ACCELEROMETER-BASED ANALYSIS OF GAIT AND PREDICTION OF FALL RISK IN OLDER ADULTS

BY

ZACHARY STOKES QUICKSALL

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioengineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Professor Bruce Schatz

ABSTRACT

Falls are the most common cause of injury in older adults with two-thirds of individuals over the age of 65 falling at least once a year. It is well known that falls represent a significant challenge to preserving quality of life as we age, but current clinical methods of screening for fall risk remain insufficient to prevent falls. This thesis summarizes the development of a modern approach to fall risk analysis and fall prevention through the use of hip-mounted triaxial accelerometers to passively monitor gait quality in free-living environments and predict risk of future falls.

Data from over 4000 individuals enrolled in the Women Health Initiative's Objective Physical Activity and Cardiovascular Health study were used for the development of an activity recognition pipeline for extraction of free-living walking bouts measured by accelerometers. A variety of measures of gait were computed from walking bout data and used as input to train statistical models which analyze gait to predict fall risk and future falls.

Results suggest that hip-mounted accelerometers are able to capture free-living gait patterns which can be used to predict measures of fall risk and physical function such as the Short Physical Performance Battery. However, these same measures of gait prove to be insufficient for direct prediction future falls.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the University of Illinois at Urbana-Champaign (UIUC) Department of Kinesiology and the facilities made available by the Carl R. Woese Institute for Genomic Biology for data analysis and software development.

I am grateful to all those with whom I have had the pleasure to work during the progression of this project. I am especially indebted to Dr. Bruce Schatz—my primary advisor, Professor of the Institute for Genomic Biology and College of Medicine—for his guidance and support of both my research, future career, and life.

Special thanks to my secondary advisor, Dr. David Buchner, for providing access to such a unique and exciting data set and making available his extensive knowledge regarding physical activity and falls in older adults. Additional thanks for his contributions during our insightful, inspiring, and lengthy meetings as well as the preparation of manuscripts.

A great many thanks to Andrew Hua for securing our publication in Nature Digital Medicine and his contributions to other manuscripts in preparation. His dedication to this project and talent for finding clever solutions to our many analytical challenges proved invaluable for maintaining progress and morale.

Thanks to Qian Cheng for giving me a running start in both my research and establishing myself as a graduate student at UIUC.

Finally, thanks to my parents who have shown me unwavering love and support during the arduous completion of this project and my degree program.

"Our greatest glory is not in never falling,

but in rising every time we fall."

- Confucius

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Falls are the most common cause of injury in older adults with two-thirds of individuals over the age of 65 falling at least once a year [1]. During 2014, approximately 2.8 million adults were treated for fall-related injuries in emergency departments, and about 27,000 older adults died due to falls or fall-related injuries [1]. The impact of falls on morbidity and mortality have made falls a top health concern in the nation. Accordingly, the U.S. Preventive Services Task Force recommends screening older adults for fall risk and implementing prevention strategies in high-risk adults, such as exercise programs [2].

There exists several methods of screening for fall risk. For example, the Centers for Disease Control and Prevention (CDC) has developed the STEADI toolkit, which includes a screening approach that combines questions about falls and functional limitations with simple physical performance tests such as the Timed Up & Go (TUG) [3]. Overall, the sensitivity and specificity of existing screening methods is modest and are performed infrequently. Additionally, the majority of tools and devices that are currently in use to address falls are reactive in nature and do little to prevent falls in the first place.

One potential approach for screening fall risk that may better reflect risk during daily life is the use of wearable devices during walking tasks to characterize gait and detect instability related to greater risk of falls. The use of triaxial accelerometers has several desirable characteristics for screening purposes as these sensors are becoming more affordable and are available in consumer devices such as smart phones, which are nearly ubiquitous. The development of automated systems for acquisition and analysis of free-living gait using these sensors present several

challenges including accurate recognition of walking activity, extraction of meaningful features to characterize gait, and the development of models that can accurately map between these features and a meaningful measure of fall risk. Ideally, this measure of risk must be clinically interpretable and timely to provide guidance in selecting proactive measures to reduce the risk of falls. As such, systems for continuous and passive monitoring of fall risk would provide the greatest opportunity to combat falls and reduce their impact on both our population and healthcare systems.

In this work I aim to develop, in conjunction with colleagues, an automated pipeline for gait analysis and fall risk prediction. First, walk test data from a small pilot study—measured via triaxial accelerometer—are used to develop preliminary methods for feature extraction and fall risk prediction models; the results from this pilot study were used to inform a larger investigation which uses free-living accelerometer data. For this more complicated project, a set of algorithms and filters were designed to analyze free-living activity data and extract bouts of smooth, clean walking which resemble those generated during a walk test. Due to the use of unlabeled data, walking bouts were confirmed via visual inspection to ensure high specificity of the filters. Finally, features extracted from these walking bouts were used to train statistical models on prospective falls data to predict measures of fall risk and future falls. Model performance was analyzed using appropriate summary measures and more detailed assessments through inspection of confusion matrices.

CHAPTER 2: BACKGROUND

As stated previously, falls represent a significant challenge for older adults; however, falls also place substantial burdens on the general population and our healthcare system. This concern is exacerbated by the aging US population of which older adults are projected to represent 20% by 2050 [4]. The physical consequences of falls such as long-term injury, disability, or diminished quality of life combined with psychological changes associated with fear of falling and confidence in mobility can lead to a negative cycle which may actually increase future fall risk [5]. Furthermore, annual costs for treatment of fall-related injuries are roughly $20 billion in the United States with costs projected to reach $32.4 billion by the year 2020 [5].

Given the prevalence and seriousness of falls in older adults, two general approaches toward falls management have emerged [5]. The first is a *reactive* approach which uses devices to detect fall events. These devices, called personal emergency response systems (PERS), provide older adults with immediate access to emergency services and have been shown to improve the general quality of life of individuals living alone [6]. These devices have substantially shortened the time between fall incidents and treatment which is an important factor in determining successful recovery. The second approach is focused on fall *prevention* through interventions including exercise, strength and balance training, assistive devices, and modifications to the home environment [5]. While multifaceted intervention programs including exercise and balance training have been shown to reduce falls, identification of at-risk individuals for placement into these programs remains a challenge [7]. In addition to managing the effects of falls, assessment of fall *risk* is needed to reduce future complications and truly improve fall prevention.

The majority of fall risk assessments take place in the clinic and use a combination of questionnaires, physical activity measurements, and gait assessments to determine fall risk. Questionnaires cover a variety of behaviors associated with risk of falls such as confidence in completing activities of daily living and engagement in outdoor activities [8]. Physical assessments, such as the Short Physical Performance Battery (SPPB) evaluate fall risk by measuring balance, gait, and muscular strength [9], [10]. While more comprehensive assessments of fall risk have been developed, their feasibility for mass screenings is questionable. For example, Lord *et al.* have developed a comprehensive fall risk assessment tool which measures physiologic capacity in each organ system related to falls, but the short version of this tool requires equipment that is not readily available, 10-15 minutes for administration, and a trained assessor [11]. Existing methods of screening for fall risk typically involve assessments in a clinic or laboratory, and therefore may not reflect fall risk during everyday life activity (i.e., real-world monitoring). Likewise, such assessments are highly subjective, depend upon observer expertise, and tend to oversimplify risk [5].

More recently, researchers have been investigating the use of instrumented fall-risk assessment and predictive tools through the use of data collected from inertial sensors worn on the body [12]. While a variety of devices exist, the use of triaxial accelerometers has several desirable characteristics for screening purposes. In addition to increasing prevalence in pervasive technologies such as smart phones, accelerometers present a cheap means of assessing gait. When sensor data are collected at 30 hertz or greater, raw data can provide precise measures of gait such as variability among gait cycles during walking tasks [13], [14]. While many other devices can be used for brief clinical gait assessments, accelerometers offer the option of basing

fall risk assessments on data collected during actual activities of daily living, including frequent longitudinal measurements from worn or carried devices.

Assessment of gait quality and fall risk prediction through the use of inertial sensors has been a popular topic over the past two decades with a substantial number of publications using accelerometer-type sensors [12]. The utility of accelerometers assumes that fall risk is consistently correlated with characteristics of body movement and gait, and that these characteristics can be accurately detected using sensors measuring body motions. Instability during movement and walking is a primary cause of actual falls, as emphasized in studies analyzing senior falls in nursing homes [15]. Several studies have demonstrated the potential of using raw data from wearable devices to predict fall risk by identifying gait-related risk factors. Some use multiple sensors across the body (head, torso, pressure insoles, etc.) [5], [16]–[19]. Other studies use specialty sensors developed in the lab (not commercially available, requiring hardware development to collect data) which limits applicability. A recent literature review by Montesinos *et al.* confirms this heterogeneity in accelerometer-based gait analysis studies and notes that differences in sensor placement, extracted features, and the type of motions or tasks performed for analysis remain a major challenge in the development of real-world tools [12]. Additionally, the majority of studies make use of retrospective fall history as their prediction target which calls into question whether these models will be able to predict *future* fall risk [12]. The absence of converging results and protocols illustrates the challenges associated with automated gait analysis and the conflicting nature of performance versus scalability with the use of low-cost wearable sensors.

Ultimately, gait analysis and fall risk prediction in free-living environments necessitates algorithms for activity recognition to extract proper inputs for predictive models. Activity recognition using body-worn accelerometers has been extensively studied with the majority of existing models using data gathered from young, healthy individuals (e.g. college students) [20]. Models trained on this population typically do not generalize well to older adults or even individuals who may be of very similar, but slightly different, demographics to those in the training set [20]. Furthermore, activity recognition models are generally trained on laboratory data using prescribed tasks which, ultimately, may not be representative of the tasks performed without observation in the real world [20]. Generalizable activity recognition models trained on older adults are necessary for automated analysis of free-living gait and must be established before models of fall risk are of any practical use.

CHAPTER 3: OBJECTIVES

The primary objective of this investigation is to develop a system for predicting fall risk (and potentially future falls) in older adults from walking data gathered in their natural free-living environment using a hip-mounted accelerometer. This larger goal requires the isolation of free-living walking data which represents the individual's typical gait characteristics and, in theory, contains information about their risk of falling. These walking data must then be converted into an appropriate format via extraction of features which summarize the signal without loss of information relevant to fall risk. Next, these features are used to train statistical models to map between the walking signal and a measure of fall risk. Finally, these three processes must be integrated into an automated pipeline for passive monitoring of risk; this was ultimately accomplished through three subsequent investigations which built upon both previous findings in the fields of activity recognition and fall prediction using inertial sensors, and the outcomes of previous investigations. In brief, a preliminary study was conducted using walk test data to verify the predictive capability of accelerometer gait measurements. The results of which informed the development of a walking activity recognition pipeline which was used to study free-living gait in relation to future falls.

CHAPTER 4: DATA SETS

The physical activity, falls, and accelerometer data used in this project were obtained from a subset of individuals initially recruited in the 1990s for the Women's Health Initiative (WHI). Individuals who consented to participate in the second extension study of WHI (2010 - 2015) and an ancillary study titled OPACH (Objective Physical Activity and Cardiovascular Health in Older Women, R01 HL105065; PI: A LaCroix) received an in-home visit for data collection and completed a physical activity questionnaire between March 2012 and May 2013.

Of the women in the OPACH study, a subset of individuals consented to participate in a calibration substudy which calibrated the accelerometers used in the OPACH in-home visits. A notable exclusion criterion for the calibration substudy was a significant change in health status affecting ability to walk or risk of walking-related injury between OPACH data collection and recruitment for the substudy. Of the N=7058 women enrolled in OPACH, N=142 participated in the calibration substudy.

The methods of both OPACH and the calibration substudy are described in detail elsewhere [21], [22]. Thus, this study is a secondary analysis of OPACH data for the purpose of exploring predictive models of fall risk. Consent to participate in OPACH was obtained by either phone or mail. After a screening phone interview, participants in the calibration substudy provided written consent at the study's clinic visit. The OPACH study and the calibration substudy were approved by the Institutional Review Boards at each clinical site and by the WHI Clinical Coordinating Center.

Both the calibration and OPACH studies utilized hip-mounted ActiGraph GT3X+ triaxial accelerometers to measure motion in laboratory and free-living environments, respectively. The axes of measurement on each accelerometer were aligned with the major directions of bodily motion as depicted below.

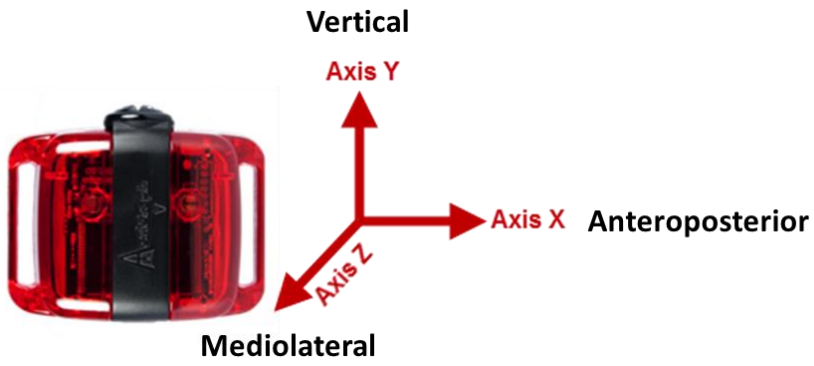

Figure 4.1: Diagram of the ActiGraph accelerometer showing alignment of the axes of measurement with the main directions of human motion.

Accelerometers were set to a sampling rate of 30 Hz and possess a dynamic range of +/- 6 G's (units of gravity). Accelerometer data in the form of RData files were converted to CSVs for subsequent processing and analysis. Structured data in the form of CSVs provided information related to demographics, physical activity, and falls for each individual. Common to both studies, demographics information including age, height, and weight were used in conjunction with the accelerometer data for the purpose of predicting measures of fall risk (i.e. physical function) and falls. For the calibration substudy, outcome measures included the composite SPPB scores and a history of falls during the prior year. While the full OPACH study also provided composite SPPB scores, a one-year prospective falls log with fall numbers and rates was available in addition to the past fall history. More detailed information concerning each of these variables can be found in the data dictionary of Appendices A and B.

CHAPTER 5: EXPERIMENTAL DESIGNS AND TOOLS

Three main projects were carried out to develop an automated system for fall risk prediction given the available sensor data detailed in Chapter 4. The first of these was a pilot study which used accelerometer data gathered during a walk test to predict past fall history. Moving toward real-world monitoring, a pipeline was developed for the identification of free-living walking bouts similar to those observed during the calibration substudy walk test. Finally, combining the outcomes of the previous two analyses, an attempt was made at automated prediction of fall risk and future falls from free-living walk data. A brief summary of the methodologies and tools used in each of these experiments are detailed here.

Fall Risk Prediction from Walk Test Bouts

A total of 142 individuals participated in the calibration substudy which generated the walk test data used in this investigation. Of this group, 69 individuals remained after exclusion for missing or inapplicable data. Accelerometer time series were segmented using preexisting timestamps and activity labels to separate walking bouts from other motion. These bouts were subsequently processed into smaller segments using a sliding window from which features were extracted for input into predictive models. Random forests were then trained via 10-fold cross validation to predict past fall history in the form of high and low risk categories.

Isolation of Good Walking Bouts from Free-living Data

Real-world use of fall risk prediction models based on gait analysis require the identification of walking activity present among the other activities of daily living. Given that good walking bouts--those which resemble walking during a walk test--are better suited for gait analysis than

the potentially chaotic bouts seen during most daily activities, care must be taken to isolate only good walking. A pipeline was developed using multiple statistical filters to separate good walking bouts from all other accelerometer activity.

Fall Risk Prediction from Free-living Walking Bouts

Unlike scripted walk tests, free-living walking bouts have greater variation in duration and quality. As such, different methodologies are required to prepare these data for gait analysis and feature extraction. Features were extracted from the full free-living bouts of each individual and averaged if multiple bouts were present for a single participant to generate a single, flat feature vector per participant. Random forest models were trained using a more robust 50/50 train-test split to predict both *future falls* and *fall risk*.

Random Forests

The random forest implementation in the scikit-learn library was chosen for model development [23]. As an ensemble modeling approach, random forests reduce the bias of their predictions through the construction of a large number (usually $\geq$ 500) of individual decision trees. A reduction in variance is obtained through the bootstrapping process which, assuming the original data set is a representative sample of the population, draws a new sample of individuals from the original set with each tree constructed. This helps guard against overfitting which is commonly seen in the real-world application of decision trees.

Going one step further, a random subset of features is selected at each split in the tree with the best split chosen as the one which maximizes entropy reduction (i.e. best improves local class separation). This additional variance would typically be expected to reduce prediction

performance. However, the effect of this additional variance is averaged across all of the trees in the forest and, in practice, usually produces a more accurate, stable, and generalizable model.

Another benefit of random forests over other modeling techniques is that they can handle categorical predictors with minimal changes to encoding. This allows for the use of continuous, ordinal, and nominal data within the same model as is often encountered in biological and epidemiologic analyses [24].

Short Physical Performance Battery

The Short Physical Performance Battery (SPPB) is a tool for evaluating lower extremity function in older adults and has been used to measure physical function status [25]. Weakness in lower limbs (measured via SPPB) has been shown to be associated with recurrent falls in older adults [9]. In addition to falls, SPPB is predictive of a variety of health measures including loss of independence, general decline in health, re-hospitalization, increase duration of hospital stay, and even mortality [26]. Moreover, the simplicity of the SPPB combined with the ability to perform the evaluation in-home without complicated equipment makes it an ideal exam for assigning fall risk to the population.

In 2010, the WHI Steering Committee selected the use of the "Look Ahead" SPPB scoring system which produces an overall SPPB score through the summation of three ratios: the "Standing Balance Ratio," "Chair Stand Ratio," and "Usual Walk Ratio"; these three values sum to a final score between 0 and 3 [27]. In more detail, Standing Balance Ratio is defined as the summation of three standing balance tests where the feet are placed (1) side-by side (2) semi-

tandem, and (3) tandem, and are evaluated based upon an individual's ability to maintain these positions for up to 10 seconds. The "Chair Stand Ratio" is obtained from the time required for the participant to rise up and sit down in a chair five times whereas the "Usual Walk Ratio" is computed from the participant's pace during a 3-4 meter walk test [27].

The "Look Ahead" SPPB scores were converted to the scoring system used in the "Established Populations for Epidemiologic Studies of the Elderly (EPESE)" project for comparison between studies. We selected the EPESE SPPB score for use in all of our SPPB-related analyses for its discrete, ordinal scale and established performance; the details of score conversion can be found elsewhere [27].

CHAPTER 6: PREDICTION OF FALL RISK FROM WALK TEST

Subject Exclusion

Several exclusion criteria were established to remove individuals with missing information and ensure that both an adequate number of participants and amount of accelerometer data were available for analysis. Individuals in the calibration substudy were somewhat self-selected against these potential pitfalls in that individuals unable to walk without an assistive device (i.e. cane or walker) were unable to participate in the substudy. This resulted in a patient population able to complete at least part of the 400 meter walk test and generate a sufficient amount of walk data for analysis. However, a number of individuals were excluded from analysis on the basis of SPPB scores and past fall histories. These criteria are described in greater detail under subsection titled “Definition of Fall Risk.”

Definition of Walk Test Bouts

Good walking refers to a steady pattern of walking similar to walking along a straight path or walking that is observed during a walk test [28], [29]. It is important to note that this does *not* include walking that is performed on a treadmill which fixes walking speed and is not equivalent to natural walking. Previous research effectively utilized good walking data as part of a pipeline for gait analysis and highly accurate prediction of pulmonary function in both laboratory and free-living environments [28], [29]. However, isolation of good walking bouts from these two environments necessitates different standards of bout quality as well as automated methods capable of handling changes in environment.

Walking data gathered in a laboratory environment is typically in the form of labeled data restricted to a set number of activities; this was the case for the calibration substudy which had individuals perform a 400 meter walk test. Specifically, participants were instructed to walk a total distance of 400 meters at their natural pace while staff monitored performance and recorded start and end times. As such, separation of the data into different categories of tasks is trivial and allowed for quick reorganization of the data into different groups or sets.

Walk Test Bout Extraction

Extraction of calibration substudy bouts was accomplished through the simple process of segmenting the data according to timestamps and activity labels already present in the data set. These resulting data were passed to a final "good walking" filter developed by Cheng *et al.* which examines the consistency of variation in a potential "good walking" sample to remove bouts which display a great amount of interruption such as frequent starting and stopping [28]. The vector magnitude of the raw accelerometer data is computed and segmented using a one-second sliding window. The standard deviation of each one-second segment is computed and compared to the dynamic threshold set by the algorithm. The resulting system generates a binary good/bad walking decision for each data point and returns a good walking example if 70% of the data are considered to be good walking [28]. Two example walking bouts which highlight the sensitivity and accuracy of the good walking filter are displayed in Figure 6.1, below.

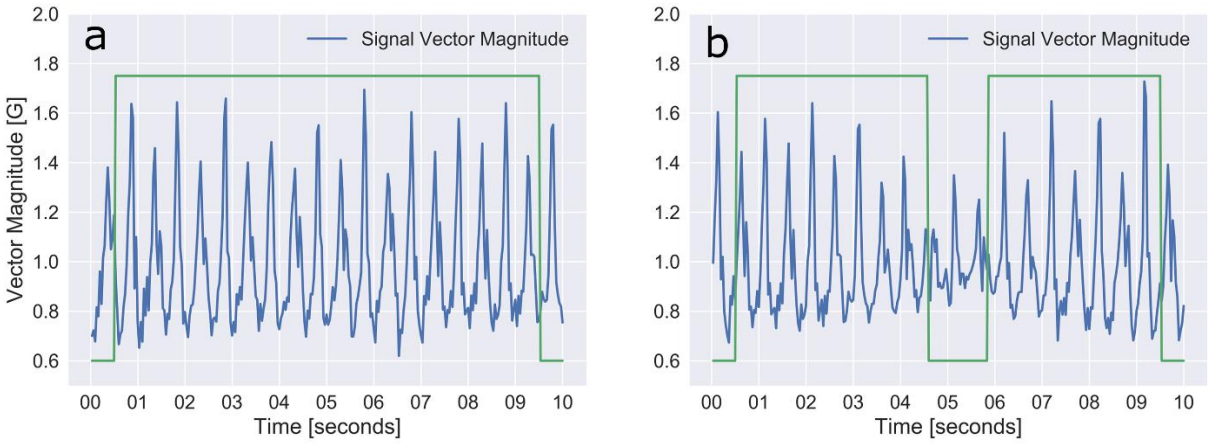

Figure 6.1: Example accelerometer vector magnitude from a single individual during the 400 meter walk test. Figure 6.1.A: A ten-second period of smooth walking without any turns. The "good walking" algorithm identifies the full segment as good walking (high green line). Figure 6.1.B: A ten-second period of smooth walking which contains a turn. The "good walking" algorithm identifies the substantial reduction in acceleration magnitude and eliminates this portion of the walk tracing (green line drops down). Note, the first and last 15 data points in any segment are always identified as "non-walking" due to insufficient information.

Definition of Fall Risk

The fall risk of calibration substudy participants was determined using two well-known predictors of fall risk: history of falls in the past year and SPPB score. The CDC also uses these predictors to classify fall risk in the STEADI Toolkit [3]. In a study of 66,134 postmenopausal women, the strongest predictor of future falls was any fall in the past 12 months [30]. A study of the SPPB and fall risk concluded that, in older women, SPPB scores of 9 or less are associated with higher fall risk in women [9]. Hence, women were classified as "high fall risk" (N=19) based on SPPB scores of $\leq$ 9 and reporting of at least one fall in the past year. Women were classified as "low fall risk" (N=47) based on SPPB scores of 10-12 and no reported falls in the past year. For the purposes of the calibration substudy which has a small number of subjects from the larger study, an attempt was not made at the more difficult classification task of distinguishing between women of high fall risk versus intermediate fall risk (i.e. women with past falls but SPPB scores of 10-12, or women with SPPB scores of 9 or less, but no past falls).

This more complicated task was reserved for exploration in the larger OPACH study and remains a significant challenge.

Feature Extraction

Using the calibration substudy walking bouts which passed the "good walking" filter, features were extracted from ten-second samples segmented using a sliding window with 50% overlap. Signal-based features in the time and frequency domains were computed for each of the individual accelerometer axes and the vector magnitude. The features selected for extraction were chosen based upon both knowledge of how gait affects fall risk and findings of research on assessing fall risk using inertial, wearable sensors [5]. Features were organized into groups to assess the relative predictive ability of traditional measures of gait (i.e. those based upon limb movement such as cadence or time between steps and strides) and signal-based features of the accelerometer data (e.g. simple statistics: mean, standard deviation, or power). The specific features in each feature group can be found in Appendix A.

Predictive Modeling

In addition to direct prediction, certain machine learning models can be used to identify a subset of features that capture the most useful content for a larger classification problem—a process called feature selection. Some models available for this purpose include Decision Trees, Random Forests, and Support Vector Machines. To determine the model likely best suited for this study's classification task, a simple spot-checking approach was used. This approach involved training each classifier with default parameters on the full feature set and evaluating performance via 10-fold cross validation. With 10-fold cross validation, the data were divided into 10 equally-sized partitions with nine partitions used for model training and one for testing. This process was

repeated such that each partition was used once for testing. Metrics averaged across all ten folds were used to compare model performance and included accuracy, precision (positive predictive value), recall (true positive rate or sensitivity), F1-score (harmonic mean of precision and recall), and area under the ROC curve (AUC). Based upon the results, it was deemed that random forests were likely the most appropriate classifier to use for the study task; this is not too surprising given the adaptability of random forests and the complex nature of the given classification task. Support vector machines, while very popular and effective in modern approaches to machine learning, require careful tuning of cost parameters and selection of kernels for projecting the data into higher dimensions before high accuracies can be achieved. By comparison, random forests demonstrate more flexibility "out-of-the-box" and require minimal parameter tuning.

Random forests of 1000 trees were trained and evaluated using 10-fold cross validation implemented in the scikit-learn library [23]. In practice, 500 trees is usually more than sufficient to maximize the benefits of averaging predictions in the forest. More modern approaches use cross-validation to select an "optimal" number of trees based upon model performance on a validation data set. This approach was not used in this investigation, however, due to the small number of participants which inhibits the creation of a reliable validation set for parameter tuning. Again, metrics including accuracy, precision, recall, F1-score, and AUC were used to assess performance. Separate forests were trained on each of eleven feature sets to obtain further insight into the usefulness of certain feature types and the combined effects of certain feature groups. These eleven feature sets were selected to obtain insight into the contribution of each accelerometer axis toward risk prediction and test possible interactions. Traditional measures of

gait were included in the feature sets to compare the utility of measures of limb movement (e.g. step and stride) with that of the hip (i.e. motion of the accelerometer).

Assessing Feature Importance

Relative feature importance in random forests was characterized by mean decrease impurity [31]. Impurity is computed by summing the weighted reduction of sample entropy for all splits that utilize the feature of interest; the resulting values are then averaged across all trees in the forest. Formally, importance for a single variable X_m is computed as the weighted sum of impurity decreases $p(t)\Delta i(s_t, t)$ for all nodes t averaged over all N_T trees [31]:

$$Imp(X_m) = \frac{1}{N_T}\sum_{T} \quad \sum_{t \in T: v(s_t) = X_m} p(t)\Delta i(s_t, t)$$

Where $p(t)$ is the proportion of instances reaching split s_t [31]. Feature importance was calculated independently for forests trained on each of the eleven feature sets. The top-ten features were identified for each forest and used to gauge feature applicability to fall risk prediction.

Results

Table 6.1 reports characteristics of the calibration substudy participants by fall risk category. Women in the high and low fall risk groups were not significantly different in age, ethnicity, and education. By design, because SPPB scores were used to define fall risk groups, the overall SPPB score and each of the three SPPB subscores differed significantly between groups. Interestingly, although the means for each subscore are different between the two risk groups,

only the “Chair Stand” subscore shows separate numeric ranges which do not overlap within one standard deviation. This is significant given that the accelerometer can directly measure gait and balance during a walk test, but not the unique motions of a chair stand. This highlights the effect of calibration substudy exclusion criteria in selecting a generally healthy sample and the possible challenges introduced by use of a walking task rather than chair stand. Average cadence during good walking was about 124 steps/min. Average values for most variables differed significantly by risk group.

Table 6.1: Calibration substudy participant characteristics by fall risk category.

Characteristic	***Total***	***High fall risk (≤9 SPPB and >0 falls)***	***Low fall risk (>9 SPPB and 0 falls)***	***p-value***
N (%)	67	19 (28.4%)	48 (71.6%)	
Age, years, mean (SD)	77.5 (6.1)	77.3 (5.9)	77.6 (6.2)	0.829
Height, inches, mean (SD)	62.9 (2.3)	63.0 (2.1)	62.9 (2.40	0.815
Weight, pounds, mean (SD)	152.3 (30.0)	156.9 (33.5)	150.5 (28.7)	0.434
Race/Ethnicity, n (%)				0.582
Non-Hispanic White	20 (29.9%)	5 (26.3%)	15 (31.3%)	
Non-Hispanic Black	19 (28.4%)	5 (26.3%)	14 (29.2%)	
Hispanic/Latina	28 (41.8%)	9 (47.4%)	19 (39.5%)	
Highest Education Level				0.793
High school diploma/GED or lower	12 (17.9%)	4 (21.0%)	8 (16.7%)	
Vocational or training school	4 (6.0%)	1 (5.3%)	3 (6.3%)	
Some college or Associate Degree	19 (28.4%)	6 (31.6%)	13 (27.1%)	
College graduate or more	32 (47.8%)	8 (42.1%)	24 (50.0%)	
EPESE SPPB Score, mean (SD)	10.1 (1.5)	8.3 (1.1)	10.8 (0.8)	<0.001
Balance Subscore , mean (SD)	3.9 (0.4)	3.7 (0.7)	4.0 (0.0)	0.004
Chair stand Subscore , mean (SD)	2.7 (1.1)	1.4 (0.8)	3.2 (0.8)	<0.001
Gait Subscore, mean (SD)	3.5 (0.7)	3.2 (0.9)	3.6 (0.6)	0.016
Number of falls in the past year				

Table 6.1: Continued.

0 Falls	*48 (71.6%)*	*0*	*48 (100%)*	
1 Fall	13 (19.4%)	13 (68.4%)	0	
2-3 Falls	6 (9.0%)	6 (31.6%)	0	
Cadence (steps/minute), mean (SD)	123.5 (16.5)	120.5 (15.5)	124.6 (16.7)	<0.001
Vector magnitude CoV, mean (SD)	0.216 (0.049)	0.205 (0.047)	0.220 (0.050)	<0.001
Vector magnitude ACC, mean (SD)	0.476 (0.202)	0.500 (0.228)	0.458 (0.191)	<0.001
Vector magnitude, mean (SD)	0.998 (0.014)	0.995 (0.012)	0.998 (0.014)	<0.001
X acceleration, mean (SD)	-0.137 (0.114)	-0.103 (0.115)	-0.149 (0.112)	<0.001
Y acceleration, mean (SD)	-0.889 (0.088)	-0.910 (0.080)	-0.881 (0.089)	<0.001
Z acceleration, mean (SD)	-0.124 (0.323)	-0.044 (0.293)	-0.153 (0.328)	<0.001
X CoV, mean (SD)	-1.0 (27.8)	-1.6 (18.2)	-0.8 (30.4)	0.221
Y CoV, mean (SD)	-0.225 (0.053)	-0.206 (0.041)	-0.231 (0.055)	<0.001
Z CoV, mean (SD)	0.0 (78.9)	0.1 (153.6)	-0.1 (13.2)	0.924
X ACC, mean (SD)	0.397 (0.193)	0.392 (0.197)	0.399 (0.191)	0.109
Y ACC, mean (SD)	0.394 (0.216)	0.415 (0.239)	0.386 (0.206)	<0.001
Z ACC, mean (SD)	0.352 (0.254)	0.374 (0.270)	0.344 (0.248)	<0.001

The results of providing various feature sets to random forest classification models are available in Table 6.2. Classifiers were trained using 10-fold cross validation to ensure proper separation of training and testing data and limit overfit. The models performed with an average accuracy of 73.7%, precision of 81.1%, and AUC of 0.706 and could discriminate between high and low fall risk classes. The best performing feature set was feature set #10 (accuracy = 79.3%, precision = 84.6%, and AUC = 0.834), and it included: data from each axis, cross-correlations between axes, and traditional measures of gait. Combining individual axes data into a vector magnitude (feature set #9) reduced the performance of the model slightly compared to the mean (accuracy = 71.4%,

precision = 78.5%, and AUC = 0.616). Traditional measures of gait alone performed the worst (accuracy = 69.0%, precision = 75.0%, and AUC = 0.545). Models including single axis data performed better though not as well as the models including data from all three axes. Of the single axis models, the model containing vertical data outperformed mediolateral and anteroposterior models. Adding traditional measures of gait to models with signal-based features had little effect in most cases. On average, accuracy and precision differed by 0.8% and AUC differed by 0.006 compared to models containing only signal-based features. In all models containing signal-based and traditional measures of gait, signal-based measures were consistently ranked above traditional measures of gait. In the top performing models, 4 out of the top 5 features were derived from the mediolateral dimension (z-axis). It is interesting to note that the use of traditional measures of gait in combination with all signal-based features (i.e. transitioning from feature set 10 to feature set 11) leads to worse performance. While it is generally the case that more information is better, it may be that the traditional measures of gait are injecting more noise than signal into the model and reducing predictive accuracy.

Table 6.2: Performance metrics from 10-fold cross validation for random forest classification of high and low function women on each of eleven feature sets.

Set	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*	*Feature Groups*
1	69.0%	75.0%	0.873	0.807	0.545	Gait
2	71.9%	79.0%	0.845	0.817	0.665	X-axis
3	72.7%	79.1%	0.858	0.823	0.661	X-axis, Gait
4	75.9%	82.6%	0.855	0.840	0.730	Y-axis
5	76.2%	82.5%	0.862	0.843	0.727	Y-axis, Gait
6	70.9%	83.3%	0.760	0.795	0.759	Z-axis
7	73.1%	84.1%	0.785	0.812	0.771	Z-axis, Gait
8	70.9%	79.3%	0.822	0.807	0.616	Vector Magnitude
9	71.4%	78.5%	0.846	0.814	0.616	Vector Magnitude, Gait

Table 6.2: Continued.

.

10	*79.3%*	*84.6%*	*0.881*	*0.863*	*0.834*	*XYZ, Cross-correlations*
11	78.9%	84.4%	0.877	0.860	0.846	XYZ, Cross-correlations, Gait
AVG	73.7%	81.1%	0.842	0.826	0.706	N/A

The top ten features used by classifiers were identified for feature set #10 and for feature set #11. With feature set #10 (all feature groups eligible), the most important features were mediolateral signal-based measures followed by anteroposterior signal-based measures (Figure 6.2.A). With feature set #11 which included traditional measures of gait, the most important features were still mediolateral and anteroposterior signal-based features (Figure 6.2.B). In both models, the top three features included mediolateral coefficient of variance, correlation coefficient between anteroposterior and mediolateral accelerations, and mean mediolateral acceleration. These statistics are measures of side-to-side sway, unsteadiness, and asymmetry, respectively, which collectively describe core instability during walking. Traditional measures of gait were of lesser importance and did not rank amongst the top ten features; again highlighting the greater value of direct, signal-based features of instability compared to the measures of variation in step patterns provide by traditional measures of gait.

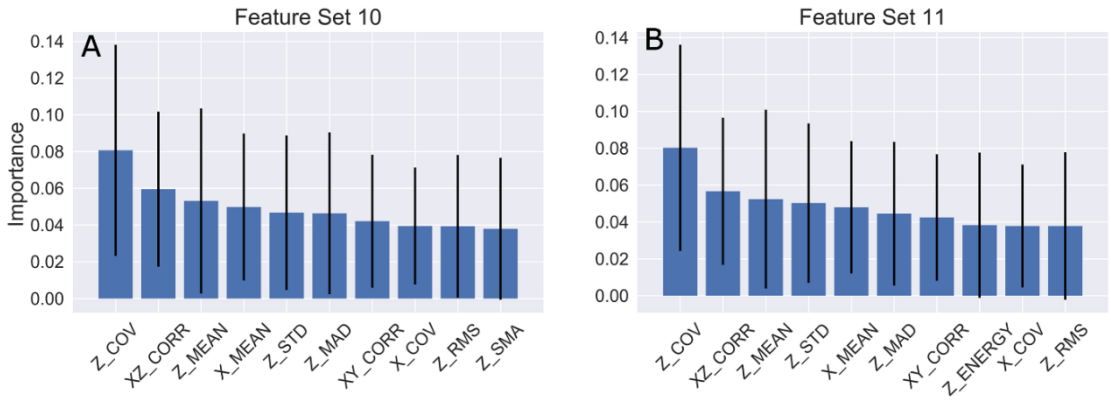

Figure 6.2: Top-ten features for two of the feature sets used in prediction of high and low function. Average importance of each feature for model prediction was computed as mean decrease impurity (see text) and is indicated by the blue bar. Black error bars represent standard deviation of importance across all trees in the forest. Figure 6.2.A. Top-ten features for a random forest model trained on features extracted from individual X, Y, and Z-axes and cross-correlations between axes. Figure 6.2.B. Top-ten features for a random forest model trained on features extracted from the individual X, Y, Z-axes, cross-correlations between axes, and traditional measures of gait.

Significance of Core Unsteadiness

The performance of predictive models developed on data from the calibration substudy suggest accelerometer-based measures of gait are potentially useful in screening older women for fall risk. Further, features derived from the accelerometer data extracted by the good walking algorithm were predictive of fall risk. Specifically, sideways (mediolateral) hip motion detected by the z-axis of a triaxial accelerometer may be a useful predictor of risk, such as the top three features in our analyses: coefficient of variance, correlation coefficient between two axes, and mean acceleration**.** The importance of features derived from z-axis data is plausible as excessive or variable sideways movement during walking, as measured by coefficient of variation, may increase fall risk [32], [33]. The sideways movement is consistent with age-related neuromuscular weakness due to slower motor unit recruitment with age [34]. That is, when motor control is diminished, gait variability increases, as older adults lack the ability of younger adults to respond to perturbations in gait by increasing neuromuscular control [35]. This may

present as chaotic accelerometer tracings since erratic muscular control can cause inconsistent accelerations. The perturbation in mediolateral movement may predispose older adults to higher risk of falling sideways by exceeding the bounds of stability and may portend greater odds of a hip fracture [36].

With the additional data, it was of little surprise that the triaxial models outperformed the single-axis models. Following the top ten features of the triaxial model, it would be expected that the mediolateral model would outperform both the anteroposterior and vertical models. However, the vertical model performed the best (accuracy = 76.2%, precision = 82.5%, and AUC=0.727). This may be explained by certain feature pairs being more predictive than any individual feature. That is, an x-axis feature may consistently adjust the instances that end up in a node further down in the tree such that the z-axis features are much better at separating the two classes. One possible biomechanical explanation is that vertical acceleration may be a correlate of primarily force production whereas anteroposterior and mediolateral acceleration are correlates of a combination of balance and force production. Therefore, the vertical model uses data that is more telling of an adult's physiologic capacity to walk safely than either the anteroposterior or mediolateral models. This is supported by the significantly lower SPPB chair stand subscore of the high fall risk group (SPPB chair stand subscore = 1.4) compared to the low fall risk group (SPPB chair stand subscore = 3.2). It is difficult to assess balance in a single plane and thus requires two axes (anteroposterior and mediolateral) to completely analyze the data. Furthermore, since musculature is a component of balance, this may explain why the mediolateral and anteroposterior features are more important than vertical features in the triaxial models.

Random Forests identified features that were reasonably predictive of fall risk, with an average accuracy of almost 73.7% and AUC of 0.706. While this level of accuracy and AUC are low compared to machine learning models that predict stable characteristics, the task of fall prediction is more challenging and the fall risk classification models met accuracy expectations. Furthermore, this average includes models that were not expected to perform well such as only traditional measures of gait and vector magnitude data. The performance of these models surpasses previous models that utilize only a single hip accelerometer during walking [37]. Multiple features of this study could explain the increased performance such as the longer one-year falls history, combination of falls history and SPPB for fall risk classification, and selection of random forests over neural networks which often suffer when provided small data sets (< 2000 instances). In-house solutions may acquire better accuracy by building higher resolution sensors or combining multiple sensors [21], [22]. Though gait quality is a strong predictor of fall risk, it is not the only risk factor. There are environmental risk factors and other host risk factors that are relatively independent of gait such as poor vision, postural hypotension, and ability of shoes to oppose slipping.

Using theoretical probabilistic models, one study estimated the maximum AUC when predicting falls within one year ranges from 0.80 to 0.89, with accuracies exceeding 80% challenging to achieve [21]. Our models are the first to nearly meet these predicted maximums with an accuracy of 79.3% and AUC of 0.834 (feature set #10). Of course, it is of interest to empirically test these theoretical maximums.

The potential of signal-based measures of acceleration as predictors of fall risk is also suggested by the fact that random forest classifiers, using only signal-based features (feature set #2, #4, #6, #8, and #10), performed similarly as classifiers including traditional measures of gait (feature set #3, #5, #7, #9, and #11**).** This may indicate the potential of machine learning to identify interactions among signal-based features that increase their predictive ability. Reduced feature-sets have also been shown to outperform full feature-sets [21].

The results of this study are consistent with the general finding of other studies that raw data from wearable accelerometers are potentially useful in fall prediction. Accelerometer-derived measures, including gait variability, can predict time to first fall in patients with Parkinson's disease [32]. The results are further consistent with other research that mediolateral and anteroposterior measures of sway and velocity are indicators of fall risk and that relatively brief gait assessments provide information on fall risk [21]. Some studies have attained greater predictive accuracy (up to 90.4%) by using a Timed Up & Go Test, rather than a simple walk test, which may better assess other risk factors including muscular strength and physiology [33], [34]. Furthermore, the models predicted fall risk based on assessments rather than actual falls history [34]. When using past falls history, reasonable accuracy, sensitivity, and specificity (80%, 74%, 96%, respectively) was achieved using accelerometer data from only a TUG test and a 20 m walk [35]. However, these studies utilized a homebrew accelerometer solution which may contain better sensors than commercial offerings but require expertise to implement [33]–[35]. Greater accuracy, sensitivity, and specificity can be achieved with multiple sensors on body parts other than the waist [21]. This finding suggests the use of accelerometers to assess characteristics of movement beyond only gait characteristics may improve predictive ability.

CHAPTER 7: EXTRACTION OF FREE-LIVING WALKING BOUTS

Definition of Free-living Walking Bouts

In free-living environments, accurate activity recognition has proven to be a substantial challenge plagued by poor performance and generalization—especially in older adults [38]. These shortcomings are further amplified in that the single activity of "walking" can vary in duration, frequency, and quality. Moreover, the inconsistency of free-living activities necessitates a specific and focused definition of free-living walking bouts for the purpose of isolating only good walking similar to that observed during scripted walk tests. In developing this definition, walking bouts were required to be at least one minute in duration for the purposes of eliminating possible shuffling performed during household tasks. This "longer" bout requirement allows sufficient time for individuals to reach their typical walking speed which eliminates transient acceleration and deceleration associated with starting and stopping motion. For bouts greater than one minute in duration, at least 70% of the bout must be "good walking" content. This constraint allows for longer walking bouts which contain short pauses during movement (e.g. pausing at a crosswalk). To guard against the unreasonable chaining together of several separate walking bouts, a cap on pauses between walking periods was defined such that a segment of continuous, non-good walking must not exceed 30 seconds.

Admittedly, these criteria are quite strict and the majority of real-world walking activity is far less than a minute in duration; indeed, the walk test used for the SPPB is only four meters in length or roughly four seconds of walking. However, it is not our goal to identify all bouts of

walking activity nor to characterize the walking profiles of individuals (which requires the capture of different types of bouts). Rather, we simply wish to identify real-world walking bouts that best mimic those seen during a walk test, but without the influence of observation during a laboratory activity or task. It is possible that shorter walking bouts may be of value for activity profiles or risk assessment, but the current focus is on long-duration, sustained walking activity.

Bout Extraction Pipeline

Unlike the calibration substudy, the bout extraction process for free-living data consists of multiple, progressive filters which extract segments of accelerometer data that meet our previous definition of a free-living good walking bout. First, a filtering process based upon the raw y-axis output from the accelerometer is used to identify segments of data that contain y-axis values between 0 and -2 which are related to the upright, vertical motion seen during walking. Depending upon sensor orientation and intensity of motion, not all periods of non-walking activity will be excluded by this initial filter as a variety of motions can generate vertical accelerations (e.g. bouncing or rocking in a chair).

Next, a cleaning process using the Activity Index (AI) developed by Bai *et al.* is applied to the sections of accelerometer data which passed through the y-axis filter [39]. AI values are computed for each one-second epoch of the vector magnitude of the raw triaxial accelerometer data according to the following formula:

$$AI_i^{rel}(t;H) = \sqrt{\max\left(\frac{1}{3}\left\{\sum_{m=1}^{3}\frac{\sigma_{im}^2(t;H)-\bar{\sigma}_i^2}{\bar{\sigma}_i^2}\right\},0\right)}$$

Formula for computing the Activity Index of accelerometer data over time period t. σ_{im}^2 is the variance of axis *m* at data point *i*. $\bar{\sigma}_i^2$ is the

natural variance (device noise) of the measurement when the device is left sitting still (e.g. on a table).

A sliding window is used to identify continuous segments of "valid" AI values within the range of 18 – 106. This AI range acts as a secondary filter to eliminate periods of inactivity where the accelerometer is not moving (i.e. AI of 0) while also limiting AI values to those previously observed during good walking activity [39]. AI values beyond the upper limit of this range were often found to be impulse-like behavior in the accelerometer signal or motion above the frequency range of walking activity. Accelerometer tracings with AI values below 18 were typically low-level motion or noise that may be generated by, for example, unconscious bouncing of the leg.

When AI values outside of the acceptable range are encountered, a decision is made about whether to continue extending the bout based upon the percentage of valid AI content and the current length of the out-of-AI-range segment. If either the 30 second invalid AI limit or 70% minimum walking activity are violated, the bout is truncated and either discarded or returned depending upon the total duration of the bout. As a final step, the previously-described "good walking" filter used in the calibration substudy is applied to all returned bouts and those with 70% or more good walking content are retained.

Validation

Without labeled data for validation, manual inspection of accelerometer tracings was necessary to confirm the accuracy of the bout extraction process; a test sample of 100 individuals were randomly selected from the larger data set for this purpose. The previously-described pipeline was applied to each individual and the vector magnitude of the returned bout tracings were

visualized first as a full plot of the bout to confirm length and then as consecutive ten-second windows for inspecting finer details such as repetition and shape. Tracings were inspected by eye to confirm the characteristic pattern produced during walking. Each bout was assigned a designation of either “good walking” or “noise” and the final percentage of valid good walking bouts was tallied.

At least one “bout” was returned for 94 of the 100 individuals. Out of a total 1637 “bouts” returned by the algorithm, 1616 were confirmed to be true good walking bouts which gives the algorithm a false positive rate of 1.28%. This is an impressively low rate given that no intelligent systems (e.g. machine learning or human input) are involved in the bout extraction process; only simple statistics and direct value comparisons and are used to identify walking bouts. It should be noted that we cannot account for false negatives or guard against them given the lack of a ground truth label for the walking bouts. As such, expert knowledge is heavily weighted in the validation process which could introduce unintentional bias.

Examples of correct and incorrect bouts can be seen in Figure 7.1 below. Of those incorrectly identified as good walking bouts, further inspection revealed that these signals do in fact meet all of the criteria defined in the bout extraction process. The y-axis values indicate upright movement while the AI of the signal falls within the acceptable range. Additionally, the signal is consistent enough to pass the final good walking filter. It is possible that these signals might be a composite of walking and other activities which overpower the “walking” portion of the signal. However, it is more likely that these signals are non-descript motion that happens to fall within the boundaries laid out for isolating walking bouts given that the intensity of walking typically

overshadows other types of motion. It is interesting to note the vertical acceleration (y-axis) appears to be the strongest contributing factor to the overall signal in both the correct and incorrect bouts.

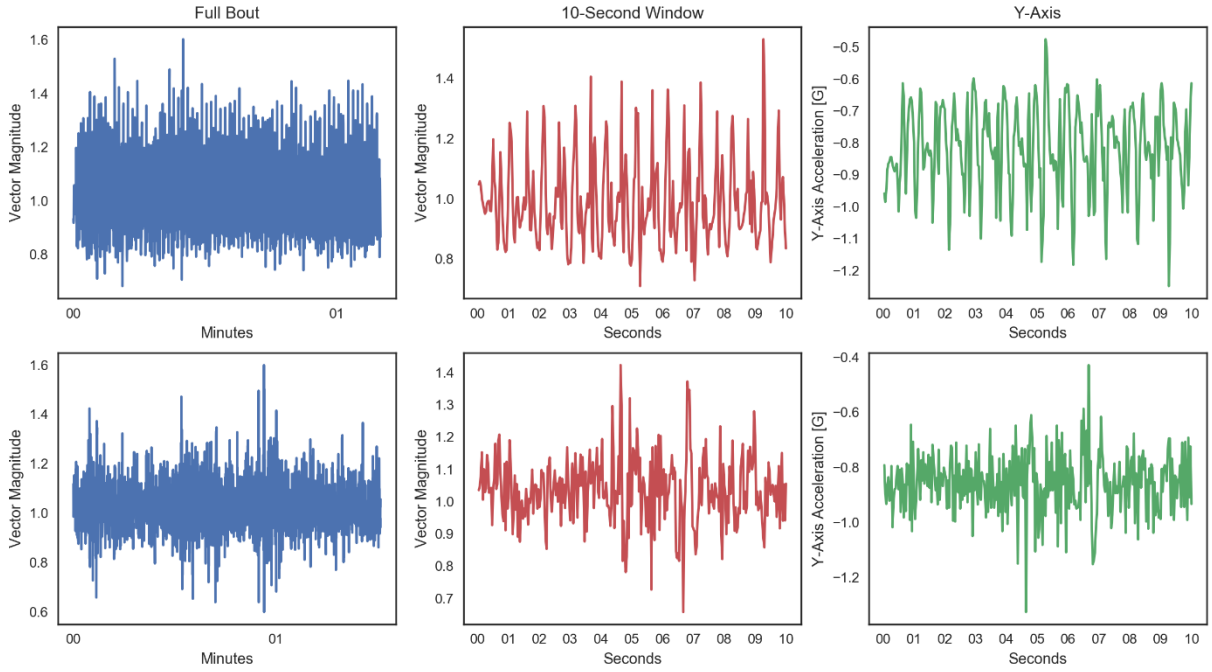

Figure 7.1: Example correctly and incorrectly identified good walking "bouts." The top row shows a true good walking bout with the characteristic repetitive shape and strong y-axis acceleration. The bottom row shows noisy motion incorrectly identified as a good walking bout. Interestingly, the y-axis component of the signal (green) appears to be the dominant motion even when compressing the signal into its vector magnitude (red).

In an effort to improve the bout extraction process and filter out these incorrectly identified bouts, an exploratory analysis of the signals was conducted. Comparison of various descriptive statistics pulled from both the correct and incorrect bouts did not reveal any properties useful for differentiating between the two signals (Figure 7.2). The feature distributions do not show a substantial mean shift which would be the easiest approach for filtering out the misidentified walking bouts. Ignoring the mean, the similar shapes across the majority of the distributions

further suggests that aggregated metrics such as descriptive statistics computed from the signal are insufficient for differentiation.

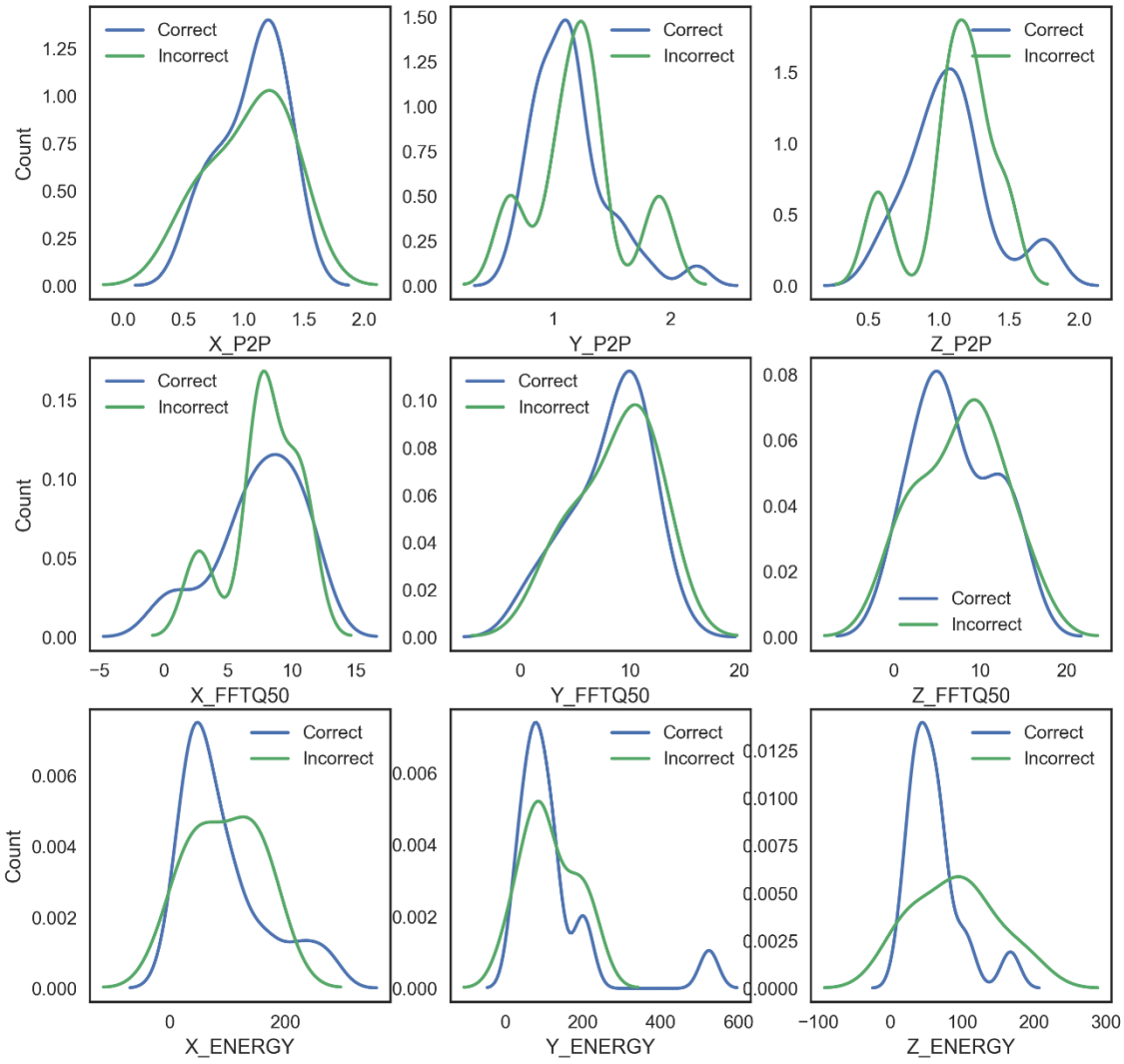

Figure 7.2: Distributions of features computed from the correctly identified good walking bouts and accelerometer tracings mistakenly identified walking bouts. All features do not show a substantial mean shift and the majority of distributions have similar shapes suggesting that these statistical measures would not be sufficient to filter out the incorrectly identified bouts.

Additional Considerations

The automated bout extraction processes exhibits strong performance with high accuracy and a low false positive rate. The similarity between the true walking bouts and the handful of incorrectly identified bouts suggests high specificity in bout identification by the pipeline. While simple statistical features did not provide a means of eliminating the incorrectly identified signals and improving the accuracy, shape-matching techniques such as discrete time warping [40] or shapelets [41] might be more effective at differentiating walking from noisy motion (not attempted). However, it is unlikely that *meaningful* improvement can be obtained without intelligent statistical methods (i.e. classification models) which would require exhaustive visual inspection and labeling of all data tracings. Alternatively, transfer learning may be especially useful in this setting if a reliable walking classification model can be found and adapted to this specific application.

Instead of changing the method of bout classification, the definition of a walking bout could be altered to change pipeline output. As previously mentioned, the one minute bout length requirement restricts walking bouts to those that best mimic walking observed during a walk test. However, this may not *necessarily* be the optimal approach. Individuals who remain in their home will rarely generate bouts of this duration, and as such would not be able to use any devices which incorporate the bout extraction pipeline. Lowering the minimum bout duration from one minute to 30 seconds could greatly increase the number of bouts for each individual. Indeed, reducing the minimum bout length to 30 seconds and the maximum pause between bouts to 10 seconds resulted in a 227% increase in number of walking bouts (from 1723 to 5653) when tested on 100 individuals; visual inspection of these bouts revealed a false positive rate of only

2.7%. A summary of the walking bouts obtained with each approach can be seen in Table C.2 and Figure C.21. Considerations must be made concerning the *quality* of these shorter bouts and the effect that they could have on model performance. It may be that these shorter bouts better-represent the typical walking behavior for an individual; they may also better capture *variation* in walking quality. However, this approach requires more robust validation than was performed in this modest spot test before comments on feasibility and effectiveness can be made.

One final consideration is the amount of pausing or stopping that is allowed during a walking bout. While walk test data should show minimal stopping during walking, we currently allow for 30% of a free-living walking bout to consist of non-walking activity considering that these individuals are likely not walking around a track. However, the physical restrictions of residing in-home challenge the reasonability of this assumption. It is probable that long-duration, in-home walking bouts consist of short periods of smooth walking punctuated by turns or pauses potentially due to the organization of the home or completion of some task. As such, placing such a strict requirement on the continuity of walking may not be possible for a large portion of the population. While reducing this requirement would lead to more noisy bouts overall, the "good walking" content of these bouts may contain a stronger signal for fall risk assessment, making this possibility an ideal investigation for future study.

CHAPTER 8: PREDICTION OF FALL RISK FROM FREE-LIVING WALKING

The full OPACH data set was used for the development and testing of free-living fall risk prediction models. Methodologies were adjusted to account for the additional variance introduced by the free-living environment and lack of scripted activities. This chapter details the complex process of curating and transforming walking bouts into the format required for proper statistical modeling.

Exclusion Criteria

Unlike individuals in the calibration substudy who were specifically instructed to perform a walk test, data gathered in the free-living environment is uncontrolled and depends upon the behavior of each individual. To maximize the likelihood of extracting good walking from free-living data and ensure reliable recording of prospective fall outcomes, a set of exclusion criteria for participants were defined as follows:

1. A minimum of six months of fall calendars were returned by the subject.

   This ensures that at least half of the available fall calendars were returned to the study and provide an accurate picture of prospective falls. This decision was informed by input from physicians and scientists familiar with this method of data collection. It should be noted that the majority of individuals (86%) returned at least 12 months of fall calendars, which is a strong return rate for survey methods of this type.

2. A minimum of 5 days of accelerometer data are available.

   Depending upon patterns of device wear, participants could have up to seven days of accelerometer data if the device was worn for the full duration of the data collection period. Given the age of the population under investigation and the suspected scarcity of good walking bouts, a minimum of 5 days of accelerometer data were required in an effort to improve the return rate of good walking bouts and hopefully guarantee a sizable sample of walking from each individual.

3. Subjects were adherent (i.e. the accelerometer was worn on the body for at least 8 hours during each of the available days).

   Subjects were able to remove the accelerometer when needed (e.g. taking a bath) and as such may not have worn the device for the full day. ActiGraph accelerometers monitor movement of the device and report whether the device was worn (i.e. recorded a sufficient amount of motion) across an eight hour time period. Accelerometers which met this criteria are assumed to have been worn during a large portion of waking hours and are labeled "adherent."

A full diagram detailing the drop-out of individuals from the study following each exclusion criteria can be seen below:

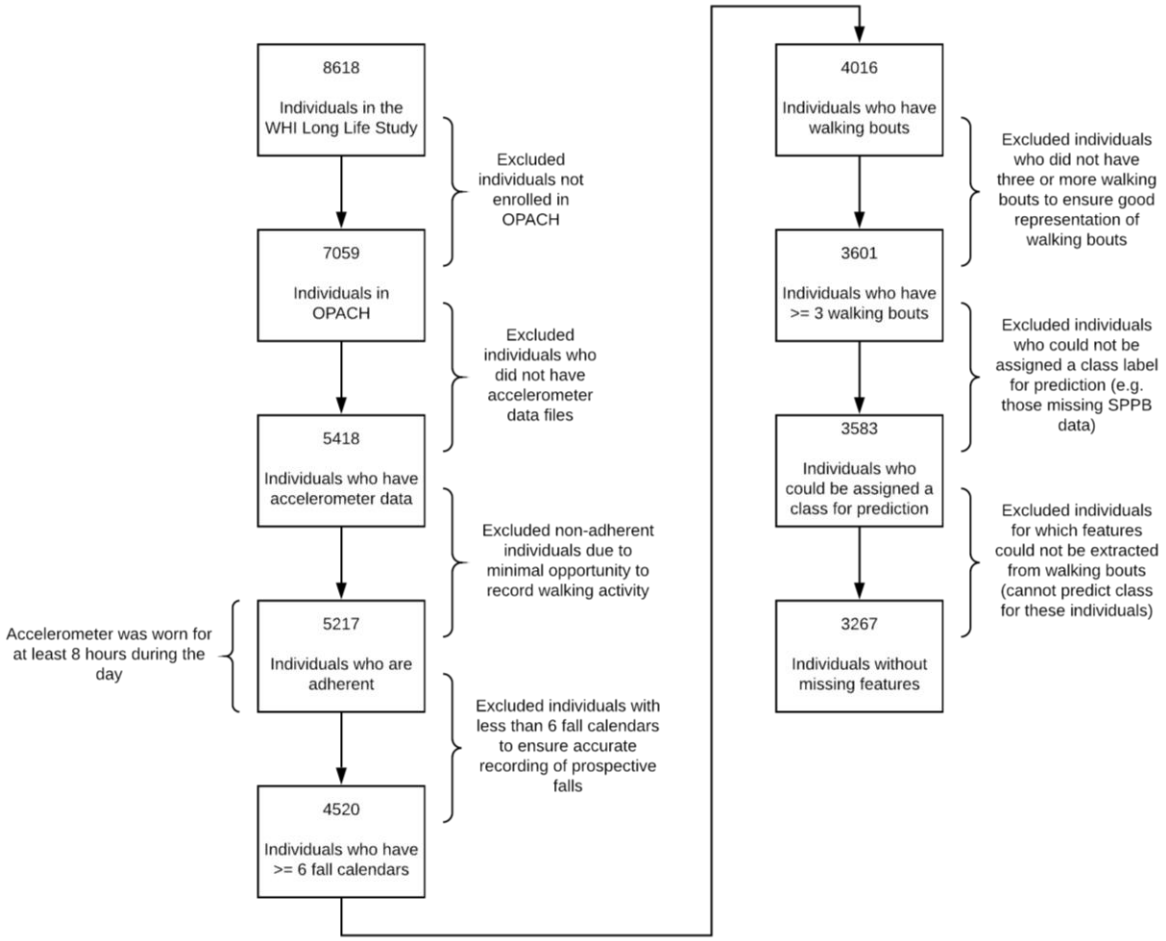

Figure 8.1: Flow chart illustrating the exclusion of individuals from the pool of data used in model development and testing. Note that 504 individuals dropped out of the participant pool for not returning a single walking bout. This does not imply that these individuals did not walk, but rather that the "good walking" bout criteria may be too strict in requiring one minute bouts.

Definition of Fall Risk

The main objective of the free-living study was to develop predictive models of *future* falls in older adults. Multiple fall risk groups were defined using fall rates calculated from the prospective fall calendars for the purpose of developing two-way and three-way classifiers. Cutoffs were applied to the fall rate variable to generate class groupings that represent the following fall risk categories:

Table 8.1: Two-way fall classification.

Fall Risk Category	Number of Future Falls
Low	0 – 1
High	2+

Table 8.2: Three-way fall classification.

Fall Risk Category	Number of Future Falls
Low	0 – 1
Medium	2 – 3
High	4+

Limitations were encountered when following the previous classification scheme and, as such, the definition of fall risk was redefined to more resemble that of the calibration substudy which better-captures the physical function component associated with fall risk. A combination of falls and SPPB scores were used to define the following new risk categories:

Table 8.3: Two-way fall risk classification.

Fall Risk Category	Future Falls	SPPB Range	Fall Rate
Low	0	10 – 12	> 0
High	> 0	0 – 6	0

Table 8.4: Three-way fall risk classification.

Fall Risk Category	Future Falls	SPPB Range	Fall Rate
Low	0	10 – 12	0
Medium	> 0	7 – 9	> 0
High	> 0	0 – 6	> 0

Bout Extraction

Good walking bouts were extracted using the pipeline detailed previously in Chapter 7.

Feature Extraction

Due to the substantial variation in bout duration and number, features extracted from the free-living walking bouts were computed across the full bout without the use of a sliding window. If individuals had multiple bouts, features were computed for each walking bout and then averaged

to generate the final feature vector. It should be noted that this process loses information about the variation of walking for individuals who returned multiple bouts and instead provides summarized measures of walking. A full list of computed features can be found in Appendix A.

Predictive Modeling

Random forests of 1000 trees were trained to classify individuals according to each of the classification schemes defined above using a 50/50 train-test split. Splits were stratified by class to ensure equal representation of each category in the training and testing sets. Confusion matrices were used for evaluating predictive performance since most of the metrics previously used for model evaluation in the calibration do not generalize well to the multiclass case. Moreover, confusion matrices provide additional insight into the specific weaknesses in prediction that can be easily missed or misunderstood when looking at summary measures such as accuracy.

Results

Out of the original 4520 individuals in the data set, 4016 returned at least one good walking bout. To ensure a representative walking sample from each individual, a minimum of three walking bouts per individual was required to remain in the analysis; this reduced the sample size to 3583. The class distributions and confusion matrices for individuals in each classification scheme (fall risk, future falls, and SPPB) are presented and discussed in the subsections below. Demographics are presented in Table 8.5 at the end of this chapter.

Falls Prediction (two-way)

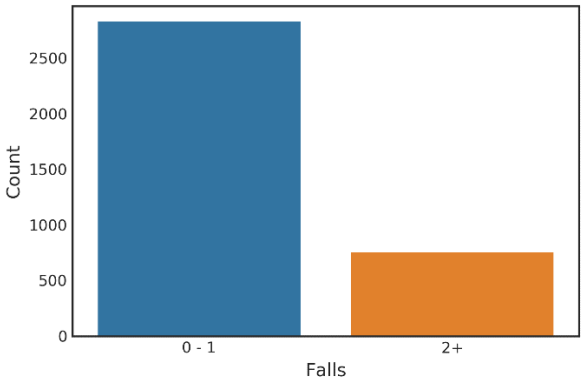

Figure 8.2.A: Falls class distribution for the full sample. The low-risk category (n=2828) corresponding to 0-1 falls has representation roughly four times that of the minority class of 2+ falls (n=754).

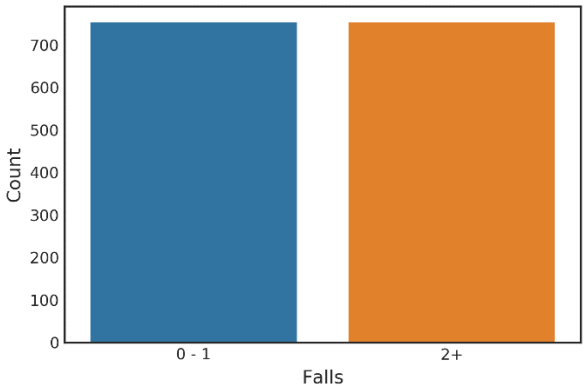

Figure 8.2.B: Falls class distribution for a balanced sample generated by under sampling the majority class. Both the low-risk category (n=754) and high-risk categories (n=754) have an equal representation of roughly 700 instances.

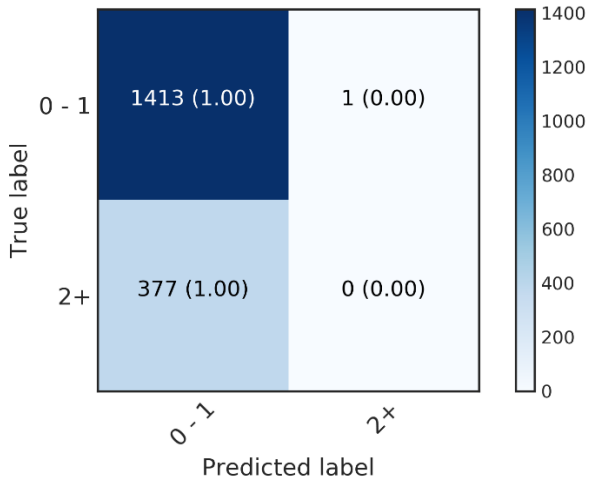

Figure 8.3.A: Confusion matrix for prediction of fall rate categories in the unbalanced testing set. All individuals were predicted as being in the low-risk category of zero or one falls.

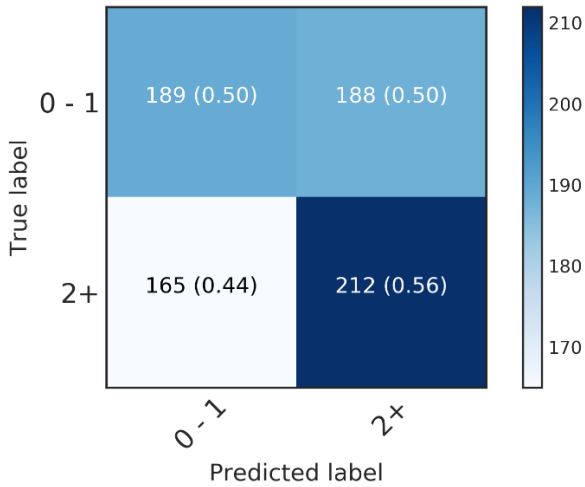

Figure 8.3.B: Confusion matrix for prediction of fall rate categories in the balanced testing set.

The results of the binary classifications of “0 – 1” falls and 2+ falls are in Figure 8.3. The binary classifier of prospective falls seemed to have high accuracy (78.76%). However, the classifier predicted every participant as “0 – 1” falls. Accordingly, the precision and sensitivity of 2+ falls were both zero. The confusion matrix shows that this bias remained even when testing on unseen individuals with all “2+ falls” cases being incorrectly classified. It is known that class imbalance

can prejudice data-driven models toward predicting the majority class. For ensemble models that utilize sampling approaches—such as bootstrapping—to train a set of smaller learners, excessive sampling of the majority class will produce a set of highly similar learners which favor prediction of the majority. This effectively negates the sought-after benefits offered by ensemble approaches.

Using a balanced data set improved the overall performance of the classifier, but the performance was still poor. Random forests trained on the balanced training set displayed a more even prediction accuracy across both classes. Figure 8.3B shows 50% test accuracy for "0 – 1" falls and 56% accuracy for 2+ falls which suggests that balancing class frequency allows for not only prediction of both classes, but actually leads to better identification of the minority class. However, this improvement in recall of 2+ falls (from 0.00 to 0.53) came at the cost of reducing overall accuracy to near random chance (only 53.00%).

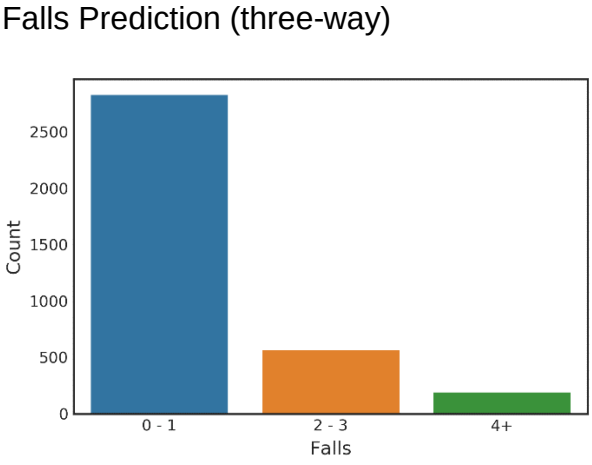

Figure 8.4.A: Falls class distribution for the full sample. The low-risk category corresponding to "0-1 falls" (n=2828) has representation roughly four times that of the middle class of "2-3 falls" (n=564) and eleven times that of the minority group with "4+ falls" (n=190).

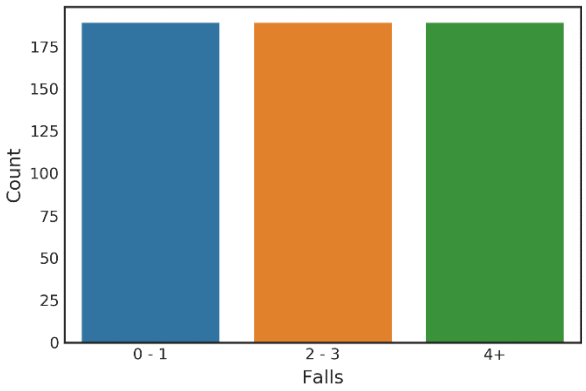

Figure 8.4.B: Falls class distribution for a balanced sample generated by under sampling the majority classes. As a result, all three classes have equal representation with 190 individuals.

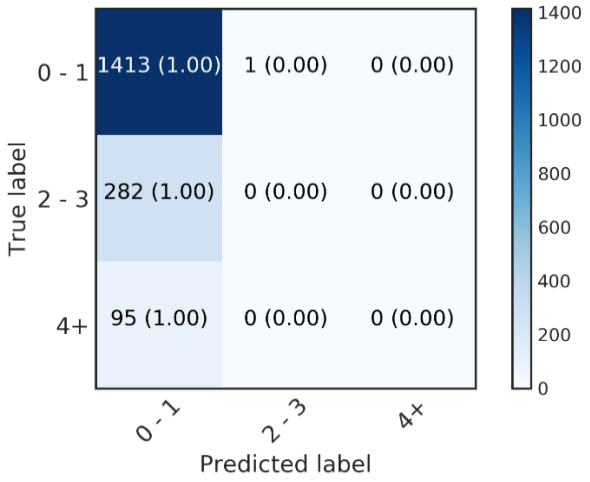

Figure 8.5.A: Confusion matrix for prediction of fall rate categories in the unbalanced testing set. Nearly all individuals were predicted as being in the low-risk category of "0 – 1 falls" with no individuals assigned to the "4+ falls" category and only a single individual predicted as having "2 – 3 falls."

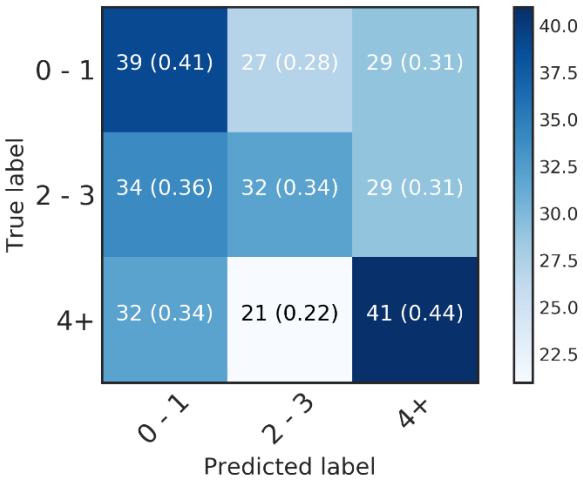

Figure 8.5.B: Confusion matrix for prediction of fall rate categories in the balanced testing set. We see individuals assigned to all classes with the greatest percentage of individuals assigned to the "0 – 1 falls" and "4+ falls" groups.

The results of the ternary classifications of "0 – 1", "2 – 3", and 4+ falls are in Figure 8.5 above. Like the binary classifier, the accuracy was deceptively high at 79.39%. However, nearly all participants were predicted to have had "0 – 1" falls with a small portion predicted to have had "2 – 3" falls. The precision and sensitivity of the were both 0.00 for the 4+ falls group and were marginally greater for the "2 – 3" falls group (0.15 and 0.02, respectively). Balancing the data set improved precision and sensitivity of the "2 – 3" (0.34 and 0.33, respectively) and 4+ falls (0.42 and 0.4, respectively) group, but classification accuracy was drastically reduced to 35.9%.

Fall Risk Prediction (two-way)

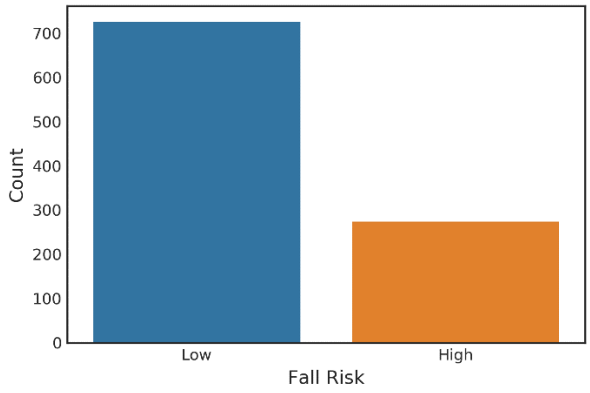

Figure 8.6.A: Fall risk class distribution for the full sample. The low-risk category (n=726) has a representation more than two times that of the high risk class (n=274).

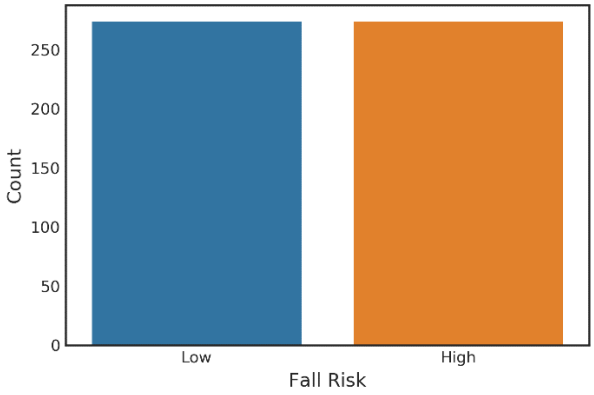

Figure 8.6.B: Fall risk class distribution for a balanced sample generated by under sampling the majority classes. As a result, both classes have equal representation with 274 individuals.

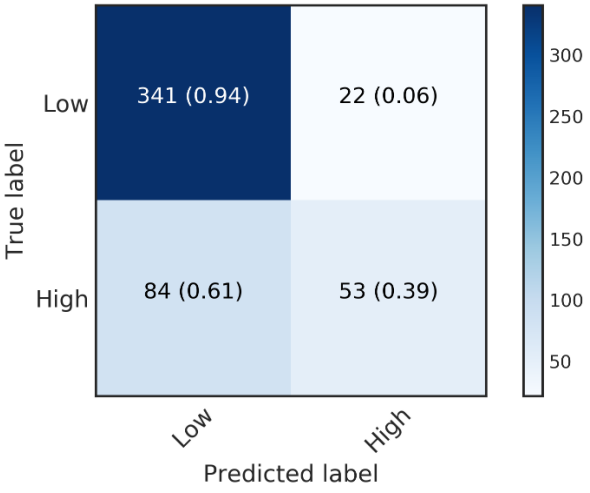

Figure 8.7.A: Confusion matrix for prediction of fall risk categories in the unbalanced testing set. The majority of individuals were predicted as having membership in the "low risk" category.

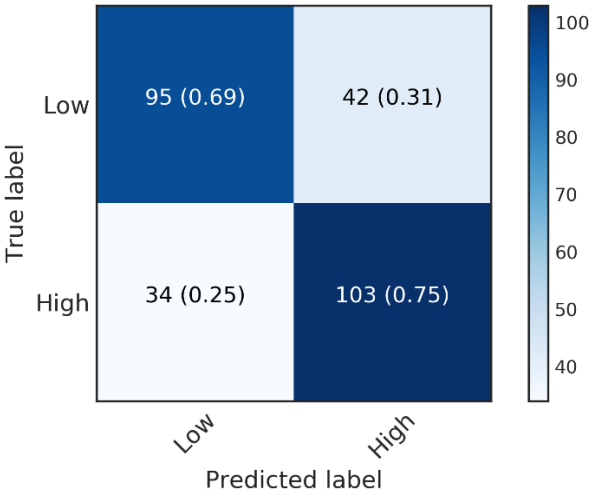

Figure 8.7.B: Confusion matrix for prediction of fall risk in the balanced testing set. We see a fairly balanced prediction accuracy for both classes with 69% of individuals correctly classified as "low risk" and 75% of individuals correctly classified as "high risk."

The results of the binary classifications of "Low" and "High" fall risk are in Figure 8.7. The binary classifier of fall risk seemed to have decent accuracy (78.80%). However, the classifier predicted the majority of participant as low risk. Like prediction of falls, the use of a balanced data set improved the overall performance of the classifier, but performance was still poor. The overall accuracy of the classifier was reduced to 72.26% with a reduction in sensitivity to the

"low" risk category (0.8 to 0.74). Interestingly, the sensitivity to the "high" risk category remained at the same value (0.71) suggesting that, in this case, the balanced data set did not improve minority class prediction, but did hurt "low" risk classification.

Fall Risk Prediction (three-way)

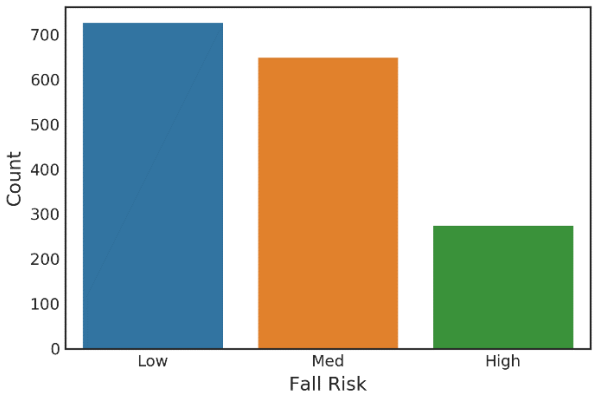

Figure 8.8.A: Fall risk class distribution for the full sample using a three-way grouping scheme. The low-risk category (n=726) has a representation pretty much equivalent to that of the middle category (n=648). However, the high risk category (n=274) is represented less than half as much as the other two classes.

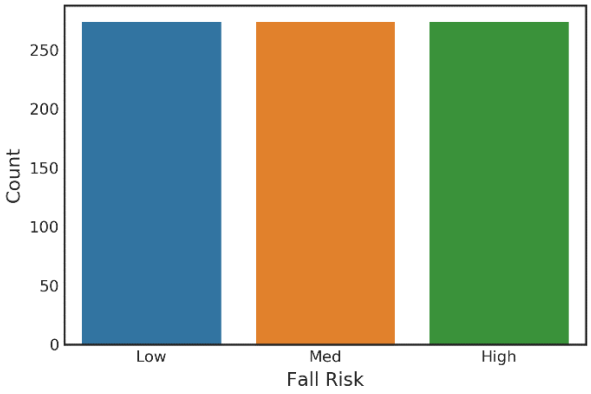

Figure 8.8.B: Fall risk class distribution for a balanced sample generated by under sampling the majority classes. As a result, all three classes have equal representation with 274 individuals.

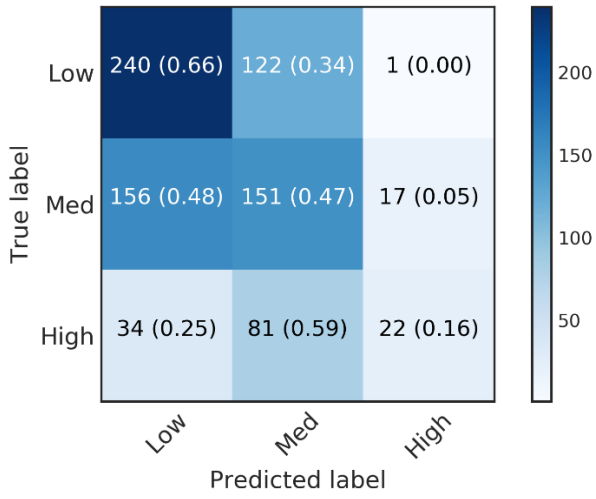

Figure 8.9.A: Confusion matrix for prediction of fall risk categories in the unbalanced testing set. The majority of individuals were predicted as having membership in the "low risk" category with some predictions in each of the three classes.

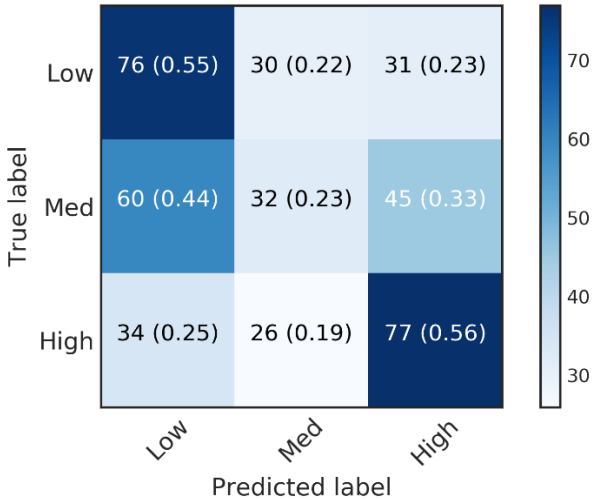

Figure 8.9.B: Confusion matrix for prediction of fall risk in the balanced testing set. We see a fairly balanced, but poor, prediction accuracy in the extreme classes of "low" and "high" risk. The middle category shows poor prediction accuracy with individuals placed in all three categories.

The results of the three-way classifications of “Low,” “Medium,” and “High” fall risk are in Figure 8.9. Unlike the binary case, the ternary classifier of fall risk exhibited very poor accuracy (50%). The balanced data set substantially improved prediction of the high risk individuals (55% correct) with only a small reduction in correct classification of low risk individuals (dropped from 66% to 55%). Medium risk individuals saw a 50% reduction in accurate classification with the majority of individuals being incorrectly classified in the extremes of low or high risk.

SPPB Prediction (two-way)

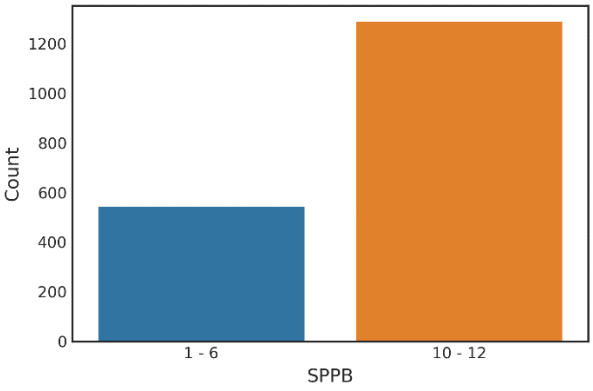

Figure 8.10.A: SPPB class distribution for the full sample using a two-way grouping scheme. The low-risk category of 10 – 12 (n=1290) has a representation more than double that of the high-risk class with SPPB scores of 1 – 6 (n=544).

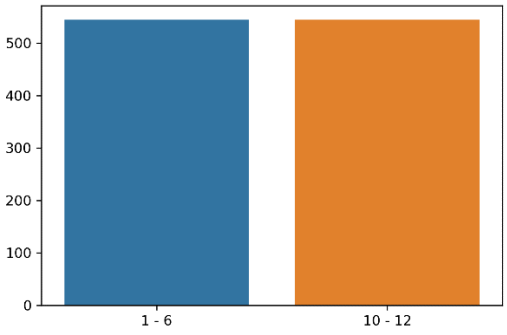

Figure 8.10.B: Fall risk class distribution for a balanced sample generated by under sampling the majority class. As a result, both SPPB categories have an equal representation of 544 individuals.

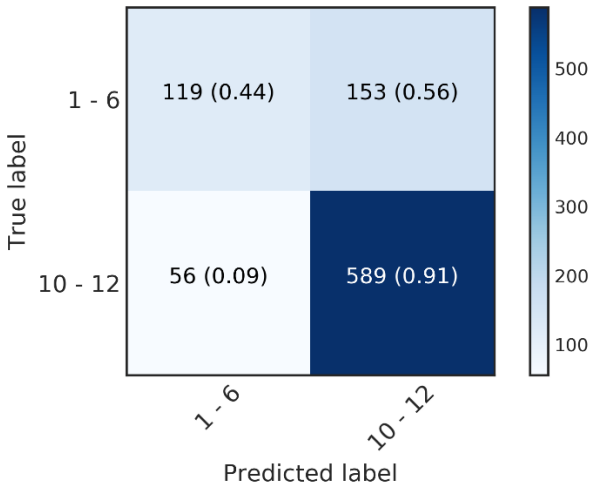

Figure 8.11.A: Confusion matrix for prediction of SPPB categories in the unbalanced testing set. The majority of individuals were predicted as having membership in the low-risk category of "10 – 12" with some correct predictions in the high-risk group.

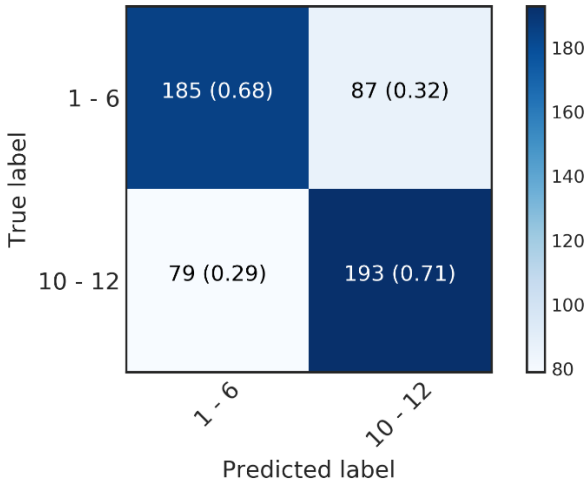

Figure 8.11.B: Confusion matrix for prediction of SPPB categories in the balanced testing set. We see a fairly balanced prediction accuracy across both classes with roughly 70% accuracy in both categories.

The results of the binary classifications of low and high physical function (SPPB) are in Figure 8.11. Performance of the binary classification of the unbalanced (complete) data set was poor in comparison to the balanced data set. The overall accuracy of the unbalanced classifier was greater (77.20%) in comparison to the balanced classifier (69.49%). However, this was accomplished by predicting most participants as high physical function (10-12 SPPB) which made up most of the sample (70.3%). After balancing the groups, sensitivity of the high physical function group improved from 0.68 to 0.70 while the sensitivity of the low physical function group decreased from 0.79 to 0.69.

SPPB Prediction (three-way)

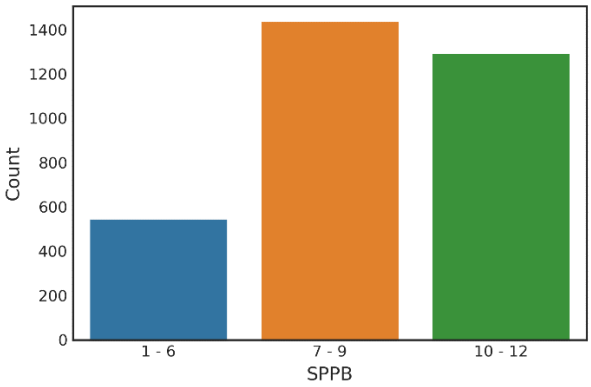

Figure 8.12.A: SPPB class distribution for the full sample using a three-way grouping scheme. The low-risk category of "10 – 12" (n=1290) has a representation fairly even with the middle category of "7 – 0" (n=1434) but is much more represented than the high-risk category of "1 – 6" (n=544).

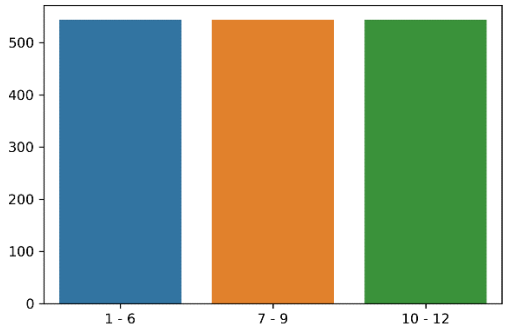

Figure 8.12.B: Fall risk class distribution for a balanced sample generated by under sampling the majority classes. As a result, all three SPPB categories have an equal representation of 544 individuals.

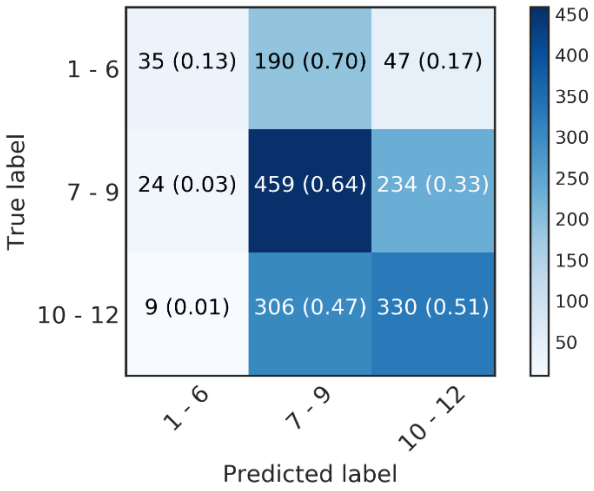

Figure 8.13.A: Confusion matrix for prediction of SPPB categories in the unbalanced testing set. The majority of individuals were predicted as having membership in the middle category of "7 – 9" with some predictions in each of the three classes.

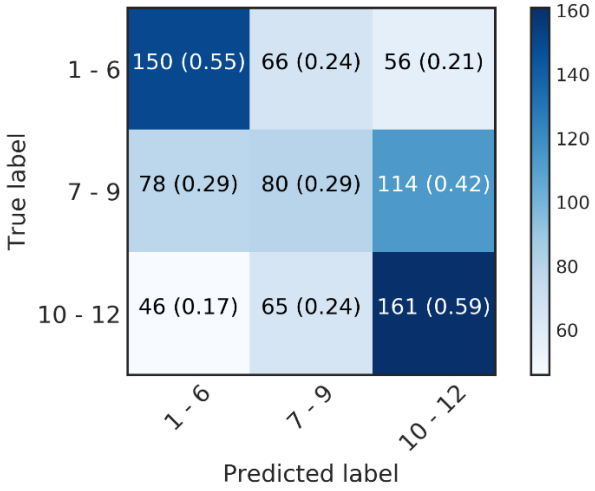

Figure 8.13.B: Confusion matrix for prediction of SPPB categories in the balanced testing set. We see a fairly balanced, but poor, prediction accuracy in the extreme classes of "1 – 6" and "10 – 12" risk. The middle category shows poor prediction accuracy with individuals placed in all three categories.

The results of the ternary classifications of low, medium, and high physical function (SPPB) are in Figure 8.13. The overall accuracies were poor (51.17% for the unbalanced data set and 43.30% for the balanced data set). Balancing the data set did not improve overall performance

with no clear improvement in precision or recall. Sensitivity for low physical function increased slightly (0.51 to 0.55) with the balanced data set while the recall for medium physical function decreased (0.48 to 0.38) and the recall for high physical function increased (0.55 to 0.49).

Table 8.5: Participant characteristics by SPPB category.

Characteristic	***Total***	***High Risk (SPPB 1 – 6)***	***Medium Risk (SPPB 7 – 9)***	***Low Risk (SPPB 10 – 12)***
N (%)	1632 (100)	544 (33.33)	544 (33.33)	544 (33.33)
Age, years, mean (SD)	78.60 (6.62)	80.89 (6.21)	78.28 (6.39)	76.76 (6.35)
BMI, mean (SD)	27.71 (5.36)	28.25 (5.81)	27.83 (5.34)	27.09 (5.14)
EPESE SPPB Score, mean (SD)	7.98 (2.60)	4.97 (1.24)	8.09 (0.81)	10.86 (0.81)
Balance Subscore, mean (SD)	3.30 (1.18)	2.28 (1.37)	3.67 (0.80)	3.96 (0.26)
Chair stand Subscore, mean (SD)	2.12 (1.21)	1.19 (0.87)	1.95 (0.97)	3.25 (0.75)
Gait Subscore, mean (SD)	2.56 (1.21)	1.50 (0.84)	2.48 (1.02)	3.66 (0.60)
Fall rate, mean (SD)	0.08 (0.21)	0.11 (0.30)	0.08 (0.17)	0.06 (0.11)
Fall count, mean (SD)	1.01 (2.35)	1.37 (3.48)	0.94 (2.01)	0.80 (1.36)
Amount of Activity				
Percent Activity	0.09 (0.03)	0.09 (0.03)	0.09 (0.03)	0.09 (0.03)
Ethnicity				
White (%)	796 (48.77)	304 (18.63)	240 (14.71)	252 (15.44)
Black (%)	551 (33.76)	165 (10.11)	212 (12.99)	174 (10.66)
Hispanic (%)	285 (17.46)	75 (4.60)	92 (5.64)	118 (7.23)
Education				
No School (%)	0 (0)	0 (0)	0 (0)	0 (0)
1-4 Years (%)	3 (0.18)	2 (0.12)	1 (0.06)	0 (0.00)
4-8 Years (%)	13 (0.80)	6 (0.37)	5 (0.31)	2 (0.12)
9-12 Years (%)	53 (3.25)	20 (1.23)	22 (1.35)	11 (0.67)
High School (%)	236 (14.46)	86 (5.27)	70 (4.29)	80 (4.90)
Vocational School (%)	166 (10.17)	63 (3.86)	58 (3.55)	45 (2.76)
Some College or Associates (%)	449 (27.51)	156 (9.56)	143 (8.76)	150 (9.19)
College Graduate (%)	169 (10.36)	43 (2.63)	68 (4.17)	58 (3.55)
Some Postgraduate or Professional School (%)	196 (12.01)	62 (3.80)	58 (3.55)	76 (4.66)
Masters (%)	303 (18.57)	91 (5.58)	106 (6.50)	106 (6.50)
Doctoral (%)	38 (2.33)	13 (0.80)	11 (0.67)	14 (0.86)
Use of Assistive Device				
Never (%)	1195 (73.22)	293 (17.95)	418 (25.61)	484 (29.66)
Occasionally (%)	261 (15.99)	124 (7.60)	89 (5.45)	48 (2.94)
Frequently or All the Time (%)	149 (9.13)	114 (6.99)	29 (1.78)	6 (0.37)
Difficulty Walking at Baseline	???	???	???	???
CHAMPS				
Expenditure All, mean (SD)	1819.36 (1494.78)	1525.52 (1294.89)	1875.79 (1695.84)	2181.40 (1652.68)

Table 8.5: Continued.

Expenditure Moderate, mean (SD)	*891.53 (1073.81)*	*662.52 (890.13)*	*918.14 (1235.98)*	*1169.77 (1237.93)*
Frequency All, mean (SD)	14.36 (12.94)	12.44 (12.50)	14.13 (12.48)	15.84 (12.89)
Frequency Moderate, mean (SD)	4.79 (5.97)	3.69 (5.60)	4.56 (5.63)	5.69 (6.20)
Characteristic	Total	High Risk (SPPB 1 – 6)	Medium Risk (SPPB 7 – 9)	Low Risk(SPPB 10 – 12)
N (%)	1632 (100)	544 (33.33)	544 (33.33)	544 (33.33)
Age, years, mean (SD)	78.60 (6.62)	80.89 (6.21)	78.28 (6.39)	76.76 (6.35)
BMI, mean (SD)	27.71 (5.36)	28.25 (5.81)	27.83 (5.34)	27.09 (5.14)
EPESE SPPB Score, mean (SD)	7.98 (2.60)	4.97 (1.24)	8.09 (0.81)	10.86 (0.81)
Balance Subscore, mean (SD)	3.30 (1.18)	2.28 (1.37)	3.67 (0.80)	3.96 (0.26)
Chair stand Subscore, mean (SD)	2.12 (1.21)	1.19 (0.87)	1.95 (0.97)	3.25 (0.75)
Gait Subscore, mean (SD)	2.56 (1.21)	1.50 (0.84)	2.48 (1.02)	3.66 (0.60)
Fall rate, mean (SD)	0.08 (0.21)	0.11 (0.30)	0.08 (0.17)	0.06 (0.11)
Fall count, mean (SD)	1.01 (2.35)	1.37 (3.48)	0.94 (2.01)	0.80 (1.36)
Amount of Activity				
Percent Activity	0.09 (0.03)	0.09 (0.03)	0.09 (0.03)	0.09 (0.03)
Ethnicity				
White (%)	796 (48.77)	304 (18.63)	240 (14.71)	252 (15.44)
Black (%)	551 (33.76)	165 (10.11)	212 (12.99)	174 (10.66)
Hispanic (%)	285 (17.46)	75 (4.60)	92 (5.64)	118 (7.23)
Education				
No School (%)	0 (0)	0 (0)	0 (0)	0 (0)
1-4 Years (%)	3 (0.18)	2 (0.12)	1 (0.06)	0 (0.00)
4-8 Years (%)	13 (0.80)	6 (0.37)	5 (0.31)	2 (0.12)
9-12 Years (%)	53 (3.25)	20 (1.23)	22 (1.35)	11 (0.67)
High School (%)	236 (14.46)	86 (5.27)	70 (4.29)	80 (4.90)
Vocational School (%)	166 (10.17)	63 (3.86)	58 (3.55)	45 (2.76)
Some College or Associates (%)	449 (27.51)	156 (9.56)	143 (8.76)	150 (9.19)
College Graduate (%)	169 (10.36)	43 (2.63)	68 (4.17)	58 (3.55)
Some Postgraduate or Professional School (%)	196 (12.01)	62 (3.80)	58 (3.55)	76 (4.66)
Masters (%)	303 (18.57)	91 (5.58)	106 (6.50)	106 (6.50)
Doctoral (%)	38 (2.33)	13 (0.80)	11 (0.67)	14 (0.86)
Use of Assistive Device				
Never (%)	1195 (73.22)	293 (17.95)	418 (25.61)	484 (29.66)
Occasionally (%)	261 (15.99)	124 (7.60)	89 (5.45)	48 (2.94)
Frequently or All the Time (%)	149 (9.13)	114 (6.99)	29 (1.78)	6 (0.37)
CHAMPS				
Expenditure All, mean (SD)	1819.36 (1494.78)	1525.52 (1294.89)	1875.79 (1695.84)	2181.40 (1652.68)
Expenditure Moderate, mean (SD)	891.53 (1073.81)	662.52 (890.13)	918.14 (1235.98)	1169.77 (1237.93)

Table 8.5: Continued.

Freq. All, mean (SD)	*14.36 (12.94)*	*12.44 (12.50)*	*14.13 (12.48)*	*15.84 (12.89)*
Freq. Moderate, mean (SD)	4.79 (5.97)	3.69 (5.60)	4.56 (5.63)	5.69 (6.20)

Insights from Analysis of Free-living Gait

The performance of predictive models developed on the free-living data further support the idea that accelerometer-based measures of gait are potentially useful in screening older women for fall risk, but that the performance of these models is heavily weighted in the specific details of model development. To start, the use of free-living data introduces a substantial amount of variation and unpredictability into the pool of available data which must be filtered and reduced to only a sample of good walking bouts. This process requires the computation of Activity Index values for the full accelerometer tracing followed by segmentation into candidate subsegments which must then pass through a series of filters to identify good walking bouts. While this process was found to be very computationally intensive using early versions of the data processing software, current runtime estimates suggest that a *week of data* for a single individual (typical size ~523 megabytes) can be processed in twenty seconds on average when run on a 64-bit system with an Intel Core i7-4610M processor at 3.00 GHz and 8 GB of RAM which represents a modest laptop by 2018 standards; the full set of 4520 individuals was processed in under two days' time. Estimates of feature extraction runtime are not currently available, but are expected to be modest given that the majority of features computed are simple statistics which make use of heavily-optimized numpy libraries.

Looking at the overall prediction performance for each of the three target variables (falls, fall risk, and SPPB), models trained on unbalanced data show higher accuracies than those trained on

balanced data sets. However, this picture was found to be misleading when looking at the confusion matrices. The high accuracy of these binary classification models (falls 78%, risk 78%, and SPPB 77%) are inflated due to the class imbalance present in the data set. Specifically, the falls classification model achieves its 78% accuracy by predicting "0 – 1 falls" for all individuals (i.e. 100% of the time). Since individuals who fell 0 – 1 times make up nearly 80% of the data set, the classifier automatically achieves this high level of accuracy by blindly assigning a single class. A similar, albeit less extreme, case can be seen for the prediction of binary SPPB categories. The classifier achieved 77% accuracy by predicting the majority of individuals as having SPPB scores of 10 – 12. However, individuals with SPPB scores of 10 – 12 make up roughly 70% of the data set. While in this scenario the classifier is not *completely* ignoring the minority class, a substantial bias in prediction can be seen. Fall risk predictions follow a trend similar to SPPB predictions and provide further evidence of the significant effect class imbalance can have in biasing statistical models. As an interesting side note, Table 8.5 shows that the "Gait Subscore" has good separation across the three SPPB classes which, in theory, should be visible in the accelerometer data. Given the classification accuracy, however, this is clearly not the case. It may be that the location of the sensor inhibits acquisition of the aspects of the gait cycle which allow for separation across the classes. Alternatively, assuming the accelerometer data are indeed capture the appropriate signal, the features extracted may not be capturing these components.

Contrary to the previous discussion, statistical models trained on balanced binary data showed lower overall accuracy but displayed an improved sensitivity to the minority class (high risk individuals). Models predicting future falls showed the greatest drop in performance when switching from unbalanced to balanced data with overall accuracy dropping from 78% to just

53%. Being just above random chance, models with this accuracy would usually be ignored. However, looking more closely we see that this reduction in overall accuracy has led to a substantial increase in sensitivity for predicting individuals in the 2+ falls group (sensitivity increased from 0 to 0.53). While this model appears to be less accurate according to summary measures, it is in fact a much better model since it actually identifies individuals in the high falls group. Fall risk predictions using the balanced data sets showed a much smaller drop in accuracy from 78% to 72%, but did not show any increase in sensitivity to the minority class (sensitivity 0.71). Surprisingly, the balanced model is likely the worse of the two given the decrease in accuracy of low-risk predictions (only 69%) for essentially no gain in detecting high-risk individuals.

Looking at the performance of the three-way classifiers, we uncover an interesting property of multi-class classification. First, like the binary classifiers, training on unbalanced data sets leads to strong bias toward predicting the majority class (e.g. solely assigning the majority class produces 80% accuracy for the ternary fall prediction model, see Figure 8.5). However, when trained on balanced data, we see an interesting shift. Rather than averaging out performance across all three classes as we might expect given their equal representation in the data, all of the models (falls, fall risk, and SPPB) show biased predictions toward the extreme classes. In greater detail, fall prediction models obtained their best accuracies in predicting the extremes "0 – 1 falls" and "4+ falls" with accuracies of 41% and 44%, respectively; all other predictions showed accuracies of 36% or less (see Figure 8.5). A similar pattern is visible in the three-way models predicting fall risk and SPPB categories with the extreme cases (risk: "Low" and "High", SPPB: "1 – 6" and "10 – 12") showing the greatest level of accuracy despite all three classes having

equal representation. At the simplest level of explanation, these patterns illustrate that it is easier to separate instances the greater they are different. We also see that individuals in the middle classes are easily confused as being members of any of the three possible classes. These behaviors may highlight the large degree of similarity in the good walking signals and the multitude of ways to achieve a mid-level SPPB score (i.e. many combinations of gait, chair stand, and balance scores can generate a mid-level total SPPB score). Alternatively, the choices made in defining class labels could have selected a set of alike individuals for which the differences in accelerometer data are minute and easy to misidentify. However, it may simply be the case that the classification problems posed in this investigation are challenging and require more advanced approaches to achieve higher degrees of sensitivity and precision.

Focusing solely on the binary classification models trained on balanced data, we can begin to infer the effect of class labels on performance and the relative usefulness of model predictions. First, we see that the best accuracy is obtained when predicting SPPB or fall risk (70% and 72%, respectively) whereas prediction of falls is no more accurate than random chance (i.e. 50% accuracy). Looking at the definitions of class labels, both SPPB and fall risk incorporate SPPB scores (see Appendix B) whereas prediction of falls is based solely on the number of future falls. It may be that the SPPB and fall risk labels are more directly related to the signal in the accelerometer data than are individual falls. This makes some intuitive sense since SPPB is a measure of physical function which incorporates a walk test and tests of balance; the movement of the accelerometer would naturally be more associated with class labels which incorporate SPPB in their definition (i.e. SPPB and fall risk) than those that do not (i.e. falls). The lower correlations between accelerometer features and fall count (Figure C.9), compared to the higher

correlations between SPPB score and these same features, provide additional evidence in support of this rationale. Moreover, falls themselves are not necessarily a result of physiologic deficiency but also include falls due to environmental effects or random chance. Such falls would have little to no association with physical function and, by association, the mediolateral movement of the accelerometer which is related to gait and balance. And, of course, individuals who walk more often increase their opportunity, but not necessarily their risk, of falling which may further cloud the relationship between falls and the accelerometer data.

With this in mind, it is reasonable to believe that accelerometer-based systems for monitoring falls will provide a more accurate measure of an individual's *risk* for future falls than the prediction of the falls *themselves*. Although this may be less ideal than the reverse case, the automated identification of changes in SPPB or fall risk status could allow for the timely application of proactive strategies, such as strength and balance training, to prevent falls. However, separate longitudinal studies would be required to confirm the performance of such a system. The accuracy of the models presented here—while impressive for the complexity of the task—will need to be improved before they can be used in any official capacity for population-level monitoring. Moreover, the models will need to be adjusted to incorporate classification across all possible categories (i.e. the full SPPB score range of 1 – 12) instead of just the extremes since the real value of these predictions are in identifying individuals who are in the middle categories of risk and are beginning to transition into high risk; we need to detect the start of this transition and attempt to mitigate the increase in risk to have a meaningful impact in reducing the burden of falls.

CHAPTER 9: CONCLUSIONS

The performance of predictive models from the calibration substudy suggests that raw data collected from a hip-worn, triaxial accelerometer during walking may be useful in assessing risk of falls. Prospective studies of the ability of accelerometer-based measures of walking to predict falls are warranted, given the potential of these inexpensive sensors to monitor walking and fall risk during activities of daily life, in large numbers of older adults, and over long periods of time. In particular, large prospective studies in older adults who vary widely in risk of falls are needed. In these studies, accelerometer-based assessments of fall risk should be based upon patterns of walking under free-living conditions, rather than only on data collected in laboratory or clinical settings. Analysis of data collected in free-living conditions may identify different gait characteristics as indicators of fall risk, in part because free-living walking occurs in a variety of environments (e.g. hills, uneven sidewalks, and wet surfaces).

Accelerometer-based fall risk models developed from the free-living OPACH data showed reduced accuracy compared to those in the calibration substudy, but maintained an accuracy rate of 70% for low vs. high fall risk classification (same categories used in the calibration substudy); this is impressive given the increased noise in the free-living data compared to the scripted and carefully timed walk test used in the calibration substudy. Although more work is required before passive screening of large populations for fall risk via smart phones becomes a feasible tool, the least expensive low-end smartphones (e.g. the LG Optimus Zone 3 which now costs $30) can measure gait as accurately as the most expensive high-end medical accelerometers (which cost $3000), while also being more accurate than fitness devices [37]. Such phones contain accelerometers of far better quality than those used in the ActiGraph and are naturally

placed near the waist (e.g. pocket or belt) which facilitates conversion of ActiGraph prediction models to work with phone data. Thus, there is a potential path toward screening and prevention of falls at population scale for the aging population, by leveraging sensor data from already carried personal phones.

CHAPTER 10: LIMITATIONS

Of course, these investigations have several limitations both in a general sense and related to specific aspects of the study. First, these investigations assume that fall risk manifests as a consistent signal in gait and that it can be measured using a hip-mounted triaxial accelerometer. For the calibration substudy, an additional assumption is made that gait was stable between the measurement of SPPB in 2012-13 and data collection of the calibration substudy up to many months later. Second, as in other laboratory studies of gait and fall risk, women may alter their gait in laboratory conditions under investigator observation [36]. Third, the calibration substudy has a small sample size with uneven numbers of women in the two risk groups. Furthermore, we did not attempt to classify fall risk in all women but only in women at upper and lower ends of risk. Fourth, the study used data on past falls rather than prospectively collected information after the calibration substudy. Consequently, older adults who have fallen in the past year might change due to fear of falling and cautious ambulation. Fifth, machine learning methods are susceptible to overfitting prediction models, though the cross-validation method of this study, combined with both the tree bagging and feature bagging used in random forests, is less likely to have overfit than the base method of using a single decision tree. Finally, because women were screened for ability to walk on a treadmill, the sample excluded women at highest fall risk for whom treadmill walking is unsafe. For example, the sample did not include any women with 4+ falls in the past year. In a study where the "high risk" group includes frequent fallers, classification accuracy might be improved and, possibly, different features or additional features could be included in predictive models.

Limitations of the free-living gait investigation overlap with those of the calibration substudy, but also posit many unique challenges due to a real-world settings. To start, like the calibration substudy, the free-living study assumes that fall risk manifests as a regular signal which can be detected using a hip-mounted accelerometer. A number of sensor locations have been used in studies of wearable sensors and fall risk with many using multiple sensors [12]. The 30 Hz sampling rate of the ActiGraph accelerometers used in this study, combined with the placement of only a single sensor at the hip, prevented full segmentation of the gait cycle and may have limited the quality of the data gathered. While 30 Hz is more than sufficient to capture walking motion, increasing the sampling rate to 100 Hz or more would allow for better resolution of the stages of motion. The use of an additional sensor placed on the head could provide information about stability while a sensor on the foot would enable not only full segmentation of the gait cycle but also the extraction of measures of rotation and potentially approximation of impact forces [5], [12], [14], [42], [43].

Unlike the calibration substudy, the use of unscripted and unlabeled activity data injects a substantial amount of noise into the bout recognition pipeline. Generalizable models for activity recognition are difficult enough to develop with the use of large, labeled data sets which contain minimal noise. The use of unlabeled activity data shifts the burden of accuracy from the data and onto the expert opinion of the modeler who is required to visually inspect output and make a decision about the correct activity label for the given instance. This can introduce bias which may limit the pool of walking activity examples to those preferred by the encoder rather than the true full sample of walking activity. Furthermore, the use of unlabeled data combined with the more casual behavior of participants who are not under observation may reduce the quality and

quantity of walking samples obtained. It is well known that, especially in the absence of high quality measurements, one of the best ways to create better statistical models is by simply obtaining *more* data. With unlabeled and unstructured data, the challenge lies in properly evaluating the effectiveness of preprocessing and modeling approaches. These limitations speak to some of the benefits of using mobile phones for monitoring gait as they offer the ability to periodically request feedback on user activity and obtain labeled, individual-specific activity data to improve monitoring and analysis.

A related, but different, limitation on walking bout extraction is the specific definition of a good walking bout used by the pipeline. To ensure the acquisition of walking behavior similar to that of a walk test, a very strict definition of walking was developed. Perhaps the strongest criterion of this definition is that bouts were required to be at least one minute in duration. This placed a considerable limitation on the pool of potential walking bouts given that minute-long bouts are unlikely to occur in the home due to spatial restrictions and are most likely to happen outside. High-risk individuals who are homebound or who greatly limit their mobility will not generate bouts of this length which immediately excludes some of the most important individuals from the population. It is also valuable to consider that shorter in-home bouts may provide greater predictive value than longer bouts since they may better capture instability and temporal inconsistency of walking. However, the use of shorter bouts would introduce their own set of challenges since short walking bouts are more likely to resemble other activities such as sweeping the floor which are not solely walking. Shorter bouts are also more susceptible to the influence of noise as less data are available to smooth out its influence. Still, the combination of both long and short walking bouts could provide data not just for better prediction of fall risk but

the creation of detailed activity profiles. One possible approach would be to first establish criteria for describing “long” and “short” walking bout durations (e.g. 30 second maximum on short walking bouts, with a minimum of 3-5 seconds to differentiate between true walking versus shuffled motion while in a standing position). Long walking bouts would simply be characterized as those greater than the maximum duration of a short bout. With regard to segmenting bouts, shorter walking bouts would most likely occur in-home punctuated by short pauses and may benefit from a shorter pause limit (e.g. 10 seconds) compared to long-duration walking bouts which most likely would be observed outside of the home with fewer reasons for stopping.

Leading from the definition of our input data, it is equally as important—if not more so—to think about the definition of class labels. Similar to the calibration study, fall and SPPB data were used to define outcome variables. However, the previously-mentioned gap in time between accelerometer data collection and SPPB measurement does not apply to the free-living study since the SPPB was performed during the accelerometer visit. That aside, the outcome variables of SPPB, fall count, and fall rate used to define class labels can have a substantial effect on model performance. The simple decision of choosing a binary or multiclass classification scheme has an immediate impact on the difficulty of the task since the probability of randomly selecting the wrong class for any given individual increases with the number of class labels. Furthermore, by segmenting the data space into more regions (i.e. classes), the decision boundaries become more strict which reduces the margin of error. One potential solution to this challenge is to train multiple classifiers which work on subsets of the larger set of data. In this way, the data are split according to some criteria (e.g. age groups) which reduces the overlap across the classes. This very approach was attempted with the three-way models but failed since we could not find a

splitting criteria which allowed for sufficient representation of all classes in each subset. On a finer level of detail, the *fall rate* and *fall count* variables have a low correlation with age (less than 5% in the full data set, less than 10% in our subset, see Figure C.5) which goes against accepted knowledge that falls occur more often as people get older. This alone suggests that even the combined power of strong epidemiological predictors, like age, with walking bout data will demonstrate poor ability to predict falls in this data set. Looking at the distribution of SPPB scores (Figure C.10), we see a low number of individuals in the 1 - 4 range (roughly 5% of individuals) which indicates a largely healthy population that could further bias predictive models to classify individuals as low risk. However, fallers make up 46% of the available data. This contradiction generates additional questions about the relationship between SPPB and falls.

In the three-way classification models, the middle SPPB category of 7 – 9 showed the highest rate of misclassification compared to the extremes of 1 – 6 and 10 – 12. This is somewhat expected given the significant overlap of classification categories as seen in the feature plots in Appendix C. Looking at Figure C.16 in particular, we can see that instances in the middle SPPB category (i.e. Class 1, SPPB 7 – 9) shows a very high degree of overlap with individuals in the other classes. Since the extremes of risk (classes 0 and 2) are more easily separable from each other than the middle class, it is expected—and indeed we see—reduced ability to accurately classify individuals in the middle category; a similar story is seen with future falls and fall risk in figures Figure C.10, Figure C.12, and Figure C.14, respectively. All together, these challenges demonstrate the limitations of the current set of features at separating individuals into fall and SPPB categories as they are currently defined. Assuming that the sensor data contains the appropriate signal for predicting falls and SPPB of individuals, new features need to be extracted

or developed to improve performance. However, it is more likely that additional data will be required to achieve better separation between the classes.

CHAPTER 11: FUTURE WORK

The objective of this study was to develop a pipeline for prediction of future falls and risk of falling in older adults through the analysis of passive walking data measured via a single, hip-worn triaxial accelerometer. Due to the limited effectiveness of this approach, several straightforward improvements—and some more complicated—could be implemented in future projects.

To begin, the use of a single 30 Hz accelerometer proved to be insufficient for segmentation of the *full* gait cycle which may have missed important information associated with fall risk. If clinical accelerometers (such as ActiGraphs) are to be used in subsequent studies, the attachment of an additional accelerometer to the ankle *may* be sufficient to segment the full gait cycle by providing much needed information about heel and toe movement; but this will need to be confirmed in practice. Alternatively, the use of sensors with a higher sampling rate (100 Hz or more) would provide much cleaner data by allowing for the application of appropriate filters without falling prey to noise artifacts. This higher rate of sampling may be enough to capture the more subtle aspects of the gait cycle if continuing the use of a single sensor mounted to the hip or back. The popularity of smart phones offers another avenue for improving sensor quality as these devices contain not only accelerometers but other sensors such as gyroscopes which may provide even more useful information concerning user motion. However, switching to a device of this type introduces new challenges such as detection of phone orientation and new preprocessing requirements for activity recognition from a sensor which is no longer in a fixed orientation or position (e.g. in pocket vs. hand). One benefit of switching to phones is the ability to determine sensor position *relative to the earth* rather than just the individual. This would allow

for the proper removal of gravity from the activity signal and possible projection of the signal to a coordinate system that takes into account the position of the subject and terrain (e.g. walking on a flat sidewalk vs. an incline). Finally, wrist-worn devices such as smart watches could be used in lieu of both ActiGraphs and cell phones given that these watches have the combined benefits of fixed orientation (can only be worn in two orientations, which simply invert the positive/negative axes) and higher-quality sensors (compared to ActiGraphs). However, wrist accelerometer data are even less similar than phone data are in comparison to hip-worn ActiGraphs. In addition to eliminating the option of applying the current bout extraction pipeline to wrist data (which may not necessarily be true if using phones), there are more complex challenges regarding activity recognition since arm and leg movement during walking—which are correlated in younger, healthy individuals—may not necessarily be true for older adults at increased risk of falls [44].

The current pipeline for good walking detection makes a number of assumptions about the nature of the "good walking" signal and what constitutes a "good walking" bout. The definition of a walking bout used in this study restricts good walking samples to those that mimic walking as seen during a walk test. While this provides a very specific signal allowing for accurate activity recognition, the strictness of this definition limits walking bouts to only those that likely occurred outside the home (minimum one minute in duration). In-home walking activity may provide a better picture of overall behavior and allow for the temporal analysis of changes in walking activity and quality. However, this would require more advanced approaches to walking detection that allow for the identification of short bouts of walking (e.g. two seconds or less) which, in the current pipeline, were found to be easily confused with noisy motion. Such a

system would likely reply upon statistical models trained on *labeled* sensor data (unlike the current pipeline which was developed on *unlabeled* data and does not make use of any statistical models) and may benefit from the use of shape-matching techniques (e.g. discrete time warping) or neural networks which are capable of automated feature extraction and have demonstrated high accuracy on time series classification problems. It's important to note that overfitting will be of greater concern taking this approach and careful selection of training data will be needed to guarantee the generalizability of the model. Alternatively, if using cell phones instead of ActiGraphs, individuals could be instructed through an application to initially—even periodically—provide example walking data to train a patient-specific activity recognizer.

A smaller—but equally important—question is how much walking data are required for accurate prediction? This study found that the majority of individuals had *fewer than three* good walking bouts over a week-long time frame. This may be the true behavior of the population or could be a result of the strict bout extraction criteria. Either way, a single bout (and even three bouts) might not be providing a sufficient representation of an individual's walking ability. However, analysis of the OPACH data set cannot provide an answer to this question since we lack *fully-labeled* periods of walking. In addition to good walking, other activity types may prove to be valuable in assessing fall risk. Activity profiles generated from accelerometer data would provide a better picture of activity levels and patterns related not just to fall risk but general health changes with age. Profile changes over time could identify fluctuations in health status and act as flags for the initiation of proactive clinical therapies to reduce risk and maintain independence.

Finally, rest of this chapter summarizes thoughts on applying the walking bout extraction, feature computation, and predictive modeling approaches discussed in this thesis for large-scale, online (i.e. active and long-term) monitoring of fall risk in older adults. Technical adjustments are detailed to address the unique challenges of processing streaming sensor data. For the following discussion, we will assume that the accelerometer device has the ability to store and send data to a remote destination and is capable of simple computations (e.g. a smartphone).

The speed and linear nature of the bout extraction pipeline suggest that it could be easily adapted for online data processing. Raw sensor data could be queried to identify vertical, upright motion which is strong and persistent during walking. The current pipeline uses non-overlapping, one-second windows and checks that sensor data are in the range of (0, -2] G's (naturally, this will have to be tuned to fit the device). For active monitoring, once vertical motion is detected consecutively for some amount of time (e.g. two/three seconds to prevent the unwanted storage of transient motion), data would be recorded until the vertical motion subsides (e.g. two/three seconds of non-vertical motion). Assuming that this recording is of sufficient length (e.g. 30 seconds or perhaps even shorter if accuracy is maintained), the Activity Index [39] would be computed and checked to verify the data are in the acceptable range for walking (i.e. AI values of [18, 106] for a one-second window). If this is true, data from the vertical axis would be passed to the "good walking" filter [29], [37] for a final decision about retaining or discarding the recording. Once a good walking bout has been obtained, the raw data should be sent to an external computer/server for long-term storage and subsequent processing and analysis.

It is important to note that this pipeline was originally developed for a fixed-position sensor at the hip with the three axes of measurement oriented in the major directions of human motion (anteroposterior, mediolateral, and vertical). As such, the use of a variable-position sensor (e.g. phone) might not be stable enough for use with the existing pipeline. It is difficult to comment further on this issue without testing the performance.

As an alternative to using the current bout recognition pipeline, individual-specific activity recognition models could be trained through the use of a phone application which allows users to provide labeled examples of walking. Ideally, multiple walking examples of varying duration should be obtained in the free-living environment (indoors and outdoors) to capture the natural variation of walking for the individual. Additionally, these walking examples could be collected periodically (e.g. monthly) to maintain an up-to-date library for model training. For active monitoring, it may be beneficial to first screen accelerometer data for vertical motion as described in the previous section before sending the data for subsequent processing and classification. This would not only greatly reduce the volume of data processed but also limit the recordings to examples which are more likely to be walking activity. The custom activity classification models would then make a final decision about the potential good walking bout.

For cross-sectional analysis of the good walking bouts, it is *easiest* to simply extract features from the full-length bouts for a given time period and take the median to obtain a single measure of walking for the individual. This approach, unfortunately, results in substantial loss of information and does not facilitate temporal analyses. If, however, a single cross-sectional view is desired (or required), bout variation could be captured by computing not only the median of

the bout features, but also the standard deviation. The features obtained using this approach are more susceptible to outliers since the calculation of standard deviation makes use of the mean. Alternatively, features could be extracted from each bout and then weighted by time to give greater pull to bouts more recently measured. These features and weights would be summed to produce a final feature vector. Ultimately, to best-preserve the unique information afforded by a set of walking bouts, features should be extracted from each bout individually. With this approach, the effect of any "extreme" bout is minimized and the contribution of each bout can be weighted equally or use a more complicated valuing scheme downstream. Moreover, this approach shifts the burden of combining the meaning of multiple bouts away from feature extraction and onto the modeling process which makes use of intelligent systems.

The prediction of current fall risk or SPPB status could be accomplished through a variety of methods. The most direct approach would be to obtain status predictions from the past "N" walking bouts and return a final status via majority vote. As previously mentioned, these predictions could also be weighted according to some reasonable factor such as recency of the bout or perhaps the length where longer bouts are valued more than shorter ones, or vice versa. Instead of considering the past "N" bouts, all of the bouts recorded during some specified period of time (e.g. one week) could be selected as a representative sample. It is difficult to select one approach, or attempt to rank these different analytical methods, without empirical evidence. Focusing on predicting fall risk and SPPB scores, it is unlikely that an individual would experience frequent fluctuations in these two health measures over the course of a week except resulting from directed therapy or acute injury. Over the course of a month, however, the health of even young individuals can change to a notable degree. On the other hand, depending upon

the patterns of bouts produced by an individual, it may be that condensing bouts with weekly or monthly averages produces a more realistic measure of risk over time by averaging out noisy fluctuations in status. Ultimately, an exploratory analysis using real-world data is needed before an optimal analytical approach can be selected.

REFERENCES

[1] G. Bergen, "Falls and Fall Injuries Among Adults Aged ≥65 Years — United States, 2014," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 65, 2016.

[2] V. A. Moyer and U.S. Preventive Services Task Force, "Prevention of falls in community-dwelling older adults: U.S. Preventive Services Task Force recommendation statement," *Ann. Intern. Med.*, vol. 157, no. 3, pp. 197–204, Aug. 2012.

[3] "Older Adult Falls | Home and Recreational Safety | CDC Injury Center," 12-Mar-2018. [Online]. Available: https://www.cdc.gov/homeandrecreationalsafety/falls/index.html. [Accessed: 25-Sep-2018].

[4] U. C. Bureau, "An Aging Nation: The Older Population in the United States." [Online]. Available: https://www.census.gov/library/publications/2014/demo/p25-1140.html. [Accessed: 25-Sep-2018].

[5] J. Howcroft, J. Kofman, and E. D. Lemaire, "Review of fall risk assessment in geriatric populations using inertial sensors," *J. NeuroEngineering Rehabil.*, vol. 10, p. 91, 2013.

[6] B. Heinbüchner, M. Hautzinger, C. Becker, and K. Pfeiffer, "Satisfaction and use of personal emergency response systems," *Z. Für Gerontol. Geriatr.*, vol. 43, no. 4, pp. 219–223, Aug. 2010.

[7] G. Feder, C. Cryer, S. Donovan, and Y. Carter, "Guidelines for the prevention of falls in people over 65," *BMJ*, vol. 321, no. 7267, pp. 1007–1011, Oct. 2000.

[8] R. J. Shephard, "Limits to the measurement of habitual physical activity by questionnaires," *Br. J. Sports Med.*, vol. 37, no. 3, pp. 197–206, Jun. 2003.

[9] N. Veronese *et al.*, "Association between Short Physical Performance Battery and falls in older people: the Progetto Veneto Anziani Study," *Rejuvenation Res.*, vol. 17, no. 3, pp. 276–284, Jun. 2014.

[10] S. Volpato *et al.*, "Predictive value of the Short Physical Performance Battery following hospitalization in older patients," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 66, no. 1, pp. 89–96, Jan. 2011.

[11] S. R. Lord, H. B. Menz, and A. Tiedemann, "A physiological profile approach to falls risk assessment and prevention," *Phys. Ther.*, vol. 83, no. 3, pp. 237–252, Mar. 2003.

[12] L. Montesinos, R. Castaldo, and L. Pecchia, "Wearable Inertial Sensors for Fall Risk Assessment and Prediction in Older Adults: A Systematic Review and Meta-Analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 573–582, Mar. 2018.

[13] W. Zijlstra, "Assessment of spatio-temporal parameters during unconstrained walking," *Eur. J. Appl. Physiol.*, vol. 92, no. 1–2, pp. 39–44, Jun. 2004.

[14] R. Lemoyne, C. Coroian, T. Mastroianni, and W. Grundfest, "Accelerometers for quantification of gait and movement disorders: a perspective review," *J. Mech. Med. Biol.*, vol. 08, no. 02, pp. 137–152, Jun. 2008.

[15] S. N. Robinovitch *et al.*, "Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study," *Lancet Lond. Engl.*, vol. 381, no. 9860, pp. 47–54, Jan. 2013.

[16] I. Bautmans, B. Jansen, B. Van Keymolen, and T. Mets, "Reliability and clinical correlates of 3D-accelerometry based gait analysis outcomes according to age and fall-risk," *Gait Posture*, vol. 33, no. 3, pp. 366–372, Mar. 2011.

[17] M. A. D. Brodie, H. B. Menz, S. T. Smith, K. Delbaere, and S. R. Lord, “Good lateral harmonic stability combined with adequate gait speed is required for low fall risk in older people,” *Gerontology*, vol. 61, no. 1, pp. 69–78, 2015.
[18] R. Senden, H. H. C. M. Savelberg, B. Grimm, I. C. Heyligers, and K. Meijer, “Accelerometry-based gait analysis, an additional objective approach to screen subjects at risk for falling,” *Gait Posture*, vol. 36, no. 2, pp. 296–300, Jun. 2012.
[19] H. B. Menz, S. R. Lord, and R. C. Fitzpatrick, “Acceleration patterns of the head and pelvis when walking are associated with risk of falling in community-dwelling older people,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 58, no. 5, pp. M446-452, May 2003.
[20] J. W. Lockhart and G. M. Weiss, “Limitations with Activity Recognition Methodology & Data Sets,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, New York, NY, USA, 2014, pp. 747–756.
[21] A. Z. LaCroix *et al.*, “The Objective Physical Activity and Cardiovascular Disease Health in Older Women (OPACH) Study,” *BMC Public Health*, vol. 17, no. 1, p. 192, 14 2017.
[22] K. R. Evenson *et al.*, “Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91 years: The Women’s Health Initiative OPACH Calibration Study,” *Prev. Med. Rep.*, vol. 2, pp. 750–756, 2015.
[23] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, p. 2825−2830, Oct. 2011.
[24] P. Blomstedt, J. Tang, J. Xiong, C. Granlund, and J. Corander, “A Bayesian Predictive Model for Clustering Data of Mixed Discrete and Continuous Type,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 489–498, Mar. 2015.
[25] “Short Physical Performance Battery (SPPB),” *National Institute on Aging*. [Online]. Available: https://www.nia.nih.gov/research/labs/leps/short-physical-performance-battery-sppb. [Accessed: 07-Oct-2018].
[26] D. Treacy and L. Hassett, “The Short Physical Performance Battery,” *J. Physiother.*, vol. 64, no. 1, p. 61, Jan. 2018.
[27] A. Z. LaCroix, “Long Life Study SPPB Scoring.” WHI, 04-Mar-2014.
[28] Q. Cheng *et al.*, “Classification Models for Pulmonary Function using Motion Analysis from Phone Sensors,” *AMIA. Annu. Symp. Proc.*, vol. 2016, pp. 401–410, Feb. 2017.
[29] J. Juen, Q. Cheng, V. Prieto-Centurion, J. A. Krishnan, and B. Schatz, “Health monitors for chronic disease by gait analysis with mobile phones,” *Telemed. J. E-Health Off. J. Am. Telemed. Assoc.*, vol. 20, no. 11, pp. 1035–1041, Nov. 2014.
[30] E. Barrett-Connor, T. W. Weiss, C. A. McHorney, P. D. Miller, and E. S. Siris, “Predictors of falls among postmenopausal women: results from the National Osteoporosis Risk Assessment (NORA),” *Osteoporos. Int. J. Establ. Result Coop. Eur. Found. Osteoporos. Natl. Osteoporos. Found. USA*, vol. 20, no. 5, pp. 715–722, May 2009.
[31] L. Breiman, *Classification and Regression Trees*. Routledge, 2017.
[32] A. Weiss, T. Herman, N. Giladi, and J. M. Hausdorff, “Objective Assessment of Fall Risk in Parkinson’s Disease Using a Body-Fixed Sensor Worn for 3 Days,” *PLoS ONE*, vol. 9, no. 5, May 2014.
[33] M. Gietzelt, G. Nemitz, K.-H. Wolf, H. Meyer Zu Schwabedissen, R. Haux, and M. Marschollek, “A clinical study to assess fall risk using a single waist accelerometer,” *Inform. Health Soc. Care*, vol. 34, no. 4, pp. 181–188, Dec. 2009.
[34] M. Marschollek, K.-H. Wolf, M. Gietzelt, G. Nemitz, H. M. zu Schwabedissen, and R. Haux, “Assessing elderly persons’ fall risk using spectral analysis on accelerometric data-a

clinical evaluation study," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 3682–3685.
[35] M. Marschollek *et al.*, "Sensor-based Fall Risk Assessment – an Expert 'to go':," *Methods Inf. Med.*, vol. 50, no. 5, pp. 420–426, Jan. 2011.
[36] J. Vickers, A. Reed, R. Decker, B. P. Conrad, M. Olegario-Nebel, and H. K. Vincent, "Effect of investigator observation on gait parameters in individuals with and without chronic low back pain," *Gait Posture*, vol. 53, pp. 35–40, 2017.
[37] Q. Cheng *et al.*, "Predicting Pulmonary Function from Phone Sensors," *Telemed. J. E-Health Off. J. Am. Telemed. Assoc.*, vol. 23, no. 11, pp. 913–919, 2017.
[38] J. E. Sasaki, A. Hickey, J. Staudenmayer, D. John, J. A. Kent, and P. S. Freedson, "Performance of Activity Classification Algorithms in Free-living Older Adults," *Med. Sci. Sports Exerc.*, vol. 48, no. 5, pp. 941–950, May 2016.
[39] "An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics." [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0160644. [Accessed: 25-Sep-2018].
[40] E. Keogh and M. Pazzani, "Derivative Dynamic Time Warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2001, pp. 1–11.
[41] L. Ye and E. Keogh, "Time Series Shapelets: A New Primitive for Data Mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2009, pp. 947–956.
[42] K.-Y. Ahn and B.-J. Ryu, "A modeling of impact dynamics and its application to impact force prediction," *J. Mech. Sci. Technol.*, vol. 19, no. 1, pp. 422–428, Jan. 2005.
[43] R. W. Bisseling and A. L. Hof, "Handling of impact forces in inverse dynamics," *J. Biomech.*, vol. 39, no. 13, pp. 2438–2444, Jan. 2006.
[44] M. P. Ford, R. C. Wagenaar, and K. M. Newell, "Arm constraint and walking in healthy adults," *Gait Posture*, vol. 26, no. 1, pp. 135–141, Jun. 2007.

APPENDIX A: CALIBRATION SUBSTUDY FEATURE SETS

Table A.1: Full list of calibration substudy features and their use in various feature sets.

Abbreviation	Feature Sets
X_MEAN	2, 3, 10, 11
X_STD	2, 3, 10, 11
X_RMS	2, 3, 10, 11
X_SMA	2, 3, 10, 11
X_COV	2, 3, 10, 11
X_PFREQ	2, 3, 10, 11
X_ENERGY	2, 3, 10, 11
X_MCR	2, 3, 10, 11
X_MAD	2, 3, 10, 11
X_P2P	2, 3, 10, 11
X_ACC	2, 3, 10, 11
Y_MEAN	4, 5, 10, 11
Y_STD	4, 5, 10, 11
Y_RMS	4, 5, 10, 11
Y_SMA	4, 5, 10, 11
Y_COV	4, 5, 10, 11
Y_PFREQ	4, 5, 10, 11
Y_ENERGY	4, 5, 10, 11
Y_MCR	4, 5, 10, 11
Y_MAD	4, 5, 10, 11
Y_P2P	4, 5, 10, 11
Y_ACC	4, 5, 10, 11
Z_MEAN	6, 7, 10, 11
Z_STD	6, 7, 10, 11
Z_RMS	6, 7, 10, 11
Z_SMA	6, 7, 10, 11
Z_COV	6, 7, 10, 11
Z_PFREQ	6, 7, 10, 11
Z_ENERGY	6, 7, 10, 11
Z_MCR	6, 7, 10, 11
Z_MAD	6, 7, 10, 11
Z_P2P	6, 7, 10, 11
Z_ACC	6, 7, 10, 11
MAG_MEAN	8, 9
MAG_STD	8, 9
MAG_RMS	8, 9

Table A.1: Continued

MAG_SMA	8, 9
MAG_COV	8, 9
MAG_PFREQ	8, 9
MAG_ENERGY	8, 9
MAG_MCR	8, 9
MAG_MAD	8, 9
MAG_P2P	8, 9
MAG_ACC	8, 9
XY_CORR	10, 11
YZ_CORR	10, 11
XZ_CORR	10, 11
CADENCE	1, 3, 5, 7, 9, 11
MEAN_STEP_TIME	1, 3, 5, 7, 9, 11
STD_STEP_TIME	1, 3, 5, 7, 9, 11
MEAN_STRIDE_TIME	1, 3, 5, 7, 9, 11
STD_STRIDE_TIME	1, 3, 5, 7, 9, 11

APPENDIX B: FREE-LIVING DATA DICTIONARY

Table B.1: Data dictionary for the traditional measures of gait computed for the free-living accelerometer study. The abbreviations for each feature used in the study are listed along with their full names and additional details about their units and computation.

Abbreviation	Full Name	Description
CADENCE	Cadence	Steps per unit time (steps/minute)
STEP_ASYMMETRY	Step asymmetry	A measure of asymmetry of the left and right step times
STEP_COUNT	Step count	Number of steps
STEP_COV	Coefficient of variation	Coefficient of variation of time between steps
STEP_MEAN	Mean step time	Mean time between steps
STEP_MEDIAN	Median step time	Median time between steps
STEP_RMS	Root mean square of step time	Root mean square of time between steps
STEP_STD	Standard deviation of step time	Standard deviation of time between steps
STRIDE_MEAN	Mean stride time	Mean time between strides
STRIDE_MEDIAN	Median stride time	Median time between strides
STRIDE_RMS	Root mean square of stride time	Root mean square of time between strides
STRIDE_STD	Standard deviation of stride time	Standard deviation of time between strides
ALPHA_BOUT_LEN	Alpha measure of bout length	Alpha is a scalar measure describing the distribution of walking bout lengths observed for an individual
BOUTS_TO_ACTIVE	Ratio of good walking bouts to all activity/motion	Ratio of "good walking" bout time to the total sum of time where the accelerometer was in motion
BOUT_DAYS	N/A	Number of days with good walking bouts
BOUT_NUMB	N/A	Number of good walking bouts
MEAN_BOUT_LEN	Mean bout length	Average good walking bout length
MEAN_START	Mean start time	Average start time of a good walking bout

Table B.1: Continued.

MEDIAN_BOUT_LEN	Median bout length	Median good walking bout length
MEDIAN_START	Median start time	Median start time of a good walking bout
PERCENT_ACTIVE	N/A	Percent of accelerometer time where the user was actively moving (this is not restricted to good walking bouts)
PERCENT_INACTIVE	N/A	Percent of accelerometer time where the user was NOT actively moving
PERCENT_BOUTS	N/A	Percent of accelerometer recording time that is good walking bouts
RMS_BOUT_LEN	Root mean square of bout length	Root mean square of good walking bout length
RMS_START	Root mean square of start time	Root mean square of good walking bout start time
STD_BOUT_LEN	Standard deviation of bout length	Standard deviation of good walking bout length
STD_START	Standard deviation of start time	Standard deviation of good walking bout start time

Table B.2: Data dictionary for the accelerometer-derived measures of gait computed for the free-living accelerometer study. The abbreviations for each feature or variable used in the study are listed along with their full names and additional details about their units and computation.

VMAG	Vector magnitude	Vector magnitude of the accelerometer data
XY_CORR	XY cross-correlation	Cross-correlation between the X-axis and Y-axis
XZ_CORR	XZ cross-correlation	Cross-correlation between the X-axis and Z-axis
YZ_CORR	YZ cross-correlation	Cross-correlation between the Y-axis and Z-axis
X_ACC	X-axis autocorrelation coefficient	Autocorrelation coefficient of the x-axis shifted by a period derived from the peak frequency of the signal
X_ENERGY	X-axis energy	Energy of the x-axis signal
X_ENTROPY	X-axis entropy	Entropy of the x-axis signal
X_SKEW	X-axis skew	Skewness of the x-axis signal
X_KURT	X-axis kurtosis	Kurtosis of the x-axis signal

Table B.2: Continued.

X_MAD	X-axis mean amplitude deviation	Average change in amplitude of the x-axis signal
X_MCR	X-axis mean crossing rate	Average rate of the x-axis signal crossing its mean
X_MEAN	X-axis mean	Average of the x-axis signal
X_STD	X-axis standard deviation	Standard deviation of the x-axis signal
X_RMS	X-axis root mean square	Root mean square of the x-axis signal
X_COV	X-axis coefficient of variation	Coefficient of variation of the x-axis signal
X_SMA	X-axis signal magnitude area	Sum of the magnitude of the absolute value of the x-axis signal
X_P2P	X-axis peak-to-peak	The difference between the maximum and minimum value of the x-axis signal
X_PFREQ	X-axis peak frequency	The peak frequency (most dominant frequency) of the x-axis signal
X_FFTQ25	X-axis FFT quartile 25	The frequency at which 25% of the data are below this frequency value
X_FFT50	X-axis FFT quartile 50	The frequency at which 50% of the data are below this frequency value
X_FFT75	X-axis FFT quartile 75	The frequency at which 75% of the data are below this frequency value
X_BIN1	X-axis bin 1	The lowest bin of x-axis values separated into ten partitions
X_BIN2	X-axis bin 2	The second-lowest bin of x-axis values separated into ten partitions
X_BIN3	X-axis bin 3	The third-lowest bin of x-axis values separated into ten partitions
X_BIN4	X-axis bin 4	The fourth-lowest bin of x-axis values separated into ten partitions
X_BIN5	X-axis bin 5	The fifth-lowest bin of x-axis values separated into ten partitions

Table B.2: Continued.

X_BIN6	X-axis bin 6	The fifth-largest bin of x-axis values separated into ten partitions
X_BIN7	X-axis bin 7	The forth-largest bin of x-axis values separated into ten partitions
X_BIN8	X-axis bin 8	The third-largest bin of x-axis values separated into ten partitions
X_BIN9	X-axis bin 9	The second-largest bin of x-axis values separated into ten partitions
X_BIN10	X-axis bin 10	The largest bin of x-axis values separated into ten partitions
X_TOPN1	X-axis top frequency 1	Second-largest dominant frequency in the x-axis signal
X_TOPN2	X-axis top frequency 2	Third-largest dominant frequency in the x-axis signal
X_TOPN3	X-axis top frequency 3	Fourth-largest dominant frequency in the x-axis signal
X_TOPN4	X-axis top frequency 4	Fifth-largest dominant frequency in the x-axis signal
X_TOPN5	X-axis top frequency 5	Sixth-largest dominant frequency in the x-axis signal
X_TOPN6	X-axis top frequency 6	Seventh-largest dominant frequency in the x-axis signal
X_TOPN7	X-axis top frequency 7	Eight-largest dominant frequency in the x-axis signal
Y_ACC	Y-axis autocorrelation coefficient	Autocorrelation coefficient of the y-axis shifted by a period derived from the peak frequency of the signal
Y_ENERGY	Y-axis energy	Energy of the y-axis signal
Y_ENTROPY	Y-axis entropy	Entropy of the y-axis signal
Y_SKEW	Y-axis skew	Skewness of the y-axis signal
Y_KURT	Y-axis kurtosis	Kurtosis of the y-axis signal
Y_MAD	Y-axis mean amplitude deviation	Average change in amplitude of the y-axis signal
Y_MCR	Y-axis mean crossing rate	Average rate of the y-axis signal crossing its mean
Y_MEAN	Y-axis mean	Average of the y-axis signal
Y_STD	Y-axis standard deviation	Standard deviation of the y-axis signal

Table B.2: Continued.

Y_RMS	Y-axis root mean square	Root mean square of the y-axis signal
Y_COV	Y-axis coefficient of variation	Coefficient of variation of the y-axis signal
Y_SMA	Y-axis signal magnitude area	Sum of the magnitude of the absolute value of the y-axis signal
Y_P2P	Y-axis peak-to-peak	The difference between the maximum and minimum value of the y-axis signal
Y_PFREQ	Y-axis peak frequency	The peak frequency (most dominant frequency) of the y-axis signal
Y_FFTQ25	Y-axis FFT quartile 25	The frequency at which 25% of the data are below this frequency value
Y_FFT50	Y-axis FFT quartile 50	The frequency at which 50% of the data are below this frequency value
Y_FFT75	Y-axis FFT quartile 75	The frequency at which 75% of the data are below this frequency value
Y_BIN1	Y-axis bin 1	The lowest bin of y-axis values separated into ten partitions
Y_BIN2	Y-axis bin 2	The second-lowest bin of y-axis values separated into ten partitions
Y_BIN3	Y-axis bin 3	The third-lowest bin of y-axis values separated into ten partitions
Y_BIN4	Y-axis bin 4	The fourth-lowest bin of y-axis values separated into ten partitions
Y_BIN5	Y-axis bin 5	The fifth-lowest bin of y-axis values separated into ten partitions
Y_BIN6	Y-axis bin 6	The fifth-largest bin of y-axis values separated into ten partitions
Y_BIN7	Y-axis bin 7	The forth-largest bin of y-axis values separated into ten partitions

Table B.2: Continued.

Y_BIN8	Y-axis bin 8	The third-largest bin of y-axis values separated into ten partitions
Y_BIN9	Y-axis bin 9	The second-largest bin of y-axis values separated into ten partitions
Y_BIN10	Y-axis bin 10	The largest bin of y-axis values separated into ten partitions
Y_TOPN1	Y-axis top frequency 1	Second-largest dominant frequency in the y-axis signal
Y_TOPN2	Y-axis top frequency 2	Third-largest dominant frequency in the y-axis signal
Y_TOPN3	Y-axis top frequency 3	Fourth-largest dominant frequency in the y-axis signal
Y_TOPN4	Y-axis top frequency 4	Fifth-largest dominant frequency in the y-axis signal
Y_TOPN5	Y-axis top frequency 5	Sixth-largest dominant frequency in the y-axis signal
Y_TOPN6	Y-axis top frequency 6	Seventh-largest dominant frequency in the y-axis signal
Y_TOPN7	Y-axis top frequency 7	Eight-largest dominant frequency in the y-axis signal
Z_ACC	Z-axis autocorrelation coefficient	Autocorrelation coefficient of the z-axis shifted by a period derived from the peak frequency of the signal
Z_ENERGY	Z-axis energy	Energy of the z-axis signal
Z_ENTROPY	Z-axis entropy	Entropy of the z-axis signal
Z_SKEW	Z-axis skew	Skewness of the z-axis signal
Z_KURT	Z-axis kurtosis	Kurtosis of the z-axis signal
Z_MAD	Z-axis mean amplitude deviation	Average change in amplitude of the z-axis signal
Z_MCR	Z-axis mean crossing rate	Average rate of the z-axis signal crossing its mean
Z_MEAN	Z-axis mean	Average of the z-axis signal
Z_STD	Z-axis standard deviation	Standard deviation of the z-axis signal
Z_RMS	Z-axis root mean square	Root mean square of the z-axis signal
Z_COV	Z-axis coefficient of variation	Coefficient of variation of the z-axis signal

Table B.2: Continued.

Z_SMA	Z-axis signal magnitude area	Sum of the magnitude of the absolute value of the z-axis signal
Z_P2P	Z-axis peak-to-peak	The difference between the maximum and minimum value of the z-axis signal
Z_PFREQ	Z-axis peak frequency	The peak frequency (most dominant frequency) of the z-axis signal
Z_FFTQ25	Z-axis FFT quartile 25	The frequency at which 25% of the data are below this frequency value
Z_FFT50	Z-axis FFT quartile 50	The frequency at which 50% of the data are below this frequency value
Z_FFT75	Z-axis FFT quartile 75	The frequency at which 75% of the data are below this frequency value
Z_BIN1	Z-axis bin 1	The lowest bin of z-axis values separated into ten partitions
Z_BIN2	Z-axis bin 2	The second-lowest bin of z-axis values separated into ten partitions
Z_BIN3	Z-axis bin 3	The third-lowest bin of z-axis values separated into ten partitions
Z_BIN4	Z-axis bin 4	The fourth-lowest bin of z-axis values separated into ten partitions
Z_BIN5	Z-axis bin 5	The fifth-lowest bin of z-axis values separated into ten partitions
Z_BIN6	Z-axis bin 6	The fifth-largest bin of z-axis values separated into ten partitions
Z_BIN7	Z-axis bin 7	The forth-largest bin of z-axis values separated into ten partitions
Z_BIN8	Z-axis bin 8	The third-largest bin of z-axis values separated into ten partitions

Table B.2: Continued.

Z_BIN9	Z-axis bin 9	The second-largest bin of z-axis values separated into ten partitions
Z_BIN10	Z-axis bin 10	The largest bin of z-axis values separated into ten partitions
Z_TOPN1	Z-axis top frequency 1	Second-largest dominant frequency in the z-axis signal
Z_TOPN2	Z-axis top frequency 2	Third-largest dominant frequency in the z-axis signal
Z_TOPN3	Z-axis top frequency 3	Fourth-largest dominant frequency in the z-axis signal
Z_TOPN4	Z-axis top frequency 4	Fifth-largest dominant frequency in the z-axis signal
Z_TOPN5	Z-axis top frequency 5	Sixth-largest dominant frequency in the z-axis signal
Z_TOPN6	Z-axis top frequency 6	Seventh-largest dominant frequency in the z-axis signal
Z_TOPN7	Z-axis top frequency 7	Eight-largest dominant frequency in the z-axis signal

Table B.3: Data dictionary for the demographic variables used in the free-living accelerometer study. The abbreviations for each variable are listed along with their full names and additional details about their units and computation.

age	Age	Age in years
eth2	Ethnicity	0: White 1: Black 2: Hispanic
EDUC	Level of education	0: No School 1: 1 – 4 Years 2: 5 – 8 Years 3: 9 – 12 Years 4: High School 5: Vocational School 6: Some College or Associates 7: College Graduate 8: Some Postgraduate or Professional School 9: Masters 10: Doctoral
bmills	Body Mass Index	Body mass index (kg/m)

Table B.4: Data dictionary for the target variables used in the free-living accelerometer study. The abbreviations for each variable are listed along with their full names and additional details about their units and computation.

EPESESPPB	Total EPESE short physical performance battery score	Score ranges from 0 – 12; scoring details can be found here
balance_epese	Balance score for the EPESE SPPB	Score ranges of 0 – 4; scoring details can be found here
gait_epese	Gait subscore for the EPESE SPPB	Score ranges of 0 – 4; scoring details can be found here
chairstand_epese	Chairstand subscore for the EPESE SPPB	Score ranges of 0 – 4; scoring details can be found here
fallcount	Fall count	Number of falls that occurred during the prospective year
fallrate	Fall rate	Rate of falls that occurred during the prospective year (number of falls / number of calendar months)
risk_binary	Binary measure of fall risk	0: No falls AND SPPB 10 – 12 1: One or more falls AND SPPB 0 – 6
risk_ternary	Ternary measure of fall risk	0: No falls AND SPPB 10 – 12 1: One or more falls AND SPPB 7 – 9 2: One or more falls AND SPPB 0 - 6
fall_binary	Binary measure of future falls	0: Zero or one falls 1: Two or more falls
fall_ternary	Ternary measure of future falls	0: Zero or one falls 1: Two or three falls 2: Four or more falls
sppb_binary	Binary measure of SPPB categories	0: SPPB 10 – 12 1: SPPB 0 – 6

Table B.4: Continued.

sppb_ternary	Ternary measure of SPPB categories	0: SPPB 10 – 12 1: SPPB 7 – 9 2: SPPB 0 – 6

APPENDIX C: EXPLORATORY DATA ANALYSIS

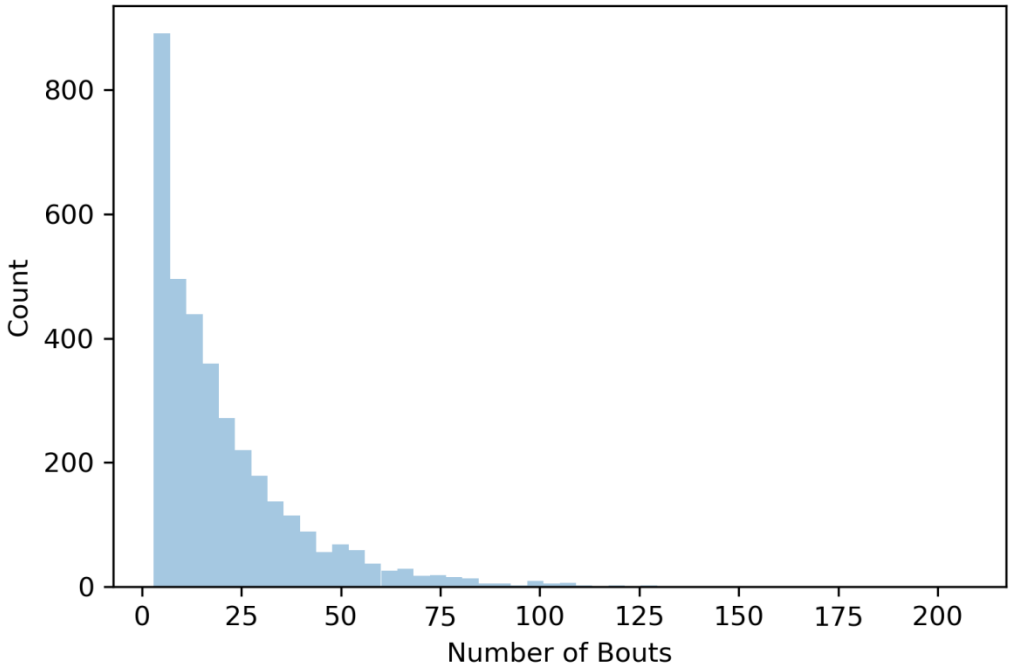

Figure C.1: Distribution of number of walking bouts returned for each individual. The distribution is dominated by individuals who had a low number of walking bouts; 10% of individuals returned fewer than three walking bouts. This substantial right-skew may be indicative of the mainly sedentary activity of the population or could perhaps be an artifact of the good walking bout definition which restricts "good walking" bouts to one minute or more in duration (this is difficult to achieve without leaving the home).

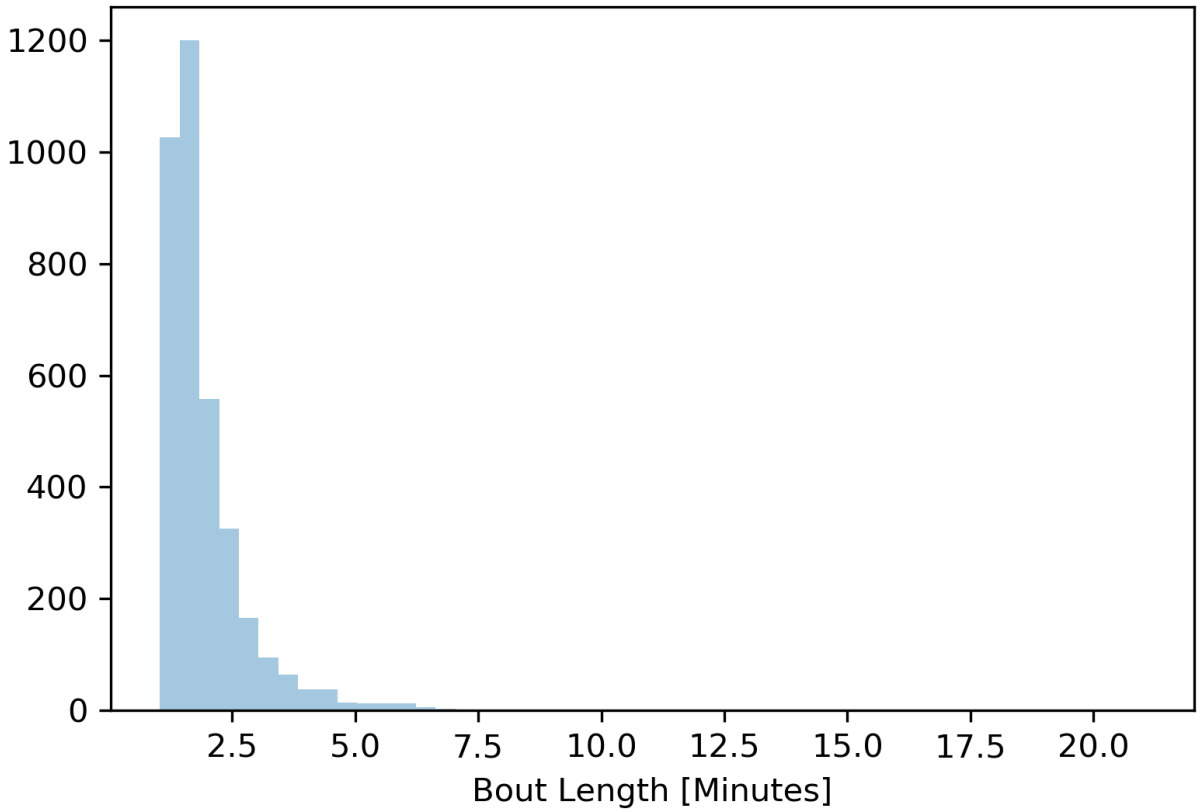

Figure C.2: Distribution of average length of walking bouts returned for each individual. For individuals who had more than one bout, the average length was computed and used in the creation of this plot. The majority of individuals have walking bouts slightly longer than one minute in duration.

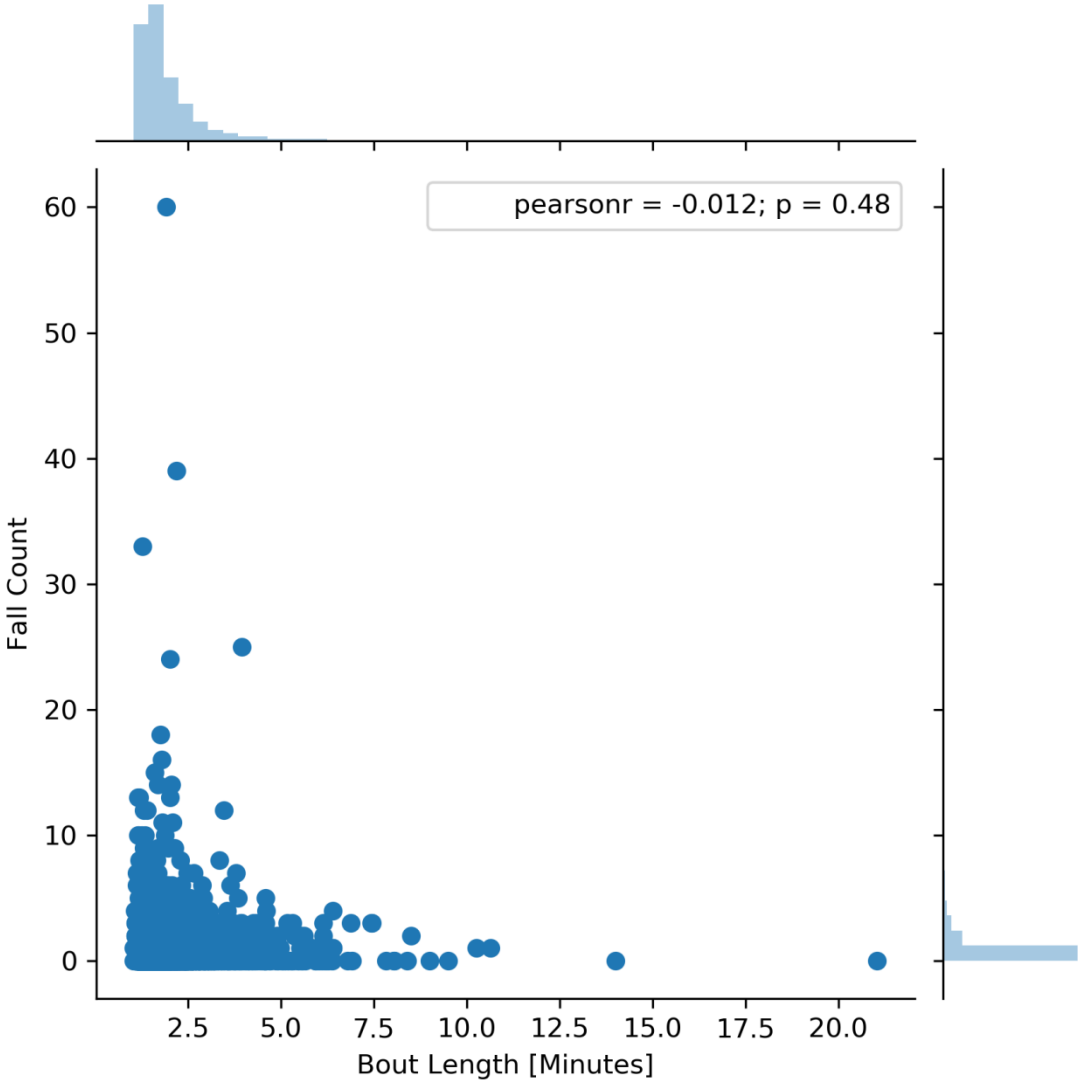

Figure C.3: Joint plot showing the relationship between bout length and falls. The very low correlation with an insignificant p-value might suggest that walking bout duration has little effect on falls. This may be related to the U-shaped curve often cited when discussing fall risk which highlights (1) individuals who walk a lot, are more stable, but have greater opportunities to fall compared to (2) individuals who walk very little but may be less stable and hence, at greater risk of falling when they do walk. This explanation somewhat fits the "corner-shaped" cloud of data points.

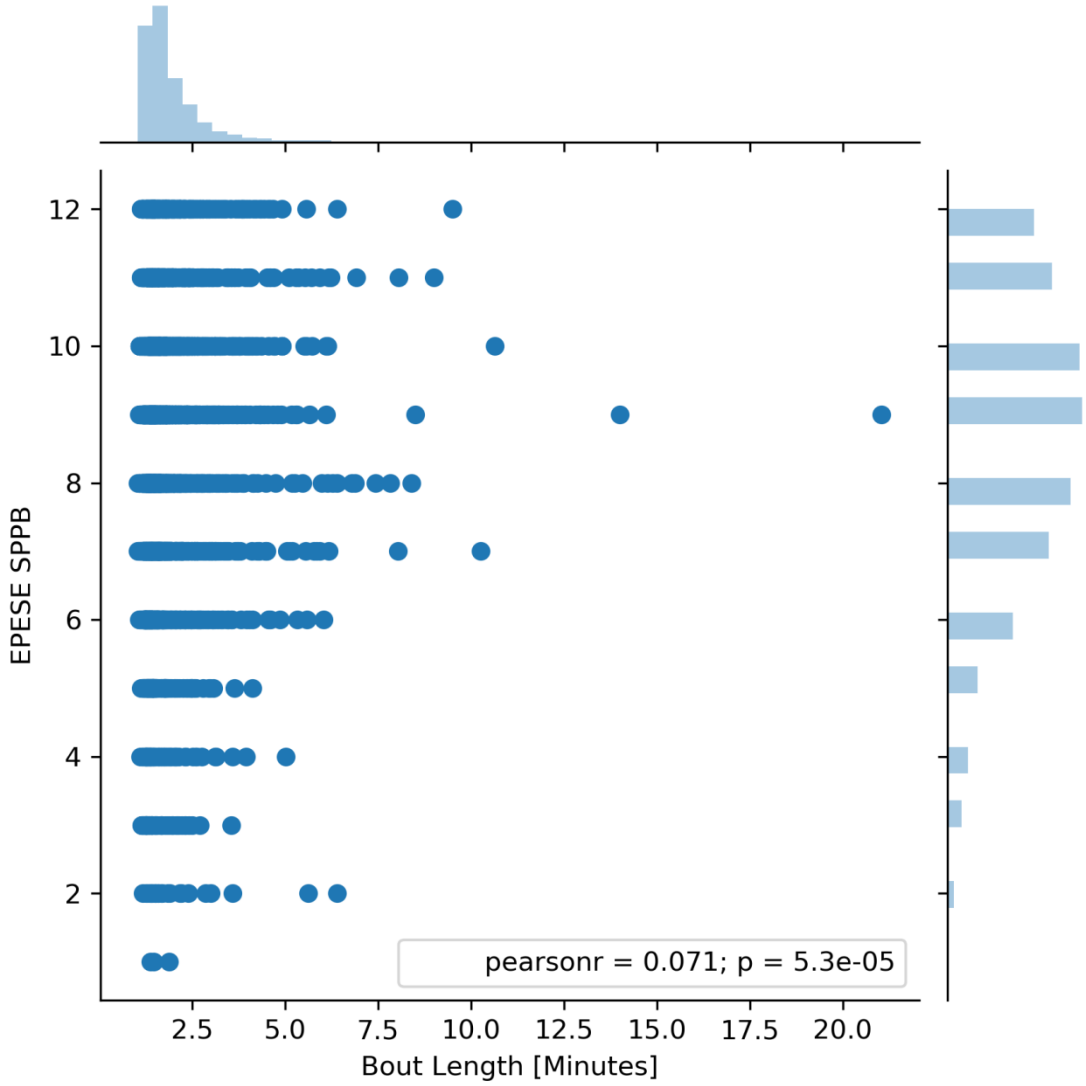

Figure C.4: Jointplot showing the relationship between bout length and SPPB score. We see an insignificant relationship between walking bout duration and physical function as measured by the SPPB. Qualitatively, there appears to be fewer individuals in the low-sppb categories who walk long distances compared to those in the high-sppb categories.

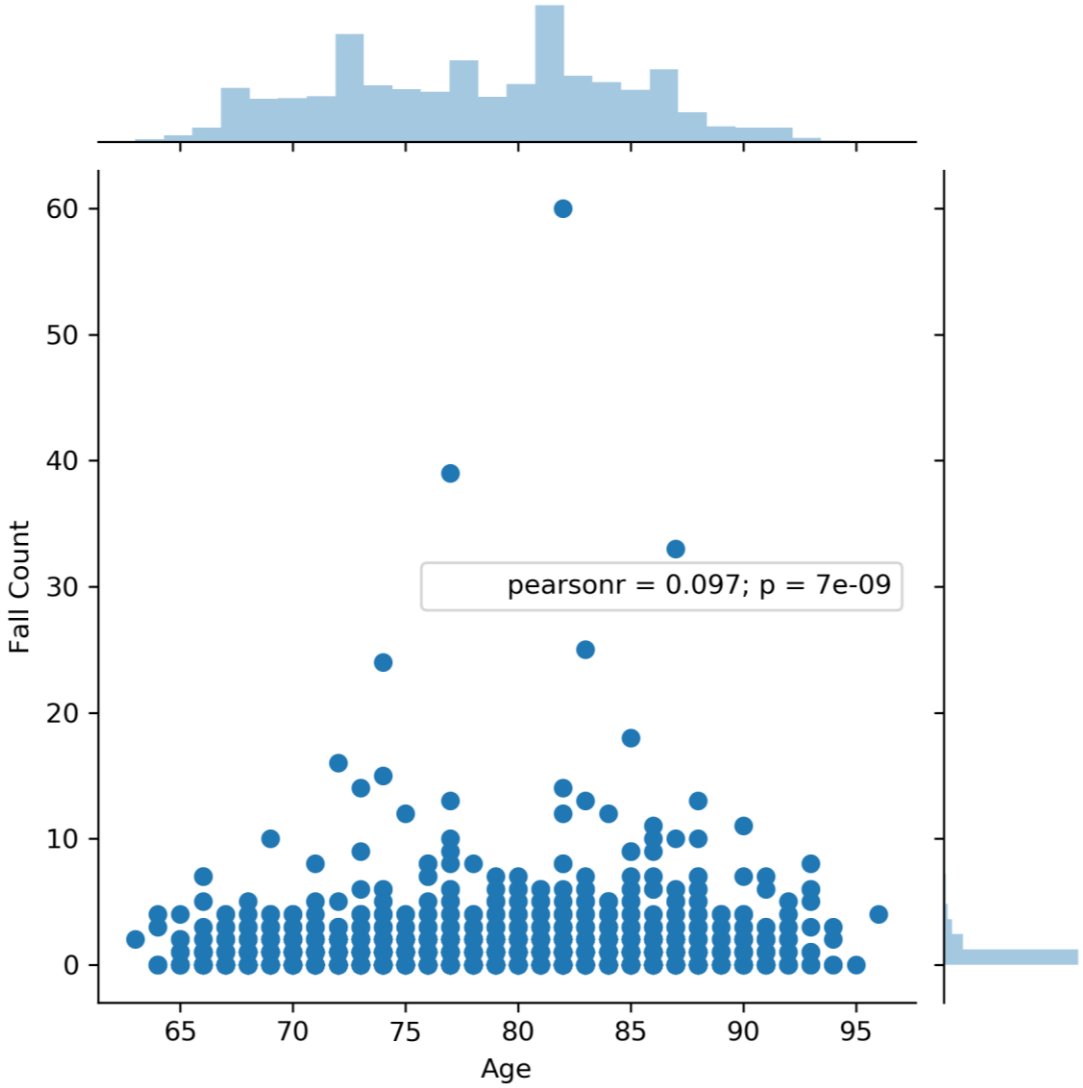

Figure C.5: Jointplot showing the relationship between falls and age. We see a very weak correlation (10%) between age and falls which goes against the general in gerontology that individuals who are older fall more often. It is unclear why this relationship is not seen in this data set.

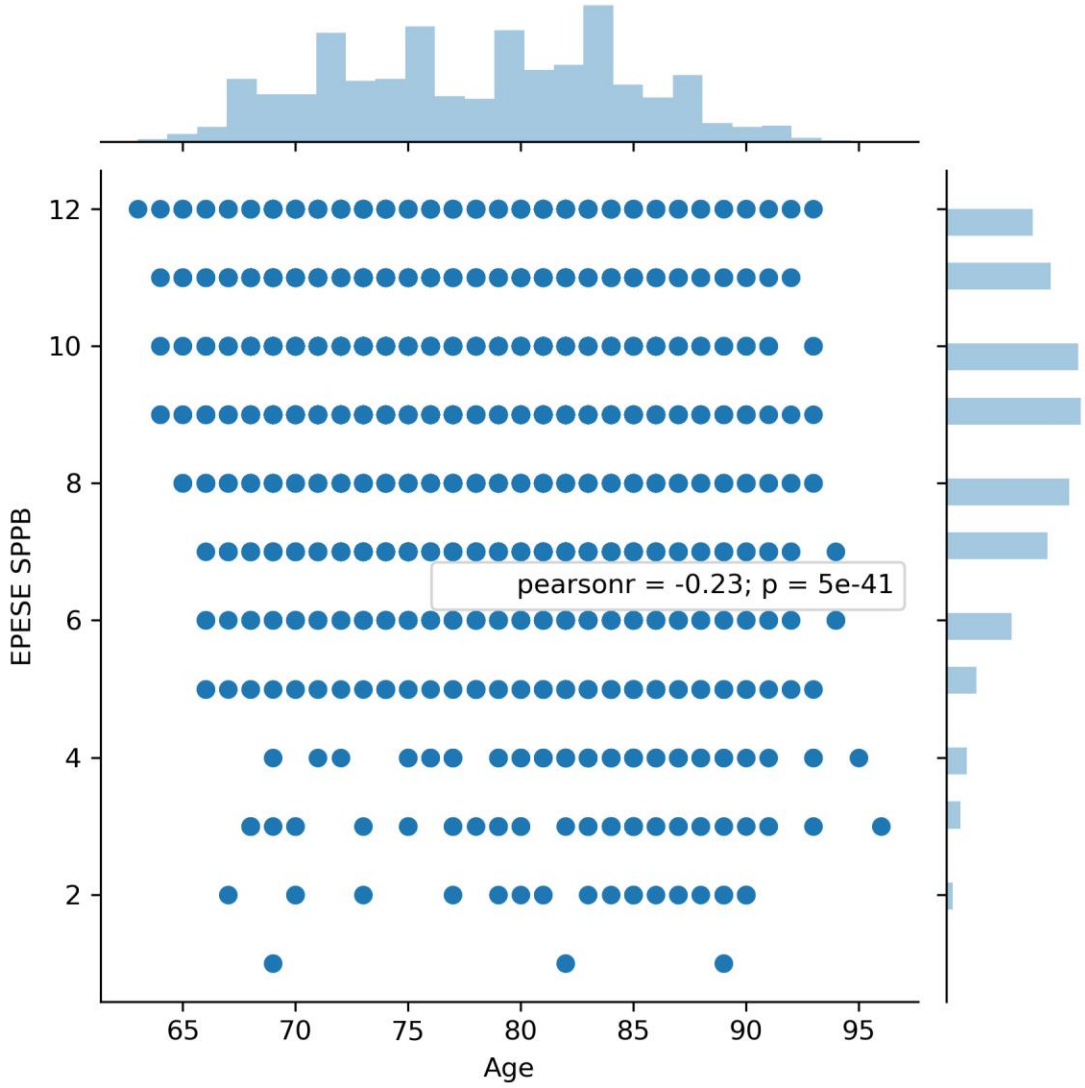

Figure C.6: Jointplot showing the relationship between age and EPESE SPPB score. A weak and negative correlation between SPPB and age (-23%) suggests that as individual's age, physical function decreases. This correlation is much stronger than that seen in age vs. falls (10%) suggesting that SPPB may be a stronger predictor of fall risk and future falls than age. This is somewhat expected given that SPPB directly measures an individual's current level of physical function, balance, and walking ability whereas age prescribes a general level of ability that may not be true for a particular individual.

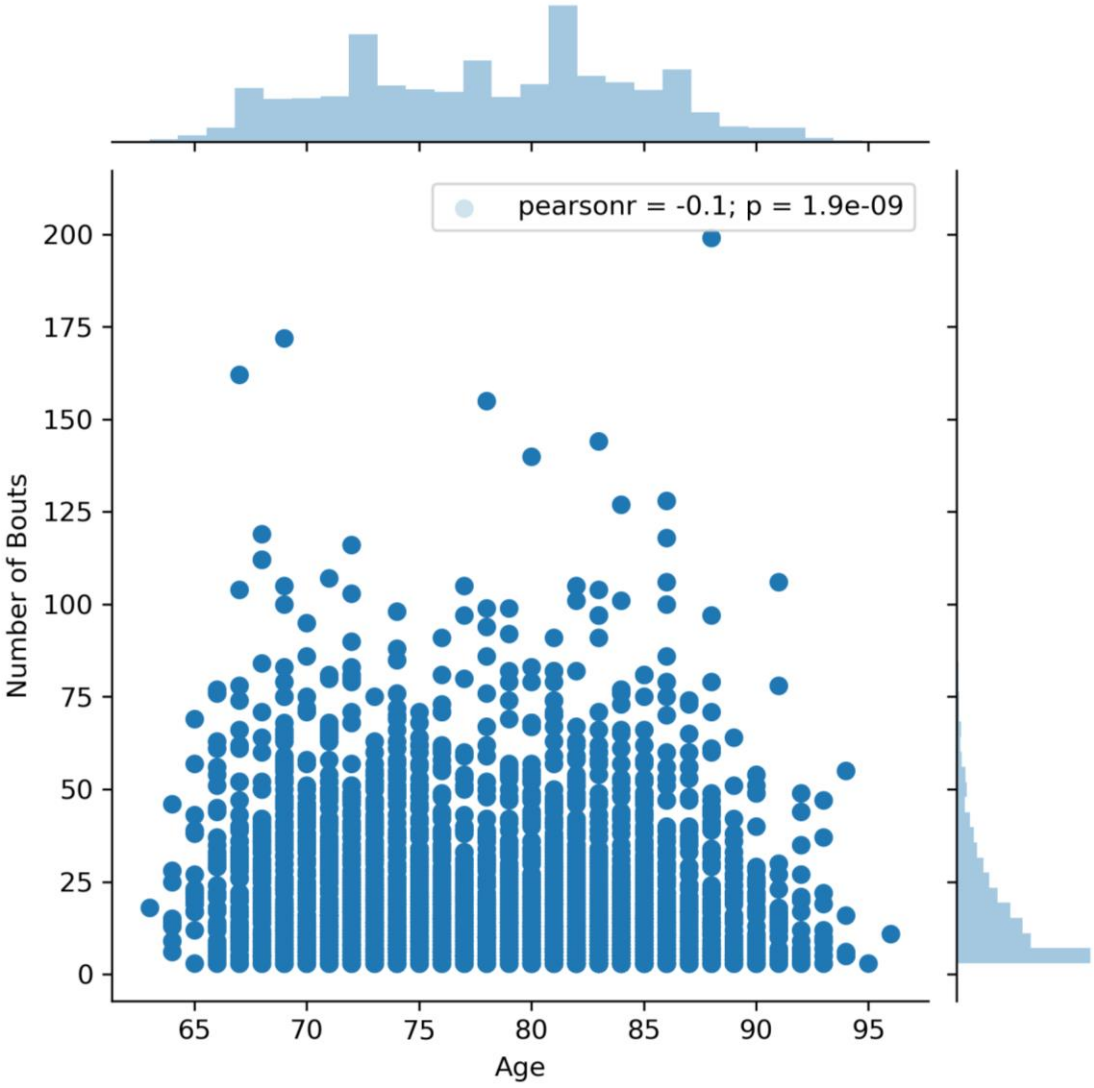

Figure C.7: Jointplot showing the relationship between age and number of good walking bouts. The weak correlation suggests that as individuals age, they generate fewer good walking bouts.

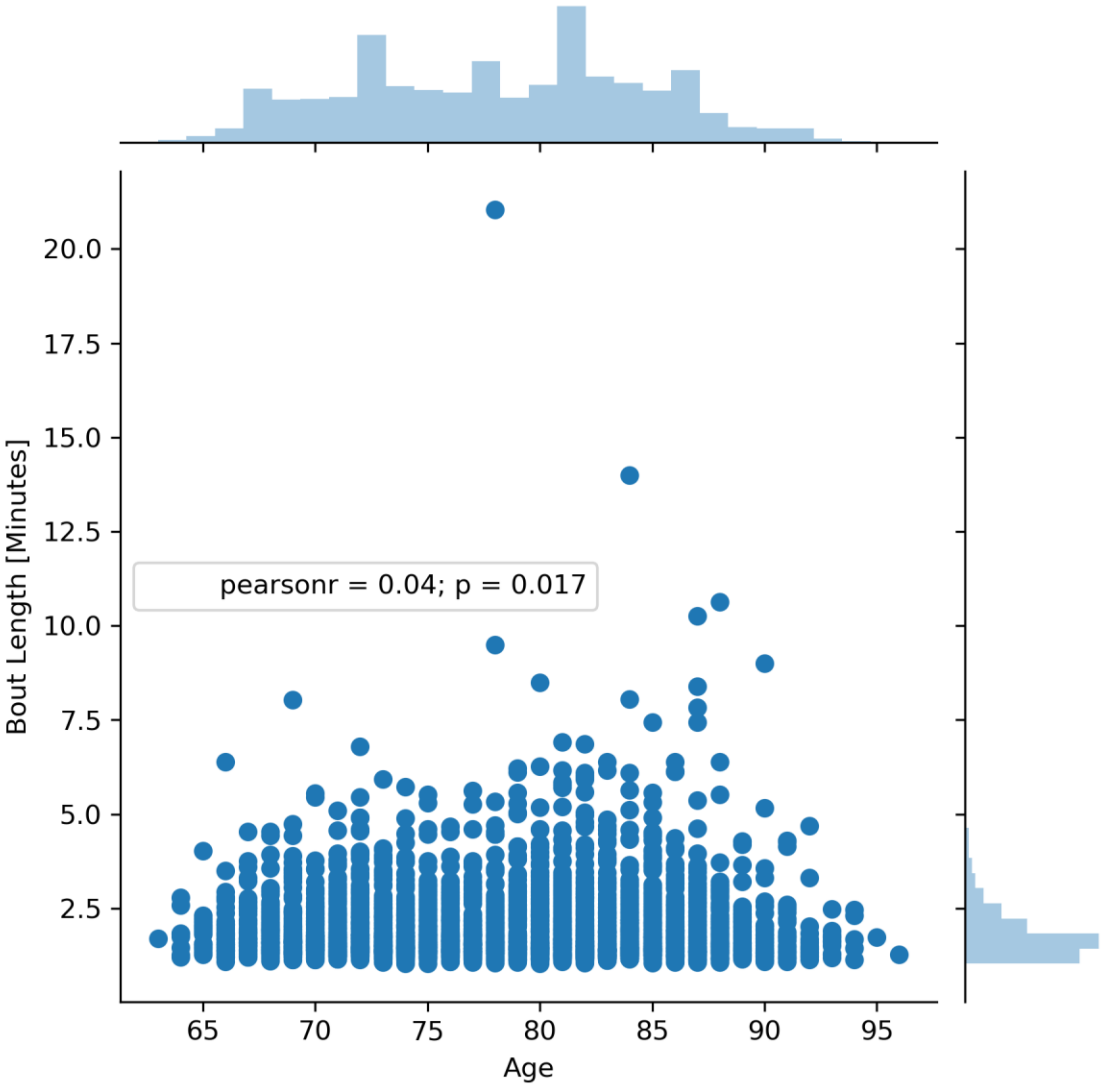

Figure C.8: Jointplot showing the relationship between age and good walking bout duration. There does not appear to be any relationship between age and the duration of good walking bouts.

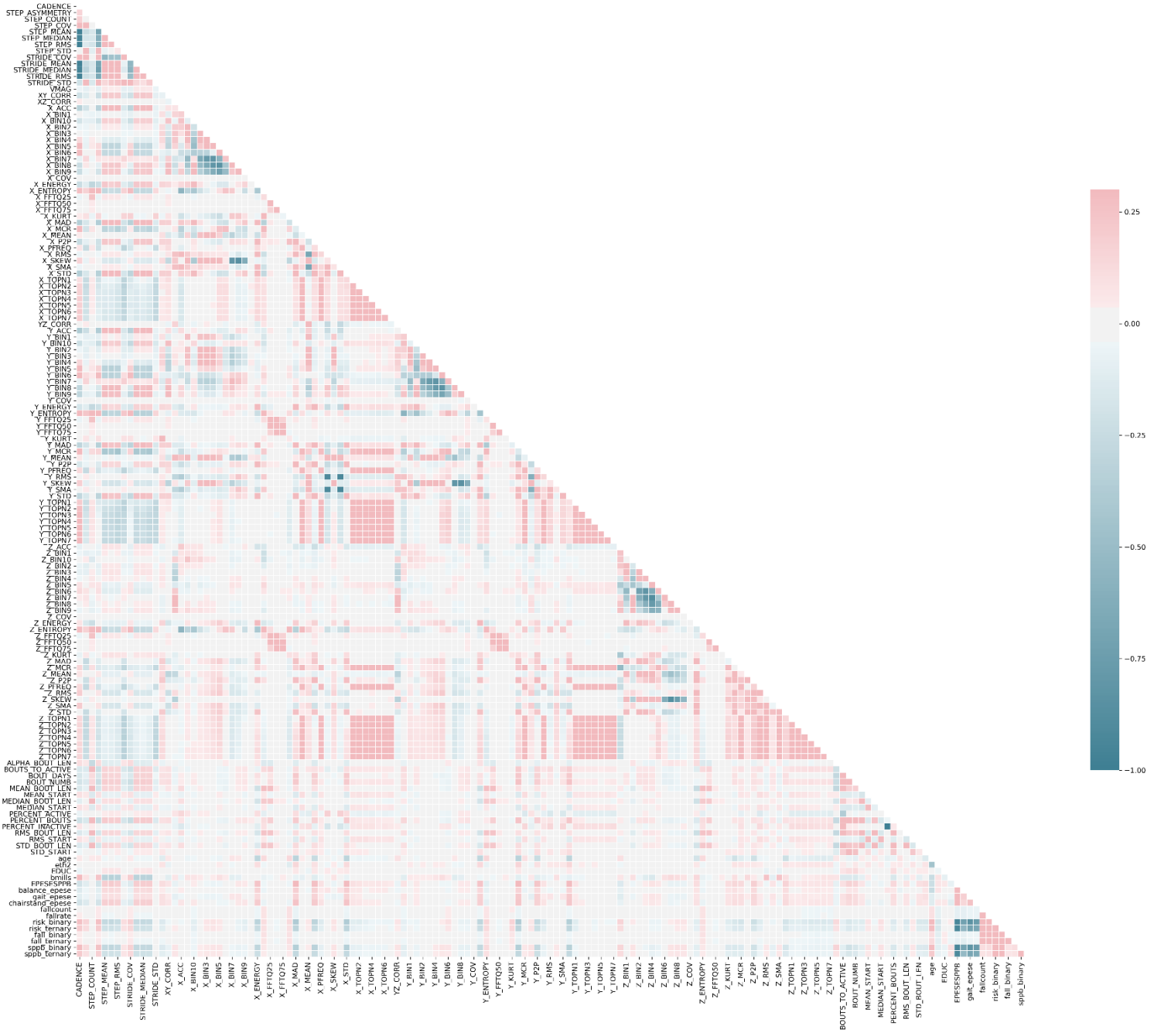

Figure C.9: Heatmap showing the pair-wise correlations between features and outcome measures. Overall, the majority of features show very weak correlations (near zero) with both other features and the outcome measures. The entire set of variables appear to have more negative correlations amongst each other than positive correlations. Blocks of variables (such as the Z-axis TOPN features) all show very similar levels of correlation with each other, but these correlations are never close to 1 suggesting that, while they contain some similar information, these variables are not really the same. The binary and ternary outcome measures show strong correlation with SPPB, fallcount, and fallrate variables as expected given that these outcome measures are derived from combinations of these three variables. To view individual variable names, zoom-in on the figure.

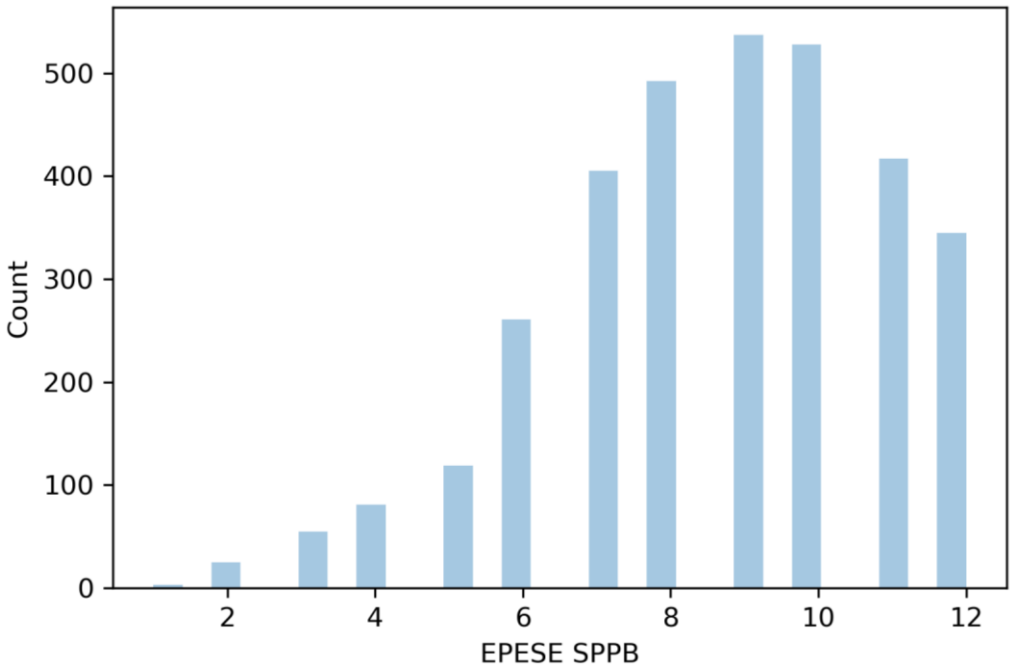

Figure C.10: Distribution of EPESE SPPB scores in the data set. The majority of individuals have scores in the "high-function" range of 8 – 12 with the fewest individuals (only 100) scoring in a "very low functioning" range of 0 – 6. This left-skewed distribution suggests that the population is largely healthy which could bias models toward predicting individuals as low risk or low falls.

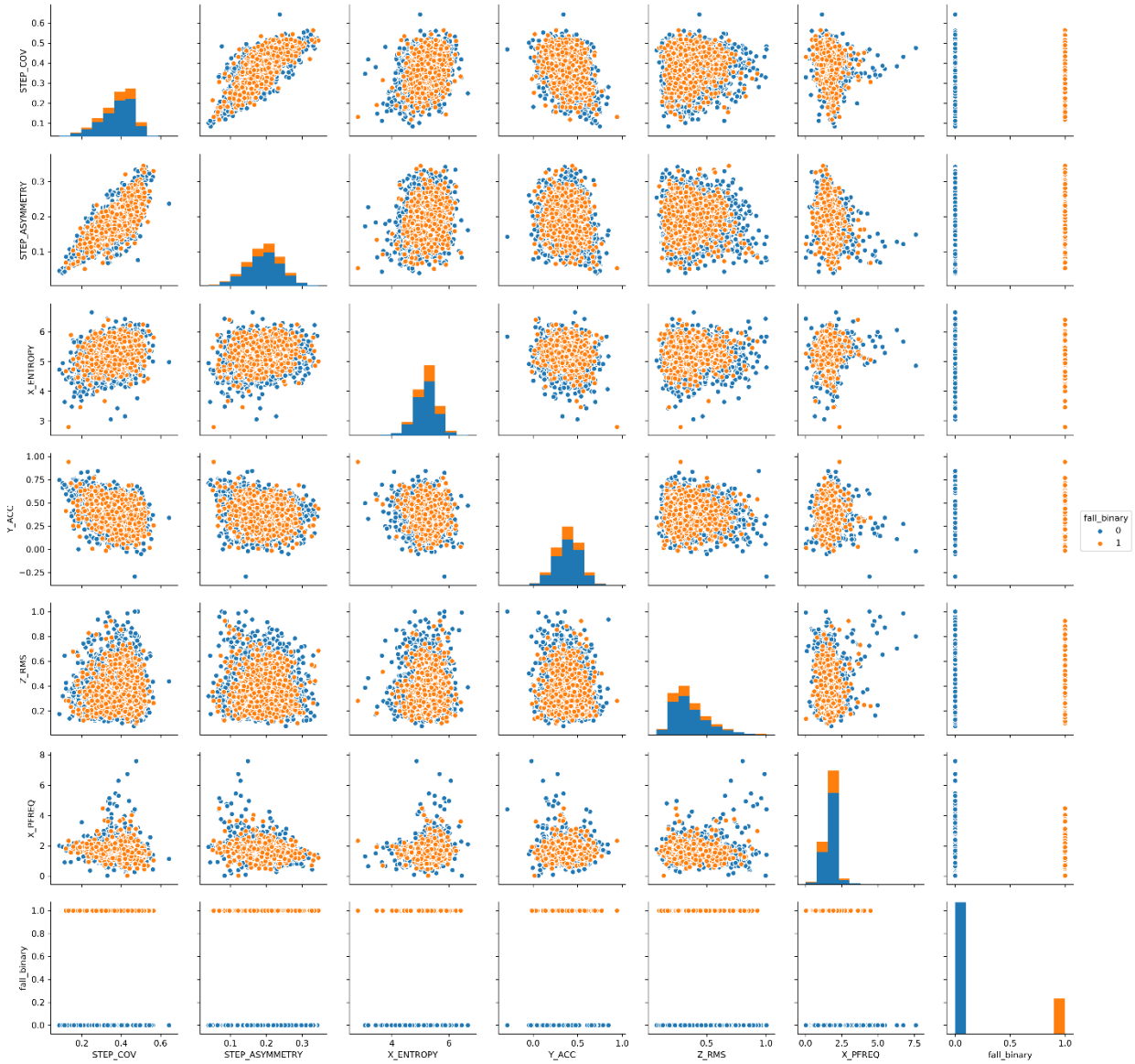

Figure C.11: Pairplot showing the distributions of a subset of features grouped by the binary future fall indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The two risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes; although, small regions of separation can be seen in a subset of instances (e.g. the blue individuals in the Z_RMS vs. X_PFREQ plot). To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = Zero or one falls
1 (Orange) = Two or more falls

Figure C.12: Pairplot showing the distributions of a subset of features grouped by the ternary future fall indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The three risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes; although, small regions of separation can be seen in a subset of instances (e.g. the blue individuals in the Z_RMS vs. X_PFREQ plot). A small degree of mean-shift can be seen in the X_PFREQ variable which could be useful for class separation when projected to higher dimensions. To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = Zero or one falls
1 (Orange) = Two or three falls
2 (Green) = Four or more falls

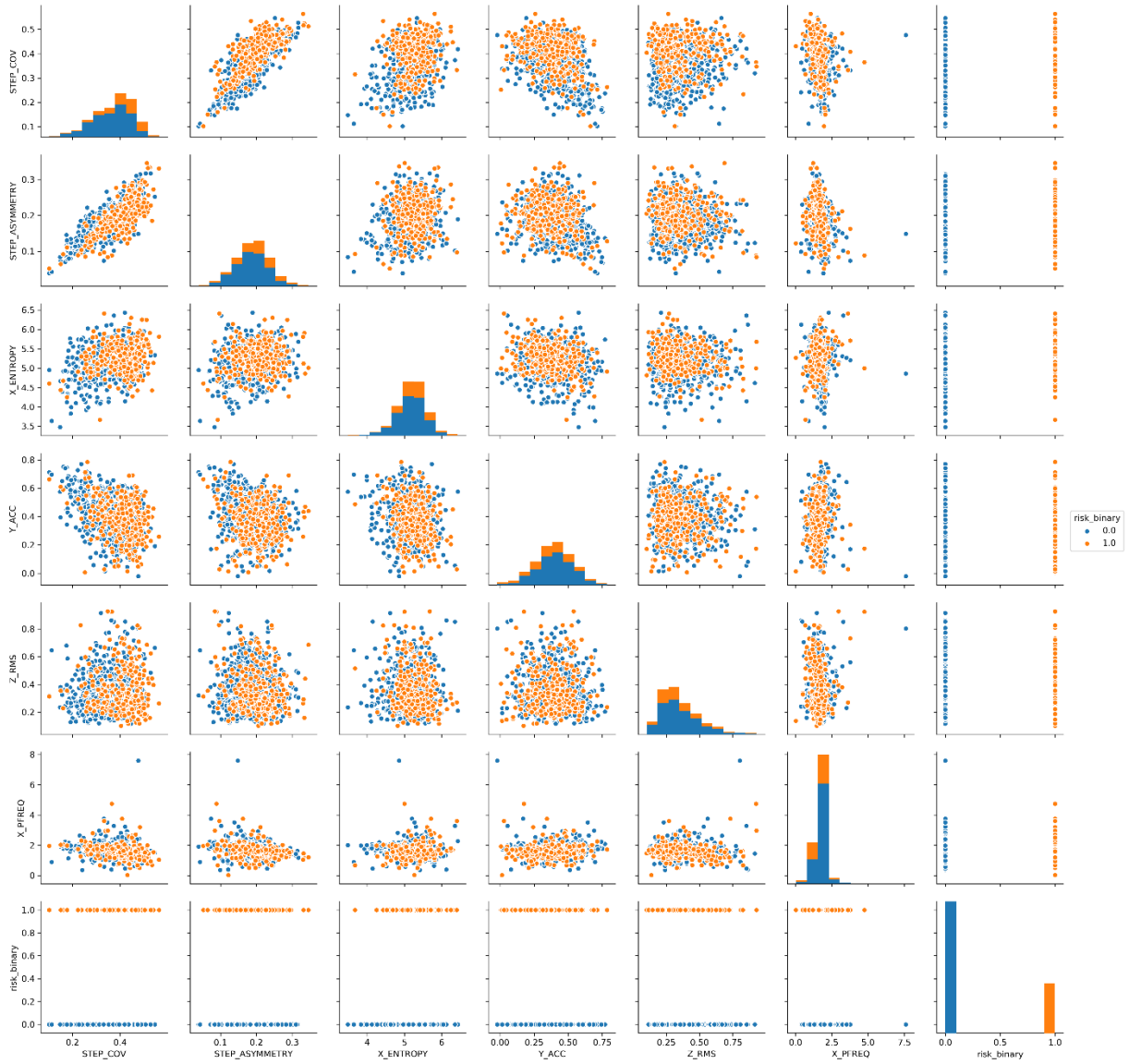

Figure C.13: Pairplot showing the distributions of a subset of features grouped by the binary fall risk indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The two risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes. A small degree of mean-shift can be seen in the X_ENTROPY variable which could be useful for class separation when projected to higher dimensions. To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = Zero falls and SPPB 10 – 12
1 (Orange) = One or more falls and SPPB 0 – 6

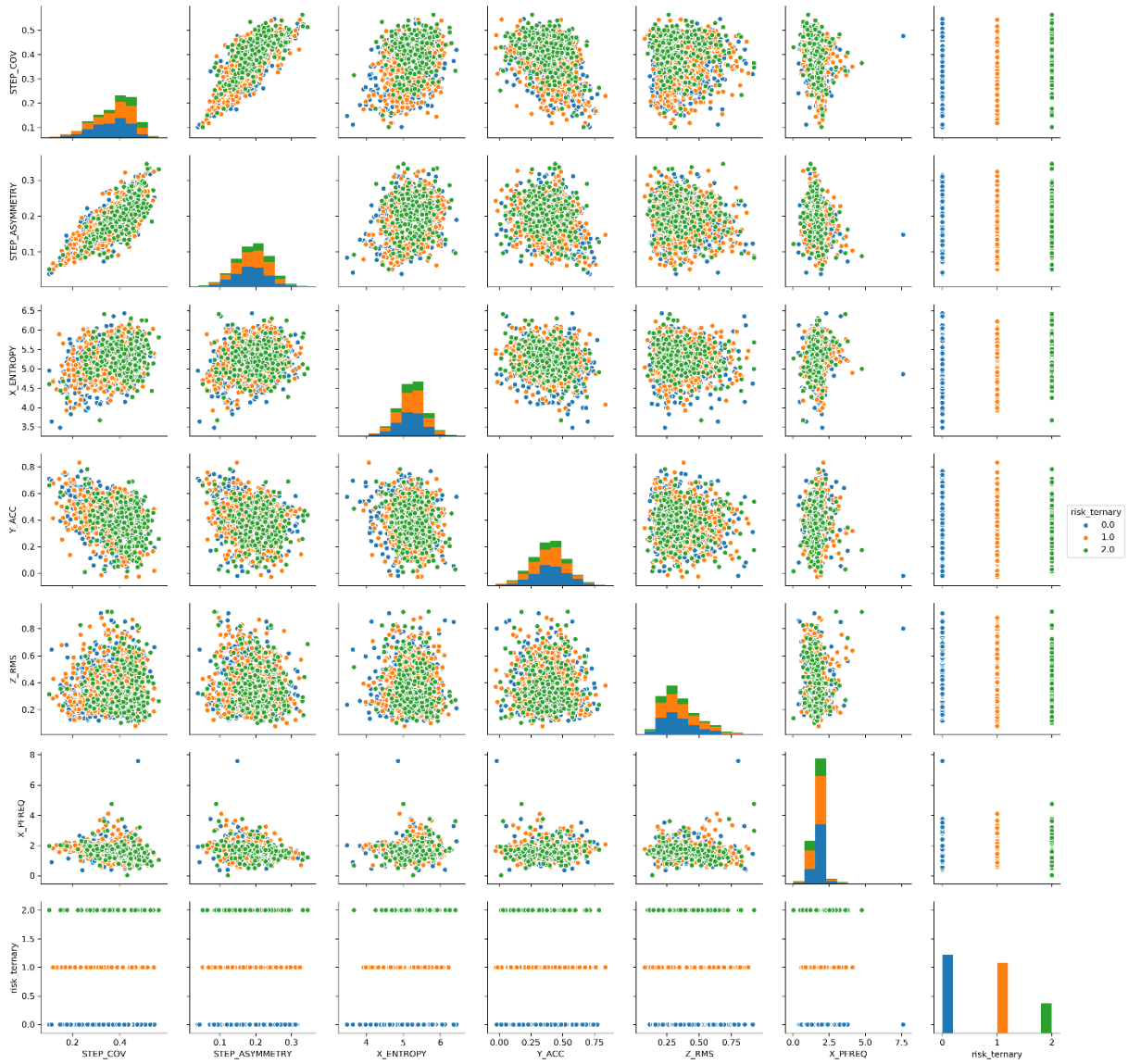

Figure C.14: Pairplot showing the distributions of a subset of features grouped by the ternary fall risk indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The three risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes. A small degree of mean-shift can be seen in the X_ENTROPY variable which could be useful for class separation when projected to higher dimensions. To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = Zero falls and SPPB 10 – 12
1 (Orange) = One or more falls and SPPB 7 – 9
2 (Green) = One or more falls and SPPB 0 – 6

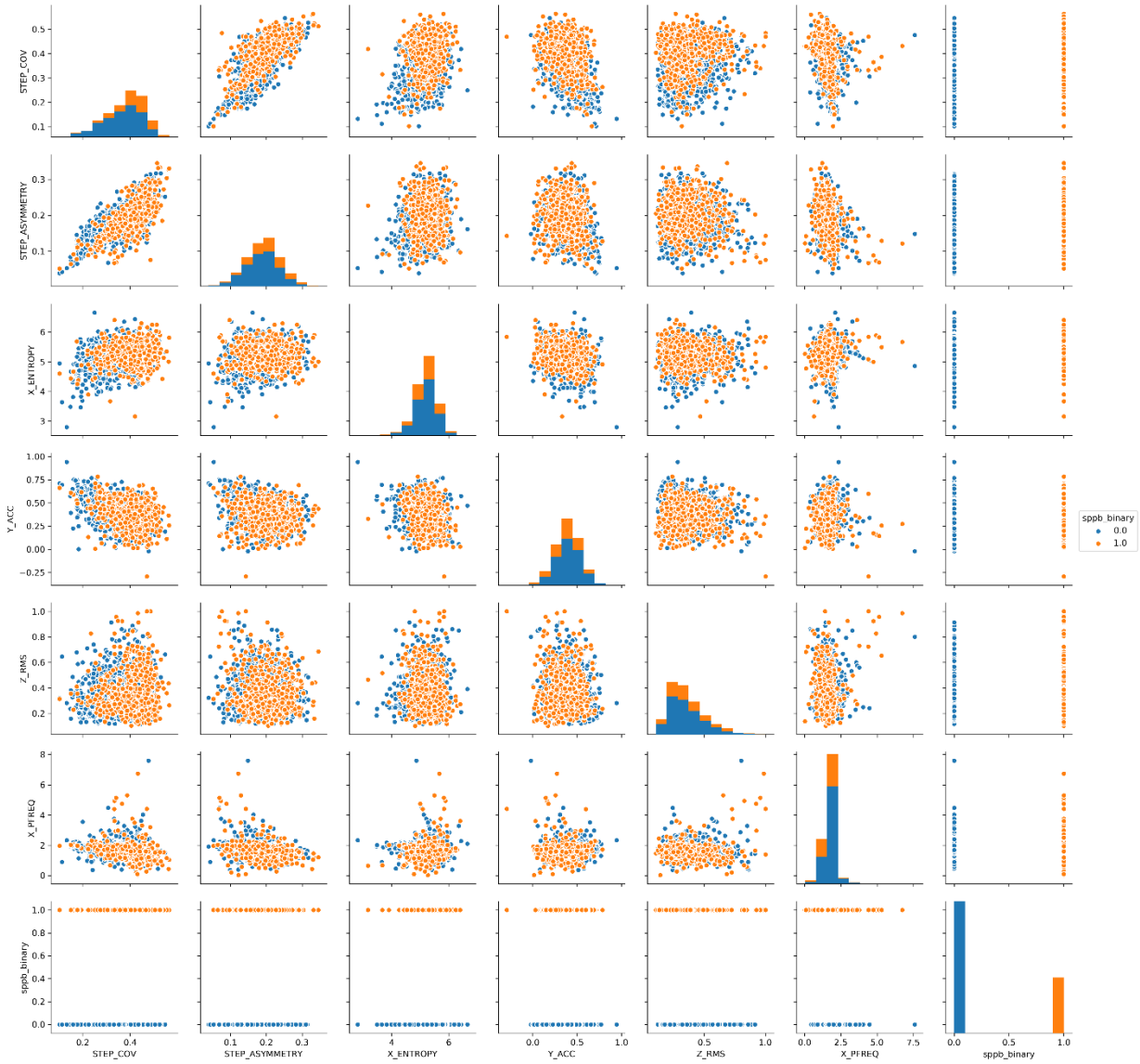

Figure C.15: Pairplot showing the distributions of a subset of features grouped by the binary SPPB indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The two risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes. Ignoring two outliers, a small degree of mean-shift can be seen in the X_PFREQ variable which could be useful for class separation when projected to higher dimensions. To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = SPPB 10 – 12
1 (Orange) = SPPB 0 – 6

Figure C.16: Pairplot showing the distributions of a subset of features grouped by the ternary SPPB indicator. Each individual plot shows instances plotted along two axes which show the values of two features. The class of each instance is indicated by the color of the marker. The three risk groups show a high degree of overlap in the subset of displayed features which suggests that these variables may not be useful for separating classes. To view individual variable names, zoom-in on the figure.

Groups:
0 (Blue) = SPPB 10 – 12
1 (Orange) = SPPB 7 – 9
2 (Green)= SPPB 0 – 6

The following figures (C.17 and C.18) illustrate classification performance for three-way "falls" and "SPPB" categories using a modified definition for assigning class labels to each individual.

Falls class 0:

a) Individuals with 0 – 1 falls
b) Individuals with 2 – 3 falls who are in the lower third of individuals by number of good walking bouts

Falls class 1:

a) Individuals with 2 – 3 falls who are in the middle third of individuals by number of good walking bouts

Falls class 2:

b) Individuals with 4+ falls
c) Individuals with 2 – 3 falls who are in the upper third of individuals by number of good walking bouts

SPPB class 0:

c) Individuals with SPPB scores of 10 – 12
d) Individuals with SPPB scores of 7 – 9 who are in the lower third of individuals by number of good walking bouts

SPPB class 1:

a) Individuals with SPPB scores of 7 – 9 who are in the middle third of individuals by number of good walking bouts

SPPB class 2:

d) Individuals with SPPB scores of 1 – 6
e) Individuals with SPPB scores of 7 – 9 who are in the upper third of individuals by number of good walking bouts

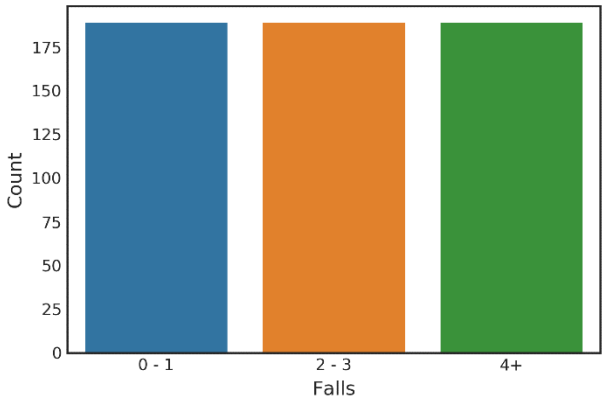

Figure C.17.A: Falls class distribution for a balanced sample generated by under sampling the majority classes. All three classes have equal representation with 188 individuals.

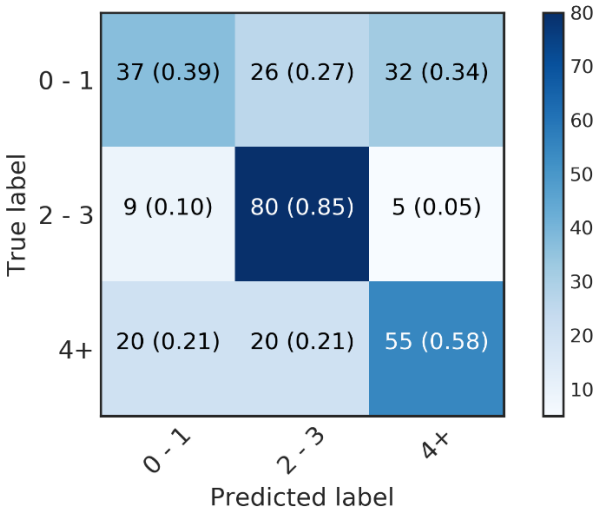

Figure C.17.B: Confusion matrix for prediction of fall categories in the balanced testing set. Very strong bias for predicting the middle class with slightly better than random chance for predicting the extremes.

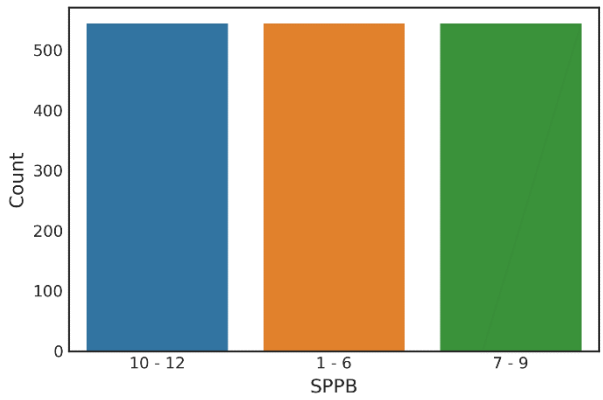

Figure C.18.A: SPPB class distribution for a balanced sample generated by under sampling the majority classes. All three classes have equal representation with 538 individuals.

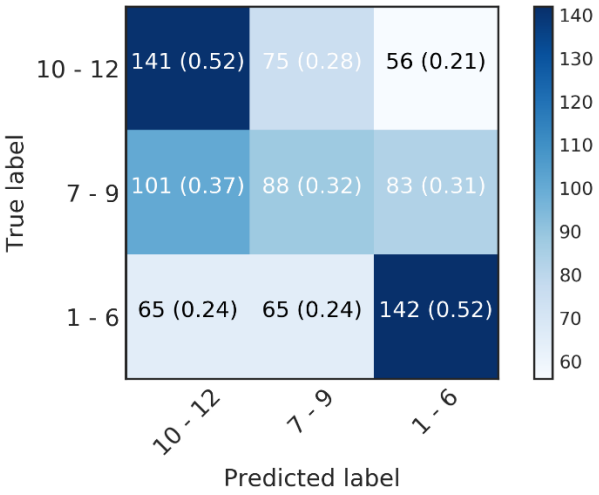

Figure C.18.B: Confusion matrix for prediction of SPPB categories in the balanced testing set. Middle class prediction remains very close to random chance with bias toward predicting the extreme classes.

The figures above summarize the results of predicting falls and SPPB using a modified approach to placing individuals in each category (see text above figures for more information). Looking at fall prediction, we find that distributing middle-falls individuals into the extremes based upon number of good walking bouts results in heavily biased prediction of the “middle” falls category even with the use of a balanced data set; compare this to the original three-way fall prediction

models (Figure 10B) which favored classifying individuals into the “low” and “high” categories. Prediction of SPPB categories using the modified labeling scheme shows bias toward predicting the extreme “10 - 12” and “1 - 6” categories which is nearly identical to the previous performance seen in Figure 18B. Ultimately, using walking bouts to adjust fall prediction does not improve accuracy but rather heavily biases the model toward classifying everyone as “2 – 3” fallers. This same adjustment does not change prediction of SPPB categories.

The following figures (C.19 and C.20) illustrate classification performance for three-way "falls" and "SPPB" categories using a modified definition for assigning class labels to each individual.

Falls class 0:

e) Individuals with 0 – 1 falls
f) Individuals with 2 – 3 falls who are in the upper third of individuals by number of good walking bouts

Falls class 1:

f) Individuals with 2 – 3 falls who are in the middle third of individuals by number of good walking bouts

Falls class 2:

g) Individuals with 4+ falls
h) Individuals with 2 – 3 falls who are in the lower third of individuals by number of good walking bouts

SPPB class 0:

g) Individuals with SPPB scores of 10 – 12
h) Individuals with SPPB scores of 7 – 9 who are in the upper third of individuals by number of good walking bouts

SPPB class 1:

b) Individuals with SPPB scores of 7 – 9 who are in the middle third of individuals by number of good walking bouts

SPPB class 2:

i) Individuals with SPPB scores of 1 – 6
j) Individuals with SPPB scores of 7 – 9 who are in the lower third of individuals by number of good walking bouts

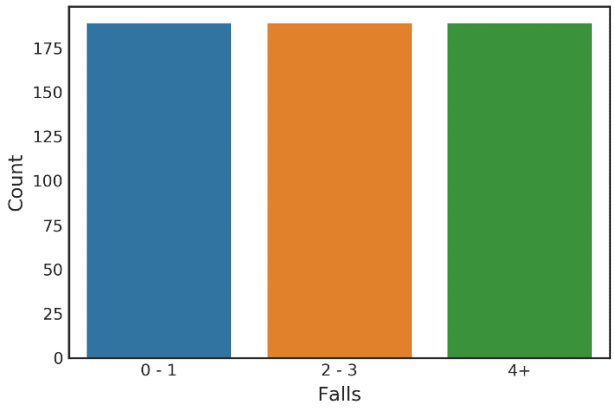

Figure C.19.A: Falls class distribution for a balanced sample generated by under sampling the majority classes. All three classes have equal representation with 188 individuals.

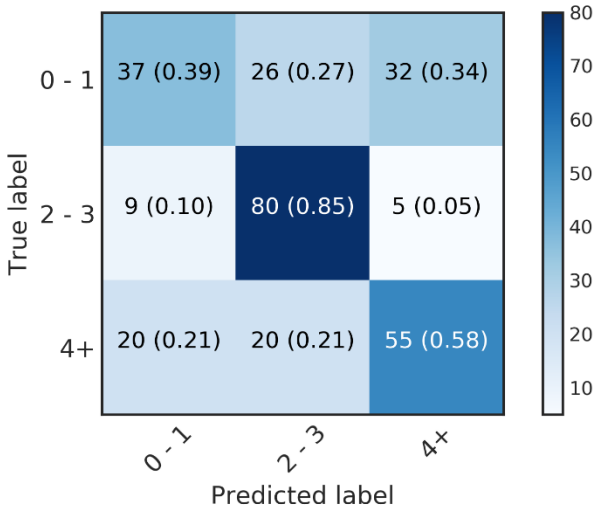

Figure C.19.B: Confusion matrix for prediction of fall categories in the balanced testing set. Very strong bias for predicting the middle class with slightly better than random chance for predicting the extremes.

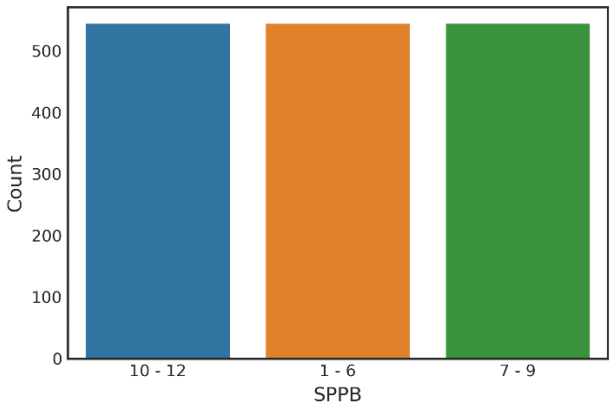

Figure C.20.A: SPPB class distribution for a balanced sample generated by under sampling the majority classes. All three classes have equal representation with 538 individuals.

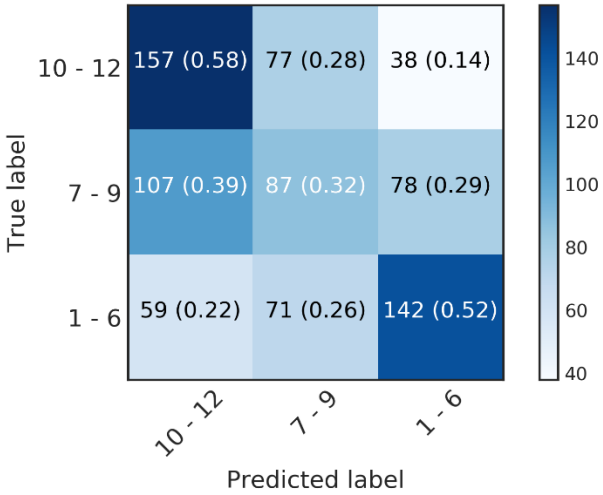

Figure C.20.B: Confusion matrix for prediction of SPPB categories in the balanced testing set. Middle class prediction remains very close to random chance with bias toward predicting the extreme classes.

The figures above summarize the results of predicting falls and SPPB using a modified approach to placing individuals in each category (see text above figures for more information). Looking at fall prediction, we find that distributing middle-falls individuals into the extremes based upon number of good walking bouts results in heavily biased prediction of the “middle” falls category even with the use of a balanced data set; compare this to the original three-way fall prediction

models (Figure 10B) which favored classifying individuals into the “low” and “high” categories. Prediction of SPPB categories using the modified labeling scheme shows bias toward predicting the extreme “10 - 12” and “1 - 6” categories which is nearly identical to the previous performance seen in Figure 18B. Ultimately, using walking bouts to adjust fall prediction does not improve accuracy but rather heavily biases the model toward classifying everyone as “2 – 3” fallers. This same adjustment does not change prediction of SPPB categories.

Table C.1: Number of individuals with N free-living "good walking" bouts.

Number of Good Walking Bouts	Number of Individuals
0	504
1	201
2	211
3	200
4	172
5	185
6	153
7	181
8	122
9	125
10	141
11	108
12	108
13	131
14	107
15	93
16	99
17	101
18	76
19	83
20	76
21	68
22	71
23	57
24	57
25	58
26	55
27	50
28	48
29	46
30	41
31	44
32	35
33	42
34	35
35	25
36	29
37	33

Table C.1: Continued.

38	22
39	31
40	25
41	26
42	19
43	19
44	11
45	13
46	13
47	19
48	21
49	17
50	16
51	14
52	17
53	13
54	12
55	9
56	8
57	13
58	8
59	6
60	10
61	6
62	8
63	7
64	5
65	5
66	7
67	7
68	10
69	3
70	4
71	9
72	2
73	3
74	6
75	6
76	4

Table C.1: Continued.

77	3
78	2
79	7
80	3
81	5
82	4
83	3
84	1
85	1
86	3
88	1
90	1
91	3
92	1
94	1
95	1
97	4
98	1
99	2
100	2
101	2
103	1
104	2
105	3
106	2
107	1
112	2
116	1
118	1
119	1
127	1
128	1
140	1
144	1
155	1
162	1
172	1
199	1
207	1

Table C.2: The number of bouts extracted for a set of 100 individuals using two different definitions of "good walking" bouts. The modified definition of walking bouts produced a false positive rate of 2.7%.

Original: Minimum bout length of one minute and a 30 second maximum pause between bouts.
Modified: Minimum bout length of 30 seconds and a 10 second maximum pause between bouts.

PID	*Number of Bouts (Original)*	*Number of Bouts (Modified)*
P10001	40	102
P10002	7	37
P10003	16	27
P10004	0	9
P10005	1	21
P10006	9	21
P10007	5	77
P10008	25	46
P10009	68	187
P10010	1	17
P10011	20	77
P10012	8	30
P10014	82	216
P10015	38	109
P10016	0	8
P10017	18	72
P10018	12	33
P10020	13	55
P10022	5	35
P10023	26	99
P10024	26	95
P10026	4	24
P10027	22	74
P10028	5	35
P10029	19	52
P10030	0	11
P10031	11	42
P10032	4	16
P10033	17	53
P10034	0	8
P10035	0	12
P10036	27	88
P10037	19	87
P10039	5	13
P10040	16	41

Table C.2: Continued.

P10043	*10*	*44*
P10045	17	43
P10046	14	33
P10047	14	75
P10049	2	26
P10050	37	91
P10052	1	14
P10053	34	152
P10054	75	237
P10055	22	63
P10057	0	2
P10058	7	29
P10059	12	37
P10060	7	34
P10061	6	38
P10062	24	64
P10063	105	183
P10064	32	83
P10065	3	23
P10066	6	29
P10067	16	68
P10068	7	50
P10069	58	87
P10070	5	58
P10071	9	61
P10072	73	155
P10073	11	61
P10074	34	94
P10075	0	11
P10076	14	101
P10077	17	52
P10078	7	34
P10079	11	24
P10080	3	18
P10081	36	147
P10082	3	28
P10083	12	35
P10084	3	17
P10085	30	85
P10086	22	67
P10087	31	71

Table C.2: Continued.

P10088	*22*	*69*
P10091	21	61
P10092	4	25
P10093	0	20
P10094	39	86
P10095	27	72
P10098	9	37
P10099	24	90
P10100	0	19
P10101	21	68
P10102	8	44
P10103	27	60
P10104	41	82
P10105	16	43
P10106	3	55
P10108	3	15
P10109	3	22
P10110	14	64
P10112	7	36
P10113	11	29
P10115	10	58
P10116	14	45
Total	1723	5653
Mean	17.58163265	57.6837
Median	12	45.5
Standard Deviation	19.05549328	44.4236
Average Increase in Bouts Per Individual	40.1020400	

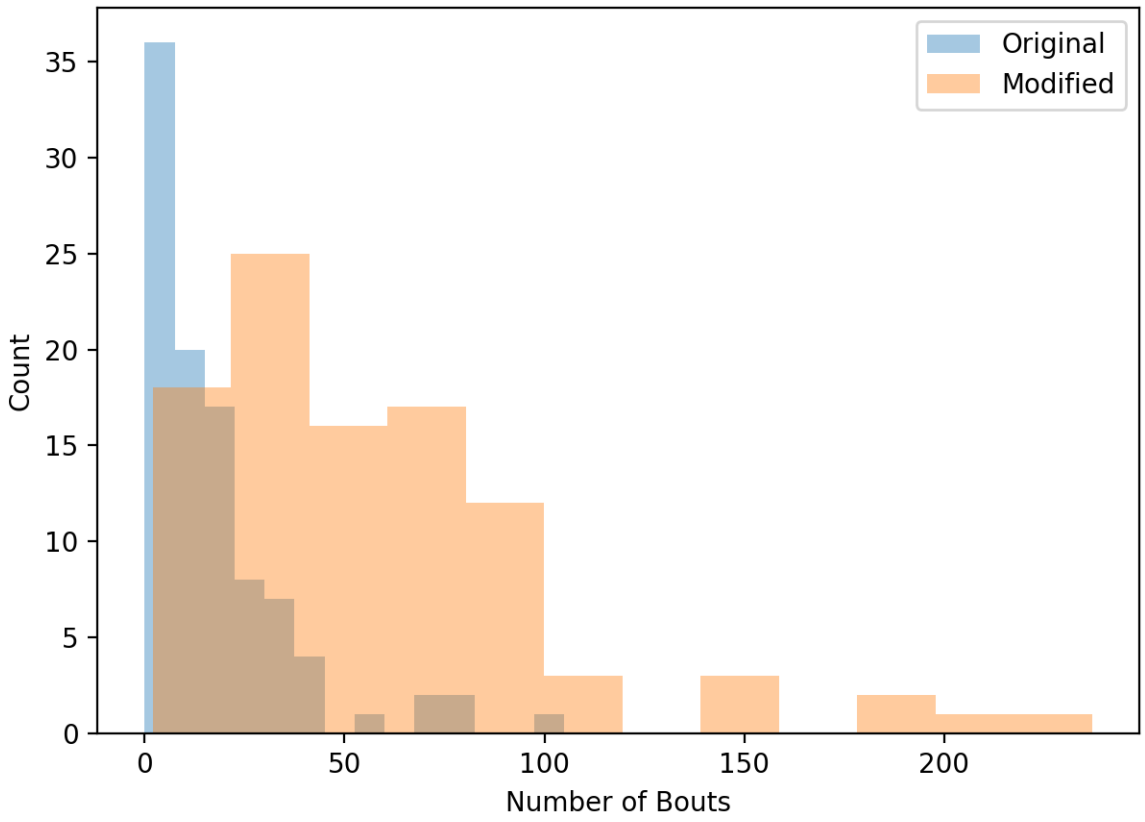

Figure C.21: Distribution of good walking bouts extracted from the 100 individuals outlined in *Table C.2* using two definitions of a "good walking" bout. *Original (blue):* Minimum bout length of one minute and a 30 second maximum pause between bouts. *Modified (orange):* Minimum bout length of 30 seconds and a 10 second maximum pause between bouts. A substantial increase in the number of good walking bouts is seen when using the modified definition which places less aggressive restrictions on the duration and continuity of a good walking bout.