapy assistants (N = 2) to collect therapy providers’ perceived clinical utility of wearable sensor data, the metrics we computed, and the visualizations we generated for mobility assessment. The interview consisted of four main components: 1) general conversation about technology, 2) rating usefulness of sensor-based metrics, 3) questions regarding sensor-based clinical assessment predictions, and 4) evaluating visualizations. We discuss the responses received from the therapists and suggest future computing research directions to potentially enhance therapy services and increase the adoption of wearable sensor systems for mobility assessment.

6.1 Prior Work on Collecting Physical Therapist Feedback

Several metrics can be derived from wearable inertial sensor data to quantify the wearer’s movements, such as the number of steps taken and walking smoothness. While sensor-based metrics are interesting from a research perspective, their utility for clinicians...
providing therapy services to patients has not been established; however, several studies have investigated acceptance of wearable technology. Therapists are open to using wearable technology, particularly when the system provides additional information about their patients [171], when shown evidence that the technology is effective [172], or when the collected health data promotes patient engagement [173]. In addition, medical professionals in hospitals are more likely to use wearable computer systems if the technology improves day-to-day work efficiency [174]. From the patient’s perspective, several studies provide evidence that patients exhibit an overall positive attitude regarding the use of wearable sensors, even in their daily life [175–178]. Several groups have generated tools that visualize sensor data [179–182]. When presented to therapists, visualizations were deemed helpful in drawing insightful conclusions from data about patients’ rehabilitation progress [182]. Recently, commercial products have been introduced for use with rehabilitation that generate movement metrics and visual representations of patient performance [183]. While the aforementioned studies have investigated the acceptance of wearable technology for healthcare, the clinical utility of wearable sensor metrics and visualizations for physical therapists working in inpatient rehabilitation has not been explored.

6.2 Interview Design

We presented data collected from the AC wearable sensor study to physical therapists (N = 5) and physical therapy assistants (N = 2) to collect their perceived clinical utility of the information. The group had a mean age of 40.14 ± 9.49 years (M = 1, F = 6) and
had been working in rehabilitation for $11.86 \pm 12.56$ years. Interviews were audio recorded and later transcribed independently by two researchers. The interview consisted of four main components:

1. Familiarity with technology: Conversation related to the following topics:
   
   - Their level of comfort with technology.
   - Their willingness to learn new technology.
   - The current technologies they use.
   - What technologies they wish they had.
   - What visualizations they use.
   - How they evaluate patient gait and transfer ability.
   - How they evaluate change in patient gait and transfer ability.

2. Metric rating: Evaluations on a scale from 1 (not useful) to 5 (very useful) for the following two ratings:
   
   - To rate the metric for how useful it is for providing therapy services for patients.
   - To rate the metric for how useful it is as an indicator of the Functional Independence Measure [7] motor score at discharge.

3. Prediction usefulness: Questions related to the utility of a system providing discharge FIM motor score predictions:
   
   - How useful would you consider the prediction?
• Would you make use of a technology-assisted prediction of FIM motor score to help you provide therapy services? Why or why not?

4. Visualization evaluation: Evaluation of three wearable sensor data visualizations: 1) task duration bar plot (see Figure 6.1), 2) gait cycle bar plot (see Figure 6.2), and 3) effect size forest plot (see Figure 6.3). Evaluations on a scale from 1 (strongly disagree) to 5 (strongly agree) for the following three ratings related to each plot:

• I think that I would use this plot frequently.

• I thought the plot was easy to understand.

• I would image most patients would learn to use this plot very quickly.

Additional questions were asked to facilitate discussion about each plot, including:

• How do you foresee using the plot to help you provide therapy services for your patients?

• What might you change about the plot?

The metrics that we presented to interviewees for evaluation of their clinical utility were selected to be representative of the metrics computed by the research community and commercial sensor systems. We grouped the presented metrics into three categories: clinical assessments of progress (see Table 3.4), whole body movements computed from the COM sensor (see Table 3.5), and gait features (see Table 3.6) derived from the shank sensors. During the interviews, the experimenter explained each metric to the therapy provider. We collected evaluations on a scale from 1 (not useful) to 5 (very useful) for
the following two ratings: 1) to rate the metric for how useful it is for providing therapy services for patients and 2) to rate the metric for how useful it is as an indicator of the Functional Independence Measure [7] motor score at discharge.

In Chapter 5, we investigated the predictive abilities of features derived from wearable inertial sensor data to predict discharge FIM scores without re-administering the FIM assessment battery [184]. The FIM is administered at admission and discharge from in-patient rehabilitation by clinical staff who are credentialed to administer the instrument. The FIM is a well-validated assessment measuring functional status on a 0-7 rating scale for 18 items representing 6 domains: self-care, sphincter control, transfers, locomotion, communication, and social cognition [7]. In addition to a total FIM score, separate scores are developed from the motor function items and cognitive function items. The results of our previous work include leave-one-out-cross-validation correlations between actual and predicted discharge FIM motor scores as high as $r = 0.95$ (NRMSE = 8.50%) for 20 AC study participants. To gather insights about the utility of such wearable sensor-based FIM predictions, during the interview therapists were instructed to consider a system that provides a highly accurate prediction of their patients’ discharge FIM motor scores. The predictions would be available at any point between admission and discharge.

We presented three wearable sensor data visualizations to interviewees for evaluation. The first presented visualization was the task duration bar plot (see Figure 6.1). Task duration plots show AC task durations for performances one week apart (S1 and S2) for an individual patient. The X-axis lists the ambulation circuit tasks and the Y-axis shows task duration, measured in seconds. A bar represents the time to perform the task. Blue
Figure 6.1: The task duration bar plot shows a patient’s ambulation circuit task durations for session 1 (S1) and 2 (S2), one week later.

bars correspond to S1 and green bars correspond to S2, one week later. This plot was chosen because of its simplicity; the amount of time to complete a task is a commonly used clinical assessment of progress, as is the case with the TUG (see Section 2.2). Bar plots are also a common visual representation of data that many people are proficient in reading.
The second presented visualization was a gait cycle bar plot (see Figure 6.2), which plots gait cycle metrics derived from sensors attached to both shanks. The gait cycle plot in Figure 6.2 shows a patient’s left and right shank peak angular velocities. On the X-axis are the gait cycle (stride) numbers during the linear walking section of the ambulation circuit. The Y-axis represents peak angular velocity, measured in degrees per second. A bar represents the peak angular velocity for a gait cycle. Blue bars correspond to the right shank peak angular velocity and green bars correspond to the left shank peak angular velocity. The dashed horizontal lines correspond to the mean peak angular velocity values. The colored horizontal band around the mean is ± one standard deviation. The plot was selected due to its similarity to graphs included in commercial reports [183]. To facilitate discussion with the therapy providers, we presented two gait cycle bar plots, each from a different patient. Figure 6.2 shows the two gait cycle bar plots, one for a participant with a left side gait impairment (participant #006, see Figure 6.2a) and one for a participant with no paresis (patient #005, see Figure 6.2b).

The last presented plot was the effect size forest plot (see Figure 6.3). The effect size forest plot displays effect sizes quantifying change after one week of physical therapy for individual participants and for the participants as a group. An effect size based on Cohen’s $d$ for repeated measures data (see Equation 4.1 in Section 4.2) is used to quantify the strength of changes in each of the computed metrics [120]. Effect size analysis and the effect size forest plot were included in the interview to facilitate conversation about visual presentations of statistically quantified performance change and to determine whether physical therapy providers consider comparisons between participants useful. The X-axis of
Figure 6.2: The gait cycle bar plots visualize shank peak angular velocities for patient #006 (top figure) with a left side impairment and patient #005 (bottom figure) with no paresis. The colored horizontal band around the dashed line mean (μ) is ± one standard deviation (σ).
Figure 6.3 represents effect sizes for the walking smoothness index metric (see Table 3.5 for metric descriptions). The effect size values are shown on the right side of the Y-axis for each individual, with the associated confidence intervals in parentheses. Individual participant IDs are on the left side of the Y-axis. The points in the plot are each individual’s effect size. The horizontal lines, or whiskers, extruding from the points depict 95% confidence intervals. The vertical red dashed line is the effect size for the group and the vertical red band around the dashed line is the 95% confidence interval for the group.

6.3 Results

All quantitative responses were on a scale from 1 (strongly disagree/not useful) to 5 (strongly agree/very useful). The therapy providers were quite comfortable with technology (4.00 ± 0.82), willing to learn new technology (4.29 ± 0.76), and interested in using wearable technology for their patients (4.43 ± 0.53). The technology therapists use to help provide therapy services for their patients included computers, the Lokomat robotic-assistive device, electrical stimulation, the Nintendo Wii, and video cameras. Of the seven interviewees, five stated a desire for technology for balance assessment and gait analysis. To evaluate patient gait and transfer ability, therapists primarily use observation and an estimate of the amount of physical assistance the patient requires to perform certain tasks. To evaluate change in patient gait and transfer ability, therapists use their memory to compare previous observations to current ones. One therapist listed several movements she looks for, “I kinda compare and contrast step lengths. I will do speed. I will do trunk
Figure 6.3: The forest plot visualizes walking smoothness index effect sizes after one week of physical therapy. The whiskers represent the 95% confidence interval (CI). The vertical dashed line is the group effect size and the surrounding band is the group 95% CI.
deviation. If there’s any toe drags. If they are using an assistive device or not. If they are using orthoses or not.” All seven therapists stated they do not currently use visualizations, plots, graphs, or drawings to describe their patients’ ambulatory ability.

Table 6.1 lists the mean and standard deviation of the interviewees’ usefulness and FIM indicator ratings for each metric. Table 6.2 contains rating responses regarding the usability of each plot presented to the therapy providers (see Section 6.2 for an overview of the questions asked and visualizations). Regarding predictions, interviewees were impartial about considering discharge FIM motor score predictions useful (3.43 ± 1.27).

6.4 Discussion

Of the three metric categories, metrics in the clinical assessments of progress group were rated highly for both usefulness (mean 3.66) and as a FIM indicator (mean 2.98) compared to whole body movement (mean 3.24, FIM mean 2.52) and gait features (mean 3.69, FIM mean 2.49) groups. Interestingly, a few of these CAP metrics (e.g. walking speed and vehicle load/unload duration) were correlated highly with discharge FIM motor score; however, recursive feature elimination did not select any CAP metrics as top-ranked features (see Section 5.6.1 for more details).

Of all the metrics, sit-to-stand duration was rated the highest for usefulness (4.14 ± 1.46). Walking speed (4.00 ± 1.41) and cadence (4.00 ± 1.00) were also highly rated. These metrics can be computed without wearable sensors, indicating a preference toward familiar metrics with previously established clinical validity. The metrics with the lowest rated
Table 6.1: Average therapy provider-rated metric usefulness for providing therapy services for patients and as an indicator of the discharge Functional Independence Measure (FIM) score. The scale was 1 (not useful) to 5 (very useful). Standard deviations are in parentheses. Horizontal lines divide the metric categories (clinical assessments of progress, whole body movement metrics, and gait features). Within each category, metrics are sorted by their usefulness rating.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Usefulness</th>
<th>FIM Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sit-to-stand duration</td>
<td>4.14 (1.46)</td>
<td>3.29 (1.70)</td>
</tr>
<tr>
<td>Walking speed</td>
<td>4.00 (1.41)</td>
<td>3.29 (1.89)</td>
</tr>
<tr>
<td>Stand-to-sit duration</td>
<td>3.86 (1.46)</td>
<td>3.29 (1.60)</td>
</tr>
<tr>
<td>Curvilinear walking duration</td>
<td>3.71 (1.38)</td>
<td>3.29 (1.38)</td>
</tr>
<tr>
<td>Total ambulation circuit duration</td>
<td>3.71 (1.38)</td>
<td>2.71 (1.11)</td>
</tr>
<tr>
<td>Floor surface speed ratio</td>
<td>3.57 (1.27)</td>
<td>3.14 (1.35)</td>
</tr>
<tr>
<td>Vehicle load duration</td>
<td>3.14 (1.07)</td>
<td>2.43 (1.13)</td>
</tr>
<tr>
<td>Vehicle unload duration</td>
<td>3.14 (1.07)</td>
<td>2.43 (1.13)</td>
</tr>
<tr>
<td>Walking smoothness</td>
<td>3.71 (1.25)</td>
<td>2.71 (1.60)</td>
</tr>
<tr>
<td>Center of mass movement intensity</td>
<td>3.14 (1.07)</td>
<td>2.57 (1.40)</td>
</tr>
<tr>
<td>Center of mass peak angular velocity</td>
<td>2.86 (1.07)</td>
<td>2.29 (1.25)</td>
</tr>
<tr>
<td>Cadence</td>
<td>4.00 (1.00)</td>
<td>2.86 (1.21)</td>
</tr>
<tr>
<td>Single support percent</td>
<td>3.86 (1.35)</td>
<td>2.71 (1.60)</td>
</tr>
<tr>
<td>Gait cycle duration</td>
<td>3.71 (1.38)</td>
<td>2.71 (1.50)</td>
</tr>
<tr>
<td>Number of gait cycles</td>
<td>3.71 (1.25)</td>
<td>2.71 (1.60)</td>
</tr>
<tr>
<td>Shank range of motion</td>
<td>3.71 (1.38)</td>
<td>2.43 (1.27)</td>
</tr>
<tr>
<td>Step length</td>
<td>3.71 (1.25)</td>
<td>2.43 (1.27)</td>
</tr>
<tr>
<td>Stride length</td>
<td>3.71 (1.25)</td>
<td>2.43 (1.27)</td>
</tr>
<tr>
<td>Stride regularity</td>
<td>3.71 (1.25)</td>
<td>2.43 (1.27)</td>
</tr>
<tr>
<td>Step regularity</td>
<td>3.57 (1.27)</td>
<td>2.29 (1.38)</td>
</tr>
<tr>
<td>Double support percent</td>
<td>3.43 (1.40)</td>
<td>2.29 (1.25)</td>
</tr>
<tr>
<td>Shank peak angular velocity</td>
<td>3.43 (1.40)</td>
<td>2.14 (1.07)</td>
</tr>
</tbody>
</table>

usefulness included vehicle load/unload duration, center of mass peak angular velocity, and center of mass movement intensity. For the vehicle duration metrics, two therapists stated a patient’s ability to complete these tasks is more important than the amount of time the patient requires. For center of mass movement intensity, one therapist stated, “I’m not sure what that relates to,” indicating she was possibly trying to map the acceleration-based metric into an assessment she was familiar with. When she was unable to produce such a mapping, she rated this metric low (2, not useful).
Table 6.2: Average therapy provider-rated responses to questions regarding the presented visualizations. The scale was 1 (strongly disagree) to 5 (strongly agree). Standard deviations are in parentheses.

<table>
<thead>
<tr>
<th>Task duration bar plot (mean rating 3.62)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that I would use this plot frequently</td>
<td>3.14 (0.90)</td>
</tr>
<tr>
<td>I thought the plot was easy to understand</td>
<td>4.43 (0.53)</td>
</tr>
<tr>
<td>I would imagine most patients would learn to use this plot very quickly</td>
<td>3.29 (0.95)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gait cycle bar plot (mean rating 2.29)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that I would use this plot frequently</td>
<td>2.57 (0.98)</td>
</tr>
<tr>
<td>I thought the plot was easy to understand</td>
<td>2.86 (0.90)</td>
</tr>
<tr>
<td>I would imagine most patients would learn to use this plot very quickly</td>
<td>1.43 (1.13)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect size forest plot (mean rating 1.76)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that I would use this plot frequently</td>
<td>1.86 (0.90)</td>
</tr>
<tr>
<td>I thought the plot was easy to understand</td>
<td>2.00 (1.00)</td>
</tr>
<tr>
<td>I would imagine most patients would learn to use this plot very quickly</td>
<td>1.43 (0.79)</td>
</tr>
</tbody>
</table>

Four of the seven therapists were enthusiastic about discharge FIM motor score prediction from wearable sensor data. One therapist stated, “It would be very useful, it could help with discharge planning if we needed to steer one way or another.” Another therapist stated, “I would make use of [the predictions] as an adjunct.” One of the three therapists who were not convinced of the utility of such a FIM prediction stated, “I probably wouldn’t [use FIM predictions], mostly because patients are really variable.” To address the mixed feelings of therapists about clinical outcome predictions, it would be best to present the system as a tool to augment the information available to therapists when making treatment decisions, and not as a replacement of current methods.

For the task duration bar plot (see Figure 6.1), all therapists noted the patient was faster for each task except for stand-to-sit. One therapist stated, “It may have been an
improvement in a safety factor, whereas they may have sat down due to a loss of balance.” This comment suggests quantitative information alone is not sufficient for therapists to determine if a reduction in a value should be classified as an improvement or regression for an individual patient. Perhaps coupling the quantitative results with a video or 3D animation of the patient performing the task would provide sufficient context for the numeric values. Only one therapist suggested changes to the task duration bar plot by recommending the tasks on the X-axis are grouped by activity instead of by location in the AC sequence (e.g., sit-to-stand next to stand-to-sit). The therapists also thought patients could understand the task duration bar plot (3.29 ± 0.95, see Table 6.2). One therapist stated, “[The plot] would be helpful to get the patient more involved in seeing their progress.” Another therapist also felt the plot could be useful for engaging patients, “I could use this as feedback for a patient as to what changes have been made and extrapolate as to why that is important that they are faster on these tasks.” The other two plots, gait cycle bar plot and effect size forest plot, were deemed too complicated for patients to learn (1.43 ± 1.13 and 1.43 ± 0.79, respectively).

For the gait cycle bar plot, all seven therapists correctly classified left side paresis for patient #006 (see Figure 6.2). They also identified several differences between the gaits of patient #006 and patient #005, such as the higher variability for patient #006’s peak angular velocity. Four of the seven therapists stated they would use the gait cycle bar plot, with one therapist stating, “It would be for my own personal measures to see where they are at. The ankle is really tough to assess when you are by their shoulders.” This comment suggests additional utility of wearable sensors could be acquired by placing
them on the body in hard to observe areas. One therapist advocated placing sensors on
the hip, knee, and ankle joints. Changes proposed by the therapists for this plot included
increasing the font size and removing the standard deviation information to reduce the
plot complexity.

The effect size forest plot (see Figure 6.3) was the lowest rated of the visualizations
presented to the therapists (see Table 6.2). Since the plot is statistically-oriented and
compares patients with other patients, the low ratings were expected. Two of the thera-
pists acknowledged the usefulness of the effect size forest plot for research, “If I got into a
research study to justify what I was doing then yes, but not for direct patient care.” Sugges-
tions for improving the effect size forest plot included grouping patients by diagnosis,
or including only one patient’s data to remove comparisons between patients. Effect size
forest plots can show single participant change for gait cycle-based features, such as stride
length. For example, each AC trial produces multiple stride length measurements, one
stride length for each gait cycle. Stride length effect sizes between two AC trials can be
computed using Equations 4.1, 4.2, and 4.3. Furthermore, multiple gait feature effect
sizes can be plotted on the Y-axis of an effect size forest plot (e.g. left/right shank range
of motion, left/right shank peak angular velocity, etc.).

A limitation of this study includes the small number of physical therapists and physi-
cal therapy assistants that were interviewed. Additional limitations include the interviewed
therapy providers were all affiliated with the same rehabilitation facility. Therapist di-
versity would yield more representative results. Also, the interviewed therapy providers
experienced unequal exposure to the ambulation circuit study prior to the interview. Five
of the seven therapists were active in the participant recruiting process and observed the data collection protocol. Future work can address these limitations by conducting a larger, multi-facility study to collect physical therapists’ feedback regarding wearable sensors for mobility assessment. We plan to incorporate the results obtained from the current preliminary study in the design process, including computing metrics that map into standard clinical assessments of progress; designing visualizations containing a single patient’s information, possibly compared to normative values for their age/etiologic group; motivating patients with simple sensor-based visualizations of their progress, especially since patient motivation and engagement is associated with positive rehabilitation outcomes [185]; and coupling wearable sensor data with videos and/or animations.

In summary, we presented wearable sensor-based computing algorithms for mobility assessment. To bridge the gap between design of mobility monitoring technology and actual use of the technology, we interviewed physical therapy providers at an inpatient rehabilitation facility to collect their opinions on the clinical utility of wearable sensor data and associated algorithms. The responses indicated providers to be interested in using wearable technology and sensor data visualizations.

This chapter closes the current research and results regarding mobility assessment for inpatient rehabilitation applications. The remainder of this dissertation focuses on detecting and analyzing changes in sensor data collected from everyday mobility, the second component of our hypothesis. The next chapter, Chapter 7, presents our Physical Activity Change Detection framework and applies it to data collected from wearable sensors to analyze changes in physical mobility. Chapter 8 adapts PACD to smart home sensor data
to track changes in behavior patterns.
In this dissertation we hypothesize that sensor data can be collected and analyzed using machine learning techniques to provide insights on rehabilitation and behavior change. In the previous chapters we focused on providing evidence in support of wearable sensors for quantifying mobility changes during rehabilitation. To address the behavior change portion of our hypothesis, we investigate changes in physical activity with wearable sensors in this chapter and changes in behavior with smart home sensors in the following chapter (see Chapter 8).

Physical activity (PA) is often defined as any bodily movement by skeletal muscles that results in caloric energy expenditure [186]. Physical activity consists of bouts of movement that are separated by periods of rest. Physical activity bouts are composed of four dimensions [186]:

1. Frequency: the number of bouts of physical activity within a time period, such as a day.
2. Duration: the length of time an individual participates in a single bout.
3. Intensity: the physiological effort associated with a particular type of physical activity bout.
4. Activity type: the kind of exercise performed during the bout.

To add exercise throughout the day, individuals can increase their number of bouts
(frequency), increase the length of bouts (duration), increase the intensity of bouts, and vary the type of physical activity performed during the bouts. These four components of PA represent four distinct types of changes that can reflect progress toward many different health goals, such as increasing physical activity or consistency in one’s daily routine.

7.1 Prior Work on Sensor-based Physical Activity Monitoring

Several studies have measured physical activity with various types of sensor technologies. For this dissertation, we focus on studies utilizing wearable, wireless sensors for monitoring physical activities. Such monitoring systems typically use a bi-axial or tri-axial accelerometer worn on the wrist [187], the upper arm [188], on the waist [189], clipped onto a belt or clothing, or carried in a bag or pocket. The number of steps taken is the most common measurement of physical activity extracted from accelerometer data. Steps are identified from accelerometer-based activity trackers by identifying episodic and cyclical body movements of vertical and anterior-posterior acceleration [190,191]. Once steps have been identified, metrics such as physical activity levels (i.e. sedentary, light intensity, moderate intensity, vigorous intensity) [192–195], estimates of energy expenditure [196–199], and distance traveled can be computed.

In the literature, research has investigated sensor-based physical activity monitoring for inpatient rehabilitation [118,200,201] and elderly populations [68,187–189,202,203]. Beginning with rehabilitation, the Stroke Inpatient Rehabilitation Reinforcement of Activity (SIRRACT) trial was the first international, multi-facility trial to deploy wearable
sensors for patients undergoing stroke rehabilitation [118]. Data were collected from triaxial accelerometers worn on each ankle from 135 participants in 11 different countries. Therapists mounted the sensors on participants in the morning and removed the sensors in the evening. From the collected acceleration signals, walking bouts were identified. Metrics related to walking bouts were computed, including: speed, duration, number of walking bouts, average walking speed, total time walking, total distance, and total steps taken. The participants were split into two groups: one receiving feedback regarding their walking speed only and one receiving, in addition to walking speed information, feedback in the form of activity graphs. The results indicated no significant differences existed between the two feedback groups in daily time spent walking (15.1 ± 13.1 minutes for walking speed feedback only compared to 16.6 ± 14.3 minutes for activity graph feedback). Additional findings of the study included 30% of participants decreased their total daily walking time over the course of inpatient rehabilitation and the majority of walking bouts only lasted between 10 and 30 seconds [118]. More recent studies investigating walking bout feedback during inpatient rehabilitation are published by Hornby et al. [200] and Mansfield et al. [201].

Outside of inpatient rehabilitation, there have been a handful of studies utilizing sensors to track physical activity for elderly populations [68, 187–189, 202, 203]. For example, Tan et al. [68] designed an indoor activity monitoring system by using a Fitbit Flex wrist-worn fitness tracker and radio frequency identification (RFID). The activity levels of participants were visualized with activity density maps (ADMs) (see Figures 7.5, 7.6, 7.7, 7.8, and 7.9 for example ADMs). An ADM is heat map proposed by Wang et al.
where the day of the week is represented on the X-axis and the hour of the day is represented on the Y-axis. The map is divided into a grid where the color of each cell represents the number of steps taken on that particular day at that particular hour. In the work by Tan and colleagues, an ADM was encoded as a 2-dimensional matrix, where each cell stores the level of activity (encoded as color on the ADM) [68]. When utilizing an ADM in their study, the first week was considered a baseline for a participants’ activity levels. Then, dissimilarity metrics were applied to subsequent weeks in the ADM to quantify changes in physical activity and health.

Furthermore, prior work determining the accuracy of physical activity measured by wearable devices is a relevant area of research. Many manufacturers do not report the accuracy of their products. The research community has aimed to address this gap by performing trials with healthy subjects to measure the accuracy of computed steps taken [204–206], energy expenditure [207, 208], energy expenditure specifically for combined accelerometer and heart rate systems [209–211], and sleep quality [212] for popular pedometers and fitness trackers. While the percentage of errors for healthy individuals appears to be tolerable for most applications, pedometers and fitness trackers have been found to not work as well with older adults who walk slowly or people with gait impairments [190, 213, 214]. A few studies have investigated the accuracy of popular consumer fitness trackers for non-healthy populations [215, 216]. Since we utilize Fitbit data to evaluate PACD, research investigating the accuracy of Fitbit devices are highly relevant. Several of the aforementioned studies evaluate the Fitbit line of products for healthy and unhealthy adults [204–207, 212, 215–217]. We summarize the Fitbit accuracy findings of these studies
here:

- When comparing the step counting accuracy of the Fitbit Flex, Fitbit One, and 8 other consumer activity trackers, the Fitbit Flex demonstrated the second lowest mean percentage error for healthy adults walking on a treadmill (-5.7%) and in free-living conditions (3.7%) [217].

- When comparing the step counting accuracy of the Fitbit (models One, Ultra, and Zip), Nike+ Fuelband, Nike+ Sportsband/Motion, Apple iPhone Moves application, Omron Steps mechanical pedometer, and SM-2000 mechanical pedometer, the Fitbit devices had the highest accuracy of approximately 1% error for healthy adults performing a 400 meter outdoor walking task [205].

- The mean difference between actual number of steps and number of steps measured by a Fitbit Ultra worn by participants with chronic stroke or traumatic brain injury performing a 2 minute walk test was found to be -9.7 steps [215].

- A high agreement (ICC = 0.88, 95% confidence interval 0.76 to 0.94) between Fitbit-measured steps (models One and Zip) and visually counted steps was reported for 32 community-dwelling older adults during a 2 minute walk test [216].

- For ten healthy young adults performing a 20 step test, the Fitbit had step count errors less than 5% [206]. This study also found Fitbits did not record spurious movement during motor vehicle tests.
7.2 Window-based Comparisons

We study the problem of detecting and analyzing change in physical activity patterns. More specifically, we introduce a framework called Physical Activity Change Detection, or PACD, to determine if a significant change exists between two windows of time series step data (data which indicates the number of steps taken by an individual over a period of time) sampled from a physical activity sensor. Algorithm 7 and Figure 7.1 outline this process. Let \( X \) denote a sample of time series step data segmented into days, \( D = \{x_1, x_2, \ldots, x_t, \ldots, x_m\} \), where \( x_t \) is a scalar number of steps taken at time interval \( t = 1, 2, \ldots, m \) and \( m \) is the number of equal-sized time intervals in a day. Let \( t_{\text{mins}} \) denote the number of minutes per time interval, \( t \). For example, if the sampling rate of the wearable sensor device is one reading per minute, \( t_{\text{mins}} = 1 \) minute and \( m = 1440 \) minutes / \( t_{\text{mins}} \) = 1400 intervals. Now, let \( W \) be a window of \( n \) days such that \( W \subseteq X \). Furthermore, an aggregate window, \( \hat{W} \), represents the average of all days within the window \( W \):

\[
\hat{W} = \frac{1}{n} \sum_{i=1}^{n} D_i, D_i \in W \tag{7.1}
\]

We can compare windows of data within time series data \( X \). These windows may represent consecutive times (e.g., days, weeks, months), a baseline window (e.g., the first week) with each subsequent time window, or overlapping windows. Let \( W_i \) denote a window starting at day number \( i \) of \( X \) (\( i \geq 1 \)) such that \( W_i = X[i : i + n - 1] = \{D_i, D_{i+1}, \ldots, D_{i+n-1}\} \). Suppose we have two windows of data, \( W_i \) and \( W_j \) (\( i \leq j \)). Windows \( W_i \) and \( W_j \) can be formed as subsets of \( X \) based on the initial value of \( i \) and a parameter \( \text{offset} \) that determines the initial value of \( j \) (\( j = i + \text{offset} \)). For change detection and analysis, a function \( F \)
computes a change score, \( CS = F(W_i, W_j) \) between \( W_i \) and \( W_j \). Iteration advancements \( adv_i \) and \( adv_j \) move windows \( W_i \) and \( W_j \) respectively for the next comparison. Two windows can be compared in either baseline or sliding window mode. For a baseline window comparison, the first window is a reference window that occurs at the beginning of the time series (\( i \) is initialized to 1) and is used in each comparison, so \( adj_i = 0 \). All subsequent windows are compared to the baseline window. Thus \( j \) is initialized to \( 1 + offset \) and is subsequently advanced by \( adv_j \). In the case of a sliding window comparison, both windows used for comparison are advanced through the time series data. Typically \( adv_i = adv_j \) for consistently spaced comparisons. In the PACD algorithm (see Algorithm 7) \( i \) is initialized to 1 and \( j \) is initialized to \( 1 + offset \). In steps 17 and 18, \( i \) is advanced to \( i + adv_i \) and \( j \) is advanced to \( j + adv_j \).

**Algorithm 7 PACD(\( X, n, offset, adv_i, adv_j \))**

1: Input: \( X = \) time series data
2: Input: \( n = \) window length in days
3: Input: \( offset = \) number of days separating windows
4: Input: \( adv_i = \) number of days to advance the first window
5: Input: \( adv_j = \) number of days to advance the second window
6: Output: \( V = \) vector of change scores
7: Initialize: \( i = 1 \) and \( j = 1 + offset \)
8: for each pair of windows to compare, \( W_i \) and \( W_j \) of time series \( X \):
    9: \( W_i = X[i : i + n - 1] \)
    10: \( W_j = X[j : j + n - 1] \)
    11: Compute \( CS = F(W_i, W_j) \)
    12: Determine if \( CS \) is significant
    13: Identify the type of change that is exhibited
    14: Manual inspection of change
    15: Unsupervised inspection (change analysis)
    16: Append \( CS \) to change score vector \( V \)
    17: \( i = i + adv_i \)
    18: \( j = j + adv_j \)
end for
19: return Change score vector \( V \)
The choice of window size, $n$, limits the algorithms that can be applied to the data. For example, the PCAR algorithm [69] is designed for longitudinal data comprising several months; consequently sensitivity decreases with small window sizes. For PACD, we categorize choices for window size $n$ into the following descriptors:

1. **Small window ($n = 1$ day).** Suitable for performing day-to-day comparisons (e.g. $D_i(\text{Monday})$ compared to $D_j(\text{Monday})$, $D_i(\text{Tuesday})$ compared to $D_j(\text{Tuesday})$, ...) or aggregate day comparisons (e.g. $\bar{W}_i$ compared to $\bar{W}_j$, $\bar{W}_{i+\text{adv}_i}$ compared to $\bar{W}_{j+\text{adv}_j}$, ...).

2. **Medium window ($2 \leq n \leq 5$ days).** Suitable for performing weekday-to-weekday comparisons (e.g. $W_i$ compared to $W_j$ where $W_i = \{D_i(\text{Monday}), D_i(\text{Tuesday}), D_i(\text{Wednesday}), D_i(\text{Thursday}), D_i(\text{Friday})\}$ and $W_j = \{D_j(\text{Monday}), D_j(\text{Tuesday}), D_j(\text{Wednesday}), D_j(\text{Thursday}), D_j(\text{Friday})\}$) or weekend-to-weekend comparisons.

Figure 7.1: An overview of Physical Activity Change Detection (PACD).
3. Large window \((n > 5 \text{ days})\). Suitable for performing week-to-week or month-to-month comparisons.

### 7.2.1 Change Detection Algorithms

In the following sections, we describe algorithmic options for the window-based change score function, \(F\). These algorithms, which include those reported in the literature and new enhancements that we introduce, are incorporated into PACD and evaluated on collected step data. A summary and comparison of the algorithms is listed in Table 7.1.

**RuLSIF**

Non-parametric approaches to change point detection include a family of methods comparing the probability distributions of two time series samples to determine the corresponding dissimilarity. A greater difference between the two distributions implies a higher likelihood that a change occurred between the two samples. Instead of estimating the probability distributions, their ratio can be estimated and used to detect changes in the underlying probability distributions. Direct density ratio estimation between two windows of time series data is substantially simpler to solve than computing the windows’ probability densities independently and then using these to compute the ratio. Unconstrained Least-Squares Importance Fitting (uLSIF) \([70]\) is one such ratio estimation approach that measures the difference between two samples of data surrounding a candidate change point. For this approach, the density ratio between two probability distributions is estimated di-
Table 7.1: Window-based change detection algorithms.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Window size</th>
<th>Window preprocessing</th>
<th>Change score</th>
<th>Change significance test</th>
</tr>
</thead>
<tbody>
<tr>
<td>RuLSIF [70]</td>
<td>Any</td>
<td>Hankel matrix</td>
<td>Probability density ratio estimation with Pearson divergence</td>
<td>Threshold learning in supervised applications. N/A for unsupervised applications</td>
</tr>
<tr>
<td>Texture-based [67,68]</td>
<td>Any</td>
<td>Grey-level co-occurrence matrix, texture features</td>
<td>Weighted normalized Euclidean distance</td>
<td>N/A</td>
</tr>
<tr>
<td>PCAR [69]</td>
<td>Large</td>
<td>$m \times N$ KL distance permutation matrix</td>
<td>Count of time intervals with significant changes (proportion of permuted KL distances greater than observed window)</td>
<td>N/A</td>
</tr>
<tr>
<td>sw-PCAR</td>
<td>Any</td>
<td>$N$ KL distance permutation vector</td>
<td>KL distance</td>
<td>Non-parametric outlier detection based on Boxplot analysis</td>
</tr>
<tr>
<td>Virtual classifier [71]</td>
<td>Large</td>
<td>Physical activity features (intra-day and inter-day if window size &gt; 1)</td>
<td>Cross validation prediction accuracy of binary classifier</td>
<td>Hypothesis testing based on prediction accuracy exceeding a threshold</td>
</tr>
</tbody>
</table>

KL = Kullback-Leibler, $m =$ number of time intervals, $N =$ number of permutations
directly with the Pearson divergence dissimilarity measure. Depending upon the data, the Pearson divergence can be unbounded. Consequently, a modification to uLSIF, relative uLSIF (RuLSIF), utilizes an alpha-relative Pearson divergence to bound the change score above by $1/\alpha$ for $\alpha > 0$ [70].

Texture-based Dissimilarity

For the texture-based approach, two windows of PA data, $W_i$ and $W_j$, are considered. From these data windows, 2-dimensional matrices are generated with rows corresponding to time intervals, columns corresponding to days, and cells containing step values measured from a PA device (see Section 7.3.1 for visualizations of PA matrices in Figures 7.5, 7.6, 7.7, 7.8, and 7.9). In order to extract texture features from the data, each matrix is converted into a grey-level co-occurrence matrix (GLCM), a histogram of co-occurring grey scale values of an image [218]. Next, texture features are computed from each co-occurrence matrix, including contrast, homogeneity, angular second moment, energy, density, and correlation features [67, 68, 218]. Figure 7.2 shows an example of computing texture features from PA step data. Parameters required to compute a GLCM include a list of pixel pair distances (e.g. $\{1\}$) and a list of pixel pair angles (e.g. $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$). The features from each window produce feature vectors $T_i$ and $T_j$. Finally, to compare two windows $W_i$ and $W_j$ for changes, a weighted normalized Euclidean distance measure is used as a change score to quantify the differences between the corresponding feature vectors $T_i$ and $T_j$. The smaller the Euclidean distance between these two vectors, the more similar the two windows of data are. Figure 7.3 shows an overview of applying the texture-based method for a comparison between two windows. The texture-based approach can operate
on small or large window sizes; however, the method lends itself more appropriately to large window sizes (Wang et al. [67] used window size of one month).

**sw-PCAR**

We propose an enhancement of PCAR to allow permutation-based change detection for any window size. Before introducing sw-PCAR, we will provide an overview of the PCAR approach. PCAR utilizes smart home sensor data to detect changes in behavioral routines with an *activity curve* model [69]. The PCAR approach assumes that an activity recognition algorithm [219] is available to label the sensor data with corresponding activity names. Using PCAR, each day within a window is broken into *m* time intervals. The

---

![Figure 7.2: Grey level co-occurrence matrix example.](image-url)
activities occurring within each time interval are modeled by a probability distribution, which form an activity curve for the corresponding window. To compute a change score $CS$ between two windows $W_i$ and $W_j$, the two corresponding activity curves are first maximally aligned with dynamic time warping (DTW). Next, the symmetric Kullback-Leibler (KL) divergence is used to compute the distance between each pair of DTW-aligned activity distributions [69]. To test significance of the distance values, $W_i$ and $W_j$ are concatenated to form a window $W$ of length $2n$ days. Next, all days within $W$ are shuffled. The first half of the shuffled days form a new first window, $W_i^*$, while the second half form a new second window, $W_j^*$. KL distances for each time interval pair in $W_i^*$ and $W_j^*$ form a vector that is inserted into a matrix. This shuffling procedure is repeated $N$ times, producing a $N \times m$ permutation matrix, $M$. If $N$ is large enough, $M$ forms an empirical distribution of the possible permutations of activity data within the two windows of time. Next, for each time interval, the number of permuted KL distances that exceed the original DTW-
aligned distance is divided by $N$ to form a p-value. After computing a p-value for each time interval, the Benjamini-Hochberg correction [220] is applied for a given $\alpha$ ($\alpha < 0.05$). Finally, the remaining significant p-values are counted to produce the change score, $CS$.

While the PCAR algorithm is intended for activity distribution data available from activity recognition algorithms, we adapt PCAR to analyze physical activity data as part of our PACD method. Instead of activity distribution vectors, we use scalar step counts. Additionally, PCAR is suitable for only large window sizes due to the requirement of permuting daily time series data. We propose a version of PCAR that is more suitable for small to medium-sized windows (sw-PCAR) as required by PACD. Finally, PCAR was originally proposed for correlating change scores with standardized clinical assessments to determine if ambient smart home sensor-based algorithms can detect cognitive decline [69]. Consequently, there is not a test for significance of PCAR change scores. In Section 7.2.2 we propose an accompanying significance test for sw-PCAR.

Algorithm 8 outlines the sw-PCAR approach. For sw-PCAR, two windows $W_i$ and $W_j$ are averaged to yield aggregate windows $\hat{W}_i$ and $\hat{W}_j$ (see Equation 7.1). A change score $CS$ is derived by computing the KL distance between the average number of steps taken in $\hat{W}_i$ and the average number of steps taken in $\hat{W}_j$. Next, $\hat{W}_i$ and $\hat{W}_j$ are concatenated to form a window $W$ of length two days. All time intervals within $W$ are shuffled. The first half of the shuffled intervals form a new first window, $W_i^*$; while the second half form a new second window, $W_j^*$. $W_i^*$ and $W_j^*$ are each averaged to produce two step values. The KL distance between the two values is computed and inserted into a vector. This is repeated $N$ times to produce a $N$-length vector $V$ of KL distances. Vector $V$ is later used
for change score significance testing (see Section 7.2.2).

**Algorithm 8** sw-PCAR($W_i, W_j, N$)

1: Input: $W_i, W_j = \text{two windows of time series data}$
2: Input: $N = \text{number of permutations}$
3: Output: $CS = \text{change score}$
4: Output: $sig = (\text{Boolean}) \text{ significance of } CS$
5: Initialize: $k = 0$
6: Initialize: $V$ as a vector of length $N$
7: Compute $\hat{W}_i, \hat{W}_j$ aggregate windows
8: Compute $CS$, the KL distance between $\hat{W}_i$ and $\hat{W}_j$
9: while $k < N$:
10: Shuffle the time intervals of $\hat{W}_i$ and $\hat{W}_j$
11: Generate new aggregate windows $\hat{W}_i^*$ and $\hat{W}_j^*$
12: Compute the KL distance between $\hat{W}_i^*$ and $\hat{W}_j^*$
13: Store resulting distance in $V$
14: $k = k + 1$
end while
15: $sig = \text{BoxplotOutlierDetection}(CS, V)$ (see Algorithm 9)
16: return $CS, sig$

**Virtual Classifier**

Change analysis, as proposed by Hido et al. [71], utilizes a virtual binary classifier to detect and investigate change. We apply the VC approach as part of PACD for large window sizes. First, a feature extraction step reduces two windows $W_i$ and $W_j$ into two $n \times z$ feature matrices, $M_i$ and $M_j$, where $n$ is the window size (in days) and $z$ is the number of features that are extracted (see Section 7.2.3 for feature descriptions). Next, each daily feature vector of $M_i$ is labeled with a positive class and each daily feature vector of $M_j$ is labeled with a negative class. VC trains a decision tree to learn the decision boundary between the virtual positive and negative classes. The resulting average prediction accuracy based on $k$-fold cross validation is represented as $p_{VC}$. If a significant change exists between $W_i$ and $W_j$, the average classification accuracy $p_{VC}$ of the learner
should be significantly higher than the accuracy expected from random noise, $p_{\text{rand}} = 0.5$, the binomial maximum likelihood of two equal length windows [71].

### 7.2.2 Change Significance Testing

Significance testing of change score $CS$ is necessary to interpret change score values. For the VC approach, Hido et al. [71] proposed a test of significance to determine if $p_{VC}$ is significantly greater than $p_{\text{rand}}$. For this test, the inverse survival function of a binomial distribution is used to determine a critical value, $p_{\text{critical}}$, at which $n$ Bernoulli trials are expected to exceed $p_{\text{rand}}$ at $\alpha$ significance. If $p_{VC} \geq p_{\text{critical}}$, a significant change exists between the two windows, $W_i$ and $W_j$.

The PCAR approach does not have an accompanying test of significance. We address this with our proposed sw-PCAR technique. sw-PCAR computes change significance by comparing $CS$ to the permutation vector $V$ with boxplot-based outlier detection (see Algorithm 9). An outlier can be defined as an observation which appears to be inconsistent with other observations in the dataset [221]. For this method, the interquartile range ($75^{\text{th}}$ percentile - $25^{\text{th}}$ percentile) of $V$ is computed. Values outside of the $Q_3 + 1.5 \cdot \text{IQR}$ are considered outliers [222]. If $CS$ is determined to be an outlier of $V$, then the change score is considered significant. There are alternative approaches to test membership of an observation (i.e. $CS$) to a sample distribution (i.e. $V$) other than boxplot outlier detection. If the sample is normal, statistical tests such as Grubb’s test for outliers [223] can be applied. However, the assumption of normality does not hold for all samples of
human behavior data. More advanced alternatives include data mining techniques relevant to outlier detection [221,224]. Exploration and testing of such data mining techniques are left as future work.

**Algorithm 9** BoxplotOutlierDetection($CS, V$)

1: Input: $CS =$ change score between two windows  
2: Input: $V =$ sample distribution vector  
3: Output: $sig =$ (Boolean) significance of $CS$  
4: Arrange $V$ in ascending order  
5: Compute $Q_1$, the $25^{th}$ percentile of $V$  
6: Compute $Q_3$, the $75^{th}$ percentile of $V$  
7: Compute the interquartile range of $V$, $IQR = Q_3 - Q_1$  
8: if $CS > Q_3 + 1.5 \cdot IQR$:  
9: $sig =$ True  
10: else:  
11: $sig =$ False  
12: return $sig$

RuLSIF does not explicitly provide a method to determine a cutoff threshold for values of the Pearson divergence function that are considered significant change scores. In supervised applications where ground truth change labels are available, a threshold parameter is typically learned by repeated training and testing with different parameter values. For unsupervised applications, domain knowledge and/or alternative data-driven approaches are necessary. Like RuLSIF, the texture-based method also does not provide a test of change significance. For the RuLSIF and texture-based approaches, we propose a large window change significance test based on intra-window variability and outlier detection.

Our proposed change significance test utilizes the existence of day-to-day variability in human behavior patterns [225]. In order to consider a change between two windows as significant, the magnitude of change should exceed the day-to-day variability within
each window. To illustrate, consider two adjacent, non-overlapping windows $W_1$ and $W_7$, each of length $n = 6$ days. Now run a pairwise sliding window change algorithm over $W_1$ concatenated with $W_7$. If there is a significant change between the windows, the magnitude of change should be higher for the inter-window comparison (between days 6 and 7) than any other intra-window comparison. Figure 7.4 shows an example plot of sw-PCAR change scores for real Fitbit data illustrating this phenomenon. There are small, noisy day-to-day changes for all comparisons except the largest maximum occurring for the inter-window comparison (6th change score).

Based on the assumption that a significant inter-window change should exceed intra-window change, we propose an intra-window change significance test (see Algorithm 10). Given a change score $CS$ between two windows, the task is to determine if $CS$ is significant. To do this, first compute a list of all possible daily change scores, $DCS$, within each
window. $DCS$ contains $2 \cdot \text{Combination}(n, 2)$ change scores (see Algorithm 11). For example, a week-to-week comparison ($n = 7$) would generate an intra-window daily change score sample of 42 day-to-day variations. Next, apply the outlier detection method (see Algorithm 9) from sw-PCAR to test if $CS$ is an outlier score when compared to the distribution of intra-window daily change scores $DCS$. Advantages of the proposed test include it is non-parametric and can be coupled with any small window change detection function, $F$. Furthermore, the candidate change score, $CS$, can be computed based on any window size (e.g. Monday-to-Monday, aggregate-to-aggregate, week-to-week, etc.).

**Algorithm 10 IntraWindowSignificance($W_i, W_j, n, CS, F$)**

1: Input: $W_i, W_j = \text{two windows of time series data}$
2: Input: $n = \text{window size}$
3: Input: $CS = \text{change score between } W_i \text{ and } W_j$
4: Input: $F = \text{change score function}$
5: Output: $\text{sig} = \text{(Boolean) significance of } CS$
6: Initialize: $DCS = \text{vector of daily change scores}$
7: Append IntraWindowChange($W_i, n, F$) to $DCS$ (see Algorithm 11)
8: Append IntraWindowChange($W_j, n, F$) to $DCS$ (see Algorithm 11)
9: Compute $\text{sig} = \text{BoxplotOutlierDetection}(CS, DCS)$ (see Algorithm 9)
10: return $\text{sig}$

### 7.2.3 Change Analysis

If a change significance test concludes that the computed change score is significant, the next step is to determine the source of change (see Algorithm 7 for an overview of the PACD process). Often this step requires the computation of features that summarize the data and provide a meaningful context for change. For example, the number of daily steps taken is an example of a simple PA feature. The change between daily steps from one
Algorithm 11 IntraWindowChange($W, n, F$)

1: Input: $W =$ window of time series data  
2: Input: $n =$ window size  
3: Input: $F =$ change score function  
4: Output: $DCS =$ vector of daily change scores  
5: Initialize: $i = 1, j = 1$  
6: while $i \leq n - 1$:  
7:      $j = i + 1$  
8:      while $j \leq n$:  
9:          $CS = F(W[i], W[j])$  
10:         Append $CS$ to $DCS$  
11:      end while  
12:      $i = i + 1$  
13: end while  
14: return $DCS$

window of time to the next can be quantified and used for an explanation of change. Several approaches exist to capture change across time in individual metrics. A straightforward method is to compute the percent change for a feature $f$ from a previous window $W_i$ to a current window $W_j$: $\Delta\% = (f_{W_j} - f_{W_i})/f_{W_i}$. Statistical approaches such as two sample tests or effect size analyses can also be applied to quantify change; however, in applying repeated statistical tests, the multiple testing problem should be accounted for with a method such as the Bonferroni or Benjamini-Hochberg correction [220].

One of the advantages of the VC approach over other change point detection algorithms is it includes an explanation of the source of change without reliance on statistical tests. Upon significant change detection, retraining a decision tree on the entire dataset and inspecting the tree reveals which features are most discriminatory in learning the differences between two windows. Naturally, this approach requires a pre-processing step to extract relevant features from the windowed PA time series data.
Features extracted from the PA data (see Table 7.2) serve two purposes: 1) as features for the VC approach (RulSIF, texture-based dissimilarity, and sw-PCAR do not make use of features for change detection), and 2) for explanation of changes discovered by change detection algorithms (see Section 7.2.1). Features are grouped together based on the number of days required for computation: 1) one day, 2) at least one day, or 3) two or more days. Daily features include PA summaries based on intensity, frequency, duration, and variability of PA bouts. Sequences of time series data with steps greater than a threshold, $S$, are considered a bout of PA. If ground truth activity labels, such as walking, biking, chores, etc., are available from the device user and/or an activity recognition algorithm [219], PA type can be inferred and $S$ can be updated dynamically for different activities. For this study, we assume such labeled information is not available and set $S = t_{mins}$, assuming physical activity is characterized by at least one step per minute. Features requiring at least two days of data summarize activity across or between days or quantify the user’s circadian rhythm (the periodicity from day-to-day [225]). Poincare-plot analysis [66] provides an additional set of useful PA features.

7.3 Physical Activity Datasets

To demonstrate the PACD approach, we utilize two datasets, hybrid-synthetic (HS) and B-fit (BF). The HS dataset comprises synthetic data and the BF dataset comprises real-world Fitbit data collected from a health intervention study. HS and BF data are subject to pre-processing prior to serving as input to PACD. Pre-processing includes down
Table 7.2: Physical activity features used for training virtual classifiers and for explanation of changes discovered by change detection algorithms. Features are categorized by time period (window size).

<table>
<thead>
<tr>
<th>Period</th>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1 Day</td>
<td>Relative amplitude</td>
<td>Normalized ratio between the most active 8 hours and the least active 4 hours</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Texture features</td>
<td>See section 7.2.1</td>
</tr>
<tr>
<td>≥ 1 Day</td>
<td>Inter-daily stability (IS) [65]</td>
<td>Quantifies stability between days</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Intra-daily variability [65]</td>
<td>Quantifies the fragmentation of rhythm and activity. Ratio of variance of consecutive time intervals and overall variance</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Circadian rhythm strength [65]</td>
<td>Ratio of average night-time activity (11pm-5am) by the average activity of the previous day (8am-8pm)</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Cosinor mesor [65, 66]</td>
<td>Time series mean from fitting a cosinor functional model with a 24 hour period to time series data via least squares method</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Cosinor amplitude</td>
<td>Difference between the mesor and peak (or trough)</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Cosinor acrophase</td>
<td>Time of day at which the peak of a rhythm occurs</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Poincare SD1 [66]</td>
<td>Standard deviation of Poincare data against the axis $x = y$</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Poincare SD2 [66]</td>
<td>Standard deviation of Poincare data against the axis orthogonal to $x = y$ and crosses this axis at the center of mass</td>
</tr>
<tr>
<td>≥ 2 Days</td>
<td>Poincare circadian rhythm preserv (CRP) [66]</td>
<td>Day-to-day circadian rhythm preservation based on dispersion values from SD1 and SD2 with delays of 24 hours and 12 hours $CRP = SD2_{24h} + SD1_{12h} \ - SD1_{24h} - SD2_{12h}$</td>
</tr>
</tbody>
</table>

SD = standard deviation
sampling the data for a given time interval length, $t_{mins}$, by summing the steps every $t_{mins}$ minutes. For the case of sw-PCAR, add one smoothing is applied to avoid division by zero during KL divergence computations. Furthermore, missing data are identified and handled for BF data. Days with zero steps taken during the day (9am - 9pm) are considered missing data. First, to fill a missing day, $D_{missing}$, the day in the opposite window, $D_{other}$, with the same day of the week as $D_{missing}$ is identified. Euclidean distance-based clustering is applied to find the $k$ nearest neighbor days, $NN_{other}$, of $D_{other}$ ($k = 3$). The days of the week for each day in $NN_{other}$ are then identified. These are used to select days, $NN_{missing}$ in the original window containing $D_{missing}$. The $k$ days of $NN_{missing}$ are aggregated (see Equation 7.1) and used to fill $D_{missing}$. By this method, three days exhibited missing data, each belonging to a different participant (11/18/15: P2; 9/28/15: P8; 12/7/15: P10, see Table 7.4 for participant information). Zero steps during the day is likely due to removing the Fitbit to charge it, then forgetting to put it back on until much later. Algorithm 12, WindowedFillMissingData, summarizes the data imputation process.

**Algorithm 12** WindowedFillMissingData($D_{missing}$, $W_{missing}$, $W_{other}$, $k$)

1: Input: $D_{missing}$ = day with missing data to fill
2: Input: $W_{missing}$ = window of time series data containing $D_{missing}$
3: Input: $W_{other}$ = window of time series data
4: Input: $k$ = number of nearest neighbors
5: Output: $D_{fill}$ day data to fill $D_{missing}$
6: $D_{fill} = D_{missing}$ // perform a copy
7: $D_{other}$ = day in $W_{other}$ with same day of week as $D_{missing}$
8: $NN_{other} = k$ nearest neighbors of $D_{other}$ in $W_{other}$
9: $NN_{missing} = k$ days in $W_{missing}$ with same days of week as days in $NN_{other}$
10: $D_{fill}[9am:9pm]$ = average of $k$ days in $NN_{missing}[9am:9pm]$
11: return $D_{fill}$
7.3.1 Hybrid-synthetic Dataset

To generate the HS dataset, we re-sampled step data collected from a volunteer wearing a Fitbit Charge Heart Rate fitness tracker and modified the data to produce five different synthetic physical activity profiles, each exhibiting a different type of change. The length of HS profiles was set to 12 days, resulting in two equal size windows of 6 days for comparison (days 1-6 compared to days 7-12). Twelve days was chosen for similarity to the BF dataset. The HS profiles with their profile identification (HS0-4) and a description are as follows (see Table 7.3 for HS profile generation details and parameters):

1. HS0: No significant day-to-day or window-to-window change. Data is subject to small daily variation. HS0 represents a baseline for “no change”.

2. HS1: Medium day-to-day change and consequently significant window-to-window change. Increased bout duration and intensity from day-to-day.

3. HS2: No significant day-to-day change but significant window-to-window change. Increased activity for days 7-12.

4. HS3: Medium day-to-day change and consequently significant window-to-window change. Increased activity variability from day-to-day.

5. HS4: No significant day-to-day change for days 1-6. Significant day-to-day activity variability for days 7-12. Consequently significant window-to-window change.
Table 7.3: Hybrid-synthetic profile generation methods and parameters. The variable $i$ represents the day index, $1 \leq i \leq 12$.

<table>
<thead>
<tr>
<th>Profile</th>
<th>Parameters</th>
</tr>
</thead>
</table>
| HS0     | intensity noise = uniform(-10%, 10%)  
displacement noise = uniform(-5 periods, 5 periods) | |
| HS1     | intensity noise = uniform($i \times 10\%$, $2 \times i \times 10\%)$  
bout extension = uniform($i + 1$ mins, $2 \times i + 1$ mins) | |

| HS2     | 1 $\leq i \leq 6$:   
intensity noise = uniform(-10%, 10%)  
bout extension = uniform(1 min, 3 mins)  
num bouts add = 3  
new bout mins = uniform(0 mins, 5 mins)  
new bout steps = uniform(0 steps, 5 steps)  
7 $\leq i \leq 12$:  
intensity noise = uniform(20%, 22%)  
bout extension = uniform(5 mins, 10 mins)  
num bouts add = 8  
new bout mins = uniform($i - 5$ mins, $i + 5$ mins)  
new bout steps = uniform($i \times 5$ steps, $i \times 5 + 20$ steps) | |

| HS3     | 12 days of different daily data is averaged to produce day 1.  
One day of the original 12 day is removed and the remaining 11 days are averaged again to produce day 2.  
Repeat to produce days 3-12. |

| HS4     | 1 $\leq i \leq 6$:   
intensity noise = uniform(-10%, 10%)  
bout extension = uniform(1 min, 3 mins)  
num bouts add = 3  
new bout mins = uniform(0 mins, 5 mins)  
new bout steps = uniform(0 steps, 5 steps)  
7 $\leq i \leq 12$:  
6 days of highly variable different daily data. |
Figure 7.5: Hybrid-synthetic activity density map for HS0.

Figure 7.6: Hybrid-synthetic activity density map for HS1.
Figure 7.7: Hybrid-synthetic activity density map for HS2.

Figure 7.8: Hybrid-synthetic activity density map for HS3.

### 7.3.2 B-fit Dataset

The BF dataset consists of data collected from 11 older adults (Male = 3, Female = 8; age $57.09 \pm 8.79$ years) participating in a 10-week health intervention (see Table...
Figure 7.9: Hybrid-synthetic activity density map for HS4.

7.4 for participant characteristics). As part of this study, participants’ PA profiles were assessed with wrist-worn Fitbit Flex fitness trackers for one week (six full 24 hour days) before and after the intervention. During weeks two through nine, the participants were educated in eight different subjects related to health and set personal goals for each subject. The subjects included: cognitive engagement, cardiovascular risk factors, stress reduction, exercise, nutrition, social engagement, sleep, and compensatory strategy use. To track goal achievement each week, individuals rated themselves on each personalized goal that they set using a 0 to 3 rating scale (0: did not meet goal, 1: partly met goal, 2: completely met goal, 3: exceeded goal). The participants’ self-ratings for the exercise category during the week of post-intervention Fitbit data collection are included in Table 7.4. After the intervention, participants’ post-intervention physical activity profiles were assessed for one week (six full 24 hour days) with the same Fitbit devices. Finally, while the Fitbit Flex
also provides distance traveled and calories burned, for this study we only consider steps
due to the high inter-metric redundancy between steps and these two Fitbit metrics. For
the BF data, Pearson correlations of $r > 0.99$ for distance and $r > 0.90$ for calories were measured.

Table 7.4: Health intervention study participant characteristics.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>Final week exercise self-rating</th>
<th>Exercise goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Female</td>
<td>65</td>
<td>5’4”</td>
<td>210 lbs</td>
<td>2</td>
<td>Brisk walking</td>
</tr>
<tr>
<td>P2</td>
<td>Female</td>
<td>68</td>
<td>5’3”</td>
<td>207 lbs</td>
<td>3</td>
<td>Walk more</td>
</tr>
<tr>
<td>P3</td>
<td>Female</td>
<td>65</td>
<td>5’2”</td>
<td>165 lbs</td>
<td>1</td>
<td>Take stairs</td>
</tr>
<tr>
<td>P4</td>
<td>Male</td>
<td>64</td>
<td>6’0”</td>
<td>190 lbs</td>
<td>1.5</td>
<td>Walking</td>
</tr>
<tr>
<td>P5</td>
<td>Female</td>
<td>57</td>
<td>5’5”</td>
<td>135 lbs</td>
<td>3</td>
<td>Yoga</td>
</tr>
<tr>
<td>P6</td>
<td>Female</td>
<td>53</td>
<td>5’6”</td>
<td>130 lbs</td>
<td>1</td>
<td>Walking</td>
</tr>
<tr>
<td>P7</td>
<td>Female</td>
<td>65</td>
<td>5’8”</td>
<td>145 lbs</td>
<td>1</td>
<td>Crossfit class</td>
</tr>
<tr>
<td>P8</td>
<td>Male</td>
<td>42</td>
<td>5’8”</td>
<td>145 lbs</td>
<td>3</td>
<td>Walking</td>
</tr>
<tr>
<td>P9</td>
<td>Male</td>
<td>48</td>
<td>6’4”</td>
<td>380 lbs</td>
<td>0</td>
<td>No exercise</td>
</tr>
<tr>
<td>P10</td>
<td>Female</td>
<td>52</td>
<td>5’6”</td>
<td>180 lbs</td>
<td>0.5</td>
<td>Walking</td>
</tr>
<tr>
<td>P11</td>
<td>Female</td>
<td>49</td>
<td>5’6”</td>
<td>262 lbs</td>
<td>1.5</td>
<td>Yoga</td>
</tr>
</tbody>
</table>

ID = identification, lbs = pounds.

7.4 Results

For PACD computations, the following algorithm parameter values are used:

- Window size $n$: 6 days.
- Window offset: 6 days.
- RuLSIF $\alpha$: 0.1.
- RuLSIF cross validation folds: 5.
- Number of sw-PCAR permutations $N$: 1000.
• VC cross validation folds: 4.

• VC prediction threshold $p_{critical}$: 0.75.

• Minimum steps in a bout $S$: $t_{mins}$.

The time interval aggregation size $t_{mins}$ is tested with values of $t_{mins} = \{1, 5, 10, 15, ..., 60 \text{ minutes}\}$. We hypothesize that PACD will find PA changes (bout frequency, intensity, duration, and variability), using each of the change detection methods. However, we anticipate that the significance of the change will vary depending on the algorithm used, the parameter value choices, and the level of change that is inherent in each dataset.

Figures 7.5, 7.6, 7.7, 7.8, and 7.9 show the associated activity density maps for HS profiles HS0-4. An ADM is a heat map proposed by Wang et al. [67] to visualize daily activity (steps for this study) as a function of 24 hour time (Y-axis) and window time (X-axis). Table 7.5 shows RuLSIF, Texture-based, sw-PCAR, and VC significant change results for each HS profile for each time interval length $t_{mins}$. To visualize the HS profile change scores, Figure 7.10 plots RuLSIF, sw-PCAR, and VC normalized change scores with time interval $t_{mins} = 5 \text{ minutes}$. Window one (days 1-6) and window two (days 7-12) values ($\text{mean} \pm \text{standard deviation}$) for the contextual features of number of bouts, bout minutes, daily steps, and sedentary minutes percent are listed in Table 7.6. Results in Table 7.6 have time interval length $t_{mins} = 1 \text{ minute}$ in order to report the most detailed feature values. For further change analysis, decision trees are shown in Figure 7.11 for HS profiles HS1-4.
Table 7.5: Hybrid-synthetic significant change detection as a function of time interval size $t_{mins}$ for each HS profile. Results are in the form count: Boolean (significant change 0: false, 1: true) \{HS0, HS1, HS2, HS3, HS4\}.

<table>
<thead>
<tr>
<th>$t_{mins}$</th>
<th>RulSIF</th>
<th>Texture-based</th>
<th>sw-PCAR</th>
<th>Virtual classifier</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2:0,0,1,0,1</td>
<td>1:0,0,1,0,0</td>
<td>3:0,1,1,0,1</td>
<td>4:0,1,1,1,1</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>3:0,1,1,0,1</td>
<td>2:0,0,1,0,1</td>
<td>3:0,1,1,0,1</td>
<td>4:0,1,1,1,1</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>2:0,0,1,0,1</td>
<td>2:0,0,1,1,0</td>
<td>2:0,1,1,0,0</td>
<td>3:0,0,1,1,1</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>3:0,1,1,0,1</td>
<td>2:0,0,1,1,0</td>
<td>1:0,1,0,0,0</td>
<td>3:0,0,1,1,1</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>3:0,1,1,0,1</td>
<td>3:0,0,1,1,1</td>
<td>1:0,1,0,0,0</td>
<td>3:0,0,1,1,1</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>1:0,0,1,0,0</td>
<td>1:0,0,1,0,0</td>
<td>1:0,1,0,0,0</td>
<td>3:0,0,1,1,1</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>3:0,1,1,0,0</td>
<td>1:0,0,1,0,0</td>
<td>1:0,1,0,0,0</td>
<td>4:0,1,1,1,1</td>
<td>9</td>
</tr>
<tr>
<td>35</td>
<td>2:0,1,1,0,0</td>
<td>1:0,0,0,1,0</td>
<td>1:0,1,0,0,0</td>
<td>3:0,1,1,0,1</td>
<td>7</td>
</tr>
<tr>
<td>40</td>
<td>3:0,1,1,0,0</td>
<td>0:0,0,0,0,0</td>
<td>1:0,1,0,0,0</td>
<td>3:0,0,1,1,1</td>
<td>7</td>
</tr>
<tr>
<td>45</td>
<td>2:0,0,1,0,1</td>
<td>1:0,0,1,0,0</td>
<td>1:0,1,0,0,0</td>
<td>2:0,0,1,0,1</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>2:0,0,1,0,1</td>
<td>0:0,0,0,0,0</td>
<td>1:0,1,0,0,0</td>
<td>4:0,1,1,1,1</td>
<td>7</td>
</tr>
<tr>
<td>55</td>
<td>3:0,1,1,0,0</td>
<td>1:0,1,0,0,0</td>
<td>0:0,0,0,0,0</td>
<td>3:0,1,1,0,1</td>
<td>7</td>
</tr>
<tr>
<td>60</td>
<td>4:0,1,1,0,1</td>
<td>0:0,0,0,0,0</td>
<td>1:0,1,0,0,0</td>
<td>4:0,1,1,1,1</td>
<td>9</td>
</tr>
</tbody>
</table>

Total 33:0,8,13,3,9 15:0,1,7,5,2 17:0,12,3,0,2 43:0,7,13,10,13 108

For the BF dataset, each participant’s change significance testing results are presented in Table 7.7. Four contextual features (number of bouts, bout minutes, daily steps, and sedentary percent) pre and post-intervention values are listed in Table 7.8. Finally, decision trees are shown in Figure 7.12 for select health intervention study participants with a significant VC change score (P2, P7, and P10). Table 7.4 shows self-ratings for the health intervention study participants for the exercise category during the week of final Fitbit data collection.
Figure 7.10: Hybrid-synthetic normalized change scores for $t_{mins} = 5$ minutes.

Table 7.6: Hybrid-synthetic feature results (mean ± standard deviation) with $t_{mins} = 1$ minute. First and second window values are separated by a comma.

<table>
<thead>
<tr>
<th>ID</th>
<th>Number of bouts</th>
<th>Bout minutes</th>
<th>Daily steps</th>
<th>Sedentary %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS0</td>
<td>70.33, 70.00</td>
<td>5.10 ± 9.91, 5.13 ± 9.92</td>
<td>20601.65, 21274.32</td>
<td>75.65, 75.56</td>
</tr>
<tr>
<td>HS1</td>
<td>34.50, 14.17</td>
<td>19.39 ± 23.93, 46.82 ± 56.78</td>
<td>36409.49, 72769.11</td>
<td>64.57, 54.59</td>
</tr>
<tr>
<td>HS2</td>
<td>71.50, 62.50</td>
<td>5.07 ± 9.83, 7.63 ± 11.13</td>
<td>20755.53, 30037.48</td>
<td>75.62, 67.44</td>
</tr>
<tr>
<td>HS3</td>
<td>54.83, 102.83</td>
<td>18.71 ± 51.74, 6.49 ± 14.49</td>
<td>14395.85, 14746.43</td>
<td>45.22, 63.72</td>
</tr>
<tr>
<td>HS4</td>
<td>53.50, 81.33</td>
<td>8.14 ± 12.61, 4.56 ± 8.48</td>
<td>22048.02, 17327.00</td>
<td>70.66, 77.86</td>
</tr>
</tbody>
</table>
Texture density ≤ 94.42
True
Value = [6,0]
Class = W1
False
Value = [0,6]
Class = W7

Rest minutes mean ≤ 39.92
True
Value = [0,6]
Class = W7
False
Value = [6,0]
Class = W1

(a) HS1
(b) HS2

Number of bouts ≤ 18.0
True
Value = [6,0]
Class = W1
False
Value = [0,6]
Class = W7

Relative amplitude ≤ 0.97
True
Value = [6,0]
Class = W1
False
Value = [0,6]
Class = W7

(c) HS3
(d) HS4

Figure 7.11: Decision trees for hybrid-synthetic profiles with significant virtual classifier change scores for $t_{mins} = 5$ minutes.
Table 7.7: Health intervention study significant change detection as a function of time interval size $t_{mins}$. Results are in the sparse form count: \{IDs\}: Boolean (significant change: 0 false, 1 true).

<table>
<thead>
<tr>
<th>$t_{mins}$</th>
<th>RuleSIF</th>
<th>Texture-based</th>
<th>sw-PCAR</th>
<th>Virtual classifier</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1:P2:1</td>
<td>0</td>
<td>10:P10:0</td>
<td>5:P2,6,7,10,11:1</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>2:P2,7:1</td>
<td>0</td>
<td>5:P1,2,3,4,8:1</td>
<td>6:P2,6,7,8,10,11:1</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>1:P2:1</td>
<td>0</td>
<td>4:P1,2,4,8:1</td>
<td>6:P2,3,6,7,8,10:1</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>1:P2:1</td>
<td>0</td>
<td>4:P1,2,4,8:1</td>
<td>4:P2,7,10,11:1</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>2:P2,7:1</td>
<td>0</td>
<td>3:P2,4,8:1</td>
<td>4:P2,7,10,11:1</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>1:P2:1</td>
<td>0</td>
<td>3:P2,4,8:1</td>
<td>2:P7,10:1</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>2:P2,8:1</td>
<td>0</td>
<td>3:P2,4,8:1</td>
<td>2:P6,10:1</td>
<td>7</td>
</tr>
<tr>
<td>35</td>
<td>3:P2,4,8:1</td>
<td>0</td>
<td>3:P2,4,8:1</td>
<td>5:P2,6,7,10,11:1</td>
<td>11</td>
</tr>
<tr>
<td>40</td>
<td>1:P2:1</td>
<td>0</td>
<td>3:P2,4,8:1</td>
<td>3:P2,7,10:1</td>
<td>7</td>
</tr>
<tr>
<td>45</td>
<td>4:P2,3,4,6</td>
<td>0</td>
<td>1:P2:1</td>
<td>2:P2,10:1</td>
<td>7</td>
</tr>
<tr>
<td>50</td>
<td>1:P2:1</td>
<td>0</td>
<td>1:P2:1</td>
<td>5:P2,6,7,9,10:1</td>
<td>7</td>
</tr>
<tr>
<td>55</td>
<td>1:P10:1</td>
<td>0</td>
<td>1:P2:1</td>
<td>3:P2,4,10:1</td>
<td>5</td>
</tr>
<tr>
<td>60</td>
<td>2:P4,9:1</td>
<td>0</td>
<td>1:P2:1</td>
<td>4:P2,6,10,11:1</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>0</td>
<td>42</td>
<td>51</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 7.8: Health intervention study feature results (mean ± standard deviation) with $t_{mins} = 1$ minute. Pre and post-intervention values are separated by a comma.

<table>
<thead>
<tr>
<th>ID</th>
<th>Number of bouts</th>
<th>Bout minutes</th>
<th>Daily steps</th>
<th>Sedentary %</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>73.50, 89.67</td>
<td>2.35 ± 1.93, 2.63 ± 2.61</td>
<td>3479.00, 4658.33</td>
<td>88.37%, 84.43%</td>
</tr>
<tr>
<td>P2</td>
<td>81.00, 15.83</td>
<td>2.57 ± 2.54, 2.75 ± 1.85</td>
<td>4275.90, 1161.44</td>
<td>86.30%, 97.44%</td>
</tr>
<tr>
<td>P3</td>
<td>88.50, 27.50</td>
<td>2.72 ± 2.86, 2.36 ± 2.08</td>
<td>5886.67, 4558.50</td>
<td>84.06%, 86.32%</td>
</tr>
<tr>
<td>P4</td>
<td>81.17, 60.33</td>
<td>3.90 ± 5.27, 3.71 ± 4.85</td>
<td>11177.00, 7399.67</td>
<td>79.11%, 85.71%</td>
</tr>
<tr>
<td>P5</td>
<td>73.67, 76.50</td>
<td>2.96 ± 4.06, 2.60 ± 3.22</td>
<td>6994.17, 5470.50</td>
<td>85.71%, 86.22%</td>
</tr>
<tr>
<td>P6</td>
<td>105.33, 88.00</td>
<td>2.63 ± 2.46, 2.66 ± 2.39</td>
<td>7127.00, 6207.67</td>
<td>81.82%, 84.85%</td>
</tr>
<tr>
<td>P7</td>
<td>64.33, 63.50</td>
<td>3.30 ± 4.20, 3.45 ± 3.61</td>
<td>7354.67, 6181.17</td>
<td>85.78%, 85.90%</td>
</tr>
<tr>
<td>P8</td>
<td>104.17, 102.50</td>
<td>5.52 ± 7.51, 3.35 ± 3.79</td>
<td>17680.78, 11440.00</td>
<td>66.66%, 77.18%</td>
</tr>
<tr>
<td>P9</td>
<td>99.50, 116.67</td>
<td>2.40 ± 2.49, 2.40 ± 2.57</td>
<td>5844.50, 6731.83</td>
<td>84.11%, 81.46%</td>
</tr>
<tr>
<td>P10</td>
<td>85.00, 80.50</td>
<td>2.99 ± 3.24, 3.58 ± 4.29</td>
<td>1136.51, 1210.85</td>
<td>82.94%, 81.62%</td>
</tr>
<tr>
<td>P11</td>
<td>83.00, 89.50</td>
<td>2.51 ± 2.73, 2.31 ± 2.23</td>
<td>5753.50, 4868.83</td>
<td>86.16%, 86.44%</td>
</tr>
</tbody>
</table>
Figure 7.12: Decision trees for the health intervention study participants with significant virtual classifier change scores for $t_{\text{mins}} = 5$ minutes.
7.5 Discussion

The goal of this study is to investigate the PACD approach for unsupervised change detection and analysis of PA time series. The abilities of four presented methods to detect change are evaluated on two original datasets: 1) the HS dataset, comprised of 5 synthetic profiles and 2) the BF dataset, comprised of 11 participants’ Fitbit data from an intervention study.

7.5.1 Hybrid-synthetic Dataset

The HS dataset reveals several insights into the change detection algorithms. First, the time interval length yielding the highest number of significant changes is $t_{\text{mins}} = 5$ minutes with 12 changes, closely followed by $t_{\text{mins}} = 1$ and $t_{\text{mins}} = 20$ minutes with 10 changes (see Table 7.5). Since HS profiles are sampled from a volunteer’s real Fitbit data, these intervals suggest movement patterns occur in 1, 5, and 20 minute chunks for this individual. For all time interval lengths, the algorithms correctly do not detect a significant change between first and second window data for the HS0 profile. HS0 is generated to exhibit small day-to-day variation in step intensity and is not characterized by large changes between windows.

For HS1-4 profiles, significant changes between windows are detected. For all time interval lengths, the VC approach picks up the most changes (43 changes), followed by RuLSIF (33), sw-PCAR (17), and texture-based (15). As a group, the algorithms’ are able
to sense changes in value (HS1, HS2) and changes in variability (HS3, HS4), with 64 and 44 changes respectively. Changes for HS2 (36) are the most frequently detected, followed by HS1 (28), HS4 (26), and HS3 (18). The lower number of changes detected for HS3 is possibly due to high intra-window daily change scores for days 7-12 (see Figure 7.8) used for Boxplot significance testing (see Algorithm 9). Window-based change (HS2, HS4) is perfectly detected for all time intervals by VC (HS2, HS4: 13) and RuLSIF (HS2: 13).

Investigating the $t_{\text{mins}} = 5$ minutes results reveal all four algorithms determine significant changes for HS2 and HS4 (see Table 7.5). For HS1, near perfect detections are made by sw-PCAR (12).

Upon inspection of the associated decision trees for HS1-4 (see Figure 7.11), the features of texture density, average daily rest minutes, number of bouts, and relative amplitude are revealed as discriminatory features. The explanatory power of the features is potentially useful for reporting to the wearable sensor user the dimensions of change in their physical activity. Features useful for such purposes are simple, common features that do not require explanation to the user. For example, texture density or relative amplitude are useful features for detecting changes in PA patterns, but are relatively unimportant to a user. More meaningful features to a user include number of bouts, minutes per bout, daily steps taken, and sedentary percent. Table 7.6 shows these features for the HS profiles. HS0 exhibits quite similar window one and window two values for all features. HS2 and HS4 both have small standard deviations due to window-based change in lieu of day-to-day change (HS1 and HS3).
7.5.2 B-fit Dataset

Analyzing the health intervention study participants’ data poses additional challenges that are not present with the HS profiles. Real-world human subject data is inherently noisy, characterized by seemingly random bouts of PA and rest periods. Furthermore, self-report and direct measurement of physical activity are often not congruent, with previous studies reporting correlations in as wide a range of -0.71 to 0.96 [226]. For the BF group, the participants demonstrated a wide spread of self-reported goal achievement ratings for the exercise category, 1.59 ± 1.05. For example, P6, P9, and P10 rated their exercise goal achievements as low (exercise rating: 1, 0, 0.5 respectively). Due to heart problems, P9’s doctor instructed him not participate in exercise-related activities. On the other hand, P2 rated her exercise goal achievement the highest (exercise rating: 3). Upon inspection of P2’s data, it is evident there is a discrepancy between the participant’s perception of her PA and the steps recorded by the Fitbit. It is not uncommon for self-reported measures of physical activity to be inconsistent with direct measures [226]; therefore, the participants’ self-ratings are used for insights into individual goal achievements, not as ground truth information for changes exhibited. The issues with self-reported PA measures exacerbate the need for unsupervised change detection and analysis methods.

Depending on the algorithm, significant changes are commonly detected for 5 out of the 11 health intervention study participants: P2: 35; P10: 14; P4: 13; P8: 13; P7: 12 (see Table 7.5). Virtual classifier and sw-PCAR detect the highest number of changes (51 and 42 changes each). The distribution of detected changes by sw-PCAR is highly influenced
by time interval length (sw-PCAR: $3.23 \pm 2.42$ number of changes detected compared to VC: $3.92 \pm 1.44$). sw-PCAR is not sensitive for small time intervals ($t_{\text{mins}} = 1$ minutes) or large time intervals ($t_{\text{mins}} = \{45, 50, 55, 60 \text{ minutes}\}$), and the number of changes detected decreases as time interval length increases. Virtual classifier does not appear to be as heavily influenced by the time interval length. The texture-based approach is the least sensitive algorithm and did not detect any changes in the BF data.

Performing change analysis and investigating the detected changes yields insights for several of the participants. P2 rated herself as completely meeting her exercise goal of walking more; however, the Fitbit data tells a different story. Several features in Table 7.6 show decreased PA for P2: average number of bouts (pre: 81.00, post: 15.83), daily steps (pre: 4279.50, post: 1161.44 steps), and percentage of time sedentary (pre: 86.30%, post: 97.44%). Additionally, P2’s decision tree (see Figure 7.12a) provides evidence that she rested more during post-intervention testing. In summary, the features suggest the changes detected by the algorithms are actually changes in the opposite direction of her goal. Contrary to P2, P10 exhibited a significant change (as detected consistently by VC) in the direction toward her goal of walking more. Inspection of P10’s features shows an increase in bout minutes and average steps per day. Average daily steps increased from 1136.51 steps pre-intervention to 1210.85 steps post-intervention testing, a 6.54% increase. The remaining participants with significant changes (P4, P7, and P8) demonstrated a decrease in average daily steps from pre to post intervention. It should be noted that during the week of post-test data collection the weather conditions were adverse and this may have partially contributed to the decrease observed in average daily steps. Research
has shown PA levels can be influenced by adverse weather conditions [227]. It is also worth noting the participants exhibited improvements in other PA features. For example, relative amplitude has been reported to decrease with worsening health [228], thus P7 and P10’s increased relative amplitude post-intervention is healthy (see Figures 7.12b and 7.12c). Also, P9 was not planning on increasing exercise; however, P9 increased his daily steps post-intervention by 15.18%.

One of the limitations of this study includes having only one week of pre-intervention Fitbit data for the health intervention study participants. With at least two weeks of pre-intervention data, change scores can be computed between week one and two of pre-intervention data to provide an estimate of inter-week variability. With a quantification of inter-week variability, we can determine if the measured change between pre and post-intervention weeks is due to the intervention or natural variability. An additional limitation includes not having a full 7 days of BF data during pre and post-intervention weeks. Finally, more sophisticated methods to fill missing data could be utilized with fitness trackers that include heart rate monitors, due to more reliable detection of sensor donned/doffed. Consequently, future work includes performing change analysis on real-world datasets from different fitness trackers, multidimensional data (e.g. heart rate, elevation, etc.), labeled activity data, and longer windows of time. With time series data longer than two years, several additional analyses could be performed including: daily/weekly/monthly/yearly period analysis and slicing along different dimensions (e.g. Mondays, weekends, holidays, or activities if labeled information is available).

In summary, we address the problem of unsupervised physical activity change detec-
tion and analysis with our proposed Physical Activity Change Detection approach. PACD is a framework we designed to detect and analyze changes in physical activity data. We compare the abilities of three change detection algorithms from the literature and one proposed algorithm, sw-PCAR, to capture different types of changes in synthetic and real-world datasets. Results indicate the approaches detect several changes in the datasets; particularly for physical activity profiles exhibiting large changes between windows instead of incremental day-to-day changes. Contextual features such as average number of daily steps, minutes per bout, and sedentary percent provide an explanation of the changes that are detected. The algorithms and analysis methods are useful data mining techniques for unsupervised, window-based change detection. Future work involves implementing our PACD method in an online, smartphone application to track users’ physical activity and motivate progress toward their health goals. In the following chapter, we demonstrate how PACD can also be utilized to track changes in behavior patterns. We do this by analyzing smart home data collected from ambient sensors.
CHAPTER 8. BEHAVIOR CHANGE DETECTION

In the previous chapter we investigated changes in mobility as measured by physical activity data collected from wearable sensors. In this chapter, we hypothesize that our change detection approach can also measure changes in mobility and physical activity based on data collected from ambient sensors in smart homes. We utilize smart home sensor data and machine learning techniques to compute results in support of this hypothesis.

Smart home sensor systems provide the capability to automatically and unobtrusively collect information about a resident’s everyday behavior. We collect data from ambient sensors placed in smart home environments and label the data with the corresponding activities using automated activity recognition. To track changes in routine behavior, we quantitatively compare two or more time periods, or windows, of activity-labeled data. If the two time windows contain significantly different activity information then this may indicate a significant behavior change, possibly due to a health event [229]. The ability to be physically active is an indicator of good health but the opposite is also true, inactivity or a change in activity can be an indicator of a change in health due to chronic or acute illness [230].
8.1 Prior Work on Activity Analysis for Health Assessment

Similar research work in this area consists of activity recognition approaches and research in behavior-based change detection. Activity recognition algorithms have been designed for wearable, phone, home, video, and other sensors using machine learning techniques that range from naïve Bayes classifiers and decision trees to more complex models including Gaussian mixture models and conditional random fields [219,231]. Most of the prior work in detecting behavior changes utilized wearable data to correlate home-based movement with health measures [232,233], although smart home data have been used to analyze mobility and time out of the home with respect to cognitive and physical health [67,234]. Earlier work has also shown that smart home data can be analyzed over time to predict performance on cognitive health assessment tests [69].

We hypothesize that the relationship between sensed behavior and health events can be observed and analyzed using smart home data. To evaluate our hypothesis, we utilize PACD to analyze smart home data collected for multiple years in the homes of older adults. Health events are identified for three of the smart home residents based on medical records review and monthly interviews with the study participants. Data surrounding the health event is compared with baseline normal data to determine if a significant behavior change has occurred and describe the nature of the change. The corresponding behavior change is then analyzed by a clinician to validate the behavior change and explain the relationship between the health event and corresponding behavior change. Results from our case studies indicate that smart home and machine learning technologies can be used to understand the
behavioral impacts of health events and to provide information indicating possible health concerns.

8.2 Smart Home Environments

We collect data in everyday home environments using the CASAS “smart home in a box” [235]. The three homes that we include in this study are single-resident apartments. These homes are equipped with combination motion/light sensors on the ceilings and door/temperature sensors on cabinets and doors. The apartment floorplans and sensor positions are shown in Figure 8.1. The sensors continuously and unobtrusively monitor daily activities of the residents by sending text message-type updates, or sensor events, whenever they sense a state change (i.e. from “door closed” to “door open” or from “no motion” to “motion”). Table 8.1 shows example sensor data. The CASAS middleware collects these sensor events and stores them in a relational database.

Once the sensor data is collected, each sensor event is labeled with the corresponding activity using the CASAS-AR activity recognition algorithm [236]. Let $A = \{a_1, a_2, ..., a_T\}$ be the set of all activities. Given features $x_t \in \mathbb{R}^d$ extracted from a sequence of sensor events ending at time $t$, the challenge of activity recognition is to map $x_t$ onto a value $a \in A$ indicating the activity that occurred at time $t$. These labels provide a vocabulary for expressing and analyzing the sensed behavioral patterns. In this study we analyze the activities of Hygiene, Sleep, Bed-Toilet, Eat/Drink, Enter/Leave Home, Relax, and Work. CASAS-AR labels sensor events with activity labels in real time as the events occur. AR is
Table 8.1: Sample raw sensor data is converted to use generalized sensor identifiers and automatically labeled by CASAS-AR activity recognition with corresponding activity labels.

<table>
<thead>
<tr>
<th>Timestamp/Identifier/Message</th>
<th>Sensor Location</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-06-15 03:38:28.094897 M009 ON</td>
<td>BedroomMotion</td>
<td>Sleep</td>
</tr>
<tr>
<td>2014-06-15 03:38:29.213955 M009 OFF</td>
<td>BedroomMotion</td>
<td>Sleep</td>
</tr>
<tr>
<td>2014-06-15 03:38:17.814393 M015 ON</td>
<td>BathroomMotion</td>
<td>Bed-Toilet</td>
</tr>
<tr>
<td>2014-06-15 03:38:58.584179 M015 OFF</td>
<td>BathroomMotion</td>
<td>Bed-Toilet</td>
</tr>
<tr>
<td>2014-06-15 03:39:17.814393 M009 ON</td>
<td>BedroomMotion</td>
<td>Sleep</td>
</tr>
</tbody>
</table>

particularly well suited for this type of analysis because it does not require that the sensor data be pre-segmented into distinct activity sequences. Instead, it labels sensor events with activity labels in real time as the events occur. To do this, a dynamic-size sliding window is moved over the sensor events and features $x_t$ describing the current window of information are extracted. The computed activity recognition features include the sensor event time of day, the size of the sliding window, the event count for each sensor within the window, time elapsed for each sensor since its most recent event, the most recent event location and sensor identifier, and the sensor generating the most events in the previous two windows.

Training data for CASAS-AR are provided by external annotators who look at one month of data and utilize both the house floorplan and resident information to generate corresponding ground truth activity labels. Using this method, CASAS-AR learns an activity model based on training data from multiple smart home sites and can thus generalize for application to new smart homes with no training data. Although CASAS-AR has been tested with a number of classifiers including naïve Bayes, decision trees, hidden Markov models, and conditional random fields, the best performance was achieved using a decision
Figure 8.1: Smart home floorplan and sensor layout for three testbeds.
For these activities in the three smart home testbeds that we analyze, CASAS-AR achieved a recognition accuracy of 98% using 3-fold cross validation.

We are interested in analyzing the behavioral impact of health events. More specifically, we want to determine if a significant change in behavior has occurred at the time of the health event and to analyze the nature of the behavior change. To do this, we introduce methods to quantify the amount of change in activity patterns between two windows of time series activity data that were sampled by smart home sensors and labeled by CASAS-AR. Let $X = \{d_1, d_2, \ldots\}$ denote a sample of time series data where each day’s data are expressed by a vector of extracted activity features $d$. Let $W$ be a window of $n$ days such that $W \subseteq X$. Daily activity features are computed for each day in $W$, including [237]:

- Amount of time spent on each activity.
- Sensor density of each activity (measured as number of sensor events).
- Total amount of movement that occurs in the home, expressed as the total distance traveled.

PACD compares two windows of data, $W_i$ and $W_j$, within time series $X$. We set the window size to one week in length ($n = 7$). PACD compares a baseline window ($i = 1$, the first week in our dataset representing normal behavior for the resident) with each subsequent window ($j = 2, 3, \ldots$). We utilize three change detection methods, RuLSIF, sw-PCAR, and virtual classifier (see Section 7.2.1 for descriptions of these algorithms), to offer three slightly different perspectives on the data comparison.
8.3 Smart Home Resident Health Events

We collected data in smart homes with older adult residents for multiple years. For each smart home resident we also recorded health events with their date and event type, based on medical records and monthly interviews with the participants. Here we describe three of these health events and utilize these case studies to illustrate the use of PACD:

- SH1 (86 year old female, see Figure 8.1a for the smart home floorplans): Three months into the data collection, the participant was diagnosed with lung cancer and started radiation treatment during week $W_{10}$. We hypothesize that radiation treatment will have an observable and quantifiable impact on her behavior.

- SH2 (91 year old female, see Figure 8.1b for the smart home floorplans): During the time that data was collected in this participant’s home, she was diagnosed with insomnia during $W_{11}$. We hypothesize this health event will have a noticeable impact on her sleep and on other routine activities.

- SH3 (80 year old female, see Figure 8.1c for the smart home floorplans): During the time that we were collecting sensor data in this home, the participant fell in her home (the fall occurred during week $W_8$). We hypothesize the weeks following the fall will demonstrated affected movement throughout the home.

We hypothesize that each of these life-changing health events will have an observable and quantifiable impact on the corresponding participant’s behavior. To validate this hypothesis, we use PACD to compare one-week baseline of smart home activity data ($W_1$)
with two other weeks. The first comparison is with another pre-event week, namely the week immediately following the baseline ($W_2$). The second comparison is with the first full week during which the individual experienced the health event ($W_{\text{event}}$).

### 8.4 Results

For PACD computations, the following algorithm parameter values are used:

- Window size $n$: 7 days.
- Window offset: 7 days.
- RuLSIF $\alpha$: 0.1.
- RuLSIF cross validation folds: 5.
- Number of sw-PCAR permutations $N$: 1000.
- VC cross validation folds: 4. VC prediction threshold $p_{\text{critical}}$: 0.71.
- Time interval aggregation size $t_{\text{mins}}$: 60 minutes.

The change score results using three PACD techniques described previously (RuLSIF, sw-PCAR, and virtual classifier) are summarized in Table 8.2. Figures 8.2, 8.3, and 8.4 illustrate weekly results from applying the three PACD change detection algorithms for SH1, SH2, and SH3, respectively. Figure 8.5 shows associated top-level rules generated from the virtual classifier decision trees for each of the smart home residents. Finally, Figures 8.6, 8.7, and 8.8 show associated activity density maps for select activities for
Table 8.2: Change scores for smart home residents SH1, SH2, and SH3. Scores are computed between two normal activity weeks (W\textsubscript{1} and W\textsubscript{2}) and between a normal activity week and a week during the health event (W\textsubscript{1} and W\textsubscript{11} for SH1 and SH2, W\textsubscript{1} and W\textsubscript{8} for SH3). For RuLSIF and sw-PCAR, larger values indicate greater change and values close to 0 indicate no change. In the case of VC, values close to 0.5 indicate no change and values close to 1.0 indicate large change. Significant results are indicated with an asterisk (*).

<table>
<thead>
<tr>
<th>Method</th>
<th>W\textsubscript{1}/W\textsubscript{2} (baseline)</th>
<th>W\textsubscript{1}/W\textsubscript{event} (health event)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RuLSIF</td>
<td>-0.017</td>
<td>2.298*</td>
</tr>
<tr>
<td>SH1</td>
<td>sw-PCAR 0.001</td>
<td>0.091*</td>
</tr>
<tr>
<td>VC</td>
<td>0.500</td>
<td>1.000*</td>
</tr>
<tr>
<td>RuLSIF</td>
<td>0.010</td>
<td>3.315*</td>
</tr>
<tr>
<td>SH2</td>
<td>sw-PCAR 0.004</td>
<td>0.042*</td>
</tr>
<tr>
<td>VC</td>
<td>0.438</td>
<td>1.000*</td>
</tr>
<tr>
<td>RuLSIF</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SH3</td>
<td>sw-PCAR 0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>VC</td>
<td>0.500</td>
<td>0.750*</td>
</tr>
</tbody>
</table>

SH1, SH2, and SH3 respectively. The darker the color, the more time was spent on the activity during that particular hour of the day in the corresponding week.

8.5 Discussion

We collected data in smart homes with older adult residents for multiple years. For each study participant we also recorded health events with their date and event type, based on medical records and monthly interviews with the participants. Here we describe three of these health events and utilize these case studies to illustrate the use of PACD.
Figure 8.2: Participant SH1 change scores for baseline week $W_1$ and health event week $W_{11}$. Values above the red line show significant changes. In each plot the green dashed line indicates the occurrence of the health event.
Figure 8.3: Participant SH2 change scores for baseline week $W_1$ and health event week $W_{11}$. Values above the red line show significant changes. In each plot the green dashed line indicates the occurrence of the health event.
Figure 8.4: Participant SH3 change scores for baseline week $W_1$ and health event week $W_7$. Values above the red line show significant changes. In each plot the green dashed line indicates the occurrence of the health event.
Figure 8.5: The top-level rules generated using a virtual decision tree classifier for smart home residents SH1-3. The top-level rules indicate the activity features that best discriminate the baseline week from the health event week.
Figure 8.6: Activity density maps for smart home participant SH1. Darker colors in the density maps indicate more time spent on the activity during that hour of the day. In each plot the green dashed line indicates the occurrence of the health event.
Figure 8.7: Activity density maps for smart home participant SH2. Darker colors in the density maps indicate more time spent on the activity during that hour of the day. In each plot the green dashed line indicates the occurrence of the health event.
Figure 8.8: Activity density maps for smart home participant SH3. Darker colors in the density maps indicate more time spent on the activity during that hour of the day. In each plot the green dashed line indicates the occurrence of the health event.
8.5.1 SH1

For participant SH1, the behavior changes during radiation treatment are evident for each of the change detection methods and the results are significant (see Table 8.2 and Figure 8.2). As the density maps show in Figure 8.6, the participant’s level of sleep decreased once treatment started and the number of times she left the home/returned home increased. Furthermore, the top-level decision tree feature for SH1 is the number of sensor events that are related to an “Enter Home” activity. Possible explanations for this are increased trips out of the home for treatment appointments or visits from family and caregivers. Another impact of the treatment is the increased number of trips this participant made to the kitchen to eat or drink. These more frequent kitchen trips are consistent with the observation that radiation treatment increases the feeling of thirst, resulting in a patient drinking more liquids throughout the day [238].

8.5.2 SH2

SH2 is a 91 year old female smart home resident diagnosed with insomnia. The change scores for SH2 reported in Table 8.2 indicate significant changes in overall routine are detected by both sw-PCAR and VC methods. In Figure 8.3 we see that changes occur not only during week $W_{11}$ but also in the days leading up to the health event and persisting to days and weeks following the insomnia diagnosis. We can also observe in the density maps (see Figure 8.7) that the amount of sleep does decrease during this period.
The change in behavior also impacts relaxation, which is time spent in a favorite chair or couch with little movement and possibly napping. These relaxation periods occur during normal sleep hours but also throughout the day. In addition, the number of trips outside the home decreases during this time. The virtual classifier actually finds the corresponding decrease in “Enter Home” events to be the main discriminating feature between baseline and health event weeks.

8.5.3  SH3

The last case study, SH3, experienced a fall during the time that we were collecting sensor data in her home. She described that her right leg hurt for several days after the fall and consequently it “slowed her down.” As the results in Table II indicate, this health event has a much more subtle impact on behaviors, at least those that can be detected by ambient smart home sensors. RuLSIF and sw-PCAR detects almost no change between weeks $W_1$ and $W_2$ or between weeks $W_1$ and $W_8$. The virtual classifier does detect the change during the health event week. The VC-generated rule indicates the difference is primarily detected based on the total distance that the individual traveled throughout the home on a daily basis (see Figure 8.5c). The decrease in movement is consistent with the observation that the hurt leg caused the resident to slow down. Furthermore, as the density maps in Figure 8.8 indicate, there appears to be less impact on other routine activities such as sleep and bed-to-toilet transitions. There is a slight decrease in trips out of the home but this is not large enough to be detected by the change detection methods.
In this chapter we apply PACD (see Chapter 7 for details regarding PACD) for behavior change detection. We describe how PACD can be used to quantify and explain changes that are detected in daily activity data. In particular, PACD can detect changes in smart home-collected behavior data that occur as a result of health events. From the three case studies we see that the ability to detect behavioral impact of health events depends on the nature of the health event itself. Some events impact multiple activities including sleep, eating, and trips out of the home. In contrast, other events have more localized impact. The ability to detect the actual health event occurrence (e.g., fall) and its impact may require additional, more sensitive sensors to be placed in the home or on the body.

The ability to detect behavior changes that are associated with health events is valuable for researchers who want to better understand the relationship between health and behavior. These insights may also help care providers respond to the needs of individuals who are experiencing changes in their health. An algorithm such as PACD can periodically look for changes in behavioral routine and alert the individual and their caregiver about these changes as they may indicate changes in cognitive or physical health. Because PACD can analyze any type of sensor data, our continued research will adapt these methods to analyze smartphone and wearable data, as well as data collected in smart homes.
CHAPTER 9. CONCLUSIONS

9.1 Summary and Conclusions

In this dissertation, we hypothesized that sensor data can be collected and analyzed using machine learning techniques to provide insights on mobility changes related to rehabilitation and everyday behavior. We validated this hypothesis by introducing sensor-based computing algorithms for mobility assessment and analyzing data from three diverse settings: wearable sensor data collected during inpatient rehabilitation, wearable sensor data collected during everyday living, and ambient sensor data collected from everyday living in smart home environments.

For facilitating sensor-based mobility assessment in rehabilitation settings, we presented the ambulation circuit, a sequence of gait and transfer tasks that patients perform while wearing inertial sensors. We designed algorithms to extract clinically-relevant information from the collected inertial sensor data, including objective metrics quantifying mobility and performance change. With our Hybrid Clinical Sensor Prediction approach, we trained models to predict Functional Independence Measure scores by combining clinical information, sensor-based metrics, and changes in metrics to achieve an even higher level of prediction accuracy. To bridge the gap between design of mobility monitoring technology and actual use of the technology, we interviewed physical therapy providers at
an inpatient rehabilitation facility to collect their opinions on the clinical utility of our wearable sensor-based metrics, visualizations, and clinical outcome predictions. The responses indicated therapy providers to be interested in using wearable technology and our associated algorithms.

For everyday mobility assessment, we proposed the Physical Activity Change Detection approach for unsupervised change detection and analysis of physical activity data. We compared the abilities of three change detection algorithms from the literature and one proposed algorithm to capture different types of changes in synthetic and real-world datasets. Results indicated the approaches detect several changes in both datasets; particularly for physical activity profiles exhibiting large changes between windows instead of incremental day-to-day changes. Since self-report and direct measurement of physical activity are often not in agreement [226], the change detection methods presented in this dissertation are potentially useful for monitoring and motivating progress towards fitness goals and healthy lifestyle changes.

Finally, we demonstrated our Physical Activity Change Detection approach can detect changes in smart home-collected behavior data that occur as a result of health events. The ability to detect behavior changes that are associated with health events is valuable for researchers who want to better understand the relationship between health and behavior. These insights may also help care providers respond to the needs of individuals who are experiencing changes in their health. An algorithm such as Physical Activity Change Detection can periodically look for changes in behavioral routine and alert the individual and their caregiver about these changes as they may indicate changes in cognitive or
physical health.

9.2 Future Work

We suggest future computing research directions in the area of mobility assessment and sensor-based change analysis. Interviews with physical therapy providers offered several insights that may increase the adoption of wearable sensor systems and visualizations. We plan to incorporate the suggestions received from the interviewees into a smart wearable system to provide useful quantitative data, visualizations, and clinical outcome predictions to aid therapists in providing therapy services for their patients.

Additional future work aims to merge sensor-based mobility assessment techniques from everyday living with rehabilitation settings. We plan to utilize our Physical Activity Change Detection algorithms to track physical activity profiles of patients receiving inpatient therapy services for the entire duration of their inpatient rehabilitation stay. Continuous mobility information available from patient-worn wearable sensors can be used in conjunction with patients' daily therapy schedules to track patients' physical activity for the duration of inpatient rehabilitation. Such dense mobility data can also be used to perform daily clinical outcome predictions and will possibly improve our prediction accuracy results. Potentially, our algorithms can track and predict physical recovery to help clinicians provide therapy services and help motivate patients during the recovery process.
The following Mini-Cog examination is adapted from Borson et al. [81, 82]. A potential study participant is provided with a pencil and a blank sheet of paper. If a table to write on is not conveniently located, a clipboard is provided with the paper.

1. Get the patient’s attention and then say: “I am going to say three words that I want you to remember. The words are banana, sunrise, chair. Please say them for me now.” (Give the patient 3 tries to repeat the words. If unable after 3 tries, go to next item.)

2. Say the following phrases in the order indicated: “Please draw a clock in the space below. Start by drawing a large circle.” (When this is done, say) “Put all the numbers in the circle.” (When done, say) “Now set the hands to show 11:10 (10 past 11).” If subject has not finished clock drawing in 3 minutes, discontinue and ask for recall items.

3. Say: “What were the three words I asked you to remember?”

For each correctly recalled word, score 1 point. Score the clock 2 points if it is normal and 0 points if it is abnormal. The total score is the 3-item recall score plus the clock score. A score of 0, 1, or 2 indicates possible impairment. A score of 3, 4, or 5 suggests no impairment.
B  Ambulation Study Questionnaire

Have you previously done any therapy in the Community at St. Lukes Rehabilitation Institute? (YES/NO)

1. (IF YES) When was the last time you were here?

2. What did you work on or do while you were here?

3. Which modules have you used in the Community?

   (a) Restaurant
   
   (b) Bank
   
   (c) Store
   
   (d) Office
   
   (e) Hotel Lobby
   
   (f) Car
   
   (g) Truck
   
   (h) Airplane
   
   (i) Stop Walk
   
   (j) Bus

Anthropometric Measurements taken with tailor’s tape to collect data about:

1. Dominant side (left/right)
2. Involved side (left/right/both/no paresis)

3. Lower leg length (ankle joint to knee joint using the following landmarks: medial malleolus to medial condyle)

4. Upper leg length (knee joint to hip joint using the following landmarks: medial condyle to greater trochanter of femur)

5. Torso length (hip joint to the base of the neck using the following landmarks: greater trochanter of femur to top of trapezius muscle)

6. Foot length (including footwear)

7. Sensor positioning for each left and right leg
   
   (a) Top of sensor to knee joint
   
   (b) Top of sensor to ankle joint

8. Assistive device type (cane/2 wheel walker/4 wheel walker/no assistive device)

9. Sensor position for assistive device (if applicable)
   
   (a) Top of sensor to floor
   
   (b) Top of assistive device to floor
C  Acronyms

- AC: Ambulation circuit
- AD: Alzheimer’s disease
- ADL: Activities of daily living
- ADM: Activity density map
- AR: Activity recognition
- ARAT: Action Research Arm Test
- BF: B-fit
- CAP: Clinical assessments of progress
- CASAS: Center for advanced studies of adaptive systems
- CI: Confidence interval
- CMG: Case mix group
- COM: Center of mass
- CP: Control point
- CS: Change score
- CV: Coefficient of variation
• DTW: Dynamic time warping

• ES: Effect size

• FIM: Functional Independence Measure

• FMA: Fugl-Meyer Assessment

• GC: Gait cycle

• GF: Gait features

• GLCM: Grey-level co-occurrence matrix

• GUG: Get Up and Go

• HCSP: Hybrid clinical sensor prediction

• HEOM: Heterogeneous Euclidean overlap metric

• HS: Hybrid-synthetic

• IC: Initial contact

• ICC: Intraclass correlation

• IMU: Inertial measurement unit

• IQR: Interquartile range

• JP: Joint prediction

• JPP: Joint patient prediction
• KINARM: Kinensiological Instrument for Normal and Altered Reaching Movements

• KL: Kullback-Leibler

• LOOCV: Leave-one-out cross validation

• LOS: Length of stay

• LS: Left shank

• MAE: Mean absolute error

• MCI: Mild cognitive impairment

• MI: Movement intensity

• MS: Mid-swing

• NAC: Non-ambulation circuit

• NN: Nearest neighbor

• NRMSE: Normalized root mean squared error

• NTBI: Non-traumatic brain injury

• NTSCI: Non-traumatic spinal cord injury

• NWS: Normalized walking speed

• PA: Physical activity

• PACD: Physical activity change detection
• PCAR: Permutation-based change detection in activity routine

• PD: Parkinson’s Disease

• PID: Participant identification

• PS: Patient similarity

• RCI: Reliable change index

• RER: Rehabilitation efficiency ratio

• RFECV: Recursive feature elimination with cross validation

• RFID: Radio frequency identification

• RIC: Rehabilitation impairment category

• RM: Repeated measures

• RMS: Root mean square

• RMSE: Root mean squared error

• ROM: Range of motion

• RS: Right shank

• RuLSIF: Relative unconstrained least-squares importance fitting

• S1: Session one

• S2: Session two
• SD: Standard deviation

• SH: Smart home

• SMD: Standardized mean difference

• SMOTE: Synthetic minority oversampling technique

• SVM: Support vector machine

• sw-PCAR: Small window adaptation of PCAR

• TBI: Traumatic brain injury

• TC: Terminal contact

• TUG: Timed Up and Go

• TUG-DT: Timed Up and Go Dual Task

• uLSIF: Unconstrained least-squares importance fitting

• VC: Virtual classifier

• WBM: Whole body movement
REFERENCES


[88] Shimmer Sensing. SDLog for Shimmer3 Firmware User Manual Revision 0.6a, 2014.


