Spring 5-15-2022

Validation, Categorizing, and Prediction of Upper Limb Outcomes after Stroke

Jessica Barth
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Interdisciplinary Program in Movement Science Program

Physical Therapy Department

Dissertation Examination Committee:
Catherine E. Lang
Linda Van Dillen
Keith Lohse
Laura McPherson
Kerri Morgan

Validation, Categorizing, and Prediction of Upper Limb Outcomes after Stroke

by

Jessica Barth

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADLs= activities of daily living

AIC= Akaike Information Criteria

ANOVA= analysis of variance

ARAT= Action Research Arm Test

CART= classification and regression tree

CCR= correct classification rate

CI= confidence interval

ICF= International Classification of Function

MANOVA= multivariate analysis of variance

MRC= Medical Research Council

MRI= magnetic resonance imaging

NIHSS= National Institutes of Health Stroke Scale

NPV= Negative predictive value

PPV= Positive predictive value

PREP= Predicting Recovery Potential

SAFE= Shoulder Abduction Finger Extension

TMS= transcranial magnetic stimulation

UEFM= Upper Extremity Fugl-Meyer

UE= upper extremity

UL= upper limb

US= United States

WHO= World Health Organization

Acknowledgments

I would like to express my deepest appreciation and gratitude to my primary mentor, Dr. Catherine Lang, for her exceptional mentorship, unparalleled support, and belief in my abilities over these past four years. Catherine has invested a substantial amount of time and effort into fostering my training and career development. Her unwavering guidance and insightful suggestions have pushed me to sharpen my thinking and have brought my research to a higher level. She has always encouraged me to take chances and explore new opportunities. Catherine, I sincerely thank you for everything, and I look forward to continuing to learn and grow with you for years to come.

I would like to thank the members of my dissertation committee for their time and their contributions to the projects in this dissertation and my training. Specifically, I would like to thank Dr. Keith Lohse for his patience and guidance as I learned the advanced statistical analysis methods for these projects. I cannot imagine navigating the complex world of machine learning with a better guide. Dr. Linda Van Dillen, Dr. Laura McPherson, and Dr. Kerri Morgan have offered ingenious feedback identifying the clinical interpretation of these projects. Thank you for helping me link these findings to improving outcomes of people with health conditions.

I would like to thank the Movement Science Program faculty, staff, and students for supporting my academic endeavors and providing an exceptional training environment, even during a global pandemic. I have a deep appreciation for the high standards of leadership and collaboration. I am sincerely thankful for the opportunity to study at Washington University in St. Louis.

I would like to thank the directors of the TL1 Clinical Research Training Program for their leadership and the opportunity to train in their interdisciplinary program. I sincerely thank

you for supporting me in completing the Masters of Science of Clinical Investigation (MSCI) and the Certificate of Dissemination and Implementation science.

To the current and former members of the Neurorehabilitation Research Laboratory: thank you for your feedback, support, research assistance, and brainstorming sessions. To the past and present students of the Movement Science Program: thank you for proofreading my documents, debugging my code, and being my friends. You all have become my second family while I have been so far from home.

Lastly, I want to thank my family and friends for their love and unconditional support throughout this entire process. This professional milestone would not have been possible without their support and guidance. Mom, thank you for being my biggest cheerleader, for always believing in me, and for encouraging me to follow my dreams. Dad, thank you for instilling a strong work ethic in me from an early age and for your unwavering confidence that I will "knock them dead." To my sister Danielle and brother-in-law Alden, thank you for setting the example for me to look up to and for your unwavering support of my academic pursuits throughout my life. To Rich, you have supported me from day one and have always pushed me to follow my dreams. I am deeply grateful for your love and encouragement. You are the best partner for me on my unconventional life journey, and I thank you for being my rock.

My dissertation research and doctoral training was made possible because of the financial support from the following sources: National Institutes of Health (TL1 TR002344, R01 HD068290, and T32 HD007434)

Jessica Barth

Washington University in St. Louis

May 2022

In memory of Dr. Alexander W. Dromerick.

ABSTRACT OF THE DISSERTATION

Validation, Categorization, and Prediction of Upper limb Outcomes after Stroke

by

Jessica Barth

Doctor of Philosophy in Movement Science

Physical Therapy

Washington University in St. Louis, 2022

Professor Catherine E. Lang, Chair

The incidence and costs of stroke in the United States are projected to rise over the next decade because of the aging population. Declining stroke mortality over the past few decades means that more people survive stroke and live with physical, cognitive, and emotional disability. Stroke remains one of the leading causes of disability in the United States because very few survivors experience a full recovery of their upper limb. Upper limb recovery after stroke is critical to performing activities of daily living and physical and occupational therapies are one of the only treatment options to address these challenges. The World Health Organization's (WHO) International Classification of Functioning, Disability, and Health Framework (ICF) informs our understanding of the importance of measuring upper limb changes across measurement levels, showing that improvements seen in one level (i.e. domain) do not directly transfer to another. Knowing this, it is important to evaluate existing prediction models of motor outcomes after stroke while simultaneously developing novel tools available to

researchers and clinicians to facilitate measurement of the upper limb across the ICF domains. This dissertation work performs an external validation of an existing prediction model of upper limb capacity (UL; capability measured in the clinic) after stroke, identifies and defines categories of UL performance (actual UL use in daily life) in people with and without neurological UL deficits, and explores how early clinical measures and participant demographic information are associated with subsequent categories of UL performance after stroke.

Recently, prediction algorithms of upper limb capacity after stroke have been developed to facilitate treatment selection, discharge planning, and goal setting for clinicians and their clients. Prediction models have tremendous clinical utility because they aid in the clinical decision making required to select the appropriate efficacious and emerging interventions that afford improvements in upper limb functional capacity, measured by standardized assessments in the therapy clinic. Prior to wide spread implementation of existing prediction algorithms into routine rehabilitation care, however, it is necessary to understand how small healthcare system differences and availability of neurophysiological assessments affect external validation of the models. In Chapter 2, we test how well an algorithm with clinical measures, developed for use in another country, applies to persons with stroke within the United States.

Knowing the importance of measurement across ICF domains, it is necessary to develop tools that facilitate clinical decision making and implementation of upper limb performance data into routine rehabilitation care. The use of wearable sensor technology (e.g., accelerometers) for tracking human physical activity have allowed for measurement of actual activity performance of the upper limb in daily life. Data extracted from accelerometers can be used to quantify multiple variables measuring different aspects of UL activity in one or both limbs. A limitation is that several variables are needed to understand the complexity of UL performance in daily life. As a

solution to the multi-variable problem, it would be helpful to form categories of UL performance in daily life. If natural groupings occur among multiple UL performance variables calculated from accelerometry data, then these groupings could facilitate clinical decision making and implementation of upper limb performance data into routine rehabilitation care. In Chapter 3 we identify and define categories of UL performance in daily life in adults with and without neurological deficits of the upper limb.

Prediction of motor outcomes after stroke have tremendous clinical utility, however there have been limited efforts to develop prediction models of upper limb performance (i.e., actual upper limb activity) in daily life after stroke. With advances in computing power, it is possible to capitalize on machine learning techniques to predict upper limb performance after stroke. These techniques allow for predicting a multivariate categorical outcome. This is important because it provides more information about the expected upper limb outcome to people with stroke, their families, and clinicians than a single continuous variable or a binary category (e.g., good or poor). Chapter 4 of this dissertation explores how different machine learning approaches can be used to understand the association between early clinical measures and participant demographics to the UL performance categories from a later post stroke time point.

Our findings provide strong support for the importance of measuring recovery of the UL across ICF domains, not just with impairment and capacity level measures. Collectively this work provides preliminary measurement tools that could eventually be available to rehabilitation clinicians following subsequent validation efforts. Additionally, this work provides a rich exploration into the strengths, weaknesses, and limitations of analytical methods and their impact on validation efforts.

Chapter 1: Introduction

This introductory chapter begins with an overview of stroke and its increasing burden on the United States (US) health care system and survivors. It then discusses measurement of upper limb (UL) outcomes using terms defined by the World Health Organization's (WHO) International Classification of Functioning Framework (ICF).[1] Emphasis is placed on the differences of measurement of UL impairment, capacity and performance after stroke along with considerations for use in stroke rehabilitation. Next there is a discussion of the development of prediction models of post stroke outcomes and their clinical utility. Finally, this chapter ends with a discussion of the importance of expanding prediction models to UL performance after stroke.

1.1 Stroke and UL impairment is a significant health problem

Stroke is one of the most significant health problems facing the country and is the leading cause of complex, long-term disability.[2-5] The US spends approximately $50 billion healthcare dollars annually on direct and indirect stroke care costs and this amount will increase because the incidence of stroke is expected to rise in the coming years due to an increased aging of the population.[4,6] In fact, the projected stroke costs in the US are expected to increase over 200% to $184 billion by 2030.[4,6] Declining stroke mortality over the past few decades[3] means that more people are surviving stroke but remain with physical, cognitive, and emotional disability.[7]

UL paresis after stroke is a leading contributor to the substantial rates of disability in the US and around the world.[8,9] Of the 795,000 people who suffer a stroke per year in the US, a staggering 80% of those individuals will incur some degree of initial impairment of the UL.[10-12] At six months post stroke, 65% of individuals will still have difficulties incorporating their UL

into their daily tasks.[13,14] In fact, only 5-20% will experience full neurological recovery of their UL following a stroke.[11,12] Post stroke UL problems may be caused by weakness, uncoordinated movements, diminished sensation, and overall reduced speed of movement, which collectively contribute to decreased independence, performance in daily tasks, and participation in life roles.[12,15-17] UL problems after stroke can lead to long-term functional deterioration, higher levels of disability, and limited community reintegration.[9,15] UL recovery after stroke is critical for performing activities of daily living (ADL). Physical and occupational therapies are one of the only treatment options to address these problems.[18-24]

Individuals with stroke have identified improving UL function as a top priority for stroke rehabilitation.[25] As a result, substantial time and research dollars have been invested to develop efficacious, in-clinic UL interventions to improve UL function early[26-29] and later[30,31] after stroke. UL impairment and loss of capacity has been a primary focus of stroke rehabilitation. Clinicians use standardized assessments within the clinic to measure UL impairment and capacity (see 1.2.1 and 1.2.2) to measure improvements made over the course of treatment.[32,33] With many efficacious therapies available, individuals make improvements in their UL capacity with the anticipation that those improvements translate into increased use of the limb during daily tasks.[34]

1.2 Using the ICF model to conceptualize UL problems post stroke

The World Health Organization's (WHO) International Classification of Functioning, Disability and Health Framework (ICF) is a comprehensive framework for measuring both individual and population health.[1,32] The ICF model is based on the biopsychosocial approach used to integrate the biological, individual, and social dimension.[35] The ICF has two parts: 1)

functioning and disability and 2) contextual factors.[35-37] In the ICF, functioning and disability are multi-dimensional concepts relating to 1) body functions and structures; 2) activities and participation; and 3) personal and environmental factors.[1,32,37] Despite these being separate concepts, there is an important interplay and influence (represented by double sided arrows in Figure 1.1) of both internal and external factors of each component that impact an individual's health status.[1,32,35,37-39] Figure 1.1 outlines the three specific ICF components that apply to rehabilitation of persons with stroke, and other health conditions including: body structures/functions, activity, and participation. Within each ICF component contains hierarchically arranged domains presented as chapters[37,38,40] which include sets of related physiological functions, anatomical structures, actions, tasks, areas of life, and external influences.[1,32,36,37] Table 1.1 provides the names of the ICF components, their definitions, and an example of the domain aligned with the UL after stroke. The ICF model emphasizes an individual's health along these domains and serves as a useful tool for understanding functional limitations following a health event, such as stroke.

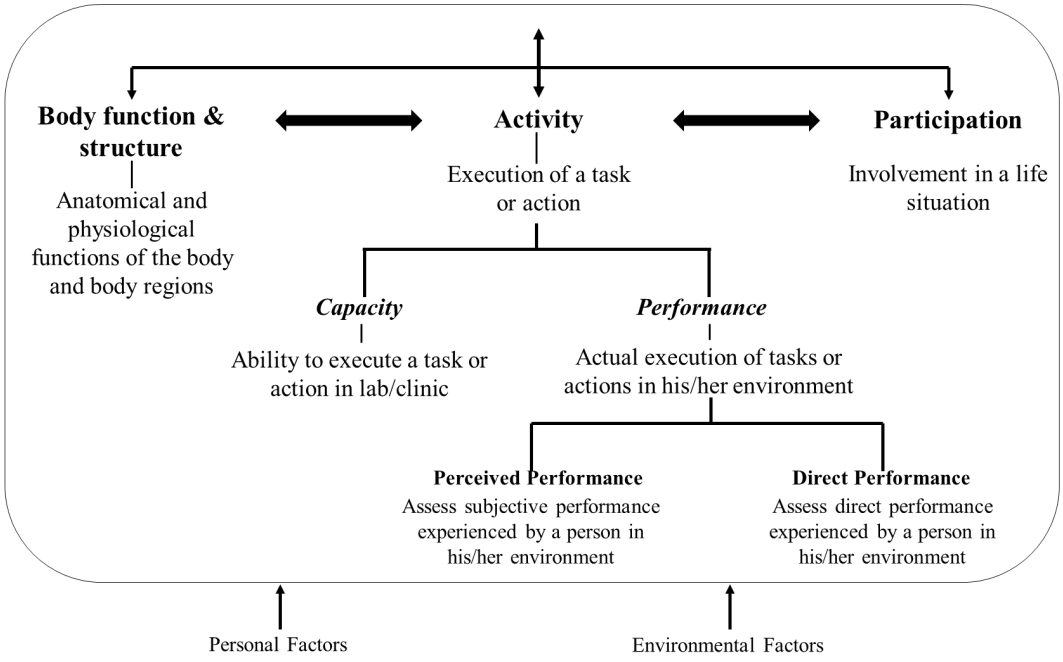


Figure 1.1 Adapted International Classification of functioning, Disability, and Health Framework (ICF)

Table 1.1. ICF components, their contents, and examples of domains with respect to the UL after stroke.

ICF Component	Definition	Domain Example
Body functions and structures	The physiological functions of body systems and the anatomical parts of the body such as organs, limbs and their components	-Sensory functions and pain including light touch discrimination and shoulder pain due to subluxation. -Neuromusculoskeletal and movement-related functions including power of muscles of one side of the body. -Structures of the nervous system such as, integrity of the cortical spinal tract. -Structures related to movement such as, the shoulder and scapula.
Impairments	Problems in body function and structure such as deviation or loss measured by clinicians with standardized processes and assessments.	-Hemiparesis or weakening of one UL -Sensory loss
Activity	The execution of a task or action by an individual that represents the individual perspective of functioning. The activity domain is divided into the capacity for activity and performance of activity.	-Handling objects such as a brush, carrying a pot, or holding a pen. -Moving around and using transportation likedriving a car. - Self-care tasks such as: bathing, dressing, and eating. -Domestic life tasks such as cleaning the kitchen or holding a child
Capacity	What someone is capable of doing in a controlled environment measured by standardized assessments in the clinic.	-UL motor ability to bathe or complete self-feeding -Functional ability of the paretic arm and hand.
Performance	What someone actually does in their free-living environment. Perceived performance is measured with self-report questionnaires and direct performance with wearable sensors.	-subjective measure of UL use in daily life. -variables calculated from wearable sensor data measuring the duration, magnitude, variability, and symmetry of UL activity in daily life.
Participation	-Involvement in a life situation and represents the societal perspective of functioning. Measures evaluate the degree to which an individual is involved with roles and relationships common to daily life that can include separate subjective and objective performance.	-Interpersonal interactions and reactions such as parent, spouse or employee. -Major life areas such as worker, student, and child caring. -Community, social, and civic life including recreation and leisure or religion and spirituality.
Activity Limitations	Difficulties an individual may have in executing activities.	
Participation restrictions	Problems an individual may experience in life situations.	
Environmental factors	The physical, social and attitudinal environment in which people live and conduct their lives. These are either barriers or facilitators of a person's functioning.	-The individual's immediate environment (workplace, home, or school). -The societal environment of social structures, services, and approaches for systems.
Functioning	Umbrella term for body function, body structures, activities and participation. Denotes the positive or neutral aspects of the interaction between a person's health condition(s) and that individual's contextual factors (environmental and personal factors).	
Disability	Is an umbrella term for impairments, activity limitations and participation restrictions. It denotes negative aspects of the interaction between a person's health condition(s) and that individual's contextual factors (environmental and personal factors).	

1.2.1 Body structures and functions

The body structures and functions components are the anatomical and physiological functions of body systems.[1,32,37] The ICF definitions clearly differentiate physiologic functions from anatomical structures of the body and two separate classification systems are offered. There are eight hierarchical domains of body structures, the sensory functions and pain and neuromusculoskeletal and movement-related functions are most considered after stroke.[37] There are also eight hierarchical domains of body functions, each can be affected by stroke however the structure of the nervous system and structures related to movement are considered when assessing the UL after stroke. Of equal importance, other domains within the body functions and structures domains such as, mental functions may also affect the UL after stroke. For example, a person's mood and emotion, self-efficacy, and confidence may impact the movement related functions of the UL after stroke demonstrating the interaction of these components during recovery from stroke.[35,41,42]

UL impairments

The term impairment is associated with the body structure/function component of the ICF (Figure 1.1 and Table 1.1) and is defined as the problems in body function or structures resulting in a significant deviation or loss.[1,37] Impairments of the UL after stroke results from stroke-related damage to the cortical and subcortical brain structures specifically, the primary motor cortex, the primary somatosensory cortex, secondary sensorimotor cortical areas, and the corticospinal tract.[43,44] UL impairments in people who have had a stroke are well documented and include paresis, loss of somatosensation, spasticity, muscle contracture, loss of dexterity or fractionation of movement, decreased active joint range of motion, and lack of movement, speed, precision, and bimanual coordination.[45,46] Clinicians typically conduct evaluations of a person's

UL movement after stroke to identify the impairments that limit normal movement.[46] These include goniometry measurement of active or passive range of motion,[47] manual muscle testing,[48] grip and pinch strength.[49,50] and somatosensory testing.[51-53] Standardized processes and assessments have been developed to measure specific stroke related UL impairments and include the Fugl-Meyer (UEFM),[54-56] the modified Ashworth,[57,58] the Motricity index,[59] the Motor Assessment Scale,[60,61] and the Shoulder Abduction Finger Extension (SAFE).[62]

1.2.2 Activity

The second ICF component is activity which defined as the execution of a task or action by an individual. Activities are characteristics of people, and they can be assessed by examining the functional performance of an individual in isolation.[36,37] There are nine hierarchical domains within the activity component, and each can be affected by stroke. The general tasks and demands, mobility, self-care, and domestic life domains are most considered when assessing the UL after stroke.[37] Activities tend to be simplistic, performed alone, and are generally results-oriented.[1,32,36] Problems completing activities after a health condition, such as stroke, are described as activity limitations. Activity limitations describe difficulties an individual may have in executing tasks and activities. Activity limitations may or may not lead to participation restrictions, depending on what activities are needed to participate in a social situation and what environmental barriers or supports exist.[35,36,63] The activity component is divided into the capacity for and the performance of activity.[1,37,63] These constructs provide a way to indicate how the environment (where measurement occurs) impacts a person's activities and how the environmental change may improve a person's functioning.[37] Capacity refers to what an individual can do in a standardized environment and performance is what someone actually does in their usual environment.

UL capacity for activity

UL capacity for activity is defined as the ability to execute a task or action in a standardized, or controlled, environment.[1,32] A standardized environment has removed the barriers within an individual's environment that can interfere with their ability to complete a task while simultaneously providing identical testing conditions for each person. There are several standardized clinical assessments used for rehabilitation that quantify UL capacity within a clinic or laboratory setting. Standardized assessments provide important information about the UL after stroke to provide a greater understanding of a person's ability to complete actions and tasks aligned with the domains (e.g. general tasks and demands and mobility).[32,37] These include the Wolf Motor Function Test,[64,65] the Box and Blocks,[66] the Nine Hole Peg Test,[67,68] the Jebsen Taylor Hand Function Test,[69] and the Chedoke Arm and Hand Activity Inventory (CAHAI).[70-72] Some of these standardized measures include aspects of measurement of compensatory UL movements in an attempt to determine their effects on a person's UL capacity.[64,65,71-73] Interestingly, the CAHAI is the only standardized measure of UL capacity that includes bilateral tasks in their design. Given that most UL tasks involve bilateral involvement,[74] it is important to include bilateral assessments when measuring UL capacity. The most common standardized assessment of UL capacity after stroke is the Action Research Arm Test (ARAT) which is a valid and reliable, criterion-rated assessment of UL activity limitations.[73,75-77]

1.2.3 UL performance of activity in daily life

Performance of activity is defined as the execution of tasks or actions in an unstructured, free-living (i.e., usual) environment.[32,37,41] Performance is a measure of what someone actually does when they are outside of the structured clinic or laboratory setting.[1] The free-living environment includes a combination of physical, social, and personal factors that may be either

barriers or facilitators to performance. Quantifying UL performance in daily life is challenging and is typically measured by assessments of perceived performance or direct, objective measures of performance.[78,79] Similar to capacity for activity, the limitations in the performance of activity after stroke is also related to participation restrictions depending on what activities are needed to participate in a social situation and what environmental barriers or supports exist.[35,36,63]

Perceived Performance

Perceived performance is generally measured with standardized self-report measures, such as, the Motor Activity Log[80,81] or the Stroke Impact Scale (activities in a typical day subscale).[82] Self-report measures of perceived UL performance have been shown to improve with rehabilitation therapies,[41,83] however self-report measures are subject to inherent biases[84,85] (e.g. social desirability, recall bias) which can compromise results.[78] Because of these limitations, objective measures of direct UL performance post stroke have been developed to provide a more objective measure of UL activity after stroke.

Direct Performance

Over the past decade, the use of wearable sensor (e.g., accelerometers) technology has emerged in rehabilitation to measure performance of activity in daily life.[86,87] Accelerometry has become an established, valid and reliable methodology to directly measure performance of UL activity in daily life in several adult populations. [74,88-93]

Data extracted from bilateral, wrist worn wearable sensors can be utilized to quantify UL performance variables measuring the duration[74,94], laterality and symmetry,[74,79,89,95,96] magnitude or intensity, and variability of one or both limbs.[79,88,97-99] These variables collectively inform clinician scientists about different aspects of real-world UL activity in daily life of nondisabled

adults[74,89] and adults with stroke.[90,92,100,101] Each UL performance variable conveys slightly different information about the collective nature of UL activity. Measuring UL performance after stroke from wearable sensor data has led to a growing recognition that there is a disparity between UL capacity and actual real-world UL use, individuals may have the ability to move their affected UL but they may not actually use it for daily activities.[9,89,102] UL capacity is a prerequisite to UL performance, however, UL performance is not a direct consequence of good or improved UL capacity.[34,103,104] This has implications for rehabilitation because currently clinicians only measure UL capacity to determine if improvements are made as a result of their interventions. Clinicians need tools to measure UL performance to provide information about what their clients actually do with their ULs, outside of the clinic to complement the in-clinic measures of UL capacity that provide information about a client's UL capability.[105]

1.2.4 Participation

Participation (Figure 1.1 and Table 1.1) is defined as involvement in a life situation and represents a societal perspective of functioning.[32,36,37,39,63,106] Unfortunately, the ICF does not clearly differentiate activity from participation, and only one classification system of domains is proposed for both.[36,37,107,108] Participation is a broad and complex concept that can have different meanings to different people.[109] It has been suggested that participation is a relational construct that can be assessed by considering other factors beyond simply the capabilities and limitations of the individual. Therefore, participation is more sensitive to the characteristics of the social, physical, and policy environments.[32,36,37] The activity and participation domains are expressed as a list over nine chapters that can be used to denote activities or participation or both.[63] The interpersonal interactions and restrictions, major life areas, and community, social, and civil life domains are most important to consider after stroke.[36,63] Problems or limitations in participation

is termed “participation restriction” and is viewed as a negative aspect and are defined as problems an individual may experience in involvement in life situations.[36,37] Measures of participation evaluate the degree to which an individual is involved with roles and relationships common to daily life that can include separate subjective and objective ratings of performance. Difficulty has been reported in applying the ICF concept of participation because of the diversity of this concept[36,39,63,107] resulting in a wide range of tools considered to measure participation. Measuring participation is difficult for conceptual and methodological reasons.[107] This is especially complicated because the domains are the same within the activity and participation concepts and it has yet to be decided if participation should be expected to be measured over all nine domains.[63] The most frequently measured domains of activity and participation belonged to the Community, Social and Civic Life, Domestic Life, and Mobility.[63] For example, a participation measure might include items related to being a productive member of society such as employment, being a student, and being a homemaker with each of these roles requiring several activities for successful participation.[37,63,107,108,110] Common validated assessments of participation after stroke include the Frenchay Activities Index,[111-114] the Activity Card Sort,[115] the Assessments of Life Habits, Stroke Impact Scale (participation subscale),[63,116] and the Reintegration to Normal Living.[117,118]

With respect to the UL after stroke, UL capacity is a predictor of participation in life roles post stroke. However, UL performance likely contributes as well because participation is a latent trait that shows its impact through a series of indicators such as employment, parenting, and churchgoing.[107] Individuals report limitations with participating in meaningful life roles even years after a stroke which can result in boredom, depression and worsening of function, affect, and quality of life.[14] Up to 50% of the community dwelling stroke population is living with the

sequalae of stroke that places them at risk for a diminished activity level and social isolation can result further compounding the negative health events.[14] While measures of UL impairment, capacity, and participation are important and convey useful information, they cannot provide information about UL activity performance, or actual UL activity, in the free-living environment.

In summary, the ICF Framework allows for the ability to view the interconnectedness and complexities of stroke related disability. Unfortunately for many years, stroke rehabilitation assessments and interventions have tended to focus on one or a few UL impairments (e.g. paresis, somatosensation) with the anticipation that changes in impairments will lead to downstream changes in other levels of measurement (e.g. activity and participation). As our understanding of these concepts grow, rehabilitation professionals are moving forward to take an integrated approach to UL rehabilitation to improve overall outcomes for people after a stroke. Additionally, as clinicians realize the importance of measurement across ICF components and domains, it is necessary to develop tools that facilitate clinical decision making and implementation of UL performance data into routine rehabilitation care.

1.3 Predicting outcomes improves post stroke care

Independence in activities of daily living and other meaningful activities after stroke depends largely on the recovery of motor impairment (primarily paresis), specifically in the UL.[17,119] In general, greater initial UL impairment is associated with worse UL capacity outcomes later after stroke.[120] The estimated recovery of motor impairment is one of several factors that influence the clinical decisions regarding the type and duration of rehabilitation, and goals set for each person after stroke. Unfortunately, these decisions are usually made without objective information about a person's likelihood of recovery of UL impairments after stroke.[119] Clinicians typically make decisions about the discharge destination from the acute care setting

based on their perception of persons post stroke prognosis for recovery.[121,122] Post stroke prognosis and discharge disposition are greatly influenced by the clinicians initial impression along with incorporating clinical and demographic factors such as initial stroke severity and age.[123] Unfortunately, even experienced clinicians have difficulty making accurate predictions of UL capacity outcomes for people with stroke.[123-125] In most US healthcare settings, clinicians have no way to know if their prognoses at the acute stage were correct, unless their system routinely conducts follow up assessments several months later. This situation creates an immediate need for standardized tools to support less variable and more equitable decision-making of post stroke prognoses for clinicians and their clients.[122,126]

1.3.1 Prediction of UL capacity

Recent efforts have explored predicting eventual UL capacity outcomes after stroke which has the potential to facilitate treatment selection, discharge planning, and goal setting for clinicians and clients.[103,125,127] To date, several factors are thought to be related to subsequent UL capacity including initial motor and sensory impairments, and measures of sensorimotor system structure and function obtained with neuroimaging[120,128-131], or neurophysiological techniques.[120,128-130,132] Several studies have demonstrated that recovery of UL capacity occurs mainly within the first three months post stroke,[133-135] and several prediction models have been developed to guide clinical decision-making.[62,136] Most of these prediction algorithms have been proposed and evaluated on participant cohorts from clinical trials, and they are mostly accurate for persons with mild to moderate UL impairment.[102,127,129,132,137-139] The most well-known algorithm was developed in New Zealand and allows for prediction of UL capacity around three months post stroke based on measures taken within the first week. PREP, the original algorithm developed in 2012, used both clinical measures and neurological biomarkers in the initial days

after stroke to determine UL capacity categories around three months with 64% accuracy.[119] PREP2, published in 2017, is the refined version, which predicts a category of UL capacity around three months with 75% accuracy using clinical measures and a single neurophysiological measure.[62] The PREP2 algorithm classifies individuals into one of four clinically meaningful categories (Excellent, Good, Limited, or Poor) in anticipation of subsequent UL capacity outcomes.[62,140] Predicting a category for an eventual UL outcome is useful because persons with stroke are more interested in what they can generally expect in recovery of the UL, not a specific numerical score on a standardized test.[123,141-143] Each of the four categories from PREP2 are defined by boundaries on the ARAT, a standardized assessment of UL capacity.[62] PREP2 categories describe an expected UL outcome and associated rehabilitation focus, providing individuals with stroke, their care givers, and clinicians with more interpretable information.[142,143]

Despite its accuracy and ease of use, there are some challenges to implementing PREP2 in the US. The rehabilitation structure of New Zealand differs somewhat from the US, with respect to who, how much, where, and when individuals receive UL therapy after stroke. A lot of young people in the US are having strokes[4,6,144-146] and many have other comorbidities that are poorly managed.[147] Another hurdle to implementation of the PREP2 in the US is the limited accessibility to the neurophysiologic measurement of corticospinal tract function, via transcranial magnetic stimulation (TMS).[148] Access to this neurophysiologic assessment may be available for research purposes and/or in some major academic medical centers, but is not available in routine rehabilitation care.[148,149] Anticipating this, the PREP2 algorithm was shown to still be 71% accurate using clinical measures only.[62,140] In the version with only clinical measures, a measure of stroke severity[150] replaces the neurophysiologic measure for participants with less

initial strength in their affected UL.[62] Prior to widespread implementation of this useful algorithm into routine rehabilitation care, it is necessary to understand how small healthcare system differences, differences in the stroke population, and lack of the neurophysiologic assessments affects the accuracy of the PREP2 algorithm in individuals who have suffered a stroke in the US. Aim 1 of this dissertation evaluates the accuracy of the PREP2 algorithm with clinical measures only here in a US healthcare system.

1.3.2 Prediction of UL performance

As seen in the ICF Framework, (Figure 1.1) UL capacity for activity is a different but related construct to UL performance of activity.[1] Data from wearable sensors are analyzed to produce variables that capture aspects of UL performance in daily life after stroke. With the ability to measure UL performance, researchers have learned that improvements in UL capacity seen in the clinic often do not lead to improved performance of the UL in daily life.[34,89,102,103] In other words, individuals improve their capability to use their arm measured by in clinic assessments, but are not transferring these improvements into better use of their limb in daily life.[9,34,90,102,103,151,152] These findings have solidified our understanding that UL functional capacity and UL performance are two separate constructs.[94,153] Knowing this information, it is important to explore prediction models of UL performance to complement existing prediction tools of UL capacity.

There have been limited efforts to develop prediction models of UL performance or actual activity of the UL in daily life after stroke. To date, two studies have explored early predictors of UL use at three months[102] and one year[103] after stroke. A longitudinal study by Rand and Eng[103] used multiple linear regression to predict daily use of each UL, measured with a single continuous variable (mean total activity counts of the paretic UL), 12 months after stroke

from motor and non-motor clinical measures and participant demographic information assessed at discharge from subacute stroke rehabilitation centers. They found that better UL capacity at discharge predicted increased UL use 12 months after stroke.[103] More recently, Lundquist et. al.[102] also used multivariate linear and logistic regression to examine the factors two weeks after stroke that could predict future UL performance, quantified by a single continuous variable (use ratio), three months after stroke.[102] They found that better UL capacity (ARAT) early after stroke predicted increased UL performance at three months. When they dichotomized UL performance as "normal" vs "non-normal", the absence of a motor evoked potential (MEP) along with the presence of visual neglect predicted "non-normal" UL use.[102] A strength of both of these studies was the inclusion of non-motor clinical assessments and participant demographic information in addition to the UL impairment and capacity measures as potential predictors of UL performance. Including these predictors was useful in determining how other post stroke problems (e.g. visual neglect, depression) influenced eventual UL use in daily life after stroke. In the Rand & Eng study the outcome of UL performance was measured one year post stroke which is a limitation because UL performance stabilizes within six weeks after stroke therefore,[105] one year post stroke is likely too late to intervene to improve the outcome.[105] There are two limitations of both studies; first was the analysis choice of multiple linear regression techniques and second was selecting a single continuous variable for the outcome of UL performance. Multiple linear regression techniques are statistical processes used for estimating the relationships among the dependent and independent variables. These techniques require that specific assumptions about the predictors and the outcome be met which limits the type of predictors that can be included.[154] Also, as more cutting edge methods emerge the term "prediction" can be misused because the results of regression models are expressed as the proportion of variance (R^2) explained, not

accuracy.[155] The second limitation was that the outcome of UL performance was a single variable (mean total activity counts of the paretic UL and the use ratio). Therefore, significant predictors can only be associated with the single variable measuring only one aspect of UL activity after stroke. From these two studies it is unknown how the included predictors were associated with other aspects of UL performance in daily life, such as the duration, magnitude or intensity, and variability of UL activity.

UL performance is likely a multivariate construct because a complete measure of UL use in daily life requires knowledge of the duration, magnitude or intensity, symmetry and variability of UL movements.[87] Currently, several variables must be calculated from wearable sensor data because each provides different information about UL activity in the free-living environment.[156] A single variable may not be not sufficient in understanding the scope of UL performance in daily life.[156] A limitation of the multivariate approach is that calculating several variables adds complexity to the interpretation of UL performance data for clinical decision making. A potential solution to the multi-variable problem would be the formation of categories of UL performance in daily life. If there were natural groupings that occur among multiple UL performance variables calculated from accelerometry data,[157] then these groupings could help to facilitate clinical decision making and implementation of UL performance data into routine rehabilitation care. Aim 2 of this dissertation investigates if categories of UL performance can be identified in adults with and without neurological UL deficits.

With the advances in computing power, it is possible to capitalize on machine learning techniques to predict UL performance after stroke. Machine learning algorithms have several advantages over regression models including 1) requiring fewer assumptions about the distributions of the data, 2) numerous options for non-parametric models, and 3) strong

predictive capabilities.[158-162] Machine learning techniques allow for the possibility of predicting a multivariate categorical outcome. This is important because categories provide more information about the expected UL outcome to the person with stroke, their families, and clinicians than a single continuous variable or a binary category (e.g. good or poor).[62,102,141,155,163] Aim 3 of this dissertation explores how different machine learning techniques can be used to understand how clinical measures and participant demographics captured early after stroke are associated with the UL performance categories from a later post stroke time point.

1.4 Summary and critical next steps

The incidence and costs of stroke in the US are projected to rise over the next decade because of the aging population. Declining stroke mortality over the past few decades[3] means that more people are surviving stroke but remain with physical, cognitive, and emotional disability.[7] Stroke remains one of the leading causes of disability in the US because very few survivors experience a full recovery of their UL. Recently, prediction algorithms of UL capacity after stroke have been developed to facilitate treatment selection, discharge planning and goal setting for clinicians and their clients. However, prior to widespread implementation of existing prediction algorithms into routine rehabilitation care, it is necessary to understand how small healthcare system differences and availability of neurophysiological assessments affect external validation of the models. Aim 1 will be the first study to externally validate an algorithm with clinical measures only on a US population of persons with first stroke.

Prediction models have tremendous clinical utility because they aid in the clinical decision making required to select the appropriate efficacious and emerging interventions that afford improvements in UL functional capacity,[32] measured by standardized assessments in the therapy clinic.[26,27,30] The ICF Framework, however, informs our understanding of the importance

of measuring UL changes across measurement levels. The literature has shown that improvements seen in one measurement level do not directly transfer to another.[34,101] Aims 2 and 3 of this dissertation investigate each of these gaps in an effort to move the field of stroke rehabilitation forward by contributing to the development of novel clinical tools that are available to stroke rehabilitation providers in the US and around the world to facilitate measurement and prediction of UL performance after stroke.

1.5 Specific Aims

Aim 1: Test how well an algorithm with clinical measures, developed for use in New Zealand, applies to persons with stroke within the US. This is a secondary analysis of data collected from a prospective, observational, longitudinal cohort tracking UL change over time.[105]

> ***Hypothesis 1a*:** UL functional capacity will be predicted, and algorithm accuracy will fall in a range of 70-80% in a cohort of stroke participants in the US.
>
> ***Hypothesis 1b**:* Those with inaccurate predictions will be within one category of their expected category at 3 months.

Recent efforts have explored predicting eventual UL capacity outcomes after stroke which has the potential to facilitate treatment selection, discharge planning, and goal setting for clinicians and clients.[102, 111, 113] Several prediction models have been developed to guide clinical decision-making,[55, 122] the most well-known algorithm is PREP2 and it was developed in New Zealand.[62] PREP2 predicts a category of UL capacity around three months after stroke with 75% accuracy using clinical measures and a single neurophysiological measure.[55] The algorithm classifies individuals into one of four clinically meaningful categories (Excellent, Good, Limited, or Poor) in anticipation of subsequent UL capacity outcomes.[55, 126] Most people with inaccurate predictions were only one category away from their actual at 90 days.[62,140] Despite the merits of

this algorithm, there are some challenges to implementing it in the US including participant factors (e.g., age of stroke and number of comorbidities)[144-146] and the rehabilitation structure.[141,149] Anticipating this, the algorithm was shown to still be 71% accurate at 90 days post-stroke using clinical measures only. Prior to widespread implementation of this algorithm into routine rehabilitation care, it is necessary to understand how small system differences and lack of the neurophysiologic tests (TMS) affects the overall accuracy of the algorithm in individuals who have suffered a first stroke in the US. The current study determines the accuracy of an algorithm on a sample of persons with first ever stroke, using clinical measures only, at time points that are most feasible in the US health care system.

Aim 2: Identify and define categories of UL performance, as quantified from accelerometer recordings. This is an analysis of data from three completed studies: 1) the same prospective, observational cohort as Aim 1;[105] 2) a sample of persons with chronic stroke who participated in a clinical trial;[31] and 3) a sample of neurologically-intact adults of similar age, race, ethnicity, and socioeconomic status.[74]

> ***Hypothesis 2a****:* Three categories of UL performance will be identified across a host of accelerometer variables, spanning the possible ranges of UL performance in daily life.
>
> ***Hypothesis 2b*****:** The categories that emerge will have clinical meaning of expected UL performance.

The use of wearable sensor technology (e.g., accelerometers) for tracking human physical activity have allowed for measurement of actual activity performance of the UL in daily life.[79,100,156,164-166] Data extracted from accelerometers can be used to quantify multiple variables measuring different aspects of UL performance in one or both limbs such as: 1) duration; 2) magnitude; 3) variability; and 4) symmetry or laterality. Each UL performance variable conveys

slightly different information about the collective nature of UL use, with a single variable providing only part of the picture.[156] One reason wearable sensor technology for measurement of UL performance has remained largely confined to rehabilitation research with limited ventures into clinical practice is because the current output from accelerometers is not easily accessible for rehabilitation professionals.[100] A potential solution to the multi-variable problem would be the formation of categories (or groups) of UL performance in daily life. Statistical analysis methods such as a *k*-means hypothesis-free cluster analysis can be used to determine categorizations of UL performance indexed by accelerometer variables in samples of persons with stroke and neurologically intact adults (adult controls). Including cohorts of people with and without stroke will capture a wide range of the variables, extracted from accelerometer data that quantify different aspects of UL performance in daily life. Thus, the emerging categories would group individuals with similar ranges of the performance variables and provide a simpler method to interpret UL performance in daily life for clinicians and persons with health conditions whom they treat. The purpose of this study is to identify categories of UL performance in daily life in adults with and without stroke using data from previously collected cohorts. If there were natural groupings that occur among multiple UL performance variables calculated from accelerometry data,[157] then these groupings could help to facilitate clinical decision making and implementation of UL performance data into routine rehabilitation care.

Aim 3: Determine if a model can be developed to predict UL performance in daily life at 3 months post stroke. This aim is an analysis of data collected from a prospective, observational, longitudinal cohort study tracking UL change over time and is the same cohort used in Aim 1.[105]

> ***Hypothesis 3a****:* A model can be derived from a collection of clinical measures to predict UL performance post-stroke.

***Hypothesis 3b**:* The developed model will predict UL performance with a minimum of 70% accuracy.

Early prediction of motor outcomes after stroke has tremendous clinical utility [136,167] because predictive knowledge of subsequent outcomes can inform the delivery and specification of individualized rehabilitation services.[162,168] Several prediction models have been developed to guide clinical decision-making,[5, 122] however, the majority predict UL impairment or capacity. [119-121] UL capacity for activity is a different but related construct to UL performance of activity.[103,105,169] Prediction models of UL performance can be informed by models of UL capacity, such as PREP2,[62] which has demonstrated that prediction of an UL capacity category provides clinically-useful information to people with stroke and their families. The PREP2 prediction model was originally built and validated with a CART which resulted in the easy to interpret decision tree with an overall accuracy of 71% to 75%.[62] Recent efforts to predict an individual's subsequent UL impairment or capacity category, measures of UL impairment and capacity emerge as the most important predictors. Advances in computing have improved upon old and led to new analysis techniques for building prediction models of UL outcomes after stroke. An alternative to creating a single decision tree is to use ensemble classifier methods.[170] which tend to have higher predictive power and reduce the risk of over-fitting relative to other CART methods, but at the expense of interpretability.[159,170] In the present study, different machine learning techniques will be used to understand how clinical measures and participant demographics captured early after stroke are associated with the UL performance categories from a later post stroke time point.[171] Using different machine learning methods to build predictive models with different input variables as predictors will determine how each method yields similar versus different results. Capitalizing on the advantages of ensemble machine

learning algorithms by applying them for prediction of UL performance outcomes could yield key insights into UL recovery post stroke.

1.6 References

1. WHO. International Classification of Functioning, Disability and Health (ICF). *World Health Organization*. 2018;(06/01/2021)

2. Benjamin EJ, Muntner P, Bittencourt MS. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. *Circulation*. 2019;139(10):e56-e528.

3. Lackland DT, Roccella EJ, Deutsch AF, et al. Factors Influencing the Decline in Stroke Mortality: A Statement From the American Heart Association/American Stroke Association. *Stroke*. Dec 5 2013;doi:01.str.0000437068.30550.cf [pii]10.1161/01.str.0000437068.30550.cf

4. Virani SS, Alonso A, Aparicio HJ, et al. Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. *Circulation*. Feb 23 2021;143(8):e254-e743. doi:10.1161/CIR.0000000000000950

5. Virani SS, Alonso A, Benjamin EJ, et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation*. Mar 3 2020;141(9):e139-e596. doi:10.1161/CIR.0000000000000757

6. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association. *Circulation*. Feb 22 2022;145(8):e153-e639. doi:10.1161/CIR.0000000000001052

7. NINDS. National Institute of Neurological Disorders and Stroke - Stroke Information Page. Accessed October 2, 2013. http://www.ninds.nih.gov/disorders/stroke/stroke.htm#skipnav

8. Katan M, Luft A. Global burden of stroke. Georg Thieme Verlag; 2018:208-211.

9. Hidaka Y, Han CE, Wolf SL, Winstein CJ, Schweighofer N. Use It and Improve It or Lose It: Interactions between Arm Function and Use in Humans Post-stroke. *Plos Comput Biol*. Feb 2012;8(2)doi:ARTN e100234310.1371/journal.pcbi.1002343

10. Wade D, Langton-Hewer R, Wood VA, Skilbeck C, Ismail H. The hemiplegic arm after stroke: measurement and recovery. *Journal of Neurology, Neurosurgery & Psychiatry*. 1983;46(6):521-524.

11. Heller A, Wade DT, Wood VA, Sunderland A, Hewer RL, Ward E. Arm function after stroke: measurement and recovery over the first three months. *Journal of Neurology, Neurosurgery & Psychiatry*. 1987;50(6):714-719.

12. Nakayama H, Jorgensen HS, Raaschou HO, Olsen TS. Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. *Arch Phys Med Rehabil*. Apr 1994;75(4):394-8. doi:10.1016/0003-9993(94)90161-9

13. Dobkin BH. Rehabilitation after stroke. *New England Journal of Medicine*. 2005;352(16):1677-1684.

14. Mayo NE, Wood-Dauphinee S, Cote R, Durcan L, Carlton J. Activity, participation, and quality of life 6 months poststroke. *Arch Phys Med Rehabil*. Aug 2002;83(8):1035-42. doi:10.1053/apmr.2002.33984

15. Demers M, Winstein CJ. A perspective on the use of ecological momentary assessment and intervention to promote stroke recovery and rehabilitation. *Topics in stroke rehabilitation*. 2020:1-12.

16. Faria-Fortini I, Michaelsen SM, Cassiano JG, Teixeira-Salmela LF. Upper extremity function in stroke subjects: relationships between the international classification of functioning, disability, and health domains. *Journal of Hand Therapy*. 2011;24(3):257-265.

17. Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW. Early prediction of outcome of activities of daily living after stroke: a systematic review. *Stroke*. May 2011;42(5):1482-8. doi:10.1161/STROKEAHA.110.604090

18. Lohse KR, Lang CE, Boyd LA. Is more better? Using metadata to explore dose-response relationships in stroke rehabilitation. *Stroke*. Jul 2014;45(7):2053-8. doi:10.1161/STROKEAHA.114.004695STROKEAHA.114.004695 [pii]

19. Veerbeek JM, van Wegen E, van Peppen R, et al. What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. *PloS one*. 2014;9(2):e87987. doi:10.1371/journal.pone.0087987PONE-D-13-44188 [pii]

20. Lubetzky-Vilnai A, Kartin D. The effect of balance training on balance performance in individuals poststroke: a systematic review. *Journal of neurologic physical therapy : JNPT*. Sep 2010;34(3):127-37. doi:10.1097/NPT.0b013e3181ef764d01253086-201009000-00002 [pii]

21. Rabadi MH, Rabadi FM, Edelstein L, Peterson M. Cognitively impaired stroke patients do benefit from admission to an acute rehabilitation unit. *Arch Phys Med Rehabil*. Mar 2008;89(3):441-8. doi:10.1016/j.apmr.2007.11.014

S0003-9993(07)01804-7 [pii]

22. Sackley C, Wade DT, Mant D, et al. Cluster randomized pilot controlled trial of an occupational therapy intervention for residents with stroke in UK care homes. *Stroke*. Sep 2006;37(9):2336-41. doi:01.STR.0000237124.20596.92 [pii]

10.1161/01.STR.0000237124.20596.92

23. Teasell R, Foley N, Salter K, Boghal S, Jutai J, Speechley M. *Evidence-Based Review of Stroke Rehabilitation*. 11 ed. University of Western Ontario; 2008.

24. Donnan GA, Davis SM, Levi CR. Strategies to improve outcomes after acute stroke. *Med J Aust*. Apr 7 2003;178(7):309-10. doi:don10111_fm [pii]

25. Pollock A, St George B, Fenton M, Firkins L. Top 10 research priorities relating to life after stroke - consensus from stroke survivors, caregivers, and health professionals. *International Journal of Stroke*. Apr 2014;9(3):313-320. doi:10.1111/j.1747-4949.2012.00942.x

26. Dromerick AW, Lang CE, Birkenmeier RL, et al. Very Early Constraint-Induced Movement during Stroke Rehabilitation (VECTORS): A single-center RCT. *Neurology*. Jul 21 2009;73(3):195-201. doi:10.1212/WNL.0b013e3181ab2b27

27. Winstein CJ, Wolf SL, Dromerick AW, et al. Effect of a Task-Oriented Rehabilitation Program on Upper Extremity Recovery Following Motor Stroke: The ICARE Randomized Clinical Trial. *JAMA*. Feb 9 2016;315(6):571-81. doi:10.1001/jama.2016.0276

28. Cramer SC, Dodakian L, Le V, et al. Efficacy of Home-Based Telerehabilitation vs In-Clinic Therapy for Adults After Stroke: A Randomized Clinical Trial. *JAMA Neurol*. Jun 24 2019;doi:10.1001/jamaneurol.2019.1604

29. Dromerick AW, Geed S, Barth J, et al. Critical Period After Stroke Study (CPASS): A phase II clinical trial testing an optimal time for motor recovery after stroke in humans. *Proc Natl Acad Sci U S A*. Sep 28 2021;118(39)doi:10.1073/pnas.2026676118

30. Wolf SL, Winstein CJ, Miller JP, et al. Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: the EXCITE randomized clinical trial. *JAMA*. Nov 1 2006;296(17):2095-104. doi:10.1001/jama.296.17.2095

31. Lang CE, Strube MJ, Bland MD, et al. Dose response of task-specific upper limb training in people at least 6 months poststroke: A phase II, single-blind, randomized, controlled trial. *Ann Neurol*. Sep 2016;80(3):342-54. doi:10.1002/ana.24734

32. Organization WH. Towards a common language for functioning, disability, and health: ICF. *The international classification of functioning, disability and health*. 2002;

33. Demers M, Levin MF. Do activity level outcome measures commonly used in neurological practice assess upper-limb movement quality? *Neurorehabilitation and neural repair*. 2017;31(7):623-637.

34. Waddell KJ, Strube MJ, Bailey RR, et al. Does Task-Specific Training Improve Upper Limb Performance in Daily Life Poststroke? *Neurorehabil Neural Repair*. Mar 2017;31(3):290-300. doi:10.1177/1545968316680493

35. Silva SM, Correa FI, Faria CDCD, Buchalla CM, Silva PFD, Correa JCF. Evaluation of post-stroke functionality based on the International Classification of Functioning, Disability, and Health: a proposal for use of assessment tools. *Journal of Physical Therapy Science*. Jun 2015;27(6):1665-1670. doi:DOI 10.1589/jpts.27.1665

36. Whiteneck G, Dijkers MP. Difficult to Measure Constructs: Conceptual and Methodological Issues Concerning Participation and Environmental Factors. *Archives of Physical Medicine and Rehabilitation*. Nov 2009;90(11):S22-S35. doi:10.1016/j.apmr.2009.06.009

37. Bruyere SM, Van Looy SA, Peterson DB. The international classification of functioning, disability and health: Contemporary literature overview. *Rehabilitation Psychology*. May 2005;50(2):113-121. doi:10.1037/0090-5550.50.2.113

38. Rauch A, Cieza A, Stucki G. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. *European journal of physical and rehabilitation medicine*. 2008;44(3):329-342.

39. Salter K, Jutai J, Teasell R, Foley N, Bitensky J, Bayley M. Issues for selection of outcome measures in stroke rehabilitation: ICF Participation. *Disability and rehabilitation*. 2005;27(9):507-528.

40. Ezekiel L, Collett J, Mayo NE, Pang L, Field L, Dawes H. Factors Associated With Participation in Life Situations for Adults With Stroke: A Systematic Review. *Archives of Physical Medicine and Rehabilitation*. 05/01/2019 2019;100(5):945-955. doi:https://doi.org/10.1016/j.apmr.2018.06.017

41. Lewthwaite R, Winstein CJ, Lane CJ, et al. Accelerating Stroke Recovery: Body Structures and Functions, Activities, Participation, and Quality of Life Outcomes From a Large Rehabilitation Trial. *Neurorehabilitation and Neural Repair*. Feb 2018;32(2):150-165. doi:10.1177/1545968318760726

42. Nicholas ML, Burch K, Mitchell JR, Fox AB, Baum CM, Connor LT. Self-Perception of Physical Function Contributes to Participation in Cognitively- and Physically-Demanding Activities After Stroke. *Frontiers in Neurology*. Jun 9 2020;11doi:ARTN 474

10.3389/fneur.2020.00474

43. Marshall RS, Perera GM, Lazar RM, Krakauer JW, Constantine RC, DeLaPaz RLJS. Evolution of cortical activation during recovery from corticospinal tract infarction. 2000;31(3):656-661.

44. Senesh MR, Barragan K, Reinkensmeyer DJ. Rudimentary Dexterity Corresponds With Reduced Ability to Move in Synergy After Stroke: Evidence of Competition Between

Corticoreticulospinal and Corticospinal Tracts? *Neurorehabilitation and Neural Repair*. 2020:1545968320943582.

45. Faria-Fortini I, Michaelsen SM, Cassiano JG, Teixeira-Salmela LF. Upper Extremity Function in Stroke Subjects: Relationships between the International Classification of Functioning, Disability, and Health Domains. *Journal of Hand Therapy*. Jul-Sep 2011;24(3):257-264. doi:10.1016/j.jht.2011.01.002

46. Lang CE, Bland MD, Bailey RR, Schaefer SY, Birkenmeier RL. Assessment of upper extremity impairment, function, and activity after stroke: foundations for clinical decision making. *Journal of Hand Therapy*. 2013;26(2):104-115.

47. Pedretti LW, Early MB. *Occupational therapy: Practice skills for physical dysfunction*. Mosby London; 2001.

48. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. *Brain*. May 2012;135:1639-1649. doi:10.1093/brain/awr318

49. Mathiowetz V, Kashman N, Volland G, Weber K, Dowe M, Rogers S. Grip and Pinch Strength - Normative Data for Adults. *Archives of Physical Medicine and Rehabilitation*. 1985;66(2):69-74.

50. Larson CC, Ye Z. Development of an updated normative data table for hand grip and pinch strength: A pilot study. *Computers in Biology and Medicine*. Jul 1 2017;86:40-46. doi:10.1016/j.compbiomed.2017.01.021

51. Suda M, Kawakami M, Okuyama K, et al. Validity and Reliability of the Semmes-Weinstein Monofilament Test and the Thumb Localizing Test in Patients With Stroke. *Front Neurol*. 2020;11:625917. doi:10.3389/fneur.2020.625917

52. Villepinte C, Catella E, Martin M, et al. Validation of French upper limb Erasmus modified Nottingham Sensory Assessment in stroke. *Ann Phys Rehabil Med*. Jan 2019;62(1):35-42. doi:10.1016/j.rehab.2018.03.004

53. Zimney K, Dendinger G, Engel M, Mitzel J. Comparison of reliability and efficiency of two modified two-point discrimination tests and two-point estimation tactile acuity test. *Physiother Theory Pract*. Jan 2022;38(1):235-244. doi:10.1080/09593985.2020.1719563

54. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J Rehabil Med*. 1975 1975;7(1):13-31.

55. Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. *Physical Therapy*. 1983;63(10):1606-1610. doi:10.1093/ptj/63.10.1606

56. Gladstone DJ, Danells CJ, Black SE. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabilitation and neural repair*. 2002;16(3):232-240.

57. Pandyan AD, Johnson GR, Price CIM, Curless RH, Barnes MP, Rodgers H. A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. *Clinical Rehabilitation*. Oct 1999;13(5):373-383. doi:Doi 10.1191/026921599677595404

58. Ansari NN, Naghdi S, Mashayekhi M, Hasson S, Fakhari Z, Jalaie S. Intra-rater reliability of the Modified Modified Ashworth Scale (MMAS) in the assessment of upper-limb muscle spasticity. *Neurorehabilitation*. 2012;31(2):215-222. doi:10.3233/Nre-2012-0791

59. Bohannon RW. Motricity index scores are valid indicators of paretic upper extremity strength following stroke. *Journal of Physical Therapy Science*. 1999;11(2):59-61.

60. Carr JH, Shepherd RB, Nordholm L, Lynne D. Investigation of a new motor assessment scale for stroke patients. *Physical therapy*. 1985;65(2):175-180.

61. Dean C, Mackey F. Motor assessment scale scores as a measure of rehabilitation outcome following stroke. *Australian Journal of Physiotherapy*. 1992;38(1):31-35.

62. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. *Ann Clin Transl Neurol*. Nov 2017;4(11):811-820. doi:10.1002/acn3.488

63. Tse T, Douglas J, Lentin P, Carey L. Measuring Participation After Stroke: A Review of Frequently Used Tools. *Archives of Physical Medicine and Rehabilitation*. Jan 2013;94(1):177-192. doi:10.1016/j.apmr.2012.09.002

64. Morris DM, Uswatte G, Crago JE, Cook III EW, Taub E. The reliability of the wolf motor function test for assessing upper extremity function after stroke. *Archives of physical medicine and rehabilitation*. 2001;82(6):750-755.

65. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. *Stroke*. 2001;32(7):1635-1639.

66. Kontson K, Marcus I, Myklebust B, Civillico E. Targeted box and blocks test: Normative data and comparison to standard tests. *Plos One*. May 19 2017;12(5)doi:ARTN e0177965

10.1371/journal.pone.0177965

67. Mathiowetz V, Volland G, Kashman N, Weber K. Nine hole peg test. *Measurement in Neurological Rehabilitation, ed DT Wade (New York, NY: Oxford University Press)*. 1992;171

68. Herren K, Radlinger L. Validity of Two Versions of the Nine Hole Peg Test in Stroke Patients Pilot Study. *Physoscience*. Jun 2008;4(2):60-64. doi:10.1055/s-2008-1027161

69. Sears ED, Chung KC. Validity and Responsiveness of the Jebsen-Taylor Hand Function Test. *J Hand Surg-Am*. Jan 2010;35a(1):30-37. doi:10.1016/j.jhsa.2009.09.008

70. Gowland C, Stratford P, Ward M, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. *Stroke*. 1993;24(1):58-63.

71. Gustafsson LA, Turpin MJ, Dorman CM. Clinical utility of the Chedoke Arm and Hand Activity Inventory for stroke rehabilitation. *Can J Occup Ther*. Jun 2010;77(3):167-173. doi:10.2182/cjot.2010.77.3.6

72. Johnson D, Harris JE, Stratford P, Richardson J. Interrater Reliability of Three Versions of the Chedoke Arm and Hand Activity Inventory. *Physiotherapy Canada*. Spr 2018;70(2):133-140. doi:10.3138/ptc.2016-70

73. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. *Neurorehabilitation and Neural Repair*. Jan-Feb 2008;22(1):78-90. doi:10.1177/1545968307305353

74. Bailey RR, Lang CE. Upper-limb activity in adults: referent values using accelerometry. *J Rehabil Res Dev*. 2013;50(9):1213-22. doi:10.1682/JRRD.2012.12.0222

75. Hsieh C-L, Hsueh IP, Chiang F-M, Lin P-H. Inter-rater reliability and validity of the action research arm test in stroke patients. *Age and ageing*. 1998;27(2):107-113.

76. Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. *Arch Phys Med Rehabil*. Jan 2001;82(1):14-9. doi:10.1053/apmr.2001.18668

77. Pike S, Lannin NA, Wales K, Cusick A. A systematic review of the psychometric properties of the Action Research Arm Test in neurorehabilitation. *Australian occupational therapy journal*. 2018;65(5):449-471.

78. Waddell KJ, Lang CE. Comparison of Self-Report Versus Sensor-Based Methods for Measuring the Amount of Upper Limb Activity Outside the Clinic. *Arch Phys Med Rehabil*. Sep 2018;99(9):1913-1916. doi:10.1016/j.apmr.2017.12.025

79. Smith BA, Lang CE. Sensor Measures of Symmetry Quantify Upper Limb Movement in the Natural Environment Across the Lifespan. *Arch Phys Med Rehabil*. Jun 2019;100(6):1176-1183. doi:10.1016/j.apmr.2019.01.004

80. Uswatte G, Taub E, Morris D, Vignolo M, McCulloch K. Reliability and validity of the upper-extremity Motor Activity Log-14 for measuring real-world arm use. *Stroke*. 2005;36(11):2493-2496.

81. Uswatte G, Taub E, Morris D, Light K, Thompson P. The Motor Activity Log-28 assessing daily use of the hemiparetic arm after stroke. *Neurology*. 2006;67(7):1189-1194.

82. Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The stroke impact scale version 2.0: evaluation of reliability, validity, and sensitivity to change. *Stroke*. 1999;30(10):2131-2140.

83. Dromerick AW, Lang CE, Birkenmeier R, Hahn MG, Sahrmann SA, Edwards DF. Relationships between upper-limb functional limitation and self-reported disability 3 months after stroke. 2006;

84. Uswatte G, Foo WL, Olmstead H, Lopez K, Holand A, Simms LB. Ambulatory monitoring of arm movement using accelerometry: an objective measure of upper-extremity rehabilitation in persons with chronic stroke. *Arch Phys Med Rehabil*. Jul 2005;86(7):1498-501. doi:10.1016/j.apmr.2005.01.010

85. Bradburn NM, Rips LJ, Shevell SK. Answering Autobiographical Questions - the Impact of Memory and Inference on Surveys. *Science*. Apr 10 1987;236(4798):157-161. doi:DOI 10.1126/science.3563494

86. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field. *Sensors*. 2020;20(20):5744.

87. David A, Subash T, Varadhan S, Melendez-Calderon A, Balasubramanian S. A Framework for Sensor-based Assessment of Upper-limb Functioning. *bioRxiv*. 2021;

88. Urbin M, Bailey RR, Lang CE. Validity of body-worn sensor acceleration metrics to index upper extremity function in hemiparetic stroke. *Journal of neurologic physical therapy: JNPT*. 2015;39(2):111.

89. Bailey RR, Birkenmeier RL, Lang CE. Real-world affected upper limb activity in chronic stroke: an examination of potential modifying factors. *Topics in stroke rehabilitation*. 2015;22(1):26-33.

90. Barth J, Geed S, Mitchell A, Lum PS, Edwards DF, Dromerick AW. Characterizing upper extremity motor behavior in the first week after stroke. *PLoS One*. 2020;15(8):e0221668. doi:10.1371/journal.pone.0221668

91. Gebruers N, Truijen S, Engelborghs S, Nagels G, Brouns R, De Deyn PP. Actigraphic measurement of motor deficits in acute ischemic stroke. *Cerebrovascular Diseases*. 2008;26(5):533-540.

92. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of accelerometry for monitoring real-world arm activity in patients with subacute stroke: evidence from the extremity constraint-induced therapy evaluation trial. *Arch Phys Med Rehabil*. Oct 2006;87(10):1340-5. doi:10.1016/j.apmr.2006.06.006

93. Uswatte G, Hobbs Qadri L. A behavioral observation system for quantifying arm activity in daily life after stroke. *Rehabil Psychol*. Nov 2009;54(4):398-403. doi:10.1037/a0017501

94. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang, C. E. Upper Limb Performance in Daily Life Improves Over the First 12 Weeks Poststroke. *Neurorehabilitation and Neural Repair*. 2019;33(10):836-847.

95. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of Accelerometry for Monitoring Real-World Arm Activity in Patients With Subacute Stroke: Evidence From the Extremity Constraint-Induced Therapy Evaluation Trial. *Arch Phys Med Rehabil*. 2006/10/01/ 2006;87(10):1340-1345. doi:https://doi.org/10.1016/j.apmr.2006.06.006

96. Bailey RR, Klaesner JW, Lang CE. Quantifying Real-World Upper-Limb Activity in Nondisabled Adults and Adults With Chronic Stroke. *Neurorehabil Neural Repair*. Nov-Dec 2015;29(10):969-78. doi:10.1177/1545968315583720

97. Urbin MA, Waddell KJ, Lang CE. Acceleration metrics are responsive to change in upper extremity function of stroke survivors. *Arch Phys Med Rehabil*. May 2015;96(5):854-61. doi:10.1016/j.apmr.2014.11.018

98. Bailey RR, Klaesner JW, Lang CE. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. *PLoS One*. 2014;9(7):e103135. doi:10.1371/journal.pone.0103135

99. Barth J, Klaesner JW, Lang CE. Relationships between accelerometry and general compensatory movements of the upper limb after stroke. *Journal of Neuroengineering and Rehabilitation*. Oct 20 2020;17(1)doi:ARTN 13810.1186/s12984-020-00773-4

100. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field. *Sensors*. Oct 2020;20(20)doi:ARTN 574410.3390/s20205744

101. Chin LF, Hayward KS, Brauer SG. Factors Influencing Paretic Upper Limb Use During First 4 Weeks After Stroke: A Cross-Sectional Accelerometry Study. *American Journal of Physical Medicine & Rehabilitation*. 2021;100(2)

102. Lundquist CB, Nielsen JF, Brunner IC. Prediction of Upper Limb use Three Months after Stroke: A Prospective Longitudinal Study. *J Stroke Cerebrovasc*. Nov 2021;30(11)doi:ARTN 106025

10.1016/j.jstrokecerebrovasdis.2021.106025

103. Rand D, Eng JJ. Predicting daily use of the affected upper extremity 1 year after stroke. *Journal of Stroke and Cerebrovascular Diseases*. 2015;24(2):274-283.

104. Doman CA, Waddell KJ, Bailey RR, Moore JL, Lang CE. Changes in Upper-Extremity Functional Capacity and Daily Performance During Outpatient Occupational Therapy for People With Stroke. *Am J Occup Ther*. May-Jun 2016;70(3):7003290040p1-7003290040p11. doi:10.5014/ajot.2016.020891

105. Lang CE, Waddell KJ, Barth J, Holleran CL, Strube MJ, Bland MD. Upper limb performance in daily life plateaus around three to six weeks post stroke. *Neurorehabilitation and Neural Repair*. 2021;35(10):903-914. doi:10.1177/15459683211041302

106. Chang FH, Coster WJ, Helfrich CA. Community Participation Measures for People With Disabilities: A Systematic Review of Content From an International Classification of Functioning, Disability and Health Perspective. *Archives of Physical Medicine and Rehabilitation*. Apr 2013;94(4):771-781. doi:10.1016/j.apmr.2012.10.031

107. Dijkers MP. Issues in the Conceptualization and Measurement of Participation: An Overview. *Archives of Physical Medicine and Rehabilitation*. Sep 2010;91(9):S5-S16. doi:10.1016/j.apmr.2009.10.036

108. Eriksson G, Baum MC, Wolf TJ, Connor LT. Perceived Participation After Stroke: The Influence of Activity Retention, Reintegration, and Perceived Recovery. *The American Journal of Occupational Therapy*. 2013;67(6):e131-e138. doi:10.5014/ajot.2013.008292

109. Foley E, Connor L, Nicholas M, Baum C. Influence of Environmental Factors on Social Participation Poststroke. *American Journal of Occupational Therapy*. Aug 2019;73(4)doi:10.5014/ajot.2019.73S1-PO7029

110. Brown M, Dijkers MPJM, Gordor WA, Ashman T, Charatz H, Cheng ZF. Participation objective, participation subjective - A measure of participation combining outsider and insider perspectives. *J Head Trauma Rehab*. Nov-Dec 2004;19(6):459-481. doi:Doi 10.1097/00001199-200411000-00004

111. Enderby PM, Wood VA, Wade DT, Hewer RL. The Frenchay Aphasia Screening Test: a short, simple test for aphasia appropriate for non-specialists. *International Rehabilitation Medicine*. 1986/01/01 1986;8(4):166-170. doi:10.3109/03790798709166209

112. Schuling J, Dehaan R, Limburg M, Groenier KH. The Frenchay Activities Index - Assessment of Functional Status in Stroke Patients. *Stroke*. Aug 1993;24(8):1173-1177. doi:Doi 10.1161/01.Str.24.8.1173

113. Post MWM, de Witte LP. Good inter-rater reliability of the Frenchay Activities Index in stroke patients. *Clinical Rehabilitation*. Aug 2003;17(5):548-552. doi:10.1191/0269215503cr648oa

114. Appelros P. Characteristics of the Frenchay Activities Index one year after a stroke: A population-based study. *Disability and Rehabilitation*. May 30 2007;29(10):785-790. doi:10.1080/09638280600919715

115. Baum CM, Edwards D. Activity card sort: test manual. *St Louis: Washington University School of Medicine*. 2001;

116. Duncan PW, Wallace D, Studenski S, Lai SM, Johnson D. Conceptualization of a new stroke-specific outcome measure: the stroke impact scale. *Topics in stroke rehabilitation*. 2001;8(2):19-33.

117. Barclay R, Ripat J, Mayo N. Factors describing community ambulation after stroke: a mixed-methods study. *Clin Rehabil*. May 2015;29(5):509-21. doi:10.1177/0269215514546769

118. Stark SL, Edwards DF, Hollingsworth H, Gray DB. Validation of the reintegration to normal living indexin a population of community-dwelling people with mobility limitations. *Archives of physical medicine and rehabilitation*. 2005;86(2):344-345.

119. Stinear CM, Barber PA, Petoe M, Anwar S, Byblow WD. The PREP algorithm predicts potential for upper limb recovery after stroke. *Brain*. Aug 2012;135(Pt 8):2527-35. doi:10.1093/brain/aws146

120. Coupar F, Pollock A, Rowe P, Weir C, Langhorne P. Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. *Clinical rehabilitation*. 2012;26(4):291-313.

121. Kennedy GM, Brock KA, Lunt AW, Black SF. Factors Influencing Selection for Rehabilitation After Stroke: A Questionnaire Using Case Scenarios to Investigate Physician Perspectives and Level of Agreement. *Archives of Physical Medicine and Rehabilitation*. Aug 2012;93(8):1457-1459. doi:10.1016/j.apmr.2011.11.036

122. Cormier DJ, Frantz MA, Rand E, Stein J. Physiatrist referral preferences for postacute stroke rehabilitation. *Medicine*. Aug 2016;95(33)doi:ARTN e4356

10.1097/MD.0000000000004356

123. Stinear CM, Smith M-C, Byblow WD. Prediction Tools for Stroke Rehabilitation. *Stroke*. 2019;50(11):3314-3322. doi:doi:10.1161/STROKEAHA.119.025696

124. Nijland RH, van Wegen EE, Harmeling-van der Wel BC, Kwakkel G, Early Prediction of Functional Outcome After Stroke I. Accuracy of physical therapists' early predictions of upper-limb function in hospital stroke units: the EPOS Study. *Physical therapy*. 2013;93(4):460-469.

125. Lundquist CB, Pallesen H, Tjørnhøj-Thomsen T, Brunner IC. Exploring physiotherapists' and occupational therapists' perceptions of the upper limb prediction algorithm PREP2 after stroke in a rehabilitation setting: a qualitative study. *BMJ Open*. 2021;11(4):e038880. doi:10.1136/bmjopen-2020-038880

126. Xian Y, Thomas L, Liang L, et al. Unexplained Variation for Hospitals' Use of Inpatient Rehabilitation and Skilled Nursing Facilities After an Acute Ischemic Stroke. *Stroke*. Oct 2017;48(10):2836-+. doi:10.1161/Strokeaha.117.016904

127. Katrak P, Bowring G, Conroy P, Chilvers M, Poulos R, McNeil D. Predicting upper limb recovery after stroke: the place of early shoulder and hand movement. *Archives of Physical Medicine and Rehabilitation*. 1998;79(7):758-761.

128. Kim B, Winstein C. Can neurological biomarkers of brain impairment be used to predict poststroke motor recovery? A systematic review. *Neurorehabilitation and neural repair*. 2017;31(1):3-24.

129. Kwah LK, Herbert RD. Prediction of walking and arm recovery after stroke: a critical review. *Brain sciences*. 2016;6(4):53.

130. Horn U, Grothe M, Lotze M. MRI biomarkers for hand-motor outcome prediction and therapy monitoring following stroke. *Neural Plasticity*. 2016;2016

131. Puig J, Blasco G, Schlaug G, et al. Diffusion tensor imaging as a prognostic biomarker for motor recovery and rehabilitation after stroke. *Neuroradiology*. Apr 2017;59(4):343-351. doi:10.1007/s00234-017-1816-0

132. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the Upper Limb Prediction Algorithm PREP2 Applied 2 Weeks Poststroke: A Prospective Longitudinal Study. *Neurorehabilitation and Neural Repair*. Jan 2021;35(1):68-78. doi:10.1177/1545968320971763

133. Kwakkel G, Kollen B, Twisk J. Impact of Time on Improvement of Outcome After Stroke. *Stroke*. 2006;37(9):2348-2353. doi:doi:10.1161/01.STR.0000238594.91938.1e

134. Winters C, Kwakkel G, Nijland R, Van Wegen E, Consortium E-S. When does return of voluntary finger extension occur post-stroke? A prospective cohort study. *PloS one*. 2016;11(8):e0160528.

135. Cortes JC, Goldsmith J, Harran MD, et al. A short and distinct time window for recovery of arm motor control early after stroke revealed with a global measure of trajectory kinematics. *Neurorehabilitation and neural repair*. 2017;31(6):552-560.

136. Kiær C, Lundquist CB, Brunner I. Knowledge and application of upper limb prediction models and attitude toward prognosis among physiotherapists and occupational therapists in the clinical stroke setting. *Topics in stroke rehabilitation*. 2020:1-7.

137. Winters C, van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. *Neurorehabilitation and neural repair*. 2015;29(7):614-622.

138. Prabhakaran S, Zarahn E, Riley C, et al. Inter-individual variability in the capacity for motor recovery after ischemic stroke. *Neurorehabilitation and neural repair*. 2008;22(1):64-71.

139. Smania N, Paolucci S, Tinazzi M, et al. Active finger extension: a simple movement predicting recovery of arm function in patients with acute stroke. *Stroke*. 2007;38(3):1088-1090.

140. Smith MC, Ackerley SJ, Barber PA, Byblow WD, Stinear CM. PREP2 algorithm predictions are correct at 2 years poststroke for most patients. *Neurorehabilitation and neural repair*. 2019;33(8):635-642.

141. Connell LA, Smith MC, Byblow WD, Stinear CM. Implementing biomarkers to predict motor recovery after stroke. *NeuroRehabilitation*. 2018;43:41-50. doi:10.3233/NRE-172395

142. Luker J, Lynch E, Bernhardsson S, Bennett L, Bernhardt J. Stroke survivors' experiences of physical rehabilitation: a systematic review of qualitative studies. *Archives of physical medicine and rehabilitation*. 2015;96(9):1698-1708. e10.

143. Luker J, Murray C, Lynch E, Bernhardsson S, Shannon M, Bernhardt J. Carers' experiences, needs, and preferences during inpatient stroke rehabilitation: a systematic review of qualitative studies. *Archives of physical medicine and rehabilitation*. 2017;98(9):1852-1862. e13.

144. Hall MJ, Levant S, DeFrances CJ. *Hospitalization for stroke in US hospitals, 1989-2009*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.; 2012.

145. Dyall L, Feigin V, Brown P, Roberts M. Stroke: A picture of health disparities in New Zealand. *Social Policy Journal of New Zealand*. 2008;33:178.

146. Kissela BM, Khoury JC, Alwell K, et al. Age at stroke: temporal trends in stroke incidence in a large, biracial population. *Neurology*. 2012;79(17):1781-1787. doi:10.1212/WNL.0b013e318270401d

147. Benjamin EJ, Muntner P, Alonso A, et al. Heart disease and stroke Statistics-2019 update a report from the American Heart Association. *Circulation*. 2019;

148. Connell LA, Smith M-C, Byblow WD, Stinear CM. Implementing biomarkers to predict motor recovery after stroke. *NeuroRehabilitation*. 2018;43:41-50. doi:10.3233/NRE-172395

149. Hoonhorst MH, Nijland RH, van den Berg PJ, Emmelot CH, Kollen BJ, Kwakkel G. Does transcranial magnetic stimulation have an added value to clinical assessment in predicting upper-limb function very early after severe stroke? *Neurorehabilitation and neural repair*. 2018;32(8):682-690.

150. Kwah LK, Diong J. National Institutes of Health Stroke Scale (NIHSS). *J Physiother*. Mar 2014;60(1):61-61. doi:10.1016/j.jphys.2013.12.012

151. Andrews K, Steward J. Sroke recovery: He can but does he? *Rheumatology*. 1979;18(1):43-48. doi:10.1093

152. Thrane G, Emaus N, Askim T, Anke A. Arm use in patients with subacute stroke monitored by accelerometry: association with motor impairment and influence on self-dependence. *Journal of Rehabilitation Medicine*. // 2011;43(4):299-304. doi:10.2340/16501977-0676

153. Waddell KJ, Tabak RG, Strube MJ, Haire-Joshu D, Lang CE. Belief, confidence, and motivation to use the paretic upper limb in daily life over the first 24 weeks after stroke. *Journal of Neurologic Physical Therapy*. 2019;43(4):197-203. doi:doi:10.1097/NPT.0000000000000287

154. Smith PF, Ganesh S, Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J Neurosci Meth*. Oct 30 2013;220(1):85-91. doi:10.1016/j.jneumeth.2013.08.024

155. Rosso C, Lamy JC. Prediction of motor recovery after stroke: being pragmatic or innovative? *Current Opinion in Neurology*. Aug 2020;33(4):482-487. doi:10.1097/Wco.0000000000000843

156. David A, Subash T, Varadhan SKM, Melendez-Calderon A, Balasubramanian S. A Framework for Sensor-Based Assessment of Upper-Limb Functioning in Hemiparesis. *Front Hum Neurosci*. 2021;15:667509. doi:10.3389/fnhum.2021.667509

157. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B*. 2001;63:411-423. doi:Doi 10.1111/1467-9868.00293

158. Lin CH, Hsu KC, Johnson KR, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput Meth Prog Bio*. Jul 2020;190doi:ARTN 10538110.1016/j.cmpb.2020.105381

159. Tozlu C, Edwards D, Boes A, et al. Machine Learning Methods Predict Individual Upper-Limb Motor Impairment Following Therapy in Chronic Stroke. *Neurorehabil Neural Repair*. May 2020;34(5):428-439. doi:10.1177/1545968320909796

160. Kuhn M, Johnson K. *Applied predictive modeling*. Springer; 2016:600.

161. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. vol 112. Springer; 2013.

162. Miller AE, Russel E, Reisman DS, Kim HE, Vu D. A machine learning approach to identifying important features for achieving step thresholds in individuals with chronic stroke. *medRxiv*. 2021;doi:doi: https://doi.org/10.1101/2022.01.27.22269966

163. Lundquist CB, Nielsen JF, Brunner IC. Prediction of Upper Limb use Three Months after Stroke: A Prospective Longitudinal Study. *Journal of Stroke and Cerebrovascular Diseases*. 2021/11/01/ 2021;30(11):106025. doi:https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106025

164. Uswatte G, Miltner WHR, Foo B, Varma M, Moran S, Taub E. Objective measurement of functional upper-extremity movement using accelerometer recordings transformed with a threshold filter. *Stroke*. Mar 2000;31(3):662-667. doi:Doi 10.1161/01.Str.31.3.662

165. Uswatte G, Taub E. Participant-centered and objective measures of real-world arm function in persons with stroke. *Circulation*. May 25 2004;109(20):E243-E243.

166. Che Bakri NA, Kwasnicki RM, Dhillon K, et al. Objective Assessment of Postoperative Morbidity After Breast Cancer Treatments with Wearable Activity Monitors: The "BRACELET" Study. *Ann Surg Oncol*. Oct 2021;28(10):5597-5609. doi:10.1245/s10434-021-10458-4

167. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*. 2009;338:b606.

168. Saltão da Silva MA, Cook C, Stinear CM, Wolf SL, Borich MR. Paretic upper extremity strength at acute rehabilitation evaluation predicts motor function outcome after stroke. *medRxiv*. 2021;

169. Rand D, Eng, J. J., Tang, P. F., Jeng, J. S., Hung, C. How active are people with stroke? Use of accelerometers to assess physical activity. *Stroke*. 2009;40(1):163-168.

170. Brownlee J. Bagging and Random Forest Ensemble Algorithms for Machine Learning. Accessed 11/15/2021, 2021. https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/?utm_source=drip&utm_medium=email&utm_campaign=Machine+Learning+Mastery+Crash+Course&utm_content=Combine+predictions+with+ensemble+methods

171. Barth J, Lohse KR, Konrad JD, Bland MD, Lang CE. Sensor based categorization of upper limb perforance in daily life of persons with and without neurological upper limb deficits. *Frontiers in Rehabilitation Science*. 2021;2doi:10.3389/fresc.2021.741393

Chapter 2: Accuracy of an algorithm in predicting upper limb functional capacity in a United States population

This chapter has been published:

Barth J, Waddell KJ, Bland MD, Lang CE. Accuracy of an algorithm in predicting upper limb functional capacity in a United States population. *Archives of Physical Medicine and Rehabilitation.* August 2021; 103(1):44-51.

2.1 Abstract

Objective: To determine the accuracy of an algorithm, using clinical measures only, on a sample of persons with first ever stroke in the US. It was hypothesized that algorithm accuracy would fall in a range of 70-80%.

Design: Secondary analysis of prospective, observational, longitudinal cohort; two assessments were done, (1) within 48 hours to 1 week post stroke and (2) at 12 weeks post stroke.

Setting: Recruited from a large acute care hospital and followed over first 6 months after stroke.

Participants: Adults with first ever stroke (N=49) with paresis of the upper limb (UL) at $\leq$48 hours who could follow 2-step commands and were expected to return to independent living at 6 months.

Intervention: NA

Main Outcome Measure(s): The overall accuracy of the algorithm with clinical measures was quantified by comparing predicted (expected) and actual (observed) categories using a correct classification rate (CCR).

Results: The overall accuracy (61%) and weighted kappa (62%) were significant. Sensitivity was high for the Excellent (95%) and Poor (81%) algorithm categories. Specificity was high for the Good (82%), Limited (98%) and Poor (95%) categories. PPV was high for Poor (82%) and NPV was high for all categories. No differences in participant characteristics were found between those with accurate or inaccurate predictions.

Conclusions: The results of the present study found that use of an algorithm with clinical measures only is better than chance alone (chance = 25% for each of the 4 categories) at predicting a category of UL capacity at 3 months post stroke. The moderate to high values of sensitivity, specificity, PPV and NPV demonstrates some clinical utility of the algorithm within healthcare settings in the US.

2.2 Introduction

Of the many people who suffer a stroke each year, a staggering 80% of those individuals will incur some degree of impairment of the upper limb (UL).[1] The ability to predict UL outcomes for an individual facilitates treatment selection, discharge planning and achievable goal setting for clinicians and persons with stroke.[2-4] Several prediction algorithms have been proposed and evaluated in clinical trials with the limitation that most are only accurate for persons with mild to moderate UL impairment.[2, 4-9] The Predict Recovery Potential (PREP2) algorithm, developed in New Zealand, allows for prediction of UL functional capacity at 90 days post stroke based on measures taken within the first week.[5] PREP2 was developed from a retrospective analysis of data from two previous studies (n=207) to improve upon a previously developed prediction model (PREP). [5, 10] The goal of PREP2 was to determine if two of the tests (Transcranial magnetic stimulation (TMS) and Magnetic resonance imaging (MRI)) could be used with fewer patients or completely eliminated, while maintaining a high level of overall accuracy.[5, 10] When developing PREP2, the authors used a hypothesis-free, data-driven analysis method (CART)[5] to determine which factors (patient demographic, clinical measures, neurobiomarker and neurophysiological measures) best predicted the outcome category.[5] The result, was a new algorithm that is sequential in nature. Variables selected were: a measure of strength in the paretic UE (SAFE score)[5], participant age, a measure of stroke severity (National

Institute of Health Stroke Scale (NIHSS))[11] and TMS to determine the presence of a motor evoked potential (MEP) for patients with less initial strength in their paretic UE.[5] The algorithm classifies individuals into one of four clinically meaningful categories (Excellent, Good, Limited or Poor) of eventual UL functional capacity that can guide rehabilitation (Table 2.1).[5, 12] With these variables, a category of expected UL outcome at 90 days after stroke is predicted with an overall accuracy (proportion with correct predictions to the total sample) of 75% using clinical measures and a single neurophysiological measure (TMS).[5] The best accuracy (78%) of the prediction model is for persons with greater initial strength in their paretic UE (SAFE $\geq$ 5) who are predicted to the top half of the model (Excellent and Good), the accuracy drops to 70% for patients with less strength in the paretic UE (SAFE <5) who are predicted to Good, Limited or Poor categories.[5] Prediction of a category rather than a test score for an individual has high clinical utility [13] and is separate from the debated issue of the proportional recovery rule of neurobiological recovery.[6, 7, 14, 15] Of critical importance for clinical practice, the predictions of the algorithm maintained accuracy for 83% (n=71/86) of individuals at 2 years post stroke.[12] External validation studies are required to test the algorithm on other populations of persons with stroke outside of New Zealand.[16] Lundquist et. al. have replicated the PREP2 at a slightly later time point and within a European country.[9]

Table 2.1. Description of categories of UL capacity

Category Name	ARAT Score Ranges[17]	Expected UL Outcome[5, 17]	Rehabilitation Focus[5, 17]
Excellent	51-57	Potential to make a complete, or near complete recovery of affected arm and hand within 3 months.	Promote normal use of the affected hand and arm with task-specific practice. Minimize compensatory and adaptive movements of the affected arm and hand during daily tasks.
Good	34-50	Potential to use the affected hand and arm with clumsiness, slowness and weakness for most daily activities within 3 months.	Promote function of the affected hand and arm. Minimize use of the other hand and trunk in task specific practice.
Limited	13-33	Potential to regain some use of the affected hand and arm, daily activities are likely to be completed with significant modification.	Promote adaptation in daily activities using the affected arm and hand as a "helper hand" whenever safely possible.
Poor	0-12	Limited return of useful movement of the affected hand and arm at 3 months.	Prevent secondary complications such as pain, spasticity and shoulder instability, learn to complete daily activities with stronger hand.

Despite the merits of this algorithm, there are some challenges to implementing it in the United States (US). The rehabilitation structure of New Zealand differs somewhat from the US, with respect to who, how much, where, and when individuals receive UL therapy after stroke. People in the US have strokes at younger ages (mean age US = 64 years; NZ = 71 years) [18-20] and many have additional comorbidities that are poorly managed.[21] Another hurdle to implementation of the algorithm is the limited accessibility to the neurophysiologic measurement (TMS) of corticospinal tract function. Access to this test in the US can be present for research purposes and/or in major academic medical centers, but is not available in routine rehabilitation care.[13, 22] It has been acknowledged that making predictions with clinical measures alone is inaccurate and difficult.[12, 23] Anticipating this, the algorithm was shown to still be 71% accurate at 90 days post-stroke using clinical measures only, with a measure of stroke severity (NIHSS)[11] being adjusted to replace the neurophysiologic measure (TMS) in participants with little or no strength (SAFE< 5) in their affected UL.[5] The algorithm with clinical measures only maintains

similar overall accuracy (95% CI = 55%-78%); however, the prediction accuracy drops to 55% for patients with a SAFE score <5.[5]

Prior to widespread implementation of this algorithm into routine rehabilitation care, it is necessary to understand how small system differences and lack of the neurophysiologic tests (TMS) affects the overall accuracy of the algorithm in individuals who have suffered a first stroke in the US. The current study is a secondary analysis (R01 HD068290) of data from a prospective, observational, longitudinal cohort tracking UL change over time. The analysis in this paper determined the accuracy of the algorithm on a sample of persons with first ever stroke, using the clinical measures only,[5] at time points that are most feasible in the US health care system. We hypothesized that algorithm accuracy would fall in a range of 70-80%.

2.3 Methods

This study was a secondary analysis of data collected from a prospective, observational, longitudinal cohort tracking UL change over time. Sources of data utilized were clinical measures and participant demographics at two time points: (1) between 48 hours to 7 days of stroke onset and (2) 12 weeks post stroke.

2.3.1 Participants

Participants were included in the prospective, observational, longitudinal cohort if the following criteria were met: (1) within two weeks of a first-ever ischemic or hemorrhagic stroke, confirmed with neuroimaging; (2) presences of UL motor deficits within the first 24 to 48 hours post stroke, as indicated by a NIHSS[11] Arm Item scores of 1 to 4 or documented manual muscle test grade of <5 anywhere on the paretic UL; (3) able to follow a 2-step command, as measured by a NIHSS Command Items score of zero; and (4) anticipated return to independent living (i.e. not institutionalized), as indicated by the acute stroke team. Participants were excluded from the

study if any of the following criteria were met: (1) history of previous stroke, other neurological condition, or psychiatric diagnoses; (2) presence of comorbid conditions that may limit recovery (e.g., end-stage renal disease or stage IV cancer); (3) lives more than 90 minutes from study location; and (4) currently pregnant by self-report. The Human Research Protection Office at Washington University in St. Louis Missouri approved this study and all participants provided written informed consent.

Cohort participants completed eight assessment sessions over the 24 weeks post stroke. This analysis used the assessment data at the first assessment (48 hours to 7 days), and at 12 weeks. Assessments were administered by trained personnel (licensed PT or OT, range of experience with measures was 2-15 years). The majority of people received their first assessments in the acute hospital setting or inpatient rehabilitation. The 12 week assessments were completed where the participant was located at that time and includes: sub-acute rehabilitation facility, research lab, or home. Since this was an observational cohort study, we did not provide or control for the amount or type of rehabilitation services delivered to enrolled participants. Participants received rehabilitation services as prescribed by their medical team.

2.3.2 Algorithm Measures

Evaluation of accuracy of the algorithm in this US cohort used early clinical measures and participant age to predict the category (Excellent, Good, Limited or Poor) of UL capacity at 12 weeks. Assessments to predict the expected UL category were: 1) the Shoulder Abduction Finger Extension (SAFE) score [5, 12] at time of consent (mean time post-stroke = 7 $\pm$ 3 days, range = 2-14 days); 2) participant age; and 3) NIHSS[11] total score captured around 48 hours post stroke. The first step in the algorithm is a calculation of the SAFE score to quantify impairment [24, 25] of the paretic UL. The total SAFE score is a sum of the Medical Research Council [26]

(MRC) strength grades (0 = no strength to 5 = full strength) for shoulder abduction and finger extension in the paretic UL. The SAFE score of 10 indicates full strength in both movements on the paretic limb. A SAFE score ≥5 places individuals into the upper half of the prediction model, with age predicting the final categorization of Excellent (< 80 years). For individuals equal to and older than 80 years, SAFE score is used again to determine predictions to either the Excellent (SAFE ≥ 8) or Good (SAFE < 8) categories.[5] A SAFE score < 5 places individuals into the lower half of the prediction model, where the NIHSS total score[11] is then used to predict the final outcome categorization. The NIHSS is a global measure of stroke severity[11], and captures stroke impairment across multiple domains.[11, 24] Scores range from 0 to 42, where higher scores indicate more severe stroke.[11] In the version of the algorithm with clinical measures only, persons with an NIHSS total score < 9 are predicted to have a Good outcome and NIHSS ≥10 a Poor outcome.[5]

The dependent variable for this study was category of UL functional capacity, as determined by the Action Research Arm Test (ARAT) score,[5, 10] a standardized measure of activity limitation.[25, 27] The ARAT is a valid and reliable measure of UL capacity[24, 28] (grasp, grip, pinch and gross motor) in adults with UL paresis; scores range from 0-57 with higher scores indicating greater functional capacity of the UL.[29-32] The actual category of UL capacity was determined from the ARAT score at 12 weeks post stroke.[5, 17] Ranges of ARAT scores dictated how people were divided into one of four clinically meaningful categories: “Excellent “(51-57), “Good” (34-50), “Limited” (13-33) and “Poor” (0-12) based on a previously published report (see Table 2.1).[17]

2.3.3 Analysis

All data were analyzed in R (version 4.0.1), an open source statistical computing program. The packages caret, gridExtra, yardstick, and vcd were used for the contingency table (CCR) and the packages ggalluvial and alluvial were used for the alluvial plot in Figure 2.2. Data were visually inspected to determine normality of the distribution. Continuous participant characteristics and 12 week clinical measure scores are summarized by mean, standard deviation (SD) and ranges when normally distributed, otherwise by median and ranges.

The algorithm using only the early clinical measures [5, 12] was used to assign predicted categories. In the process, we had to modify the algorithm slightly to account for participants with an NIHSS of 9, which was not included in the PREP2 algorithm, where the only choices were > 9 or < 9.[5] Here, individuals with an NIHSS total score = 9 were predicted to have a "Limited" outcome. The remaining decision points of PREP2 were unaltered.

The overall accuracy of the algorithm with clinical measures was quantified by comparing the predicted (expected) and actual (observed) ARAT categories using a correct classification rate (CCR). CCR calculates accuracy of the algorithm as a whole and other statistics for each category. Overall accuracy in this analysis is the proportion of individuals with correct predictions with respect to the total sample. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for each category. Additionally, a weighted kappa was calculated to determine the classification accuracy of the whole algorithm. In order to interpret the sensitivity, specificity, PPV and NPV, we classified values from 0.75 and above as being "high", 0.50-0.74 as "moderate" and anything 0.49 and below as "low". We chose these words to avoid confusion with the names on the categories. The weighted kappa was selected instead of the non-weighted version because there was not an

equal, random chance (chance = 25% for each of the 4 categories) of prediction for each category. Following categorical analyses, data were explored further to see if any participant characteristics distinguished between persons with accurate versus inaccurate predictions. Due to the nonparametric nature of the samples, a Wilcoxin signed-rank test was calculated to explore differences in age, initial SAFE score and NIHSS total score, number of days post stroke to SAFE score, and number of comorbidities.

2.4 Results

Overall, the sample of persons with first ever stroke were generally in their sixties and had mild to moderate stroke (90% with NIHSS 0-15). At the time of this analysis, a total of 69 subjects had been enrolled in the study. Enrollment in the prospective, observational, longitudinal cohort study was suspended and then closed due to the on-going COVID-19 pandemic. Only participants with complete initial and 12 week assessment time points were included here (n=49). Reasons for exclusion include: missing initial or week 12 clinical measures (n=14) and data collection suspended due to COVID (n=6). Demographics of the participants are provided in Table 2.2.

Table 2.2 Participant Characteristics. Values are median (range) or number (%).

	n=49
Demographic Characteristics	
Age (years)	
Median age (range)	66 (43-81)
< 80 years n (%)	47 (96%)
Sex	
Male	28 (57%)
Female	21 (43%)
Ethnicity	
Non-Hispanic/ Non-Latino	49 (100%)
Race	
White	27 (55%)
African American	21 (43%)
Asian	1 (2%)
Common Co-morbidities (self-report)	
Diabetes (% yes)	16 (33%)
High Blood Pressure	37 (76%)

Heart Disease	15 (31%)
Stroke Characteristics	
First stroke	49 (100%)
Stroke Type	
Ischemic	43 (88%)
Hemorrhagic	6 (12%)
Stroke Location	
Cortical	31 (63%)
Subcortical	17 (35%)
Cortical & Subcortical	1 (2%)
Post. Circ./ Cerebellar	0
Handedness	
Right	42 (86%)
Left	6 (12%)
Both	1 (2%)
% Concordant	39%
Stroke Severity	
Mild (NIHSS score 0-4)	19 (39%)
Moderate (NIHSS score 5-15)	25 (51%)
Severe (NIHSS score > 16)	5 (10%)
Paretic Upper Limb Measures	
Initial SAFE score	
Excellent outcome median (range)	8 (2-10)
Good outcome median (range)	7 (2-8)
Limited outcome median (range)	2 (0-8)
Poor outcome median (range)	1 (0-2)

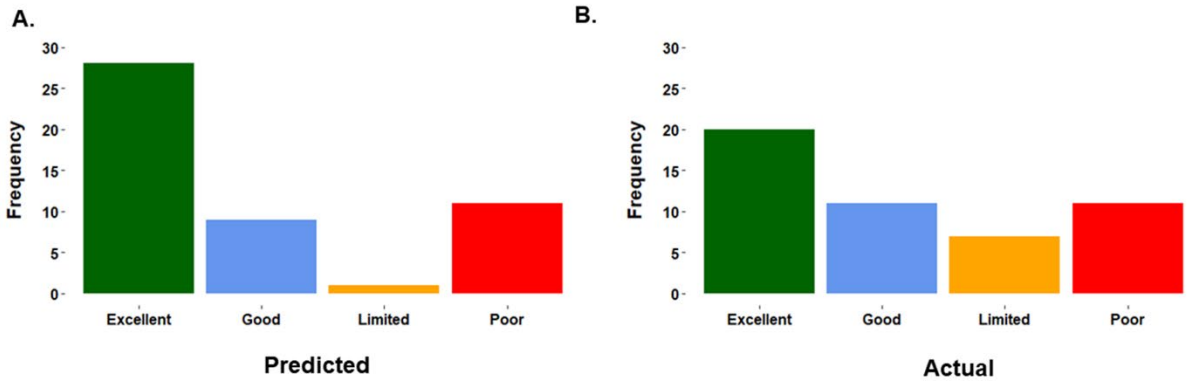


Figure 2.1 Frequency of participants predicted to each category (A) and then actual category observed (B).

Table 2.3: Contingency table of predicted and actual categories

Contingency Table Showing Numbers					
	Actual Category				
Predicted Category	Excellent	Good	Limited	Poor	*Sum*
Excellent	19	8	1	0	*28*
Good	0	2	5	2	*9*
Limited	0	1	0	0	*1*
Poor	1	0	1	9	*11*
Sum	*20*	*11*	*7*	*11*	*49*

Contingency Table Showing Percentages					
	Actual Category				
Predicted Category	Excellent	Good	Limited	Poor	*Sum*
Excellent	39%	16%	2%	0%	*57%*
Good	0%	4%	10%	4%	*18%*
Limited	0%	2%	0%	0%	*2%*
Poor	2%	0%	2%	18%	*22%*
Sum	*41%*	*22%*	*14%*	*22%*	*100%*

Frequencies of participants predicted and confirmed within each category are presented graphically (Figure 2.1) and as a 4x4 contingency table (Table 2.3). The colors of each bar graph are the colors used in the PREP2 analysis: [5] Excellent (green); Good (blue); Limited (orange); and Poor (red). The predicted categories were weighted to the extremes (Excellent and Poor) where the actual categories at 12 weeks were more balanced among the groups.

The algorithm is overall better than chance (chance = 25% for each of the 4 categories) and most useful in determining the category someone will not end up in at 12 weeks. Table 2.4 presents the accuracy and other calculated statistics. The overall accuracy (0.61) and the weighted kappa (0.62) were significant, but lower than hypothesized. Sensitivity, specificity, PPV, and NPV differed across categories. Note that the small numbers within some cells, e.g. only one person was predicted into the Limited category have a large effect on the calculation of these values. Further, a total of 20 people in this sample had an initial SAFE score <5 and the model was accurate for 50% of them.

Table 2.4. Overall statistics (top) and per category statistics (bottom)

Statistic	Value	95% CI Range		p-value
Overall Accuracy	0.61	0.46	0.75	0.003
Weighted Kappa	0.62	0.46	0.78	<.0001
Per Category Values				
Statistic/ Category	Excellent	Good	Limited	Poor
Sensitivity	0.95	0.18	0	0.81
Specificity	0.69	0.82	0.98	0.95
PPV	0.67	0.22	0	0.82
NPV	0.95	0.77	0.85	0.95

People with inaccurate predictions were typically inaccurate within one category of their prediction. Figure 2.2 is an alluvilial plot providing a visual representation of participants who were accurately and inaccurately predicted. The top axis is the predicted category and the bottom axis is the actual category, with the width of the category box representing the frequency of individuals who were within that category for the predicted and actual time points. The colored alluvium in the center of the plot are filled with the predicted category colors: Excellent (green), Good (blue), Limited (orange) and Poor (red). The most movement between categories is seen within the middle of the plot. One can see that the majority of inaccuracies were only off by one category (n=15/19), the remaining were inaccurate by two (n=3) or three (n=1) categories. Those inaccurate by two categories included: (a) two people who were predicted to the Good and ended up in the Poor category (ARAT at 90 days = 0 & 11); and (b) one person who was predicted to the Excellent and ended up in the Limited category (ARAT at 90 days = 26). The one person inaccurate by three categories ended up with a better outcome; predicted to the Poor and ended up in the Excellent category (ARAT at 90 days = 53). This is a good example of an individual where correct classification would likely have been enhanced with TMS. Of those inaccurately predicted, the algorithm was too optimistic for 16 people and too pessimistic for 3 people. The most movement between predicted to actual categories was seen between those predicted to be in Excellent and ended up in Good (n=8).

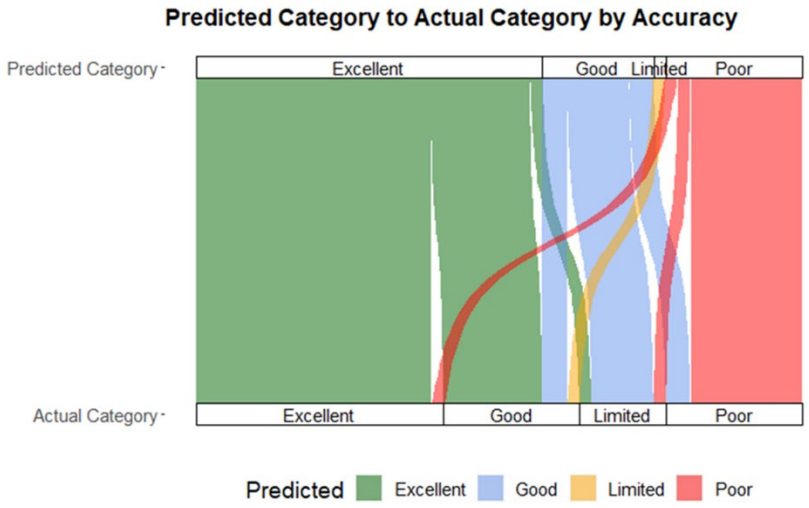


Figure 2.2: Alluvial plot displaying accurate and inaccurate categorizations. The predicted (top) and actual (bottom) axes are scaled such that the width of the box represents the frequency of each category. Note that the narrow width of the Limited category on the predicted axis (top) is due to only one person being predicted to end up in this category. The fill of the color bands are the predicted category. Inaccuracies tended to be one category off, with the exception of a few individuals who had better than expected (e.g. thin red line, Poor to Excellent) or worse than expected (e.g. thin green line, Excellent to Limited) outcomes.

Further exploration of the data revealed that participants with inaccurate predictions were not different from those with accurate predictions on clinical or demographic measures. We found no significant differences between inaccurately vs. accurately categorized individuals with respect to strength in the paretic UL (SAFE score median: inaccurate=5, accurate= 8, p=0.18), the day the SAFE score was captured (day median: inaccurate= 5.5, accurate= 6.3, p=0.33), stroke severity (NIHSS median: inaccurate=5, accurate = 6, p=0.45), age (median: inaccurate: 65 yrs, accurate: 67 yrs, p=0.69), or number of reported comorbidities (mean number ± SD: inaccurate=2.35 ±1.9, accurate = 3.03 ±1.8, p=0.11).

2.5 Discussion

The algorithm, with clinical measures only, on a sample of persons with first ever stroke in the US had an overall accuracy of 61%, which was less than hypothesized. The predicted categories were more weighted to the best and worst categories than the actual categories at 12 weeks. Some of the values of sensitivity, specificity, PPV and NPV were high, but not all.

Participants with inaccurate predictions were most often only one category away from their actual category and no differences were found between those with accurate or inaccurate predictions on days of assessments, clinical or demographic measures.

Implementation of prediction models into routine clinical practice requires external validation studies such as this one.[13, 33] Statistical methods used to evaluate predictive models arise from evaluating diagnostic medical tests,[34] referenced to a gold standard.[35] Medical diagnostic tests must be highly accurate [35], due to the potentially serious downstream consequences of inaccuracy on outcomes (e.g. mortality).[32] In the field of psychology, accuracy statistics of diagnostic tests have been identified as: > 0.90 are considered "excellent", 0.80-0.89 are "good", 0.70 to 0.79 are "adequate" and < 0.70 "may have limited applicability".[36] The rehabilitation field has not yet identified values considered to be good enough for implementation of predictive models into standard of care. Accuracy statistics of rehabilitation prediction models could theoretically be looser than values used to diagnose a life threatening condition or interpret lab values, since the consequences of inaccuracy are not as serious. While the present accuracy values were lower than desirable (61%, 95%CI: 46%-75%), the confidence intervals overlap with the PREP2 models (CI 45%-84%) [5, 10] and with the other external validation study (accuracy: 60%, with CI 50%-71%).[9] These overlapping confidence intervals occurred despite the use of neurophysiologic test (TMS) included in other analyses [5, 9] and the timing differences of the initial assessment ranging from 2-3 days [5] out to two weeks[9]. Thus, the current study and others confirm that use of this algorithm is better than chance alone (chance = 25% for each of the 4 categories) at predicting a category of UL capacity at 3 months from measures taken in the early days after stroke. Given these results, it may be realistic to implement the algorithm, with clinical measures only, into US clinical practice. Implementation

would require the caveat that for persons with less initial strength in their paretic UE (SAFE<5), the prediction is likely less accurate[5] and repeat assessments and predictions will be required.

In the sample used to develop PREP2, 67% of the people were younger than 80 years old,[5] while 96% of people in this US sample were younger than 80. This is likely a function of the average age of stroke in the US vs. New Zealand[18-20] (see Introduction). Having fewer people older than 80 years of age reduced the number of people who could be predicted to end up in the Good category. It is therefore not surprising that in this analysis the most inaccurate predictions were those predicted to Excellent but ended up in Good. Another potential reason for inaccuracies in this analysis could be because of our decision that an NIHSS=9 would be predicted to the Limited category. One person in this sample had an NIHSS=9 and was predicted to the Limited category but ended up in Good (ARAT=39 at 90 days), illustrating an added benefit of TMS, if that had been available and the person had positive motor evoked potentials. In PREP2, no one in their sample had an NIHSS=9, and thus it was not an option in the decision tree.[5] The variables and their ranges in PREP2 were selected by an atheoretical approach that arrives on an algorithm that is dependent on the sample.[5] Lundquist et. al.[9] had similar values of overall accuracy as those observed in this analysis, however with the same data driven analysis method (CART) the same variables were selected but with different ranges for the NIHSS total score.[9] It is anticipated that within the US, different ranges of the variables (SAFE, age and NIHSS) may be needed. It is possible that future prediction models could increase the predictive accuracy, as more is known about differences around the world in populations of persons with stroke.

In the present study, the PPV for all categories was lower than PREP2 but the NPV was higher.[5] Additionally, the categories of Excellent and Poor demonstrated greater overall

accuracy, sensitivity and NPV. The high NPV means the model does well at predicting which categories someone will not end up in. In this analysis 10/19 people with inaccurate predictions had an initial SAFE score <5, contributing to the high PPV and NPV observed in the Poor category. Similar to the PREP2 papers,[5, 12] and other replications[9] of the algorithm, most people with inaccurate predictions were only one category away in the current study. Having a patient who winds up one category away from the original prediction is much easier to manage clinically than one who is two or three categories off.

Predictive models or algorithms are not designed to take over the job of the clinician, but are intended to assist in the clinical decision-making process by providing more objectivity.[37, 38] PREP2 was developed as a tool to provide clinicians, caregivers and persons with stroke more information in the early days after a stroke of the anticipated UL outcome.[33] Predicting a category of the UL outcome rather than a score on a test has clinical utility, as persons with stroke are more interested in what they can generally expect in recovery of the UL.[13, 38-40] A challenge with category boundaries is that a one point difference in the ARAT score at 3 months might be the difference between an accurate vs. inaccurate prediction. In this analysis, four people had a total ARAT score between 47 to 49 points at 90 days, placing them in the Good category when they were predicted to Excellent (bottom ARAT limit = 51). Experienced clinicians might not consider this problematic because a point on a standardized test likely will not change the persons UL ability at 3 months. The predictive algorithm with clinical measures only is clinically efficient and does not require expensive equipment or extensive training, making it possible to use within the confines of the US healthcare system. Additionally, the true PREP2 provides resources[41] to aide clinicians about providing prediction information to persons with stroke. These resources will be extremely important for clinicians in the US to deliver

prediction information from this algorithm, especially to people with an initial SAFE score<5. Even with less than optimal accuracy, we speculate that using this algorithm with information routinely captured during stroke rehabilitation can provide useful information for clinicians and persons with stroke. In the US, discharge disposition is largely dictated by patient insurance, availability of support upon discharge and current mobility status.[42] The predictions provide clinicians with an objective tool and language to efficiently evaluate potential UL outcome, communicate results to the person with stroke and begin the appropriate therapy interventions to meet the outcome.[5, 13, 17] Since the therapy process is fluid, clinicians can modify therapy content appropriately as the individual's progress unfolds.[17]

2.5.1 Study limitations

There are two key limitations to consider when interpreting the results of this analysis. First, our sample size is small which limits generalizability and produces wide confidence intervals. This sample size, however, is similar to that in the initial algorithm (n= 40).[10] A large sample size of 200-300 would be needed to produce enough data within each cell and understand the difference between those who are accurately or inaccurately classified. Second, the inclusion criteria used to enroll participants with first ever stroke excludes persons with substantial cognitive and language deficits. This is similar to the samples used to develop PREP2 which excluded people with impaired UL somatosensation, vision, visuospatial attention and cognition.[5] We only included people with first stroke, but one of the samples used in PREP2 included persons with second stroke. We recognize that future studies will need to include people with these impairments and with second stroke. While it is unknown how the algorithm would work on the full spectrum of stroke survivors, the population enrolled here is typical of those who receive rehabilitation services in the US.

2.6 Conclusions

The present study found that the algorithm, with clinical measures only, is better than chance alone at predicting a category of UL capacity at 3 months post stroke. The values of sensitivity, specificity, PPV and NPV of this study and others demonstrate the potential clinical utility of this predictive algorithm. Until there are agreed upon values of statistics calculated from predictive models for use in the rehabilitation field, we would continue to advocate for the clinical utility of this tool.

2.7 Acknowledgements

We thank research physical therapist Christine Gordon, BS, PT, for her efforts with data collection that made this study possible.

2.8 References

1. Wade D, Langton-Hewer R, Wood VA, Skilbeck C, Ismail H. The hemiplegic arm after stroke: measurement and recovery. Journal of Neurology, Neurosurgery & Psychiatry 1983;46(6):521-4.

2. Katrak P, Bowring G, Conroy P, Chilvers M, Poulos R, McNeil D. Predicting upper limb recovery after stroke: the place of early shoulder and hand movement. Archives of Physical Medicine and Rehabilitation 1998;79(7):758-61.

3. Rand D, Eng JJ. Predicting daily use of the affected upper extremity 1 year after stroke. Journal of Stroke and Cerebrovascular Diseases 2015;24(2):274-83.

4. Kwah LK, Herbert RD. Prediction of walking and arm recovery after stroke: a critical review. Brain sciences 2016;6(4):53.

5. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. Ann Clin Transl Neurol 2017;4(11):811-20.

6. Winters C, van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. Neurorehabilitation and neural repair 2015;29(7):614-22.

7. Prabhakaran S, Zarahn E, Riley C, Speizer A, Chong JY, Lazar RM et al. Inter-individual variability in the capacity for motor recovery after ischemic stroke. Neurorehabilitation and neural repair 2008;22(1):64-71.

8. Smania N, Paolucci S, Tinazzi M, Borghero A, Manganotti P, Fiaschi A et al. Active finger extension: a simple movement predicting recovery of arm function in patients with acute stroke. Stroke 2007;38(3):1088-90.

9. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the Upper Limb Prediction Algorithm PREP2 Applied 2 Weeks Poststroke: A Prospective Longitudinal Study. Neurorehabilitation and Neural Repair 2021;35(1):68-78.

10. Stinear CM, Barber PA, Petoe M, Anwar S, Byblow WD. The PREP algorithm predicts potential for upper limb recovery after stroke. Brain 2012;135(Pt 8):2527-35.

11. Kwah LK, Diong J. National Institutes of Health Stroke Scale (NIHSS). J Physiother 2014;60(1):61-.

12. Smith MC, Ackerley SJ, Barber PA, Byblow WD, Stinear CM. PREP2 algorithm predictions are correct at 2 years poststroke for most patients. Neurorehabilitation and neural repair 2019;33(8):635-42.

13. Connell LA, Smith MC, Byblow WD, Stinear CM. Implementing biomarkers to predict motor recovery after stroke. NeuroRehabilitation 2018;43:41-50.

14. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity. Annals of neurology 2015;78(6):848-59.

15. Stinear CM, Byblow WD, Ackerley SJ, Smith M-C, Borges VM, Barber PA. Proportional motor recovery after stroke: implications for trial design. Stroke 2017;48(3):795-8.

16. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

17. Stinear CM, Byblow WD, Ackerley SJ, Barber PA, Smith M-C. Predicting recovery potential for individual stroke patients increases rehabilitation efficiency. Stroke 2017;48(4):1011-9.

18. Hall MJ, Levant S, DeFrances CJ. Hospitalization for stroke in US hospitals, 1989-2009. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.; 2012.

19. Dyall L, Feigin V, Brown P, Roberts M. Stroke: A picture of health disparities in New Zealand. Social Policy Journal of New Zealand 2008;33:178.

20. Kissela BM, Khoury JC, Alwell K, Moomaw CJ, Woo D, Adeoye O et al. Age at stroke: temporal trends in stroke incidence in a large, biracial population. Neurology 2012;79(17):1781-7.

21. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation 2020;141(9):e139-e596.

22. Hoonhorst MH, Nijland RH, van den Berg PJ, Emmelot CH, Kollen BJ, Kwakkel G. Does transcranial magnetic stimulation have an added value to clinical assessment in predicting upper-limb function very early after severe stroke? Neurorehabilitation and neural repair 2018;32(8):682-90.

23. Nijland RH, van Wegen EE, Harmeling-van der Wel BC, Kwakkel G, Early Prediction of Functional Outcome After Stroke I. Accuracy of physical therapists' early predictions of upper-limb function in hospital stroke units: the EPOS Study. Physical therapy 2013;93(4):460-9.

24. Rauch A, Cieza A, Stucki G. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. European journal of physical and rehabilitation medicine 2008;44(3):329-42.

25. Salter K, Jutai J, Teasell R, Foley N, Bitensky J, Bayley M. Issues for selection of outcome measures in stroke rehabilitation: ICF Participation. Disability and rehabilitation 2005;27(9):507-28.

26. James MA. Use of the Medical Research Council Muscle Strength Grading System in the Upper Extremity. Journal of Hand Surgery 2007;32(2):154-6.

27. Cieza A, Kostansjek N. The International Classification of Functioning, Disability and Health: the first 20 years. Developmental Medicine & Child Neurology 2021;63(4):363-.

28. Organization WH. Towards a common language for functioning, disability, and health: ICF. The international classification of functioning, disability and health 2002.

29. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. Neurorehabilitation and Neural Repair 2008;22(1):78-90.

30. Lang CE, Wagner JM, Dromerick AW, Edwards DF. Measurement of upper-extremity function early after stroke: properties of the action research arm test. Archives of physical medicine and rehabilitation 2006;87(12):1605-10.

31. Platz T, Pinkowski C, van Wijck F, Kim I-H, Di Bella P, Johnson G. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. Clinical rehabilitation 2005;19(4):404-11.

32. Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001;82(1):14-9.

33. Rosso C, Lamy JC. Prediction of motor recovery after stroke: being pragmatic or innovative? Current Opinion in Neurology 2020;33(4):482-7.

34. van Stralen KJ, Stel VS, Reitsma JB, Dekker FW, Zoccali C, Jager KJ. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney International 2009;75(12):1257-63.

35. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Frontiers in Public Health 2017;5(307).

36. Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 1993;78(1):98-104.

37. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. Bmj 2009;338:b606.

38. Stinear CM, Smith M-C, Byblow WD. Prediction Tools for Stroke Rehabilitation. Stroke 2019;50(11):3314-22.

39. Luker J, Lynch E, Bernhardsson S, Bennett L, Bernhardt J. Stroke survivors' experiences of physical rehabilitation: a systematic review of qualitative studies. Archives of physical medicine and rehabilitation 2015;96(9):1698-708. e10.

40. Luker J, Murray C, Lynch E, Bernhardsson S, Shannon M, Bernhardt J. Carers' experiences, needs, and preferences during inpatient stroke rehabilitation: a systematic review of qualitative studies. Archives of physical medicine and rehabilitation 2017;98(9):1852-62. e13.

41. Stinear C, Byblow, Winston and colleagues PRESTO Predict Stroke outcomes. Available from: URL: https://presto.auckland.ac.nz/about-us/.

42. Deom C, Carpenter, J. Bodine, A., Taylor S. A Mobility Measure for Inpatient Rehabilitation Using Multigroup, Multidimensional Methods. Journal of Neurologic Physical Therapy 2021;45(2):101-11.

Chapter 3: Sensor-based categorization of upper limb performance in daily life of persons with and without neurological upper limb deficits

This chapter has been published:

Barth J, Lohse KR, Konrad JD, Bland MD, Lang CE. Sensor-based categorization of upper limb performance in daily life of persons with and without neurological upper limb deficits. *Frontiers in Rehabilitation Science.* October 2021; 2: 1-14.

3.1 Abstract

Background: The use of wearable sensor technology (e.g., accelerometers) for tracking human physical activity has allowed for measurement of actual activity performance of the upper limb (UL) in daily life**.** Data extracted from accelerometers can be used to quantify multiple variables measuring different aspects of UL performance in one or both limbs. A limitation is that several variables are needed to understand the complexity of UL performance in daily life.

Purpose: To identify categories of UL performance in daily life in adults with and without neurological UL deficits.

Methods: This study analyzed data extracted from bimanual, wrist-worn, triaxial accelerometers from adults from three previous cohorts (N=211), two samples of persons with stroke and one sample from neurologically intact adult controls. Data used in these analyses were UL performance variables calculated from accelerometer data, associated clinical measures, and participant characteristics. A total of twelve cluster solutions (3-, 4- or 5-clusters based with 12, 9, 7, or 5 input variables) were calculated to systematically evaluate the most parsimonious solution. Quality metrics and principal component analysis of each solution were calculated to arrive at a locally-optimal solution with respect to number of input variables and number of clusters.

Results: Across different numbers of input variables, two principal components consistently explained the most variance. Across the models with differing numbers of UL input performance variables, a 5-cluster solution explained the most overall total variance (79%) and had the best model-fit.

Conclusion: The present study identified five categories of UL performance formed from five UL performance variables in cohorts with and without neurological UL deficits. Further validation of both the number of UL performance variables and categories will be required on a larger, more heterogeneous sample. Following validation, these categories may be used as outcomes in UL stroke research and implemented into rehabilitation clinical practice.

3.2 Introduction

The use of wearable sensor technology (e.g., accelerometers) for tracking human movement has allowed for efficient measurement of activity of the upper limb (UL) in daily life.[1-6] Accelerometry has become an established, valid and reliable methodology to directly measure performance of UL activity in daily life in neurologically intact adults[7, 8] and adults with stroke.[9-13] Per the World Health Organization International Classification of Functioning, Disability and Health (ICF) model,[14] activity *performance*, defined as what a person does in the unstructured, free-living environment, is a different but related construct to the *capacity* for activity (i.e. functional capacity), which is measured by standardized assessments in the structured clinical or laboratory setting. Clinicians and researchers typically assess a person's functional capacity for activity in the structured clinic or laboratory environments with standardized assessments. However, people seek out rehabilitation services because they want to be able to perform better in their daily lives,[15] and improvements in UL capacity seen in the clinic do not necessarily translate to improvements in UL performance in daily life.[13, 16-19] Therefore, assessment of UL activity performance in an individual's unstructured, free-living environment is critical to evaluating effectiveness of rehabilitation services and determining if the services provided have achieved the goal of improving performance in daily life.

Data extracted from bilateral, wrist-worn accelerometers can be used to quantify variables measuring different aspects of UL performance in one or both limbs. These variables collectively inform clinician scientists about the real-world activity performance. The numerous variables calculated from accelerometers measure different aspects of UL performance, such as: 1) duration;[7, 20] 2) magnitude;[12, 21, 22] 3) variability;[12, 23] 4) symmetry or laterality;[3, 7, 10] and 5) quality of movement.[5, 24-26] Each UL performance variable conveys slightly different information about the collective nature of UL use, with a single variable providing only part of the picture.[5] Furthermore, some variables are narrowly distributed in neurologically-intact (adult controls) individuals (e.g. use ratio, an index of duration of activity of one limb versus the other), while other variables are widely distributed (e.g. bilateral magnitude, a measure of magnitude of bilateral UL activity).[3] Thus, multiple variables quantifying different aspects of movement along with heterogeneous distributions of those variables can make it difficult to interpret UL performance data for clinical decision-making.

One reason wearable sensor technology (e.g. accelerometry) for measurement of UL performance has remained largely confined to rehabilitation research with limited ventures into clinical practice is because the current output from accelerometers is not easily accessible for rehabilitation professionals.[4] A potential solution to the multi-variable problem would be the formation of categories (or groups) of UL performance in daily life. If there were natural groupings that occur among multiple UL performance variables calculated from accelerometry data,[27] then these groupings could help to facilitate clinical decision making and implementation of UL performance data into routine rehabilitation care. In other biomedical science fields, formation of categories which encompass multi-dimensional measures have facilitated clinical

decision making for persons with health conditions such as, spinal cord injury,[28] heart failure,[29, 30] and chronic obstructive pulmonary disease.[31]

The purpose of this study, therefore, was to identify categories of UL performance in daily life in adults with and without stroke using data from previously collected cohorts. Cluster analyses were performed with variables of UL performance calculated from 24 hour accelerometer recordings from three cohorts, two samples of persons with stroke and one from neurologically-intact adult controls. We hypothesized that at least three categories (low, medium, and high) of UL performance would be identified across the UL performance variables quantified by accelerometer data, spanning the possible ranges of UL performance in daily life. We also anticipated that the emerging categories would group individuals with similar ranges of the performance variables and provide a simpler method to interpret UL performance in daily life for clinicians and persons with health conditions whom they treat.

3.3 Methods

This study analyzed accelerometer data from adults from three previous cohorts, using the same accelerometry methodology.[32] Data used in these analyses were UL performance variables calculated from accelerometer data over one day, associated clinical capacity measures, and participant characteristics.

3.3.1 Participants

The three cohorts in this analysis include; 1) people with stroke (*stroke cohort 1, n=57*) from a prospective, observational, longitudinal cohort tracking UL change over time;[19] 2) people with chronic stroke (*stroke cohort 2, n=78*) who participated in a clinical trial; [33] and 3) a sample of neurologically-intact adults (*adult controls, n=76*) of similar age, race, ethnicity, and socioeconomic status of persons in the clinical trial (*stroke cohort 2*).[7] All participants provided

signed informed consent to participate in the individual studies. Inclusion and exclusion criteria for each sample are described elsewhere (*stroke cohort 1*,[19] *stroke cohort 2,*[33] and *adult controls*[7]). In general, persons in the stroke cohorts had documented UL motor impairments and diminished functional capacity as measured by the Action Research Arm Test (ARAT)[34, 35] at the time of the study enrollment. UL motor severity ranged from mild to severe, as indicated by the National Institute of Health Stroke Scale (NIHSS)[36] arm item scores of 1 to 4. Persons with stroke had to be able to follow two-step commands to enroll, and were enrolled even if they had other, mild, stroke-induced, non-motor deficits such as hemispatial neglect, aphasia, or mild cognitive impairment. Neurologically intact community-dwelling older adults had to be willing to participate and be able to follow two-step commands. Combining the three cohorts provided a broad sampling of UL performance variables. With respect to power analyses, there is no agreed upon sample required for a cluster analysis,[37, 38] however the combined cohorts yield a sample size of over 200 individuals, which was deemed sufficient to proceed with a cluster analysis.[39]

3.3.2 Data collection

UL performance was captured using data from bilateral, wrist-worn accelerometers.[7, 8, 40-42] A single time point was chosen for participants in each of the three cohorts. In *stroke cohort 1* (assessments from two – 24 weeks post stroke), data from the latest assessment time point available between weeks six and 24 were used in the analysis, since UL performance appears to stabilize between three and six weeks post stroke.[19, 43] In *stroke cohort 2* (assessments at baseline and weekly for eight or more weeks), data from the earliest available assessment time point was used in the analysis. Data points later than the baseline (when baseline was unavailable) were included because UL performance did not change as a result of this treatment.[18, 33] The *adult*

control cohort completed a single assessment in the cross-sectional study and this time point was used.[8]

3.3.3 Upper limb performance variables

Participants wore the Actigraph GT3X-BT or GT9X-Link accelerometers on both wrists for the three cohorts, with methods described previously.[32] Briefly, tri-axial acceleration data are sampled at 30 Hz for 24 or more hours continuously. Once the accelerometers were returned to the lab, data were uploaded, visually inspected, and processed using Actilife 6 (Actigraph Corp., Pensacola, FL) proprietary software. For most variables, data were band-pass filtered (0.25 and 2.5 Hz) and down sampled into 1-second epochs with ActiLife proprietary software, where each second is the sum of the 30 Hz values in that second and converted to activity counts (1 count = 0.001664g). For a few variables, (see Table 1) calculations were done directly on the 30 Hz data.[5, 24-26] Similar to previous work,[7, 12, 19-21] accelerometery data was processed using custom written software in MATLAB (Mathworks, Inc, Natick, MA) to calculate UL performance variables which qualify various aspects of UL activity in everyday life. Table 1 presents the twelve UL performance variables included in the analysis along with their description and the source of accelerometer data for calculation (1 Hz versus 30 Hz). The variables independently measure duration, magnitude, variability, symmetry and quality of movement of one or both ULs.

Table 3.1 Upper limb performance variables.

Upper limb performance variable name	Description	Data source	*Included in final solution*
Duration			
Hours of paretic/non-dominant limb activity[7, 8]	Time, in hours, that the paretic/non-dominant limb is moving.	1 Hz	✓
Hours of non-paretic/ dominant limb activity[7, 8]	Time, in hours, that the non-paretic/dominant limb is moving.	1 Hz	✓
Isolated paretic/non-dominant limb activity[8]	Time, in hours, that the paretic/non-dominant limb is moving, while the non-paretic/dominant limb is still.	1 Hz	
Isolated non-paretic/dominant limb activity[8]	Time in hours that the non-paretic/dominant limb is moving, while the paretic/non-dominant limb is still.	1 Hz	
Magnitude			
Median acceleration of paretic/non-dominant limb* [3, 8, 21]	Magnitude of accelerations of the paretic/non-dominant limb, in activity counts or gravitational units.	1 Hz	✓
Bilateral Magnitude* [3, 8, 21]	Intensity, or magnitude of accelerations of movement across both arms, in activity counts.	1 Hz	
Variability			
Acceleration variability of paretic/non-dominant limb activity* [12, 23]	Standard deviation of the magnitude of accelerations across the paretic/non-dominant limb, reflecting the variability of paretic/non-dominant limb movement, in activity counts.	1 Hz	✓
Symmetry			
Use Ratio† [7, 10, 40]	Ratio of hours of paretic/non-dominant limb movement, relative to hours of non-paretic/dominant limb movement.	1 Hz	✓
Magnitude Ratio† [8, 12, 23]	Ratio of the magnitude of paretic/non-dominant UL accelerations relative to the magnitude of the non-paretic/dominant UL accelerations. This ratio reflects the contribution of each limb to activity, expressed as a natural log.	1 Hz	
Quality of Movement			
Jerk asymmetry index [26]	Ratio of the average jerk magnitude between the paretic/non-dominant limb and the non-paretic/dominant limb. Higher jerk represents less smooth movement, and an index of 0 represents similar smoothness of movement in the paretic/non-dominant and non-paretic/dominant limbs. Values are bounded between -1 to +1.	30 Hz	
Spectral arc length of paretic/non-dominant and non-paretic/dominant limb [5, 24, 25]	A measure of movement smoothness that quantifies movement intermittencies independent of the movement's amplitude and duration. Longer spectral arc lengths are reflective of less smooth or less coordinated movement in either the paretic/non-dominant or non-paretic/dominant limb respectively	30 Hz 30 Hz	

* Variables that are quantified in activity counts, computed by the Actilife proprietary software such that 1 activity count = 0.001664g.
† For persons with stroke, ratios are paretic to non-paretic, while for neurologically-intact adults, ratios are non-dominant to dominant.

3.3.4 Analysis

All data were analyzed in R (version 4.0.1), an open source statistical computing program. A *k*-means hypothesis-free cluster analysis was used to determine categorizations of UL performance indexed by accelerometer variables in samples of persons with stroke and neurologically intact adults (adult controls). A cluster analysis is a robust statistical algorithm that groups similar objects into sub-groups called clusters,[27, 44, 45] with identified clusters becoming the categories of UL performance. The end point is a set of clusters where individuals within each cluster are more similar to each other, on average, than they are to other members of the other clusters formed.[44] A *k*-means method was chosen over other methods (e.g. hierarchical clustering or partial around the medeoid) to use an iterative approach to qualitatively explore the effect of adding more input variables and increasing the number of clusters on the dataset used in the analysis.[45, 46]

First, several steps were completed prior to the cluster analysis. The dataset of UL performance variables were standardized (using z-scores) as each variable is on a different measurement scale (e.g., hours, counts, and ratios). Then, a Hopkins statistic was calculated to determine if ursing a cluster analysis on these data was appropriate. The Hopkins statistic ranges from 0 to 1, and values >0.5 indicate clusters exist in the dataset.[47] The distributions of all twelve UL performance variables and pairwise spearman scatterplots of variables with both strong and weak relationships were examined using the GGally package.[48] Distributions and scatterplots were used to understand the relationships between UL performance variables in preparation for additional analyses and for later simplification of the cluster solutions that emerged.

Second, a principal component analysis (PCA) was conducted using the factoextra package on datasets that included 12, 9, 7 or 5 of the UL performance variables.[49] Principal

components can be thought of as the underlying dimensions of the individual UL performance variables.[45] PCAs were calculated including all 12 performance variables, then variables were systematically eliminated to exclude the variables that are complex to calculate (e.g. used 30 Hz vs 1 Hz data) and the variables with less straight forward clinical interpretation. Scree plots were examined for each of the models (5, 7, 9, and 12 UL performance variables) to determine how many principal components explained variance in the UL performance variables. Further, we examined the loadings of the input variables on each of the resulting PCs.

Third, different numbers of clusters were evaluated and the solutions were calculated using the NbClust and clusertend packages.[50, 51] A *k*-means cluster analysis expects the number of clusters to be specified prior to the analysis. Thus, we started with 3-clusters as a reasonable solution to produce clusters of low, medium and high UL performance. There are multiple statistical methods for determining the optimal number of clusters. We evaluated potential solutions using: 1) the elbow method,[52] 2) the silhouette method,[53], and 3) the gap statistic.[27] Although there was no clear single "elbow" where adding clusters led to diminishing returns in variance explained, these methods indicated that 3-,4-, and 5-cluster solutions were progressively better explanations of the data (see Results). Thus in the interests of parsimony, we focused on these three different cluster sizes in subsequent analyses.

A total of twelve cluster solutions (3-, 4- or 5-clusters with 12, 9, 7, and 5 input variables) were calculated to systematically eliminate UL performance variables to create the most parsimonious solution.[50, 51] The most complex model was calculated first (including all 12 performance variables) for a 3-, 4-, and 5-clusters. The second most complex model included 9 UL performance variables, excluding the three variables calculated from the 30Hz data that are proposed to measure quality of UL activity[5, 24-26] (see Table 3.1). These variables were removed

because they are more complex to calculate, have not been validated in clinical populations[54], and did not add relevant information to the analysis. For the seven and five input variable models, the decision was made to maintain at least one performance variable from each of the other four aspects of UL performance (duration, magnitude, variability and symmetry) to capture the dimensionality of UL performance in daily life. Variables that were simpler to calculate (1 Hz versus 30 Hz) and interpret were retained over those that required more complex calculations and/or are more difficult to interpret for ease of eventual integration into rehabilitation clinics.[4] For example, both the bilateral magnitude and the median acceleration of the paretic/non-dominant limb activity quantify the magnitude or intensity of UL activity. These two variables are highly correlated to each other and the loadings from the PCA indicate that these two variables had moderate, positive loadings on PC1, primarily. For the five variable solution, the median acceleration of the paretic/non-dominant limb was selected to remain because it had a higher contribution to PC1 than the bilateral magnitude and it is a simpler variable to calculate and interpret.

Fourth, we examined model fit metrics for each of the 12 solutions calculated to avoid overfitting as additional variables and clusters were added. The total variance explained by the models were extracted for each of the cluster-variable solutions (3-, 4- or 5-clusters with 12, 9, 7, and 5 input variables). Models that had a higher % of total explained variance were deemed to have a better model-fit.[45] Additionally, a multivariate analysis of variance (MANOVA) was calculated to re-fit the cluster classifications (3-, 4-, and 5-clusters) to the multi-dimensional space of all the UL performance variables (5, 7, 9, and 12 variables). This allowed for the Akaiki information criterion (AIC) to be extracted to compare the model-fit for each of the cluster solutions with respect to the variables included.[45] As the AIC imposes a penalty for additional

model parameters, selecting the model with the lowest AIC value helps avoid overfitting and improve generalizability.

Fifth, the means and ranges of the UL performance variables, concordance, and UL capacity (e.g. ARAT score) were computed for each cluster in the final solution. Given statistically significant omnibus effects from the multivariate analyses described above, univariate ANOVAs were computed to determine how the means of the UL performance variables differed from each other across the clusters (alpha = 0.05).[55, 56] Post-hoc comparisons (using a Tukey HSD correction) of each cluster to other clusters for five different performance variables were calculated (alpha = 0.05). Additionally, we looked at how the input cohorts (*stroke cohort 1, stroke cohort 2, adult controls*) were distributed across the cluster solutions.

Finally, coxcomb charts were created. Coxcomb charts are a two-dimensional chart type designed to plot one or more series of values over multiple quantitative variables. The 5 UL performance variables are divided into equally segmented wedges on the radial chart. The area of each individual wedge is proportional to the magnitude of the score on that dimension. Coxcomb charts were created from the standardized performance variables to provide a visual representation of the UL performance variable scores in each cluster both at the group and individual level.

3.4 Results

A sample of 211 participants were included in the analyses. Demographic and participant characteristics for the three cohorts are provided in Table 3.2. UL capacity was measured by the ARAT and indicated that both stroke cohorts had moderate deficits in UL functional capacity.

Table 3.2. Demographics and participant characteristics of the three cohorts

Variable	Stroke Cohort 1 (n=57)	Stroke Cohort 2 (n=78)	Adult Controls (n=76)
Age, years	66.5 ± 8.8	59.7± 10.9	54.3 ± 11.3
Sex, female	42% (24)	35% (27)	51% (39)
Race			
African American	40% (23)	47% (36)	59% (44)
Caucasian	58% (33)	51% (40)	41% (30)
Asian	2% (1)	1% (1)	--
Other	--	1% (1)	--
Time post-stroke, weeks	12 (7-12)	52 (21-960)	NA
Hand dominance, right	82% (47)	88% (68)	82% (62)
Concordance*	42% (24)	51% (40)	NA
Action Research Arm Test †	22.46 ± 20.76	31.3 ± 11.9	NA

Values are Mean ± SD or Percentage (n) except for Time-post stroke which are median (range).
** Concordance is where dominant limb = paretic limb.*
† Action Research Arm Test is a measure of UL functional capacity. Higher scores are better, with a maximum total score of 57 indicating normal performance.

The Hopkins statistic was H=0.78, indicating that clusters exist in the sample. Table 3.3 summarizes the range of solutions evaluated including 12, 9, 7, and 5 UL performance variables in either a 3-, 4-, or 5-cluster solutions. Across the different numbers of input variables, two principal components explained the majority of the variance, PC1 and PC2. There were similar loadings of the input variables onto these principal components, regardless of the number of variables entered. Interestingly, adding more performance variables (e.g.12 versus 5) was associated with both PC1 and PC2 explaining less of the total variance (see the first column of Table 3.3). Thus, across different numbers of input dimensions, the number of principal components was relatively stable. PC1 and PC2 appeared to be explaining similar variance in all models. We therefore proceeded with including only 5 input variables. When including 5 UL performance variables, the first principal component (PC1) explained the most variance (76.4%)

and was comprised of variables that all had moderate to strong, positive loadings, including; paretic/non-dominant hours, median acceleration of paretic/non-dominant limb activity, acceleration variability of paretic/non-dominant limb activity and the use ratio. The second principal component (PC2) explained less variance (17.6%) and was comprised of primarily the non-paretic/dominant hours, a single variable that had a strong, negative loading. See Appendix A for the loadings of all factors of PC1 and PC2 for the final chosen solution.

Across the models with differing numbers of UL performance variables, a 5-cluster solution explained the most overall total variance when compared to a 3- or 4-cluster solution as seen in the middle portion of Table 3.3 and visually in Figure 3.1A and 3.1B (including five performance variables). We then examined several metrics to determine how many clusters were appropriate for the 5-variable solution. Figure 3.1A supports that there are $\geq$ at least two clusters in this dataset and the flattened slope on Figure 3.1A indicates that the reduction of within-cluster variance is minimal and there are no further improvements after 5-clusters for this dataset. We therefore explored a 3-, 4-, and 5- cluster solutions. Figure 3.1B displays the effect of increasing numbers of clusters on the total explained variance when including 5 UL performance variables and confirms that a 5-cluster solution explains more total variance than the 3- or 4-cluster solutions. Examining the AIC values seen in the last three columns of Table 3.3, also confirmed that a 5-cluster solution produced the best model fit compared to the 3- and 4-cluster solutions across the different number of input variables (5, 7, 9, and 12 UL performance variables). Although each solution was statistically feasible, the chosen final solution was 5-clusters, from 5 UL performance variables including: 1) hours of use of paretic/non-dominant limb; 2) hours of use of non-paretic/dominant limb; 3) median acceleration of paretic/non-dominant limb; 4) acceleration variability of paretic/ non-dominant limb activity; and 5) use

ratio. Figure 3.1C presents the location of the 5-clusters across the two dimensional space. Dimension 1 (x-axis) is the first principal component and dimension 2 (y-axis) is the second principal component. The two clusters with the lowest overall UL performance are represented by clusters numbered 1 and 2 with the highest in number 5. Figure 3.2 shows a scatterplot matrix of how the five input variables relate to each other and to the 5-clusters.

Table 3.3 Selection of clusters based on variance explained and model-fit

Number of Variables	**Variance explained by each PC**		**Total variance by # of clusters**			**AIC value by # of clusters**		
	PC1	PC2	3	4	5	3	4	5
12	57.4%	13.1%	53%	58%	62%	2683.7	2442.2	2149.8
9	68.5%	16.5%	64%	69%	73%	1426.4	1114.7	879.2
7	75.6%	14.1%	70%	76%	79%	734.3	452.9	228.0
5	76.4%	17.6%	68%	75%	79%	475.6	229.1	44.7

Explained variance is presented in %. Values closer to 100% indicate greater variation explained.
AIC = Akaike's Information Criterion. A lower AIC value indicates a better model when the clusters were used as predictor variables in multivariate ANOVAs based on the different outcome variables (of 12, 9, 7, and 5 dimensions).

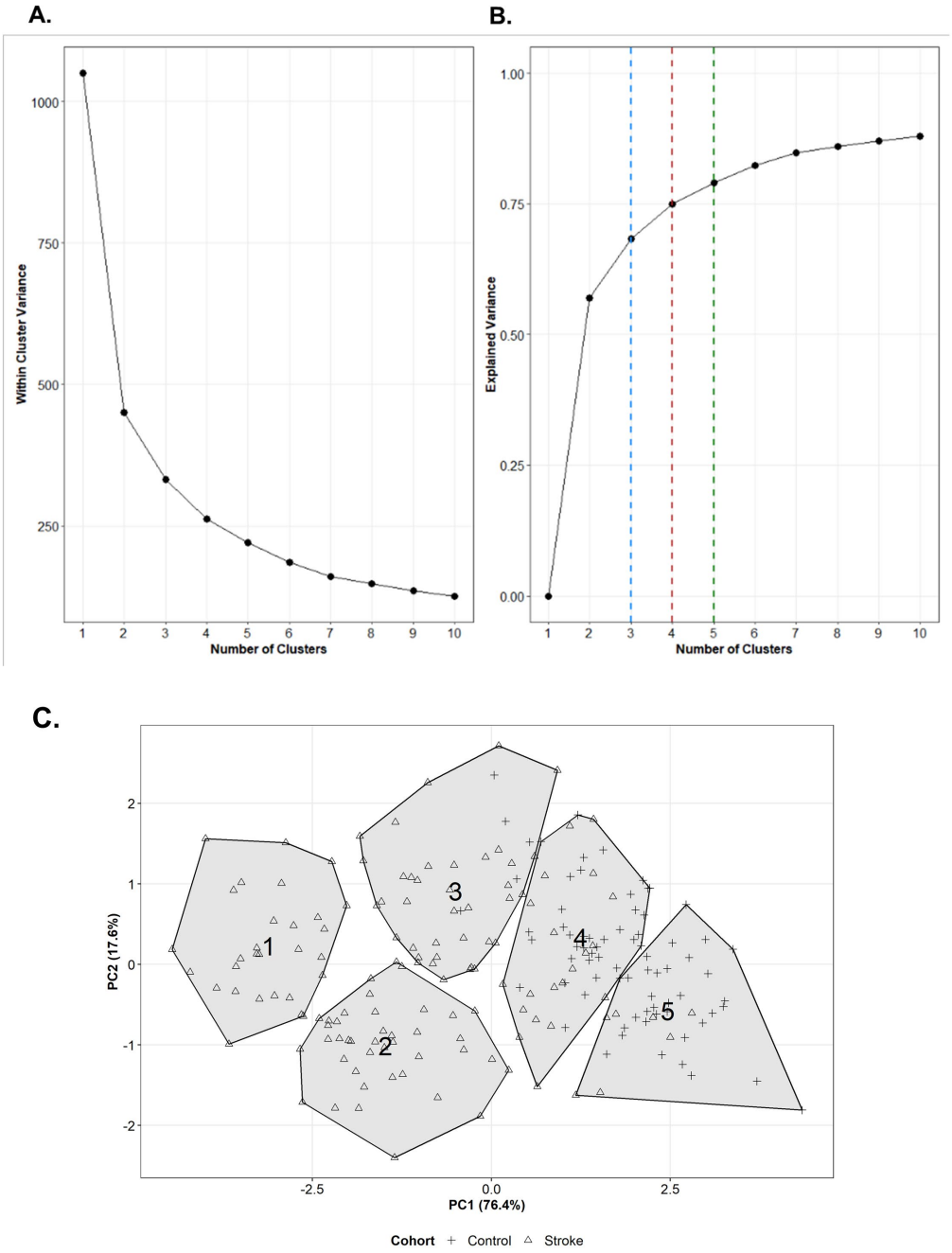


Figure 3.1. (A) Scree plot representing how the within-cluster variance changes as increasing numbers of clusters are formed with 5 UL performance variables. **(B)** Line plot representing how the total explained variance changes with increasing numbers of clusters on dataset including 5 UL performance variables. The dashed lines represent the total variance explained for a 3- (blue), 4- (red), or 5- (green) cluster solution. **(C)** Visual representation of the 5-clusters with 5 UL performance variables across dimension 1 (x-axis) and dimension 2 (y-axis). The cluster number is presented in the location of the centroid of each cluster. The shape of the point within the cluster represents if a participant was from a stroke (triangle) or control (+ sign) cohort.

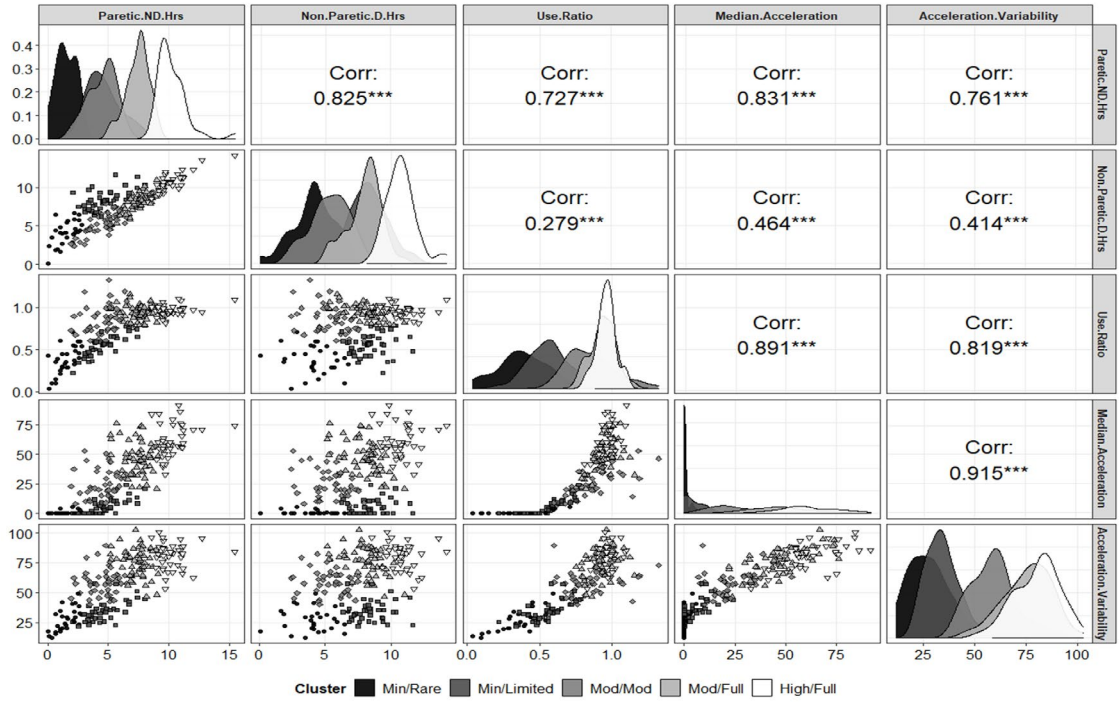


Figure 3.2: Scatterplot matrix of the 5 input variables as a function of the 5 different clusters. The diagonal shows density plots (i.e., the univariate distribution) of each input variable as a function of the different clusters. The lower left panels' show the bivariate distributions for each pair of variables with the point shapes and gray scales corresponding to the different clusters (see legend). The upper right panels show the Spearman rank order correlations for each pair of variables (on the whole, ignoring clusters). *** = $p<0.001$

The means and ranges of each UL performance variable, percentage concordant, and UL capacity for each of the 5-clusters in the final solution are presented in Table 3.4. The clusters are presented with the "lowest" overall UL performance within the first column and the "highest" overall UL performance in the last column. The 5-clusters become categories of UL performance and are named based on a synthesis of information from other publications that have described UL performance in daily life,[45, 56-58] not on the underlying PCA dimensions. The cluster names were chosen as intuitively as possible and represent the overall amount of UL activity and integration of the ULs into daily life activities (see Discussion for further interpretation). We refer to these clusters/categories as: 1) *Minimal Activity/Rare Integration; 2) Minimal Activity/Limited Integration; 3) Moderate Activity/Moderate Integration; 4) Moderate*

Activity/Full Integration; and 5) High Activity/Full Integration. The cluster with the lowest UL performance is the *Minimal Activity/Rare Integration,* this cluster has the lowest mean values on variables that quantify duration, magnitude and variability of UL activity. People in this cluster use their non-paretic UL approximately 2.5 times more than their paretic UL and have little to no magnitude or variability of their paretic UL activity in daily life. People in the *Minimal Activity/Limited Integration* cluster use both the paretic and non-paretic limb for more overall hours than the *Minimal Activity/Rare Integration* cluster, but the non-paretic limb is still active twice as much as the paretic UL. Additionally, people in this cluster have slightly higher mean values on performance variables that quantify both the magnitude and variability of the paretic limb when compared to the *Minimal Activity/Rare Integration* cluster. Both of these clusters have little integration of the ULs into activity, as suggested by a mean use ratio below 0.50 the *Minimal Activity/Rare Integration* cluster and a mean use ratio just above 0.50 in the *Minimal Activity/Limited Integration* cluster. The cluster with overall, moderate UL performance is the *Moderate Activity/Moderate Integration* cluster. In this cluster, people have more symmetrical UL use compared to the two lower clusters, which is reflected in the in the use ratio (0.85) and the mean values of both duration variables (4.5 hrs vs. 5.3 hrs). People in this cluster have moderate values on variables that quantify both the magnitude and variability of paretic/non-dominant limb activity. The two clusters with the highest overall UL performance are the *Moderate Activity/Full Integration and the High Activity/Full Integration* clusters. These clusters have progressively higher mean values of variables quantifying duration, magnitude and variability of UL activity with those in the *High Activity/Full Integration* cluster having the highest mean values compared to the other clusters. Both of these clusters however, have similar mean values of the use ratio, which is approaching 1.0 indicating that people in these two

clusters have relatively equal contributions of both ULs. Interestingly, if only the use ratio was used to examine these two clusters it could be assumed that they are relatively equal, but the other variables show they are not. The two clusters with the highest overall UL performance also had the highest % of people with concordant stroke. It is also noteworthy that participants within each of the 5 clusters have wide, overlapping ranges of UL capacity, as indicated by the mean and ranges of ARAT scores in the bottom row of Table 3.4, consistent with the premise that UL capacity and UL performance are different, but related constructs. Figure 3.3 presents how the three included cohorts separated into the 5-clusters. The two clusters with the lowest overall UL performance (*Minimal Activity/Rare Integration* and *Minimal Activity/Limited Integration)* are comprised of only persons from the stroke cohorts. The cluster with moderate UL performance (*Moderate Activity/Moderate Integration)* contains mostly people with stroke but there are also a few neurologically intact adult controls in this cluster too. The two clusters with the highest overall UL performance (*Moderate Activity/Full Integration* and *High Activity/Full Integration)* contains the neurologically intact adult controls and some persons with stroke. Finally, there was a statistically significant omnibus effect of cluster in each of the univariate ANOVAs for the five UL performance variables (p values for each variable <0.001). Note that not all clusters were statistically different from all other clusters in each variable, based on post-hoc t-tests. However, this speaks to the multivariate nature of the cluster analyses; across all dimensions, these clusters group similar observations together, but along any single dimension there will likely be overlap in the neighboring clusters.

Table 3.4. Means (ranges) of UL performance and capacity variables by cluster.

Variable name	Minimal Activity/ Rare Integration (N=29)	Minimal Activity/ Limited Integration (N=41)	Moderate Activity/ Moderate Integration (N=43)	Moderate Activity/ Full Integration (N=57)	High Activity/ Full Integration (N=41)
Duration					
Paretic/ND Hrs	1.5 (0.0-2.8)	4.6 (2.1-8.0)	4.5 (1.9-6.5)	7.4 (5.2-9.1)	10.2 (8.6-15.5)
Non-paretic/D Hrs	4.1 (0.1-6.7)	8.4 (6.2-11.6)	5.3 (2.4-8.0)	8.0 (5.1-11.0)	10.7 (8.5-14.2)
Magnitude					
Median acceleration paretic/ND (counts)*	0 (0-6)	5 (5-24)	25 (7-53)	47 (21-76)	61 (33-92)
Variability					
Acceleration variability of paretic/ND (counts)*	27.3 (11.9-49.4)	34.8 (21.6-57.3)	58.9 (40.0-89.3)	75.9 (46.5-102.6)	80.3 (53.0-100.8)
Symmetry					
Use Ratio	0.38 (0.04-0.70)	0.55 (0.22-0.78)	0.85 (0.60-1.32)	0.94 (0.75-1.15)	0.96 (0.81-1.10)
Additional data about the clusters					
Concordance †	38% (11)	39% (16)	50% (19/38)	70% (14/20)	57% (4/7)
Action Research Arm Test ‡	18.5 (0-43)	27.8 (6-57)	45.3 (22-57)	48.4 (33-57)	44.1 (24-55)

*Data are reported in activity counts computed by the Actilife proprietary software, such that 1 activity count = 0.001664 gravitational units (g).

† Dominant limb= paretic limb, computed for persons in stroke. Percentage is expressed relative to only persons with stroke, not controls, in the upper three categories.

‡ Action Research Arm Test is a measure of UL functional capacity. Higher scores are better, with a maximum total score of 57 indicating normal capacity.

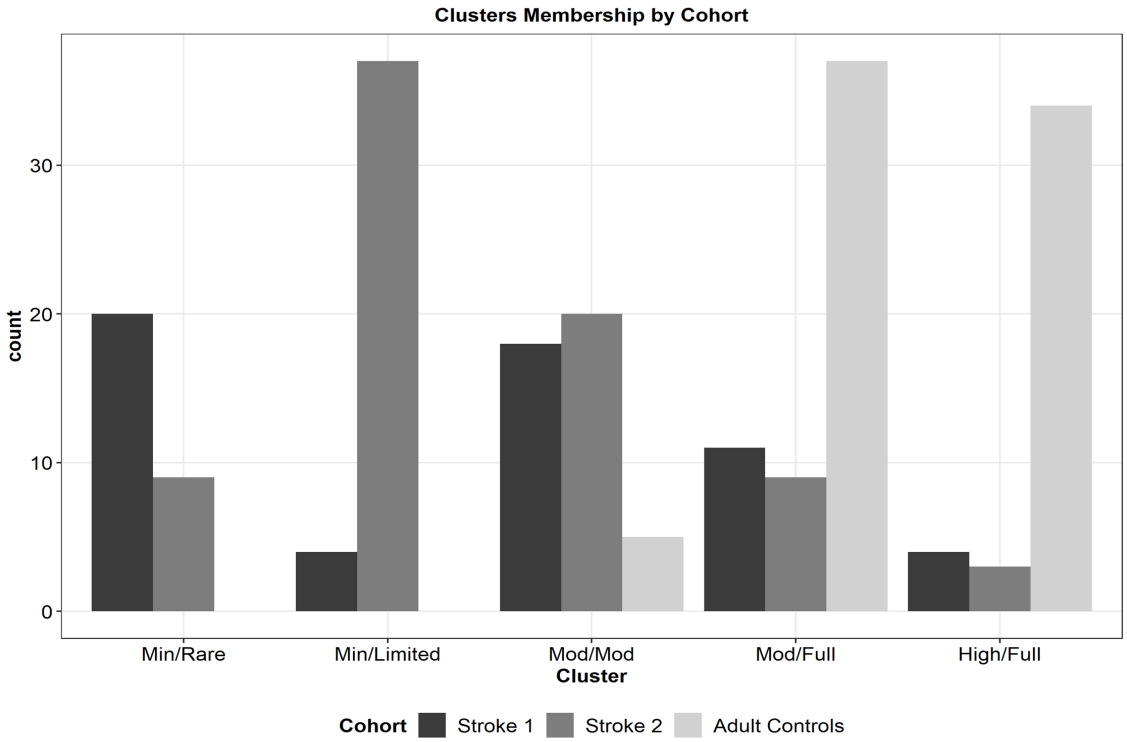
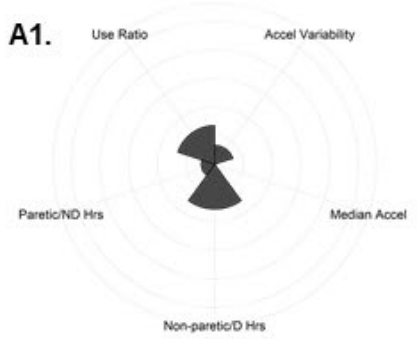


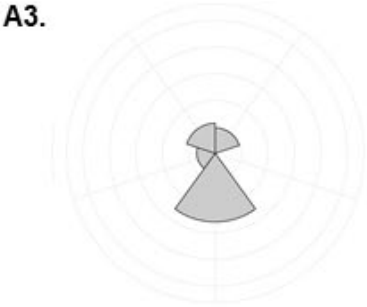
Figure 3.3: Bar plot of the counts of participants from each of the 3 cohorts that separated into the 5-clusters. The two clusters with the lowest overall UL performance are comprised of persons from the stroke cohorts only. The cluster with moderate UL performance contains primarily persons with stroke and a few neurologically intact adult controls. The two clusters with the highest overall UL performance include primarily neurologically intact adult controls, as well as persons with stroke.

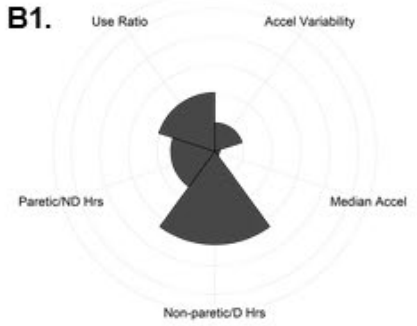
Figure 3.4 presents the group and individual coxcomb charts for each of the 5-clusters. The rows (A, B, C, D, E) are presented in order of increasing overall UL performance, with group data in the first column in dark gray and then individual examples of people in that cluster in columns two and three (A-E, numbered 2 and 3 respectively) in light gray. Each of the five standardized UL performance variables are represented by wedges within the plot, and the area of the wedge reflects the standardized value on that single variable. Figure 3.4A and 3.4B present the two clusters with the lowest overall UL performance (*Minimal Activity/Rare Integration and Minimal Activity/Limited Integration)*, the wedges in these two clusters are small with the exception of the non-paretic/D hours of use, indicating that people in these two clusters use their non-paretic UL out of proportion to their paretic UL. As you move down each row from *Minimal*

Activity/Rare Integration (3.4A) to *High Activity/Full Integration* (3.4E) one can see that the wedges get larger and begin to fill more area of the radial plot, however some variables are still out of proportion to the others as seen in 3.4C-3.4D. By the final group plot in Figure 3.4E1, the wedges for each variable span the largest area and almost form a perfect circle, compared to the clusters with lower UL performance (3.4A and 3.4B), indicating people in this cluster have the highest values across all five performance variables.

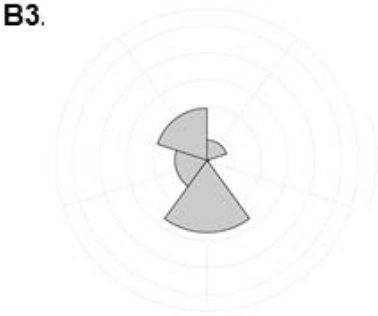

A1.
Use Ratio
Accel Variability
Paretic/ND Hrs
Median Accel
Non-paretic/D Hrs

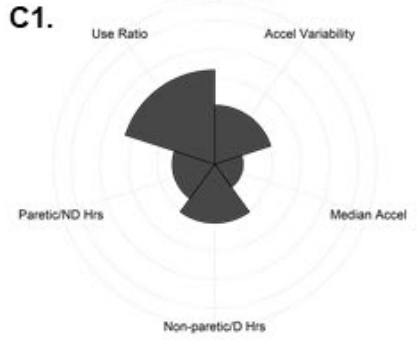

A2.

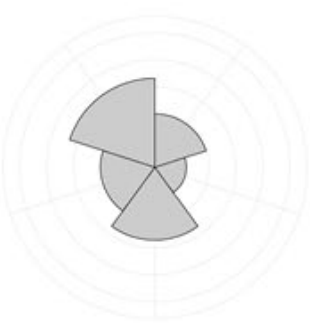

A3.

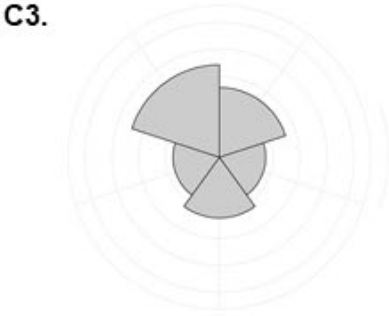

B1.
Use Ratio
Accel Variability
Paretic/ND Hrs
Median Accel
Non-paretic/D Hrs

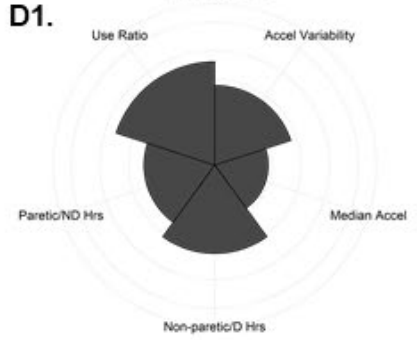

B2.

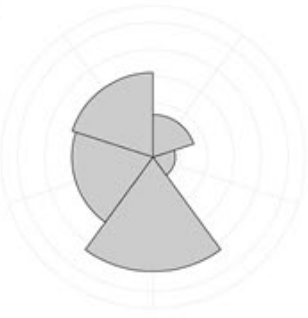

B3.

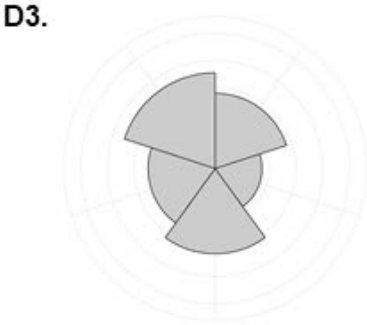

C1.
Use Ratio
Accel Variability
Paretic/ND Hrs
Median Accel
Non-paretic/D Hrs

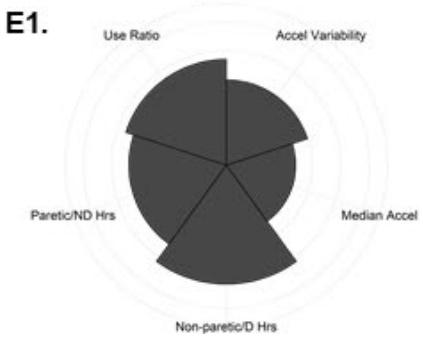

C2.

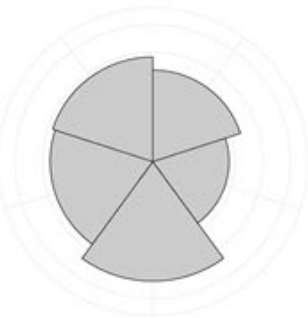

C3.


D1.
Use Ratio
Accel Variability
Paretic/ND Hrs
Median Accel
Non-paretic/D Hrs


D2.


D3.


E1.
Use Ratio
Accel Variability
Paretic/ND Hrs
Median Accel
Non-paretic/D Hrs


E2.

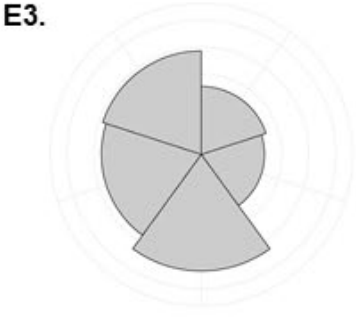

E3.

Figure 3.4: Coxcomb charts of the five clusters, illustrating the contribution of the UL performance variables on a standardized scale. The first column plots group data, while the 2nd and 3rd columns plot individual participant examples.
(A) *Minimal Activity/Rare Integration* cluster; (**A1)** group chart of people within this cluster; **(A2)** is a person from stroke cohort 1, ARAT=4; and **(A3)** is a person from stroke cohort 2, ARAT=10.
(B) *Minimal Activity/Limited Integration* cluster; **(B1)** group chart of people within this cluster; **(B2)** a person from stroke cohort 2, ARAT=10; and **(B3)** a person from stroke cohort 1, ARAT=6.
(C) *Moderate Activity/Moderate Integration* cluster. **(C1)** group chart of people within this cluster; **(C2)** a person from stroke cohort 1, ARAT=36; and **(C3)** a person from the adult controls.
(D) *Moderate Activity/Full Integration* cluster; **(D1)** group plot for this cluster; **(D2)** a person from stroke cohort 2, ARAT=42; and **(D3)** a person from the adult controls.
(E) *High Activity/Full Integration cluster;* **(E1)** group chart of people within this cluster; **(E2)** a person from stroke cohort 1, ARAT=55; and **(E3)** a person from the adult controls.

3.5 Discussion

In a large sample of persons with and without neurological UL deficits, we used a *k*-means cluster analysis with multiple UL performance variables, captured via accelerometry, to derive a 5-cluster categorization that included 5 UL performance variables. Two principal components explain most of the variance in the input variables and 5-clusters explained the most total variance and had the best model fit. In this 5-cluster solution, two groups with what might be considered "normal" UL performance (*Moderate Activity/Full Integration and High Activity/Full Integration*) emerged, as indicated by the presence of many neurologically intact adult controls in those categories. One category in the middle had moderate UL performance (*Moderate Activity/Moderate Integration*), while two categories had low, overall UL performance (*Minimal Activity/Rare Integration* and *Minimal Activity/Limited Integration)*. The names of each of the 5 categories were chosen for their overall UL activity and integration, with the future goal that these categories could be evaluated for their application to other clinical populations, not just persons with stroke.

The 5-category solution from five UL performance variables, derived from this statistical analysis, leads to a clinically-logical interpretation of UL performance in daily life. In this analysis we purposefully included three cohorts of persons with and without stroke in order to

capture a wide range of the variables, extracted from accelerometer data that quantify different aspects of UL performance in daily life. In Figure 3.3, the two categories with the highest overall UL performance (*Moderate Activity/Full Integration* and *High Activity/Full Integration*) contain most of the neurologically intact adult controls indicating that people without neurological impairments display a wide range of UL activity that can be considered unimpaired or normal. This is important because these people have integrated their ULs, as indicated by the use ratio variable, but people in these categories have different levels of overall UL activity, ranging from moderate to high UL activity. This is not unusual when we consider the wide range of activities and behaviors of people.[59-61] For example, when walking performance is quantified by pedometers, neurologically-intact adults walk symmetrically but present with a wide range of variability in the total number of steps-per-day that can all be considered "normal" walking performance.[59, 62-66] Based on the current results, it appears that people without neurological UL impairments similarly display a wide range of UL activity that can also be considered unimpaired or normal. For example, two neurologically intact older adults may have very different activities of daily living and leisure activities (e.g. swimming versus knitting) but would both be considered to have "normal" UL performance. In other efforts to categorize UL activity, some groups have found four categories,[55, 56, 58] and others have found six.[57] These analyses however tended to examine only the separation of UL activity of persons with stroke. In this analysis, the goal was <u>not</u> to form categories to differentiate between those who had a stroke and those who did not. Instead, the goal was to categorize people based on their overall UL use in daily life. In the 5-category solution here, we see that the two categories with the lowest UL activity and integration are comprised of only persons with stroke, but there are also people with stroke in the three categories with the highest overall UL performance too. This is a positive

finding, showing that some people with stroke use their ULs similarly to neurologically intact adults. Persons with stroke who ended up in the two categories with the highest overall UL performance have likely experienced either full recovery of their ULs following their stroke, or have figured out how to use the wide range of capacity that they have to integrate their paretic limb and be active in daily life.[19] An example of this is shown in Figure 4E2 which is an individual from *stroke cohort 1* who ended up in the *High Activity/Full Integration* category.

Categories of UL performance have tremendous research and clinical potential. Within other biomedical science fields, formation of categories which encompass multi-dimensional measures have facilitated clinical decision making for persons with health conditions (See Introduction). Specific to rehabilitation, categories of ambulation (based on the capacity measure of walking speed) have been validated, shown to be sensitive to change,[67-69] used to set goals in clinical practice, and have been used as a primary outcome in a Phase III clinical trial.[70] In that trial, the primary outcome was the percentage of people who changed (leaped) to a higher ambulation category after the intervention. The identified categories of UL performance that emerged in this analysis could be useful for future trials of persons with UL impairments following subsequent, future validation studies. Categories that emerged in this analysis have stratified participants into groupings with similar overall UL performance, representing a profile of arm activity in daily life.[38, 56, 71] Individuals within each category have similar ranges of each performance variable included (e.g., duration, magnitude, variability and symmetry) that formed the 5-clusters. Interestingly, in this analysis people with stroke within each of the five clusters display a wide range of UL capacity across the clusters. Additionally, more people in the two clusters with highest overall UL performance have concordant stroke compared to the three clusters with lower UL performance. These findings are consistent with prior work indicating

that people with concordant stroke (dominant limb=paretic limb) tend to have differences in the patterns of UL use [57, 72] and experience better recovery.[19] One can envision that these categories could be used in future trials to analyze smaller subsets of individuals based on their UL category and to better understand how UL performance variables quantify change during rehabilitation therapy.

From a clinical perspective, the categories that emerged offer the future opportunity to transition measurement of UL performance in daily life for persons receiving UL rehabilitation away from the current confines of rehabilitation research labs, and into standard of care.[4, 73] The results of this analysis are a first step in simplifying measurement of UL performance in daily life by exploring the underlying structure in the set of observed variables.[74] A future option could be to offer a user-friendly, software package to rehabilitation clinicians that would calculate the 5 UL performance variables included in this analysis from data extracted from bilateral wrist-worn accelerometers. Based on a person's values across the variables, a category of UL performance could be determined and used to communicate current UL performance and used to set goals for future UL performance. Based on the aspects of movement (duration, magnitude, variability, symmetry) selected to form the categories, it is possible that these categories could be highly relevant for many clinical conditions affecting UL performance in daily life, not just those with stroke. Just as with mobility, there are plenty of biological and psychological reasons why people could have limited UL performance in daily life.[59, 75, 76] Thus, the names selected for each category might be applicable to other clinical populations that have similar or different UL impairments and capacity limitations, beyond the typical asymmetrical deficit which is a major aspect of stroke UL movement.[3]

3.5.1 Limitation

There are a few limitations to consider when interpreting the results of this study. First, the three cohorts used in this analysis generated a sample of over 200 people with stroke and neurologically intact adult controls. While our sample size was large and had wide distributions of each UL performance variable, validation on another large, independent sample is needed for generalization and implementation into clinical practice. Future studies, including people with other clinical diagnoses beyond stroke are needed in order to understand how the number of UL performance variables and subsequently the number of clusters generalize to other populations. Second, the *Moderate Activity/Moderate Integration* category is less straightforward to understand than the other four categories that emerged in this analysis. This category is comprised primarily of persons from both stroke cohorts, however there are a few neurologically intact adult controls who ended up in this category as well. Unfortunately, we do not have enough information about other cognitive, socioeconomic, physical, emotional or behavioral reasons why these few people without neurological UL impairments ended up in this category with reduced overall UL activity and integration. This category specifically will need to be externally validated in a larger sample.

3.6 Conclusions

The present study identified five categories of UL performance in a combined cohort of neurologically impaired and unimpaired adults. These categories can be formed with a minimum of 5 UL performance variables, extracted from bilateral wrist-worn accelerometers that span the possible ranges of UL activity and integration. Further validation of both the number of UL performance variables and categories will be required on a larger, more heterogeneous sample. Following validation, these categories may be used as outcomes in UL stroke research and implemented into rehabilitation therapies.

3.7 References

1. Uswatte G, Miltner WHR, Foo B, Varma M, Moran S, Taub E. Objective measurement of functional upper-extremity movement using accelerometer recordings transformed with a threshold filter. Stroke 2000;31(3):662-7.

2. Uswatte G, Taub E. Participant-centered and objective measures of real-world arm function in persons with stroke. Circulation 2004;109(20):E243-E.

3. Smith BA, Lang CE. Sensor Measures of Symmetry Quantify Upper Limb Movement in the Natural Environment Across the Lifespan. Arch Phys Med Rehabil 2019;100(6):1176-83.

4. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field. Sensors 2020;20(20).

5. David A, Subash T, Varadhan SKM, Melendez-Calderon A, Balasubramanian S. A Framework for Sensor-Based Assessment of Upper-Limb Functioning in Hemiparesis. Front Hum Neurosci 2021;15:667509.

6. Che Bakri NA, Kwasnicki RM, Dhillon K, Khan N, Ghandour O, Cairns A et al. Objective Assessment of Postoperative Morbidity After Breast Cancer Treatments with Wearable Activity Monitors: The "BRACELET" Study. Ann Surg Oncol 2021;28(10):5597-609.

7. Bailey RR, Lang CE. Upper-limb activity in adults: referent values using accelerometry. J Rehabil Res Dev 2013;50(9):1213-22.

8. Bailey RR, Klaesner JW, Lang CE. Quantifying Real-World Upper-Limb Activity in Nondisabled Adults and Adults With Chronic Stroke. Neurorehabil Neural Repair 2015;29(10):969-78.

9. Uswatte G, Taub E, Morris D, Light K, Thompson P. The Motor Activity Log-28 assessing daily use of the hemiparetic arm after stroke. Neurology 2006;67(7):1189-94.

10. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of accelerometry for monitoring real-world arm activity in patients with subacute stroke: evidence from the extremity constraint-induced therapy evaluation trial. Arch Phys Med Rehabil 2006;87(10):1340-5.

11. Gebruers N, Truijen S, Engelborghs S, Nagels G, Brouns R, De Deyn PP. Actigraphic measurement of motor deficits in acute ischemic stroke. Cerebrovascular Diseases 2008;26(5):533-40.

12. Urbin MA, Bailey RR, Lang CE. Validity of body-worn sensor acceleration metrics to index upper extremity function in hemiparetic stroke. Journal of neurologic physical therapy : JNPT 2015;39(2):111-8.

13. Barth J, Geed S, Mitchell A, Lum PS, Edwards DF, Dromerick AW. Characterizing upper extremity motor behavior in the first week after stroke. PLoS One 2020;15(8):e0221668.

14. WHO. International Classification of Functioning, Disability and Health (ICF). World Health Organization 2018(06/01/2021).

15. Waddell KJ, Birkenmeier RL, Bland MD, Lang CE. An exploratory analysis of the self-reported goals of individuals with chronic upper-extremity paresis following stroke. Disabil Rehabil 2016;38(9):853-7.

16. Andrews K, Steward J. Sroke recovery: He can but does he? Rheumatology 1979;18(1):43-8.

17. Hidaka Y, Han CE, Wolf SL, Winstein CJ, Schweighofer N. Use It and Improve It or Lose It: Interactions between Arm Function and Use in Humans Post-stroke. Plos Computational Biology 2012;8(2).

18. Waddell KJ, Strube, M. J., Bailey, R. R., Klaesner, J. W., Birkenmeier, R. L., Dromerick, A. W., Lang, C. E. Does task-specific training improve upper limb performance in daily life poststroke? Neurorehabil Neural Repair 2017;31(3):290-00.

19. Lang CE, Waddell KJ, Barth J, Holleran CL, Strube MJ, Bland MD. Upper limb performance in daily life plateaus around three to six weeks post stroke. Neurorehabilitation and Neural Repair 2021;35(10):903-14.

20. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang, C. E. Upper Limb Performance in Daily Life Improves Over the First 12 Weeks Poststroke. Neurorehabilitation and Neural Repair 2019;33(10):836-47.

21. Bailey RR, Klaesner JW, Lang CE. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. PLoS One 2014;9(7):e103135.

22. Barth J, Klaesner JW, Lang CE. Relationships between accelerometry and general compensatory movements of the upper limb after stroke. Journal of Neuroengineering and Rehabilitation 2020;17(1).

23. Urbin MA, Waddell KJ, Lang CE. Acceleration metrics are responsive to change in upper extremity function of stroke survivors. Arch Phys Med Rehabil 2015;96(5):854-61.

24. Balasubramanian S, Melendez-Calderon A, Burdet E. A robust and sensitive metric for quantifying movement smoothness. IEEE Trans Biomed Eng 2012;59(8):2126-36.

25. Balasubramanian S, Melendez-Calderon A, Roby-Brami A, Burdet E. On the analysis of movement smoothness. Journal of Neuroengineering 2015;12(1):112.

26. de Lucena DS, Stoller, O., Rowe, J. B., Chan, V., & Reinkensmeyer, D. J. Wearable sensing for rehabilitation after stroke: Bimanual jerk asymmetry encodes unique

information about the variability of upper extremity recovery. . Proceedings of the In 2017 International Conference on Rehabilitation Robotics (ICORR); 2017.

27. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Roy Stat Soc B 2001;63:411-23.

28. Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A et al. International standards for neurological classification of spinal cord injury (revised 2011). J Spinal Cord Med 2011;34(6):535-46.

29. Ovbiagele B, Goldstein LB, Higashida RT, Howard VJ, Johnston SC, Khavjou OA et al. Forecasting the future of stroke in the United States: a policy statement from the American Heart Association and American Stroke Association. Stroke 2013;44(8):2361-75.

30. Schlotter F, Orban M, Rommel KP, Besler C, von Roeder M, Braun D et al. Aetiology-based clinical scenarios predict outcomes of transcatheter edge-to-edge tricuspid valve repair of functional tricuspid regurgitation. Eur J Heart Fail 2019;21(9):1117-25.

31. Pauwels R. Global initiative for chronic obstructive lung diseases (GOLD): time to act. Eur Respir J 2001;18(6):901-2.

32. Lang CE, Waddell KJ, Klaesner JW, Bland MD. A Method for Quantifying Upper Limb Performance in Daily Life Using Accelerometers. Jove-J Vis Exp 2017(122).

33. Lang CE, Strube MJ, Bland MD, Waddell KJ, Cherry-Allen KM, Nudo RJ et al. Dose response of task-specific upper limb training in people at least 6 months poststroke: A phase II, single-blind, randomized, controlled trial. Ann Neurol 2016;80(3):342-54.

34. Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001;82(1):14-9.

35. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. Neurorehabilitation and Neural Repair 2008;22(1):78-90.

36. Kwah LK, Diong J. National Institutes of Health Stroke Scale (NIHSS). J Physiother 2014;60(1):61-.

37. Dolnicar S. A review of unquestioned standards in using cluster analysis for data-driven market segmentation. 2002.

38. Varley BJ, Shiner CT, Johnson L, McNulty PA, Thompson-Butel AG. Revisiting Poststroke Upper Limb Stratification: Resilience in a Larger Cohort. Neurorehabil Neural Repair 2021;35(3):280-9.

39. Skalski JR, Richins SM, Townsend RL. A statistical test and sample size recommendations for comparing community composition following PCA. PLoS One 2018;13(10):e0206033.

40. Uswatte G, Taub E, Morris D, Vignolo M, McCulloch K. Reliability and validity of the upper-extremity Motor Activity Log-14 for measuring real-world arm use. Stroke 2005;36(11):2493-6.

41. Rand D, Eng, J. J., Tang, P. F., Jeng, J. S., Hung, C. How active are people with stroke? Use of accelerometers to assess physical activity. Stroke 2009;40:163-8.

42. Bailey RR, Birkenmeier RL, Lang CE. Real-world affected upper limb activity in chronic stroke: an examination of potential modifying factors. Topics in stroke rehabilitation 2015;22(1):26-33.

43. Waddell KJ, Tabak RG, Strube MJ, Haire-Joshu D, Lang CE. Belief, confidence, and motivation to use the paretic upper limb in daily life over the first 24 weeks after stroke. Journal of Neurologic Physical Therapy 2019;43(4):197-203.

44. Beauchaine TP, Beauchaine RJ. A comparison of maximum covariance and k-means cluster analysis in classifying cases into known taxon groups. Psychol Methods 2002;7(2):245-61.

45. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.

46. Davisdson I. Understanding K-means non-hierarchical clustering. SUNY Albany Technical Report 2002;2:2-14.

47. Banerjee A, Dave RN. Validating clusters using the Hopkins statistic. Proceedings of the 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat No04CH37542); 2004.

48. GGally: Extension to 'ggplot2'. R package version 2.0.0.; 2020.

49. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7.; 2020.

50. NbClust: an R package for determining the relevant number of clusters in a data set. . 2014.

51. clustertend: Check the Clustering Tendency. 2015.

52. Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications 2014;105(9).

53. Rousseeuw PJ. Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J Comput Appl Math 1987;20:53-65.

54. Barth J, Klaesner JW, Lang CE. Relationships between accelerometry and general compensatory movements of the upper limb after stroke. J Neuroeng Rehabil 2020;17(1):1-10.

55. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. Ann Clin Transl Neurol 2017;4(11):811-20.

56. Woytowicz EJ, Rietschel, J. C., Goodman, R. N., Conroy, S. S., Sorkin, J. D., Whitall, J., Waller, S. M. . Determining levels of upper extremity movement impairment by applying a cluster analysis to the Fugl-Meyer assessment of the upper extremity in chronic stroke. Archives of physical medicine and rehabilitation 2017;98(3):456-62.

57. Simpson LA, Eng JJ, Backman CL, Miller WC. Rating of Everyday Arm-Use in the Community and Home (REACH) scale for capturing affected arm-use after stroke: development, reliability, and validity. PLoS One 2013;8(12):e83405.

58. Demartino AM, Rodrigues LC, Gomes RP, Michaelsen SM. Manual dexterity is associated with use of the paretic upper extremity in community dwelling individuals with stroke. Journal of Neurologic Physical Therapy 2021.

59. Bassett DR, Jr., Wyatt HR, Thompson H, Peters JC, Hill JO. Pedometer-measured physical activity and health behaviors in U.S. adults. Med Sci Sports Exerc 2010;42(10):1819-25.

60. Hayward KS, Eng JJ, Boyd LA, Lakhani B, Bernhardt J, Lang CE. Exploring the Role of Accelerometers in the Measurement of Real World Upper-Limb Use After Stroke. Brain Impair 2016;17(1):16-33.

61. Waddell KJ, Patel MS, Clark K, Harrington, T. O., Greysen SR. Leveraging insights from behavioral economics to improve mobility for adults with stroke: Design and rationale of the BE Mobile clinical trial. Contemporary Clinical Trials 2021;107.

62. Dwyer T, Pezic A, Sun C, Cochrane J, Venn A, Srikanth V et al. Objectively Measured Daily Steps and Subsequent Long Term All-Cause Mortality: The Tasped Prospective Cohort Study. PLoS One 2015;10(11):e0141274.

63. Fukushima N, Inoue S, Hikihara Y, Kikuchi H, Sato H, Tudor-Locke C et al. Pedometer-determined physical activity among youth in the Tokyo Metropolitan area: a cross-sectional study. Bmc Public Health 2016;16.

64. Tudor-Locke C, Schuna JM, Jr., Han HO, Aguiar EJ, Green MA, Busa MA et al. Step-Based Physical Activity Metrics and Cardiometabolic Risk: NHANES 2005-2006. Med Sci Sports Exerc 2017;49(2):283-91.

65. Kraus WE, Janz KF, Powell KE, Campbell WW, Jakicic JM, Troiano RP et al. Daily Step Counts for Measuring Physical Activity Exposure and Its Relation to Health. Med Sci Sports Exerc 2019;51(6):1206-12.

66. Holleran CL, Bland MD, Reisman DS, Ellis TD, Earhart GM, Lang CE. Day-to-Day Variability of Walking Performance Measures in Individuals Poststroke and Individuals With Parkinson Disease. Journal of neurologic physical therapy : JNPT 2020;44(4):241-7.

67. Fritz S, Lusardi M. White Paper: "Walking Speed: the Sixth Vital Sign". Journal of Geriatric Physical Therapy 2009;32(2):2-5.

68. Nadeau SE, Wu SS, Dobkin BH, Azen SP, Rose DK, Tilson JK et al. Effects of task-specific and impairment-based training compared with usual care on functional walking ability after inpatient stroke rehabilitation: LEAPS Trial. Neurorehabil Neural Repair 2013;27(4):370-80.

69. Nadeau SE, Dobkin B, Wu SS, Pei Q, Duncan PW, Team LI. The Effects of Stroke Type, Locus, and Extent on Long-Term Outcome of Gait Rehabilitation: The LEAPS Experience. Neurorehabil Neural Repair 2016;30(7):615-25.

70. Duncan PW, Sullivan KJ, Behrman AL, Azen SP, Wu SS, Nadeau SE et al. Body-weight-supported treadmill rehabilitation after stroke. N Engl J Med 2011;364(21):2026-36.

71. Rosso C, Lamy JC. Prediction of motor recovery after stroke: being pragmatic or innovative? Current Opinion in Neurology 2020;33(4):482-7.

72. Rinehart JK, Singleton RD, Adair JC, Sadek JR, Haaland KY. Arm use after left or right hemiparesis is influenced by hand preference. Stroke 2009;40(2):545-50.

73. Lang CE, Cade WT. A step toward the future of seamless measurement with wearable sensors in pediatric populations with neuromuscular diseases. Muscle Nerve 2020;61(3):265-7.

74. Suhr DD. Exploratory or confirmatory factor analysis? Proceedings of the SUGI 31; 2006.

75. Peel C, Sawyer Baker P, Roth DL, Brown CJ, Brodner EV, Allman RM. Assessing mobility in older adults: the UAB Study of Aging Life-Space Assessment. Phys Ther 2005;85(10):1008-119.

76. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med 2007;357(4):370-9.

Chapter 4: Predicting later categories of upper limb activity from earlier clinical assessments following stroke: An exploratory analysis

This chapter is under has been submitted and is in review at the *Journal of NeuroEngineering and Rehabilitation*

4.1 Abstract

Background: Accelerometers allow for direct measurement of upper limb (UL) activity. Recently, multi-dimensional categories of UL performance have been formed to provide a more complete measure of UL use in daily life. Prediction of motor outcomes after stroke have tremendous clinical utility and a next step is to determine what factors might predict someone's subsequent UL performance category.

Purpose: To explore how different machine learning techniques can be used to understand how clinical measures and participant demographics captured early after stroke are associated with the subsequent UL performance categories.

Methods: This study analyzed data from two time points from a previous cohort (n=54). Data used was participant characteristics and clinical measures from early after stroke and a previously established category of UL performance at a later post stroke time point. Different machine learning techniques (a single decision tree, bagged trees, and random forests) were used to build predictive models with different input variables. Model performance was quantified with the explanatory power (in-sample accuracy), predictive power (out-of-bag estimate of error), and variable importance.

Results: A total of seven models were built, including one single decision tree, three bagged trees, and three random forests. Measures of UL impairment and capacity were the most important predictors of the subsequent UL performance category, regardless of the machine learning algorithm used. Other non-motor clinical measures emerged as key predictors, while participant demographics predictors were less important across the models. Models built with the bagging algorithms outperformed the single decision tree.

Conclusions: UL clinical measures were the most important predictors of the subsequent UL performance category in this exploratory analysis regardless of the machine learning algorithm used. Interestingly, cognitive and affective measures emerged as important predictors when the number of input variables was expanded. These results reinforce that UL performance, *in vivo*, is not a simple product of body functions nor the capacity for movement, instead being a complex phenomenon dependent on many physiological and psychological factors. Utilizing machine learning, this exploratory analysis is a productive step towards prediction of UL performance.

Trial Registration: NA

4.2 Introduction

Wearable movement sensors allow for direct measurement of upper limb (UL) activity in daily life, i.e. performance.[1] Performance is operationally defined in the World Health Organization's (WHO) International Classification of Function (ICF) model as activity in the unstructured, free-living environment, and is distinguished from capacity, operationally defined as the capability for activity in a structured or standardized environment. [2,3] The most common wearable sensors used are accelerometers, from which numerous clinically relevant variables about UL activity can be computed to provide insight into how people with or without neurological impairment use their ULs in daily life.[4-7] Data extracted from bilateral, wrist worn wearable sensors can be used to quantify UL performance variables measuring the duration,[8,9] symmetry,[6,10-12] magnitude,[5,7,13] and variability of one or both limbs.[4,5,7,13] Each UL performance variable conveys slightly different information about the collective nature of UL use; multiple variables may provide a fuller understanding of the scope of UL performance in daily life.[14] As a solution to the multi-variable problem, we recently categorized UL performance in adult cohorts

with and without stroke.[15] The most parsimonious solution was five categories of UL performance formed from five UL performance variables. The UL performance categories are multi-dimensional, with each category providing information about UL activity with respect to the different movement characteristics in adults with and without neurological UL deficits. Thus, the five categories of UL performance may provide a more complete measure of UL use in daily life.[14,15]

Early prediction of motor outcomes after stroke has tremendous clinical utility.[16,17] Our next step, therefore, was to determine what factors might predict someone's subsequent UL performance category. Predictive knowledge of subsequent outcomes can inform the delivery and specification of individualized rehabilitation services.[18,19] This effort to predict an individual's subsequent UL performance category is informed by the development of the PREP 2 algorithm, which has demonstrated that prediction of an UL capacity (i.e. activity a person has the capability to do) category provides clinically-useful information to people with stroke and their families. Advances in computing have improved upon old and led to new analysis techniques for building prediction models of UL outcomes after stroke. Recently, machine learning techniques of support vector machines (SVM) and tree-based methods (e.g., Classification and Regression Trees [CARTs]) have been used to classify people with stroke into categories with different ranges of UL capacity.[18,20-23] The PREP 2 prediction model was originally built and validated with a CART which resulted in the easy to interpret decision tree.[21] Machine learning techniques have the advantages of: 1) requiring fewer assumptions about the distributions of the data, 2) numerous options for non-parametric models, and 3) strong predictive capabilities.[19,22-25] There are strengths and weaknesses to each machine learning technique. For example, the CART algorithm yields a single, easy to interpret decision tree

(strength), but lower predictive accuracy on new, external samples because of high variance (weakness).[26] An alternative to creating a single decision tree is to use ensemble classifiers like *b*ootstrap *agg*regation (called "bagging") or random forests.[27] These ensemble techniques rely on the collective judgement of many decision trees (hundreds or even thousands) in order to make a classification. These ensemble methods tend to have higher predictive power and reduce the risk of over-fitting relative to other CART methods, but at the expense of interpretability (as there is no longer one single decision tree to follow, but a whole forest of trees).[22,27] Capitalizing on the advantages of ensemble machine learning algorithms by applying them for prediction of UL performance outcomes could yield key insights into UL recovery post stroke.

The purpose of this study, therefore, was to explore how different machine learning techniques can be used to understand how clinical measures and participant demographics captured early after stroke are associated with the UL performance categories from a later post stroke time point. We utilized the same data set from which we had previously predicted the trajectory of single, continuous UL performance variables with regression techniques.[28] In this analyses, we attempt to predict the subsequent multivariate categories of UL performance that people with stroke fell into. We explicitly tested different machine learning methods to build predictive models with different input variables as predictors (also called feature sets) to determine how each method yields similar versus different results. We hypothesized that measures of UL impairment,[21] UL capacity,[18,28] and other non-motor clinical measures, would be the most important predictors of the subsequent UL performance category.

4.3 Methods

This study was a secondary analysis of data collected from a prospective, observational, longitudinal cohort tracking UL change over time.[28] Sources of data from two time points were participant characteristics, clinical measures from early after stroke, and subsequent categories of UL performance (from a previous report)[15] later after stroke.

4.3.1 Participants

Participants were included in the prospective, observational, longitudinal cohort if the following criteria were met: (1) within two weeks of first-ever ischemic or hemorrhagic stroke, confirmed with neuroimaging; (2) presence of UL motor deficits within the first 24-48 hours post stroke, as indicated by a NIHSS[29] Arm Item scores of one to four or documented manual muscle test grade[30] of <5 anywhere on the paretic UL; (3) ability to follow a two-step command, as measured by a NIHSS[29] Command Items score of 0; and (4) anticipated return to independent living (i.e., not institutionalized), as indicated by the acute stroke team. Persons with stroke were excluded if any of the following criteria were met: (1) history of previous stroke, other neurologic condition, or psychiatric diagnoses; (2) presence of comorbid conditions that may limit recovery (e.g., end-stage renal disease or stage IV cancer); (3) lived more than 90 minutes from the study location; and (4) currently pregnant by self-report. The Human Research Protection Office at Washington University in St. Louis approved this study, and all participants provided written informed consent.

4.3.2 Data Collection

Cohort participants completed eight assessment sessions over the first 24 weeks post stroke. This analysis used data from the first assessment (within two weeks of stroke onset) to predict the subsequent category of UL performance[15] from the latest time point between six and

24 weeks post stroke. We retained any person in the cohort whose last measurement was between six and 24 weeks post stroke because UL performance appears to stabilize between three and six weeks.[28,31] Participants were excluded from this analysis if they were missing any of the predictor variables from the first assessment point (see Table 4.1). Assessments were administered by trained personnel (licensed physical therapists or occupational therapist, range of experience with measures was two-15 years). Since this was an observational cohort study, we did not provide nor control for the amount or type of rehabilitation services delivered to enrolled participants. Participants received rehabilitation services as prescribed by their medical team.

4.3.3 Dependent variable used for the models

The dependent variable (outcome or class in machine learning) in this analysis was a category of UL performance established in previous report.[15] These were derived from UL performance variables quantified via accelerometer data.[15] Participants in the prospective, longitudinal, observational, cohort wore Actigraph GT9X-Link accelerometers on both wrists at each time point with methods previously described.[1] Briefly, tri-axial acceleration data are sampled at 30 Hz for 24 or more hours continuously. Once the accelerometers were returned to the lab, data were uploaded, visually inspected, and processed using ActiLife 6 (Actigraph Corp., Pensacola, FL) proprietary software. For most variables, data were band-pass filtered (0.25-2.5 Hz) and down sampled into one-second epochs with ActiLife proprietary software, where each second is the sum of the 30 Hz values in that second and converted into activity counts (1 count = 0.001664 g). Similar to previous work,[7,11,28,32] accelerometry data was processed using custom written software in MATLAB (Mathworks, Inc., Natick, MA) to calculate UL performance variables which quantify various aspects of UL activity in everyday life. The variables measure the duration, magnitude, variability, and symmetry of one or both ULs. Five UL performance

variables were used to create the five categories of UL performance; the participant's category assignment in the prior report was the outcome in this analysis.[15] The names of each of the five categories were chosen for their overall level of UL activity and the integration of both ULs into activity in daily life and are named: A) *Minimal Activity/Rare Integration*; B) *Minimal Activity/Limited Integration*; C) *Moderate Activity/Moderate Integration*; D) *Moderate Activity/Full Integration*; and E) *High Activity/Full Integration*, see Figure 4.1 for visual representation of the categories for the participants included in this analyses. The categories are presented in order of increasing overall UL performance.[15]

4.3.4 Independent predictor variables

The input variables (also known as feature sets in machine learning) were participant demographics and clinical measures. In the prospective, longitudinal, observational cohort, 15 demographic variables and nine clinical measures were administered at the first assessment time point, within two weeks post stroke. Of the 24 variables available, seven were excluded because of multi-collinearity and extremely low (or no) variability.[31] In the case of multi-collinearity, we retained the variables that were more likely to be available in routine post-stroke clinical care.[33] Table 4.1 presents the 17 predictors selected for this analysis organized into three main categories: 1) UL clinical measures; 2) non-motor clinical measures; and 3) participant demographics.

Table 4.1. Predictors included in the analysis

Predictor Name	Description	Construct	Scoring
UL Measures			
Action Research Arm Test (ARAT)[34-36]	Standardized measure assessing UL functional ability for activity	UL capacity	Scores range from 0-57, higher values indicate greater UL function
Shoulder Abduction Finger Extension (SAFE)[21]	Sum of two Medical Research Council strength grades from the shoulder abductors and the finger extensors.	UL impairment	Scores are whole numbers and range from 0-10, higher values indicate less impairment in the affected UE
Upper Extremity Fugl-Meyer (UEFM)[37,38]	Standardized measure assessing movement in and out of synergies of the affected UL	UL impairment	Scores range from 0-66, higher values indicate less impairment in the affected UL
Non-motor clinical measures			
Center for Epidemiological Studies Depression Scale (CES-D)[39,40]	Questionnaire asking about the frequency and severity of symptoms associated with mood	Depression screen	Scores range from 0-60, higher scores indicative of greater depressive symptomatology
Unstructured Mesulam[41,42]	Paper, pencil test measuring visual spatial ability	Hemispatial neglect	Scores are calculated by subtracting the omissions from the total score of 60 (0-30 on both sides), >4 omissions on one side are considered pathological
Montreal Cognitive Assessment (MOCA)[43]	Brief tool to screen for cognitive impairment across multiple domains	Cognitive screen	Scores range from 0-30, scores < 26 indicate cognitive impairment
National Institute of Health Stroke Scale (NIHSS)[29]	Standardized measure of global stroke severity	Stroke severity	Scores range from 0-42, lower scores indicate less severe stroke overall
Demographics			
Area deprivation index (ADI)[44-46]	Multi-dimensional evaluation of a region's socioeconomic conditions	socioeconomic disadvantage	Scores are % rankings and range from 1 to 100, lower values indicate lowest level of "disadvantage"
Age	Participant age at time of testing	--	Minimum age of 18
Concordance	Affected UL is dominant UL	--	Categorized as: Yes/No
Ethnicity	Participant report of ethnicity	--	Categorized as: Non-Hispanic/Non-Latino or Latino
Living status pre-stroke	Participant report of prior living situation	--	Categorized as*: Living alone, independent ADLs/ Living alone, assist ADLs/ Living with others, independent ADLs/ Living with others, assist ADLs
Living status 2-weeks post stroke	Participant location at 2-week time point	--	Categorized as: Inpatient / Skilled-nursing facility / Assisted living /, Home / Other
Race	Participant reported racial identification	--	Categorized as: White/ Black or African-American/ Asian/ American Indian or Alaska Native/ Hawaiian or other Pacific Islander
Sex	Participant report of sex	--	Categorized as: Male/ Female
Stroke type	Cause of disruption of blood flow, from medical record	--	Categorized as: Ischemic/ Hemorrhagic/ Unknown
Time-post stroke	Number of days from stroke-onset to 2-week testing, from medical record	--	Difference in days between date of testing and stroke onset

Abbreviations: * Activities of daily living (ADL)

4.3.5 Statistical Analysis

All data were analyzed in R (version 4.1.2), an open source statistical computing program.[47] Distributions and pair-wise scatterplots of the correlations between the variables were

examined to understand the variability in the sample and the relationships among the variables. We tested a series of supervised machine learning algorithms with different numbers of input variables to predict subsequent activity of the UL measured via accelerometry as a function of clinical and demographic variables collected within two weeks of stroke. These algorithms were a single decision tree[48], bagged trees, and random forests.[49] We present several different measures of classification accuracy and the importance of different predictor variables.[50]

Classification Using Supervised Learning Algorithms

In this analysis, different machine learning techniques were explored to understand how clinical variables captured early after stroke best predicted an subsequent category of UL performance. Given the smaller sample size (n=54) we lacked the capability to partition the data into training, validation, and testing sets. As such, our focus is on the accuracy of in-sample prediction (how well the model explains the data on which it was trained), cross-validation accuracy (the out-of-bag error estimate, defined below), and measures of variable importance (to identify the most important predictors).[24,25] The models were built with different machine learning algorithms as described in the steps below.

First, a single unpruned classification tree was built[48] using the CART algorithm.[26,48] If one thinks of the dataset as a matrix, each person (or observation) is a row and each predictor (or feature) is a column. The algorithm looks at all predictors and selects the one that best explains the outcome, creating a "branch" in the growing tree (if the predictor is > a certain value, go left, otherwise, go right). Moving down each branch, the algorithm then looks at all remaining predictors, selecting the one that explains the most variance (i.e., creates the most separation in predictions based on the Gini-index, a measure of node purity).[25,51] In our analysis, we built a single decision tree based on all 17 predictors (Table 4.1) to predict the UL performance

categories (outcomes).[25,51] The process then repeats, creating a tree made up of many branches that ends in the "leaves", the final prediction at the end of that branch.[24,26]

Second, we used *b*ootstrap *agg*regation (*bagging*) as an ensemble method to reduce the likelihood of overfitting the data with a single tree.[24] Bagging works identically to the single tree algorithm above, but rather than building one tree out of all available data, samples are bootstrapped: made by randomly sampling individuals (rows) from data with replacement. Each sample then gets its own tree based on the individuals who made it "in the *bag*". Critically, this also means that the accuracy of each individual tree can be cross-validated against the observations left "out-of-bag", yielding out-of-bag error as measure of cross-validation accuracy.[24] The bagged model is thus the aggregated vote of all of the different trees when given input data for classification.

Finally, we also used random forests as a slightly more complicated ensemble method.[24,51] The random forest is constructed similarly to the bagged trees, building bootstrapped samples and fitting trees within each sample. However, in order to avoid potential bias and correlations between trees (i.e., a dominant predictor always being selected first), the random forest only considers a random subset of predictors (columns) at each node.[51] Thus, the random forest model allows for a similar calculation of cross-validation accuracy with the out-of-bag error, but generally leads to trees being less similar than in bagging, because only a subset of predictors are considered at each node. This more diverse forest of uncorrelated trees can then be used to get an aggregate vote when given input data for classification.[24,51]

Table 4.2. Model names and specifications of the six ensemble models.

Model	# Trees	Predictors Included in Input Data Set			# Predictors considered at each node (m=)*
		UL clinical measures	Non-motor clinical measures	Demographics	
Small bagged	2,000	✓	X	X	3
Small random forest	2,000	✓	X	X	√3 = 2
Medium bagged	2,000	✓	✓	X	7
Medium random forest	2,000	✓	✓	X	√7 = 3
Large bagged	2,000	✓	✓	✓	17
Large random forest	2,000	✓	✓	✓	√17= 4

* m is a tuning parameter for the bagged trees and random forest models

A total of six models were built with the two bagging algorithms and by systematically changing the model specifications, known as tuning parameters. Table 4.2 presents the model names and specifications for the six models built. Each of the six models were built with the number of trees held constant at 2,000. A high number of trees was chosen to ensure that all of the models would stabilize regardless of the data set used.[25] Three different input data sets were formed from the list of predictors (Table 4.1). The small data set included the UL clinical measures, the medium included the small data set + other non-motor clinical measures, and the large data set included the medium data set + demographic predictors. The bagged and random forest models were built by changing the tuning parameter m, which is the number of predictors available at each split. In the bagged models, m is equal to the total number of predictors in the data set whereas in random forests, m is equal to the square root of the number of predictors in the data set (last column Table 4.2).

Model performance and variable importance

An iterative process was used to quantify the explanatory power (in-sample accuracy), predictive power (out-of-bag estimate of error), and variable importance of the single decision tree, bagged models, and random forests. For all seven models, the full data set was fed back into the fitted model and the "in-sample" accuracy was quantified by comparing the predicted category and the actual UL performance categories.[24,51] As in-sample accuracy uses the same

data the model is trained on, it is best thought of as the "explanation" rather than "prediction" (because prediction requires an independent test data set). Second, for the bagged and random forest models the average out-of-bag error was used as a measure of cross-validation accuracy. This out-of-bag error is genuine prediction because each individual tree is independent of its out-of-bag data.[25] (Note that out-of-bag error cannot be calculated for the single tree, as all data are included in the training set for that tree.) Finally, the importance of each predictor for the six models was evaluated in two ways: 1) mean change in accuracy, and 2) mean change in the Gini index.[24,25] The mean change in accuracy is the improvement or decrease in the in-sample accuracy when each predictor is included in the model, predictors with higher accuracy values are more important for the successful classification (accuracy) of the outcome. Predictors with negative accuracy values decrease the model performance (accuracy), and are considered unimportant in predicting the outcome. The mean change in the Gini index is a measure of how each predictor contributes to the purity of the nodes and leaves in the models. The mean change in the Gini index is a positive integer, higher values of the mean change in the Gini index indicate greater importance of that predictor for the models.

4.4 Results

Table 4.3. Participant characteristics and demographics: Total sample and UL performance category

Characteristic	Total Sample N=54	A: Min Activity/ Rare Integration N=20	B: Min Activity/ Limited Integration N=4	C: Mod Activity/ Moderate Integration N=16	D: Mod Activity/ Full Integration N=10	E: High Activity/ Full Integration N=4
Age	66.3 ± 8.8	69.0 ± 7.8	63.3 ± 8.1	65.4 ± 10.1	65.6 ± 8.6	61.3 ±10.4
Sex, n (%)						
Male	31 (57)	10 (50)	1 (25)	10 (63)	7 (70)	3 (25)
Female	23 (43)	10 (50)	3 (75)	6 (11)	3 (30)	1 (25)
Ethnicity, n (%)						
Non-Hispanic/Non-Latino	57 (100)	20 (100)	4 (100)	16 (100)	10 (100)	4 (100)
Race, n (%)						
White	32 (59)	12 (60)	4 (100)	6 (38)	7 (70)	3 (75)
African-American	21 (39)	8 (40)	-	9 (56)	3 (30)	1 (25)
Asian	1 (2)	-	-	1 (6)	-	
Stroke type, n (%)						
Ischemic	48 (89)	20 (100)	4 (100)	11(69)	9 (90)	4 (100)
Hemorrhagic	6 (11)	-	-	5 (31)	1 (10)	-
Concordance, n (%)	23 (43)	8 (40)	1 (25)	4 (25)	7 (70)	3 (75)
Time post stroke in days	13 (12,15)	13 (12,14)	14 (13,15)	13 (12,14)	15 (13,16)	16 (13, 18)
Living status pre-stroke n (%)						
Alone, independent	1 (20)	5 (25)	1 (25)	5 (31)	-	-
Others, independent	43 (80)	15 (75)	3 (75)	11 (69)	10 (100)	4 (100)
Living status 2-weeks post stroke, n (%)						
Inpatient	47 (87)	20 (100)	3 (75)	14 (87)	7 (70)	3 (75)
Home	7 (13)	-	1 (25)	2 (13)	3 (30)	1 (25)
ADI	75 (39, 86)	80 (41, 88)	66 (48, 78)	76 (45, 87)	67 (31, 84)	31 (21, 47)
Upper limb measures						
ARAT	20 (0, 43)	0 (0, 3.3)	4 (0, 20)	37 (23, 48)	45 (30, 53)	37 (28, 41)
SAFE	7 (1, 8)	1 (1, 4)	5 (1, 8)	8 (7, 8)	8 (8, 8)	8 (7, 8)
UEFM	37 (10, 57)	10 (8, 21.3)	23 (9, 43)	54 (36, 56)	59 (48, 61)	54 (43, 57)
Non-motor clinical measures						
CES-D	14.0 ± 9.5	18.2 ± 10.3	14.0± 9.1	15.5 ± 8.2	7.0 ± 6.1	4.8 ± 1.7
Mesulam	0 (0, 3)	1 (0, 7)	1 (0, 4)	1 (0, 4)	0 (0, 1)	-1 (-1, 1)
MOCA	17.6 ± 7.1	14.6 ± 7.9	20.3 ± 8.3	18.3 ± 7.1	20.5 ± 4.1	20.3 ± 4.4
NIHSS	6 (4, 10)	10 (6, 15)	5 (4, 9)	5 (3, 8)	4 (4, 6)	3 (3, 4)

Summary statistics for demographic information and the predictors as means and standard deviations when normally distributed, otherwise by medians and the 1st and 3rd inter-quartile values. Categorical variables are presented as counts (n) and % of the total sample and by category. Abbreviations: Area deprivation index (ADI), Action Research Arm Test (ARAT), Shoulder Abduction Finger Extension (SAFE), Upper Extremity Fugl-Meyer (UEFM), Center for Epidemiological Studies Depression Scale (CES-D), Montreal Cognitive Assessment (MOCA), National Institute of Health Stroke Scale (NIHSS).

Overall, the sample of persons with first-ever stroke were generally in their 60's and had mild to moderate stroke (83% with NIHSS 0-15). Of the 67 participants enrolled in the prospective, observational, longitudinal cohort study,[28] 54 had the necessary data to be included in this secondary analyses. Participant assignment into the five UL performance categories was pulled from our previous report.[15] Demographics of the 54 participants and for the subsets

assigned to the UL performance categories are provided in Table 4.3. Figure 4.1 is included for descriptive purposes as a visual representation of the five UL performance categories. In Figure 4.1, categories (4.1A-4.1E) are represented by Coxcomb plots, where the individual variables used to define the categories are wedges. As one moves from Figure 4.1A to Figure 4.1E, the wedges take on different relative proportions, generally getting larger, with the best UL performance represented by category E, *High Activity/Full Integration*.

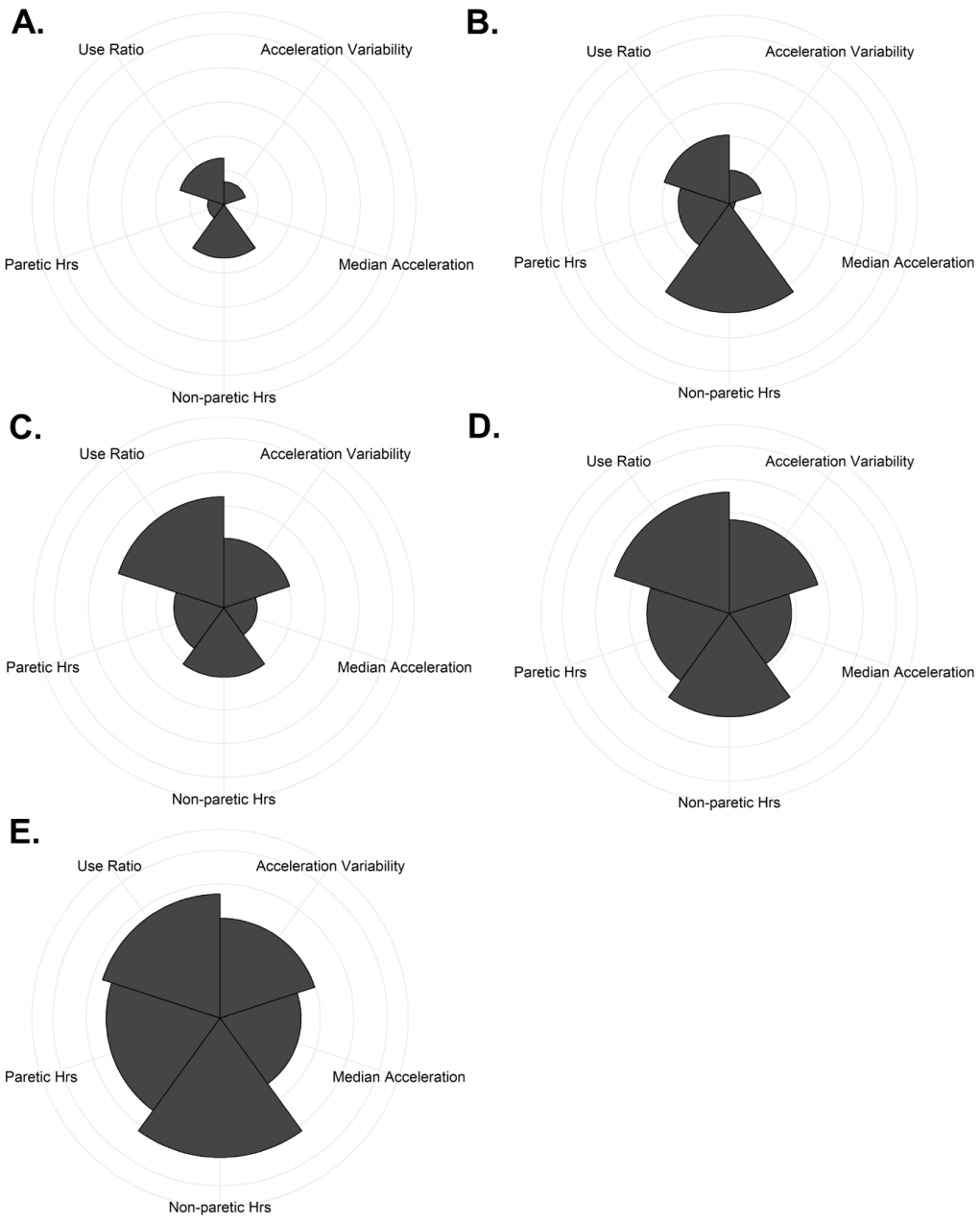


Figure 4.1. Coxcomb charts of the five UL performance categories of the 54 participants in this analysis (categories assigned in Barth et.al 2021). The five UL performance variables are divided into equally segmented wedges on the radial chart and the area of each wedge is proportional to the magnitude of the score on that dimension relative to the sample that created the categories. Each chart illustrates the contribution of the five UL performance variables on a standardized scale and are anchored to the minimum and maximum value of each variable in the prior analysis used to establish the categories. The categories are presented in order of increasing overall UL performance and are named: (A) *Minimal Activity/Rare Integration*; (B) *Minimal Activity/Limited Integration*; (C) *Moderate Activity/Moderate Integration*; (D) *Moderate Activity/Full Integration*; (E) *High Activity/Full Integration*.

The single, unpruned decision tree (Figure 4.2) allocated participants into only three of the five UL performance categories. The predictors that were selected for this tree included all three UL clinical measures (SAFE, ARAT, and UEFM) and two non-motor clinical measures (CES-D and Mesulam). This tree has a misclassification rate of 29%, meaning that 16/ 54 people were misclassified into a different category than their actual. In this tree, the SAFE score is the root node. For participants with less overall strength in their paretic UL (SAFE <7.5), the left side of the tree is used, with the ARAT, Mesulam, and UEFM scores used to assign people into either category A (Minimal activity/rare integration) or C (*Moderate Activity/Moderate Integration*). For participants with more overall strength in their paretic UL (SAFE >7.5), the right side of the tree is used, with participants assigned to either category C (*Moderate Activity/Moderate Integration*) or D (*Moderate Activity/Full Integration*) based on their scores on the depression scale (CES-D).

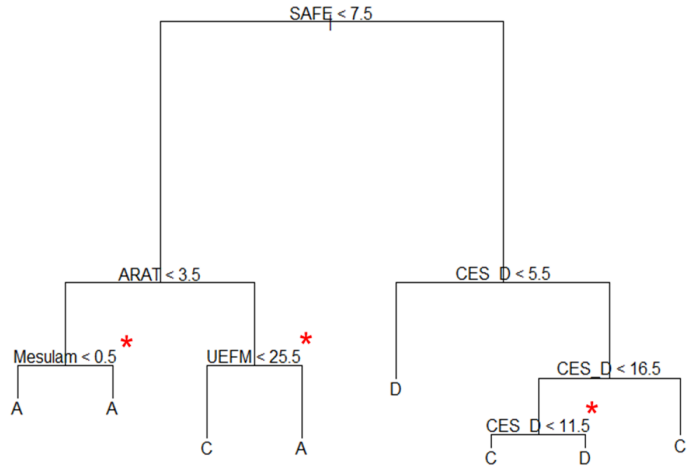


Figure 4.2. Single unpruned decision tree predicting a category of UL performance from all 17 predictors. In reading the tree, if the argument is true, go left; if the argument is false, go right. This model predicts three out of five UL performance categories (A, C, and D) from 3 UL clinical measures and 2 other non-motor clinical measures. The categories are: (A) *Minimal Activity/Rare Integration*; (B) *Minimal Activity/Limited Integration*; (C) *Moderate Activity/Moderate Integration*; (D) *Moderate Activity/Full Integration*; (E) *High Activity/Full Integration*.

* Measures with a red asterisk are counter-intuitive results. For example, UEFM score above 25.5 indicates less impairment in the paretic UL and with this tree, people with less UL impairment would be predicted to category A , the lowest UL performance category. Abbreviations: Shoulder Abduction Finger Extension (SAFE), Action Research Arm Test (ARAT), Upper Extremity Fugl-Meyer (UEFM), Center for Epidemiological Studies Depression Scale (CES-D)

Table 4.4. Model performance of models built with different machine learning algorithms.

Model Performance Statistic	Model Name						
	Single Decision Tree	Small		Medium		Large	
		Bagged model	Random forest	Bagged model	Random forest	Bagged model	Random forests
In-sample Accuracy*							
Mean	0.70	0.96	0.96	1.0	1.0	1.0	1.0
IQR	(0.56, 0.82)	(0.87, 1.0)	(0.87, 0.99)	(0.93, 1.0)	(0.93, 1.0)	(0.93, 1.0)	(0.93, 1.0)
Out-of-bag estimate of error†							
Mean	*na*	0.55	0.52	0.47	0.46	0.48	0.48
95% CI		(0.54, 0.56)	(0.52,0.54)	(0.46, 0.48)	(0.44, 0.46)	(0.46, 0.48)	(0.48, 0.48)

*In-samle accuracy is a measure of the explanatory power of the model and was quantified for all seven models by comparing the predicted category and the actual UL performance categories. Values closer to 1.00 indicate better model performance.

† Out-of-bag estimate of error is a measure of the predictive power of the models and was quantified for the six bagging models as cross-validation accuracy. Lower error-rate values indicate better model performance.

Abbreviations: confidence interval (CI), inter-quartile range (IQR)

Our next step was to explore the use of bagging methods to build the six models (Table 4.2). The statistics used to evaluate model performance of the single decision tree, bagged models, and random forest models are presented in Table 4.4. The in-sample accuracy of all the models is better than chance (chance=0.20 for each of the five categories) alone. The single decision tree has an in-sample accuracy of 0.70 whereas the in-sample accuracy of the bagged and random forest models ranges from 0.96 to 1.00, indicating better performance. Predictive power of the six models was better than chance, with the medium and large models being better than the small models and having mostly overlapping 95% confident intervals.

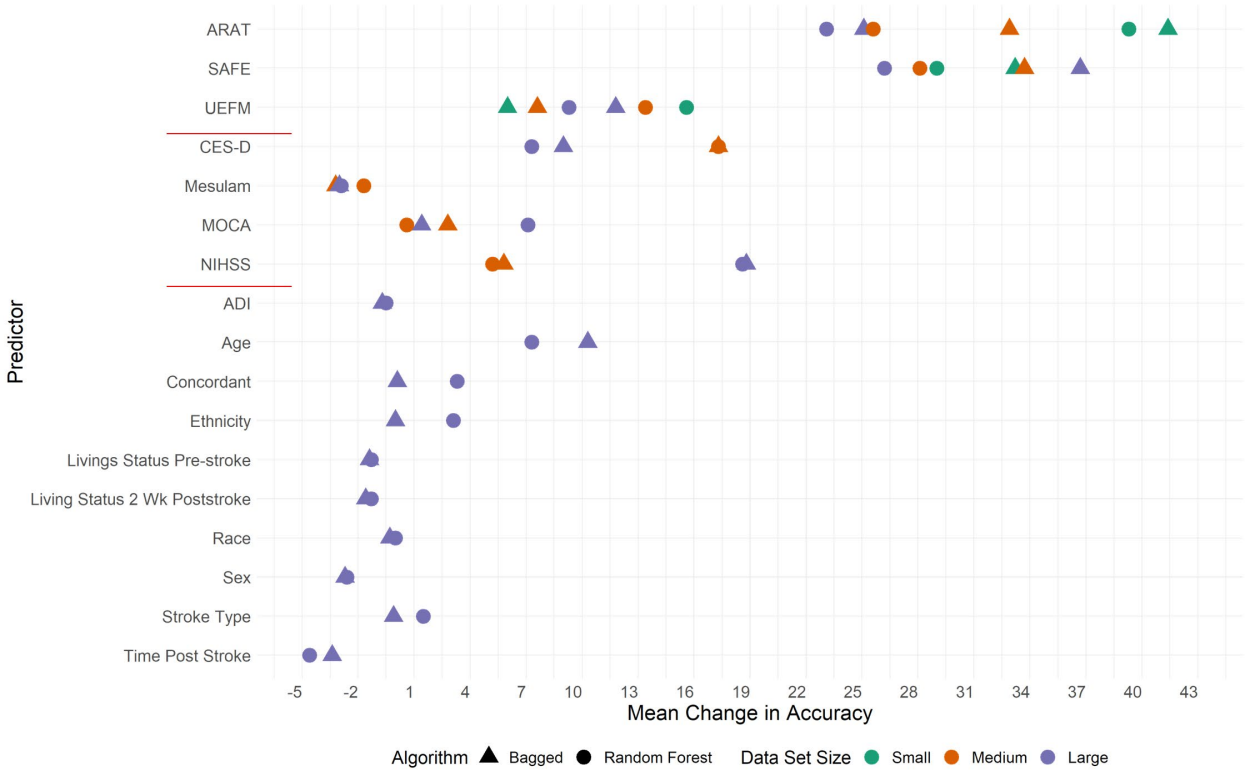


Figure 4.3. Variable importance computed as the mean change in accuracy. Variable importance plot for the six models built with bagging algorithms from different input datasets and tuning parameters. Variable importance is computed using the mean change in accuracy, and is expressed relative to the maximum. Higher values indicate greater importance of the specific predictor in the model and values ≤ 0 indicate these predictors decrease the overall accuracy of the model. The shape represents the algorithm used and color represents the size of the input dataset. The small data set that includes UL clinical measures, the medium sized data set includes UL clinical measures + non-motor clinical measures, and the large sized data set includes UL clinical measures + non-motor clinical measures + demographics. The bagged models were built with all predictors available in the data set and random forests were built with the square root of the number of predictors.

Figures 4.3 and 4.4 present the variable importance plots for the six models using the mean change in accuracy (Figure 4.3) and the mean change in the Gini index (Figure 4.4). The two red lines on the y axis are placed to separate the predictors relative to the data sets with UL clinical measures first, non-motor clinical measures second, and the demographics last. In these plots, shape represents the algorithm, which was either bagged (triangle) trees or random forests (circle) and the three colors represent the size of the input data set as: small (green), medium (orange), or large (purple). In Figure 4.3, one can see that two UL clinical measures (ARAT and

SAFE) are the most important predictors regardless of the algorithm used or input data set size. Additionally, the UEFM and CES-D emerge as important predictors to maintain accuracy of the models, specifically with the medium and large input data sets. Interestingly, a few of the demographic predictors (sex, living status pre-stroke, living status 2-weeks post stroke, and time post stroke) all have negative values indicating including these predictors decrease the accuracy of the models. In Figure 4.4, the UL clinical measures and non-motor clinical measures are most important with respect to the mean change in the Gini index regardless of the algorithm or input data set size. All UL clinical measures (ARAT, SAFE, and UEFM) have the highest values of the mean change in the Gini index for the small data set only compared to the models built with the medium and large data sets. For most of the predictors, the circles and triangles of the same color are close together indicating similar values for mean change in the Gini index regardless of if the bagged or random forest algorithm was used. Additionally, some of the demographic predictors have a mean change in the Gini index close to 0 indicating less importance of these predictors similar to the mean change in accuracy (Figure 4.4).

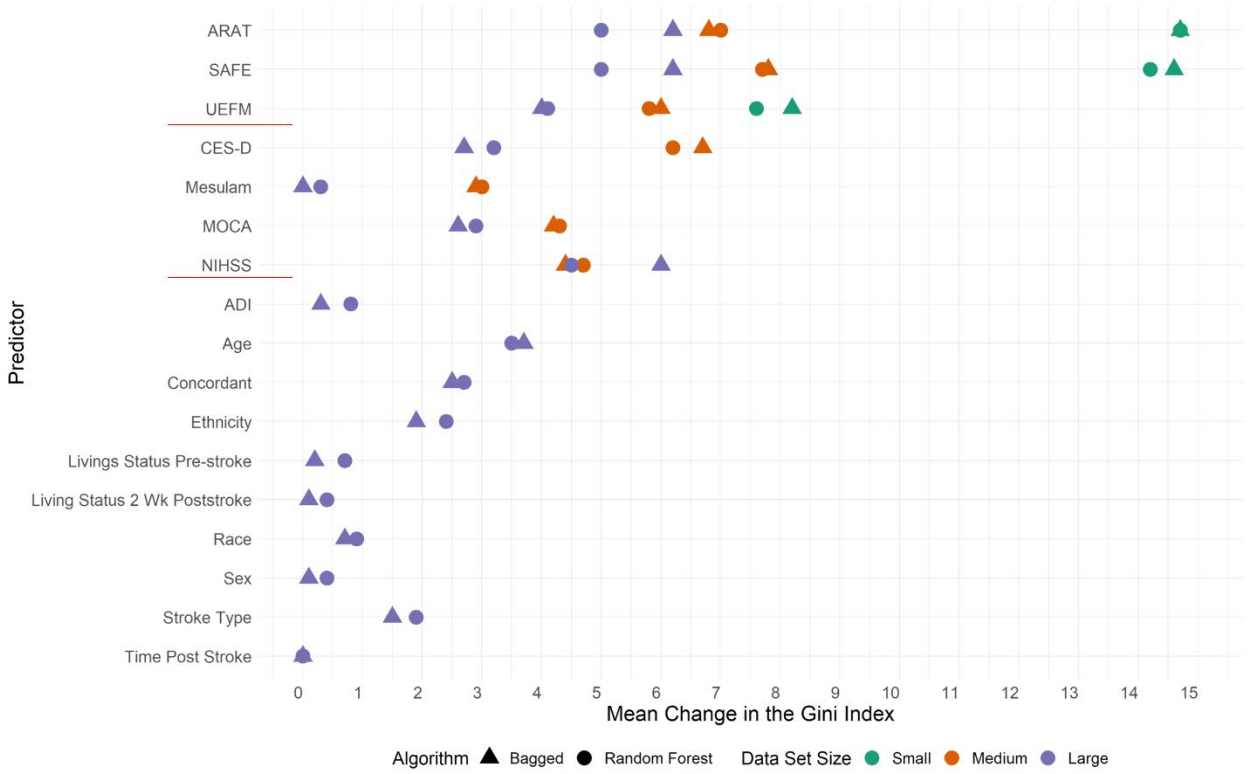


Figure 4.4. Variable importance computed as the mean change in the Gini Index. Variable importance plot for the six models built with different input datasets and tuning parameters. Variable importance is computed using the mean change in the Gini index, and is expressed relative to the maximum. The shape represents the algorithm used and color represents the size of the input dataset. The small data set that includes UL clinical measures, the medium sized data set includes UL clinical measures + non-motor clinical measures, and the large sized data set includes UL clinical measures + non-motor clinical measures + demographics. The bagged models were built with all predictors available in the data set and random forests were built with the square root of the number of predictors.

4.5 Discussion

The purpose of this study was to explore how different machine learning techniques could be used to understand the association between clinical measures and participant demographics captured early after stroke and the subsequent UL performance category. Our hypothesis was supported, such that measures of UL impairment and capacity were the most important predictors of subsequent UL performance category, regardless of the machine learning algorithm used. Other non-motor clinical measures emerged as key predictors, while participant demographics predictors were less important across the models. Models built with the bagging algorithms had better in-sample accuracy compared to the single decision tree. The models had

moderate out-of-bag errors, indicating that there are likely unevaluated, missing predictive factors. There are two novel contributions of the present study: 1) different machine learning techniques were used to allow for comparison of the results and 2) the outcome of the models was a multi-dimensional category of UL performance. These findings contribute to the understanding of what factors early after stroke may partially influence the subsequent UL performance categories.

Using different machine learning techniques provided information about which predictors were most important to the outcome for this sample and those that generalize to the population. Consistent with efforts to predict an individual's subsequent UL impairment or UL capacity category,[21,52-56] these results point to the importance of measures of UL impairment and capacity for predicting subsequent UL performance category. The UL impairment (SAFE score, UEFM) and capacity (ARAT) measures were generally the most important predictors regardless of the algorithm or input data set used. Depressive symptomology (CES-D) and overall stroke severity (NIHSS) increased the overall predictive ability of the models. The NIHSS is a measure of global stroke severity and it is possible that the non-motor aspects of this measure are driving the added value of this predictor. Collectively these results indicate that the subsequent UL performance categories were most influenced by UL impairment, capacity, presence of depression, and overall stroke severity. These are different secondary predictors than were identified in a previous study predicting a single UL performance variable, where non-motor clinical measures of hemispatial neglect and cognitive impairments along with participant demographic information were found to be important.[20,57] It is reasonable that different factors emerged as predictors of a single UL performance variable vs. a multivariate category of UL performance. A key finding in the current analysis is the substantial out-of-bag estimate of error

(0.48 to 0.55), indicating that there other factors (predictors) that likely contribute to UL performance that were not assessed in the cohort studied here. Other possible factors that could influence UL performance include: biopsychosocial, cognitive constructs (e.g. apraxia), neurobiology (e.g. motor evoked potential [MEP]), and other demographics (e.g. employment status). Future research should explore how these factors captured early after stroke are associated with subsequent UL performance.

The purpose of the present analysis was not to make perfect predictions, but rather to explore the associations between these variables using different machine learning techniques. The first model we explored was a single decision tree built with the CART algorithm. While the graphical representation of this decision tree (Figure 4.2) may be easy to interpret, the output is somewhat counterintuitive. For example, on the left side of the tree for people with a SAFE score < 7.5 and an ARAT score > 3.5, it is counterintuitive that a UEFM score < 25.5 (more UL impairment) would put one in a better category (C, Moderate activity/moderate integration), and an UEFM > 25.5 would put one in a worse category (A, Minimal activity/rare integration). This likely occurred because the tree is constructed with all of the input data and the algorithm assigned people to the categories based on the probability of ending up in each node. It is possible that there were a few people in this data set that had this unusual pattern of scores with respect to UL impairment and capacity. Likewise, the tree selected the Mesulam, a non-motor clinical measure of hemispatial neglect, for the most leftward node, but inclusion of this node does not change the categorization. Then, two different bagging algorithms were used to build predictive models with different sized input data sets to determine if they yielded similar versus different results. A benefit of the bagged and random forest is the relative increase in the explanatory and predictive power of these models because of cross-validation, even with the

smaller sample size of our data set, as seen in Table 4.4. The in-sample accuracy for these six models was improved compared to the single decision tree, indicating that these models are doing a good job explaining the data we had. The out-of-bag estimate of error, however, is more important with respect to the predictive power of these models. While the out-of-bag estimate of error does decrease (indicating more accurate predictions) with the addition of the non-motor clinical measures, the outcome of the models remain largely unchanged and the out-of-bag error remains substantial. These data illustrate the point that, ultimately, prediction models are only going be as "good" as the input data available; simply switching to a different machine learning method with different tuning parameters may not substantially change the predictive ability of the model. While we did not try all possible machine learning algorithms (e.g. SVM or neural networks), this could be an important consideration for future research. The single decision tree is a transparent model because one can see how the decisions are made in the tree where the bagged and random forest models are harder to interpret, because the classification is based on thousands of trees. These ensemble classifiers, however, can still be clinically useful. As we move to a world where electronic health records are integrated into advanced information management systems with data visualization and machine learning capability,[58] the possibilities are endless to imagine how clinical measures and participant demographics early after stroke could be used to predict meaningful outcomes for people with stroke and their families. Implementation of these techniques into routine care will require extremely large data sets to build and then to validate models. Sample sizes will need to be at least an order of magnitude bigger than the larger data sets available today (i.e. in the thousands, not tens or hundreds of participants).[24,25] For these machine learning methods to be clinically-available in the future,

research groups need to start now on pooling participant data, data sharing, and/or using common data elements across studies.

There are a few limitations to consider when interpreting our findings. First, these results should be interpreted as exploratory or "hypothesis-generating". Additional studies are required to validate these results. Second, due the small sample size the data could not be split into test and training sets. Nonetheless, we were still able to capitalize on the computing power of these techniques to provide additional information that contributes to our understanding of how typical information captured early after stroke was associated with a subsequent UL performance category. Finally, the predictor sets only included clinical measures and participant demographic information because we were limited by the data collected for the prospective, observational, longitudinal cohort study and the variability of the predictors across the cohort.[28] One example of a potential predictor variable not collected here is a positive motor evoked potential (MEP), which has been identified as an important factor predicting UL capacity for persons with greater UL impairment in their paretic UL.[20,21,52] As an example of lack of variability in potential predictors, we also collected a survey quantifying self-perception of UL performance recovery. Scores on these measures were highly homogenous across participants, making it impossible for this factor to contribute to the variance in the outcome. Future studies will need to be designed with a more comprehensive set of potential predictors, including neurobiological and psychosocial factors.

4.6 Conclusion

Machine learning techniques can be used to understand how clinical measures and participant demographics captured early are associated with subsequent post stroke UL performance category. UL clinical measures were the most important predictors of the

subsequent UL performance category in this exploratory analysis regardless of the machine learning algorithm used. Other non-motor clinical measures emerged as important predictors to maintain the accuracy of the models, but including these measures had little impact on out-of-bag estimate of error. These results reinforce that UL performance, *in vivo*, is not a simple product of body functions nor the capacity for UL movement, instead being a complex phenomenon dependent on many physiological and psychological factors. Utilizing machine learning, this exploratory analysis is a productive step towards prediction of UL performance. Future research is required to explore other factors associated with UL performance along with the role of predictive models in rehabilitation after stroke.

4.7 References

1. Lang CE, Waddell KJ, Klaesner JW, Bland MD. A Method for Quantifying Upper Limb Performance in Daily Life Using Accelerometers. *Jove-J Vis Exp*. Apr 17;(122)doi:ARTN e55673 10.3791/55673

2. WHO. International Classification of Functioning, Disability and Health (ICF). *World Health Organization*. 2018;(06/01/2021)

3. Demers M, Levin MF. Do activity level outcome measures commonly used in neurological practice assess upper-limb movement quality? *Neurorehabilitation and neural repair*. 2017;31(7):623-637.

4. Urbin MA, Waddell KJ, Lang CE. Acceleration metrics are responsive to change in upper extremity function of stroke survivors. *Arch Phys Med Rehabil*. May 2015;96(5):854-61. doi:10.1016/j.apmr.2014.11.018

5. Barth J, Klaesner JW, Lang CE. Relationships between accelerometry and general compensatory movements of the upper limb after stroke. *Journal of Neuroengineering and Rehabilitation*. Oct 20 2020;17(1)doi:ARTN 13810.1186/s12984-020-00773-4

6. Smith BA, Lang CE. Sensor Measures of Symmetry Quantify Upper Limb Movement in the Natural Environment Across the Lifespan. *Arch Phys Med Rehabil*. Jun 2019;100(6):1176-1183. doi:10.1016/j.apmr.2019.01.004

7. Urbin MA, Bailey RR, Lang CE. Validity of body-worn sensor acceleration metrics to index upper extremity function in hemiparetic stroke. *Journal of neurologic physical therapy : JNPT*. Apr 2015;39(2):111-8. doi:10.1097/NPT.0000000000000085

8. Bailey RR, Lang CE. Upper-limb activity in adults: referent values using accelerometry. *J Rehabil Res Dev*. 2013;50(9):1213-1222. doi:10.1682/JRRD.2012.12.0222

9. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang CE. Upper Limb Performance in Daily Life Improves Over the First 12 Weeks Poststroke. *Neurorehabil Neural Repair*. 2019;33(10):836-847.

10. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of Accelerometry for Monitoring Real-World Arm Activity in Patients With Subacute Stroke: Evidence From the Extremity Constraint-Induced Therapy Evaluation Trial. *Arch Phys Med Rehabil*. 2006/10/01/ 2006;87(10):1340-1345. doi:https://doi.org/10.1016/j.apmr.2006.06.006

11. Bailey RR, Lang CE. Upper-limb activity in adults: referent values using accelerometry. *J Rehabil Res Dev*. 2013;50(9):1213-22. doi:10.1682/JRRD.2012.12.0222

12. Bailey RR, Klaesner JW, Lang CE. Quantifying Real-World Upper-Limb Activity in Nondisabled Adults and Adults With Chronic Stroke. *Neurorehabil Neural Repair*. Nov-Dec 2015;29(10):969-78. doi:10.1177/1545968315583720

13. Bailey RR, Klaesner JW, Lang CE. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. *PLoS One*. 2014;9(7):e103135. doi:10.1371/journal.pone.0103135

14. David A, Subash T, Varadhan SKM, Melendez-Calderon A, Balasubramanian S. A Framework for Sensor-Based Assessment of Upper-Limb Functioning in Hemiparesis. *Front Hum Neurosci*. 2021;15:667509. doi:10.3389/fnhum.2021.667509

15. Barth J, Lohse KR, Konrad JD, Bland MD, Lang CE. Sensor based categorization of upper limb perforance in daily life of persons with and without neurological upper limb deficits. *Frontiers in Rehabilitation Science*. 2021;2doi:10.3389/fresc.2021.741393

16. Kiær C, Lundquist CB, Brunner I. Knowledge and application of upper limb prediction models and attitude toward prognosis among physiotherapists and occupational therapists in the clinical stroke setting. *Topics in stroke rehabilitation*. 2020:1-7.

17. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*. 2009;338:b606.

18. Saltão da Silva MA, Cook C, Stinear CM, Wolf SL, Borich MR. Paretic upper extremity strength at acute rehabilitation evaluation predicts motor function outcome after stroke. *medRxiv*. 2021;

19. Miller AE, Russel E, Reisman DS, Kim HE, Vu D. A machine learning approach to identifying important features for achieving step thresholds in individuals with chronic stroke. *medRxiv*. 2021;doi:doi: https://doi.org/10.1101/2022.01.27.22269966

20. Lundquist CB, Nielsen JF, Brunner IC. Prediction of Upper Limb use Three Months after Stroke: A Prospective Longitudinal Study. *Journal of Stroke and Cerebrovascular Diseases*. 2021/11/01/ 2021;30(11):106025. doi:https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106025

21. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. *Ann Clin Transl Neurol*. Nov 2017;4(11):811-820. doi:10.1002/acn3.488

22. Tozlu C, Edwards D, Boes A, et al. Machine Learning Methods Predict Individual Upper-Limb Motor Impairment Following Therapy in Chronic Stroke. *Neurorehabil Neural Repair*. May 2020;34(5):428-439. doi:10.1177/1545968320909796

23. Lin CH, Hsu KC, Johnson KR, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput Meth Prog Bio*. Jul 2020;190doi:ARTN 105381 10.1016/j.cmpb.2020.105381

24. Kuhn M, Johnson K. *Applied predictive modeling*. Springer; 2016:600.

25. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. vol 112. Springer; 2013.

26. Galarnyk M. Understanding Decision Trees for Classificaton (Python). 2022. https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952

27. Brownlee J. Bagging and Random Forest Ensemble Algorithms for Machine Learning. Accessed 11/15/2021, 2021. https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/?utm_source=drip&utm_medium=email&utm_campaign=Machine+Learning+Mastery+Crash+Course&utm_content=Combine+predictions+with+ensemble+methods

28. Lang CE, Waddell KJ, Barth J, Holleran CL, Strube MJ, Bland MD. Upper limb performance in daily life plateaus around three to six weeks post stroke. *Neurorehabilitation and Neural Repair*. 2021;35(10):903-914. doi:10.1177/15459683211041302

29. Kwah LK, Diong J. National Institutes of Health Stroke Scale (NIHSS). *J Physiother*. Mar 2014;60(1):61-61. doi:10.1016/j.jphys.2013.12.012

30. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. *Brain*. May 2012;135:1639-1649. doi:10.1093/brain/awr318

31. Waddell KJ, Tabak RG, Strube MJ, Haire-Joshu D, Lang CE. Belief, confidence, and motivation to use the paretic upper limb in daily life over the first 24 weeks after stroke.

Journal of Neurologic Physical Therapy. 2019;43(4):197-203. doi:doi:10.1097/NPT.0000000000000287

32. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang, C. E. Upper Limb Performance in Daily Life Improves Over the First 12 Weeks Poststroke. *Neurorehabilitation and Neural Repair*. 2019;33(10):836-847.

33. Hoonhorst MH, Nijland RH, van den Berg PJ, Emmelot CH, Kollen BJ, Kwakkel G. Does transcranial magnetic stimulation have an added value to clinical assessment in predicting upper-limb function very early after severe stroke? *Neurorehabilitation and neural repair*. 2018;32(8):682-690.

34. Hsieh C-L, Hsueh IP, Chiang F-M, Lin P-H. Inter-rater reliability and validity of the action research arm test in stroke patients. *Age and ageing*. 1998;27(2):107-113.

35. Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. *Arch Phys Med Rehabil*. Jan 2001;82(1):14-9. doi:10.1053/apmr.2001.18668

36. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A standardized approach to performing the action research arm test. *Neurorehabilitation and Neural Repair*. Jan-Feb 2008;22(1):78-90. doi:10.1177/1545968307305353

37. Fugl-Meyer AR JL, Leyman I, Olsson S, Steglind S. FUGL-MEYER ASSESSMENT UPPER EXTREMITY (FMA-UE)Assessment of sensorimotor function. 1975;

38. Duncan PW, Propst M, Nelson SG. Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. *Physical Therapy*. 1983;63(10):1606-1610. doi:10.1093/ptj/63.10.1606

39. Gillen R, Eberhardt TL, Tennen H, Affleck G, Groszmann Y. Screening for depression in stroke: relationship to rehabilitation efficiency. *Journal of Stroke and Cerebrovascular Diseases*. 1999;8(5):300-306.

40. Hamre BK, Pianta RC. Self-reported depression in nonfamilial caregivers: Prevalence and associations with caregiver behavior in child-care settings. *Early Childhood Research Quarterly*. 2004;19(2):297-318.

41. Mesulam M-M. Dementia: its definition, differential diagnosis, and subtypes. *JAMA*. 1985;253(17):2559-2561.

42. Rengachary J, d'Avossa G, Sapir A, Shulman GL, Corbetta M. Is the Poster Reaction Time Test More Accuracy Than Clinical Tests in Detecting Left Neglect in Acute and Chronic Stroke? *Arch Phys Med Rehabil*. 2009;90(12):2081-2088. doi:https://doi.org/10.1016/j.apmr.2009.07.014

43. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*. 2005;53(4):695-699.

44. Health UoWSoMaP. Area Deprivation Index v2.0. https://www.neighborhoodatlas.medicine.wisc.edu

45. Kind AJH, Buckingham W. Making Neghiborhood Disadvantage Metrics Accessible: The Neghiborhood Atlas. *New England Journal of Medicine*. 2018;378:2456-2458. doi:DOI: 10.1056/NEJMp1802313

46. Maroko AR, Doan TM, Arno PS, Hubel M, Yi S, Viola D. Integrating Social Determinants of Health With Treatment and Prevention: A New Tool to Assess Local Area Deprivation. *Prev Chronic Dis*. Sep 15 2016;13:E128. doi:10.5888/pcd13.160221

47. *A language and environment for statistical computing*. R Foundation for Statistical Computing; 2021. https://intro2r.com/citing-r.html#:~:text=URL%20https%3A//www.R%2Dproject.org/.

48. *tree: Classification and Regression Trees*. 2021. https://CRAN.R-project.org/package=tree

49. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22.

50. *caret: Classification and Regression Training*. Version R package version 6.0-90. 2021. https://CRAN.R-project.org/package=caret

51. Yiu T. Understanding Random Forest: How the algorithm works and why it is so effective. Accessed 1/16/2022, 2022. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

52. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the Upper Limb Prediction Algorithm PREP2 Applied 2 Weeks Poststroke: A Prospective Longitudinal Study. *Neurorehabilitation and Neural Repair*. Jan 2021;35(1):68-78. doi:10.1177/1545968320971763

53. Buch ER, Rizk S, Nicolo P, Cohen LG, Schnider A, Guggisberg AG. Predicting motor improvement after stroke with clinical assessment and diffusion tensor imaging. *Neurology*. May 17 2016;86(20):1924-5. doi:10.1212/WNL.0000000000002675

54. Winters C, van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. *Neurorehabilitation and neural repair*. 2015;29(7):614-622.

55. Gebruers N, Truijen S, Engelborghs S, De Deyn PP. Prediction of upper limb recovery, general disability, and rehabilitation status by activity measurements assessed by

accelerometers or the Fugl-Meyer score in acute stroke. *American Journal of Physical Medicine & Rehabilitation*. 2014;93(3):245-252. doi:10.1097/phm.0000000000000045

56. Byblow WD, Stinear CM, Barber PA, Petoe MA, Ackerley SJ. Proportional recovery after stroke depends on corticomotor integrity. *Annals of neurology*. 2015;78(6):848-859.

57. Chin LF, Hayward KS, Brauer SG. Factors Influencing Paretic Upper Limb Use During First 4 Weeks After Stroke: A Cross-Sectional Accelerometry Study. *American Journal of Physical Medicine & Rehabilitation*. 2021;100(2)

58. Lindberg D, West RT, Corn M. IAIMS: an overview from the National Library of Medicine. *Bulletin of the Medical Library Association*. 1992;80(3):244.

Chapter 5: Summary of Major Findings

5.1 Summary of specific Aims

In Chapter 2, we tested how well an algorithm with clinical measures, developed for use in New Zealand, applied to persons with stroke within the US. Our *first hypothesis* was that UL capacity would be predicted and algorithm accuracy would fall within a range of 70-80% in a cohort of US participants with first ever stroke. Our *second hypothesis* was that that participants with inaccurate predictions would be within one category of their expected at three months. Our first hypothesis was partially supported; the algorithm successfully predicted participants into the four UL capacity categories, however overall accuracy of the algorithm with clinical measures only was 61% (CI: 46%-75%). Our second hypothesis was supported; 39% (19/49) of the total number of participants had inaccurate predictions, 79% (15/19) were one category away from their actual category at three months post stroke. The findings of the present study indicate that use of this algorithm, with clinical measures only, is better than chance alone (chance =25% for each of the four categories) at predicting the category of UL capacity three months after stroke. The moderate to high values of sensitivity, specificity, PPV, and NPV demonstrates some clinical utility of the algorithm within the US healthcare setting.

In Chapter 3, we explored if categories of UL performance, as quantified from accelerometer recordings, could be identified and defined. Our *first hypothesis* was that three categories of UL performance would be identified across a host of accelerometer variables, spanning the possible ranges of UL performance in daily life. Our *second hypothesis* was that the categories that emerged would have clinical meaning of expected UL performance in daily life. Our first hypothesis was refuted; five categories of UL performance were identified from five UL performance variables in cohorts of adults with and without neurological UL deficits. Our second hypothesis was supported; the five categories that emerged have stratified people into

groupings with similar overall UL performance, representing a profile of arm activity in daily life. Individuals within each category have similar ranges on each performance variable included (e.g. duration, magnitude, variability, and symmetry) in each of the five clusters. As such, the names of the five categories were chosen for their overall activity and integration of the ULs into their daily activity, and are named: *Minimal Activity/Rare Integration, Minimal Activity/Limited Integration, Moderate Activity/Moderate Integration, Moderate Activity/Full Integration,* and *High Activity/Full Integration.*

In Chapter 4, we explored if a model could be developed to predict categories of UL performance in daily life at three months post stroke. Our *first hypothesis* was that a model could be derived from a collection of clinical measures to predict a category of UL performance post stroke. Additionally, our *second hypothesis* was that the developed model would predict UL performance with a minimum of 70% accuracy. Our first hypothesis was supported; the categories of UL performance identified in Aim 2 were predicted using different machine learning techniques. Measures of UL impairment and capacity were the most important predictors of the subsequent UL performance category, regardless of the machine learning algorithm used. Other non-motor clinical measures emerged as key predictors, while participant demographic predictors were less important across the models. Our second hypothesis was partially supported; the in-sample accuracy across the seven models ranged from 70% to 100%. The out-of-bag estimate of error, however, ranged from 48% to 55%. The out-of-bag estimate of error is a better assessment of model accuracy with respect to the predictive power because it is calculated from data not used to build the model. These values of the out-of-bag estimate of error can be compared to overall accuracy with the inverse values, such that, the overall accuracy of these models ranged from 42% to 55%.

5.2 Significance of Findings

In the validation and development of novel clinical tools available to stroke rehabilitation providers, we report multiple findings that add to the current scientific body of knowledge.

5.2.1 Importance of quantifying UL performance

Chapters 2-4 highlight the value of measurement of UL recovery after stroke across the ICF domains, because of the differences observed in UL impairment, capacity, and performance. Wearable sensor data has provided evidence to challenge the status quo, because clinicians and researchers have seen that improvements in UL capacity observed in the clinic do not necessarily translate to improvements in UL performance in daily life.[1-6] In Chapter 3, five categories of UL performance were formed from five UL performance variables in persons with and without neurological UL deficits. Each of the five categories differ with respect to the duration, magnitude, variability, and symmetry of UL activity in daily life. In the five category solution, two categories emerged that could be considered to have "normal" UL performance (*Moderate Activity/Full Integration and High Activity/Full Integration)* as indicated by the presence of many neurologically-intact adult controls in those categories. People assigned to these two categories engage both ULs into activities to a similar degree, as evidenced by similar mean values of the use ratio. The use ratio values in both categories approach 1.0, indicating that people in these two categories have relatively equal contributions of both ULs to daily activity. The differences between the two categories (and why they are two separate categories) is that people in these categories have different values on variables measuring the duration, magnitude, and variability of UL movement. Taken together, people in these categories have different levels of overall UL activity, ranging from moderate to high UL activity. People assigned to the *High*

Activity/Full Integration category having the highest values on these variables, while those assigned to the *Moderate Activity/Full Integration* have lower values.

Of the participants with stroke included in this analysis, 20% (27/135) of people ended up in the two categories with the highest overall UL performance. This is a positive finding, as it indicates these individuals have either experienced a full recovery of their ULs following their stroke, or they have figured out how to use their available capacity for activity in order to integrate their paretic limb and be active in daily life.[7] What is highly interesting is that people assigned to these categories do not necessarily have the highest values of scores on measures of UL impairment and capacity. One can see from Table 5.1 that people in these two categories had similar median values of UL capacity (ARAT) at the same time point post stroke time point as when the accelerometer data were collected for the two studies.[8] Interestingly, only 26% (7/27) of the participants with stroke in these two categories achieved a score of 57, the highest possible score on the ARAT. Even more interesting is the wide range of values of UL capacity of people in these performance categories. Stinear et. al.[9, 10] formed categories of UL capacity based on ranges of ARAT scores (see Table 2.1) and only 41% (11/27) of people in these top two performance categories would have been in the highest UL capacity category of Excellent. Of the remaining participants 52% (14/27) would have been in the Good category and 7% (2/27) in the Limited category. This is important because all of these people are using their ULs with similar duration, magnitude, variability, and symmetry as adults without neurological UL deficits. A similar finding emerged when examining the subset of people in the top two categories with impairment measures (N = 20), i.e. SAFE and UEFM scores captured between six and 12 weeks post

stroke. Again, one can see that people assigned to both of these categories displayed similar median values and wide ranges of scores on the UL impairment measures (UEFM and SAFE). These findings taken together contribute to our understanding that UL capacity and impairment measures are not a direct indicator of UL performance after stroke.

Table 5.1. Median (ranges) of UL capacity and impairment measures

Measure name and ICF domain*	Total† N=27 or N=15	Moderate Activity/ Full Integration N=20	High Activity/ Full Integration N=7
UL Capacity			
ARAT	47 (24-57)	47 (33-57)	46 (24-55)
UL Impairment			
UEFM	62 (42-66)	63 (59-66)	61 (42-66)
SAFE	9 (7-10)	10 (8-10)	9 (7-10)

*UL capacity and impairment measures were collected at the same time point used to form the UL performance categories in Aim 2

† Both stroke cohorts completed the UL capacity measure (N=27) and only stroke cohort completed the UL impairment measures (N=15)

5.2.2 Lessons learned

The first lesson learned is about the importance of carefully defining study aims and hypotheses to match the intended purpose of the analysis. Upon reflection, to simply state "develop a prediction model" as written in Aim 3 was too simplistic. We quickly learned that with the plethora of analysis choices, it is certainly possible to develop a model. It is now apparent that one must consider how that model will be used given the available data. For example, we lacked the sample size to split our data into training, validation, and test sets that are typically required of machine learning methods, therefore our results and thus prediction models were exploratory in nature and should be interpreted as a first step to understanding factors associated with UL performance after stroke.[11] The novelty of this Aim was the comparison of the results from three unique machine learning algorithms, the single decision tree, the bagged tree, and the random forest. In this comparison, we realized that our second hypothesis was lacking specificity because the in-sample prediction (e.g. explanation) of the models was well

above 70% however, the out-of-bag accuracy (e.g. predictive power) was much lower. An unexpected finding from this Aim was the identification of which variables measured early after stroke are most important to the five UL performance categories. The in-sample prediction and the above chance (20% for each of the five categories) out-of-bag estimate of error revealed that we in fact included *some* variables important to the UL performance categories, but we failed to build a predictive model with 70% accuracy as our hypothesized benchmark.

The second lesson learned was that there is a lack of detail in the quality of reporting the "whys" and "how's" in stroke rehabilitation prediction studies that is limiting science in the field. Historically, methodological approaches to prediction rely on linear and logistical regression models.[12, 13] These models are by nature inflexible, because they yield temporally fixed predictions from temporally fixed input data.[12, 14, 15] Another limitation of these models is the assumption of linearity between the dependent variable and the independent variables. There is rapid growth with respect to UL recovery the first three months after stroke that may not be appropriately modeled with linear regression techniques.[7, 15-17] For example, a ten-point gain observed on the lower end of an UL impairment or capacity measure may not be the same as a ten-point gain at the higher end of that same scale.[17] Other non-linearity that might not be detectable by linear regression models include how the influence of different non-motor impairments (e.g. visual neglect) impact UL motor recovery.[12, 15] These non-linearities are likely not detectable by linear regression models, regardless of the amount or quality of data available.[17] As in other fields, unfortunately, the majority of the prediction modeling in rehabilitation has employed these analytic techniques.[14, 15, 18-24] To overcome the limitations of regression techniques, others have used the CART algorithm for prediction modeling.[12, 13] This method was used to develop the PREP2 algorithm that we attempted to externally validate in

Aim 1.[9, 25] External validation studies of prediction models are critical because a model's predictions may not be reproducible on samples external to the model.[26-29] Certainly poor predictions of the model in a new data set of persons with stroke could arise from differences between the settings of the new and derivation samples, including differences in health-care systems, methods of measurement, and participant characteristics.[30] However, another source that contributes to poor predictions is due to the CART algorithm itself. To build a single decision tree, the CART algorithm used the entire data set and therefore only the exact values of scores on clinical measures and participant characteristics are considered. Lundquist et.al.[25] replicated PREP2 within two weeks of stroke onset to accommodate differences in the timing of transitions of care between Denmark and New Zealand. The overall accuracy (accuracy: 60%; CI: 50%-71%)[25] was lower than they hypothesized. Similar to our analyses in Aim 1 (Chapter 2), most people in their study with inaccurate predictions were still within one category of their predicted at three months post stroke. Because of the low overall accuracy values, a separate CART analysis was conducted on this sample from Denmark and the same measures selected in PREP2 emerged as important predictors in the decision tree but the values of participant age and stroke severity changed.[25] We noted similar issues with the cut-off values in our sample as well. If we had run a separate CART analysis too, it is likely that the cut-off values would again change, since they are completely dependent on the input data. While using the CART algorithm is appealing because it results in the easy to interpret single decision (e.g. Figure 4.2), the limitations of the CART analytic method itself may not be able to generate a single UL capacity prediction algorithm for use around the world. In Chapter 4, the CART algorithm was one of the machine learning techniques used to predict the categories of UL performance from early clinical measures. Different predictors emerged in the single decision tree that did not emerge when

using the bagging algorithms. For example, the Mesulam was selected in the single decision tree but had little to no improvement in accuracy or node purity of the bagged trees or random forest models (Figure 4.3 and 4.4 in Chapter 4). As prediction models of post stroke outcomes continue to be explored, it will be important for researchers to explore the full range of analytic options available. Additionally, researchers should be transparent when reporting how and why decisions were made to build the prediction model including highlighting the strengths and weaknesses of each analytic choice.

The third lesson learned from this work is that there are important factors that influence UL performance after stroke that are not currently being measured and therefore have not been included in recent prediction efforts. All of the attempts to predict UL performance in daily life after stroke have explained a portion of the variance (48% – 55%), alluding to the fact that there are missing variables in our models and others.[19, 31] In one previous study, a multivariate model that included an UL impairment measure (UEFM), the presence or absence of a motor evoked potential (MEP) measured by transcranial magnetic stimulation (TMS), and the presence or absence of visual neglect within two weeks of stroke accounted for 55% of the variance in a single UL performance variable, the use ratio, three months post stroke.[31] In Chapter 4, models built with the medium sized data set and the bagged tree or random forest algorithm had an out-of-bag estimate of 0.47 and 0.46, respectively when predicting a multivariate category of eventual UL performance. While variance explained is not the same as accuracy these results highlight that there are missing predictors important to UL performance in all of these studies. Other groups have begun exploring how various cognitive constructs (e.g. apraxia, spatial neglect) may contribute to UL performance after stroke.[32] Our data sets did not include any thorough cognitive assessments, nor did it include thorough social or emotional assessments –

which also may impact UL performance.[7] For example, we did not have access to a measure of apraxia, characterized by spatiotemporal deficits in imitation, pantomime of tool use movements, and/or tool use, even with the non-paretic UL.[32, 33] It is possible that the presence and severity of apraxia and other cognitive and emotional impairments could substantially limit UL performance in daily life.

The fourth lesson learned was that we do not yet know whether the same factors that influence daily performance of walking or other activities influence UL performance after stroke. Despite recent efforts, UL performance after stroke is a relatively unexplored area and it is undetermined if the factors that affect lower limb performance are similar to those affecting the UL. With respect to lower limb performance, measures of walking capacity, [34-39] depressive symptomology,[40-42] the environment,[34, 35, 43] participant demographics,[43-45] and biopsychosocial factors[35, 43, 46, 47] contribute to walking performance after stroke in varying degrees. Age and other demographic factors did not emerge as important predictors in Chapter 4, whereas they have in models predicting UL capacity.[9, 19, 22] In the present analysis, superficial measures of the physical and social (e.g. ADI, pre- and post-stroke living status) environment were included as demographic predictors, but neither were important to the accuracy or node purity of the models. It is plausible that other aspects of the social environment, such as marital status or employment status may influence UL performance, as they have been shown to influence walking performance after stroke.[35, 43] Balance self-efficacy is a biopsychosocial factor important to walking performance after stroke[43, 48-51] and in the prospective, observational, longitudinal cohort study (data used in Chapter 3 and Chapter 4) a survey was administered to quantify a persons' belief, confidence, and motivation to use the paretic UL in daily life.[7] We did not include these data in the analysis in Chapter 4 because of the high, unwavering levels of these three

psychosocial constructs.[52] Perhaps the influence of belief, confidence, and motivation is different for UL activity than for walking activity because of the low risk of using the paretic UL in everyday life. There is a much higher risk associated with poor walking performance (e.g. falls, fracture) and therefore balance self-efficacy may matter more in what people choose to do every day.[35]

5.3 Future directions

In Chapter 2, the algorithm with clinical measures predicted a category of UL capacity with lower overall accuracy than hypothesized and people with inaccurate predictions were typically within one category of their predicted category at three months post stroke. There are a few possible solutions that could be explored to combat the limitations of the systematic differences in the healthcare systems, differences in stroke rehabilitation populations, and limitations of the algorithms used to develop prediction models. First, it could be possible to pool data from persons with stroke across the world, including clinical measures and participant demographic information along with the geographical location and time post stroke.[53] A new model predicting UL capacity could be created using the same CART algorithm with pruning and cross validation to create a model that could potentially be used around the world. Others have advocated for the importance of including the neurophysiologic assessment of the integrity of the corticospinal tract (via Transcranial Magnetic Stimulation) especially for people with less initial strength in their paretic UL.[25, 54, 55] Unfortunately, access to this test in the US and other countries is not possible outside of academic medical centers.[54, 56] Therefore it may be necessary to develop two models, one that uses this assessment for use in countries that have access to this test early after stroke (e.g. New Zealand) and another for countries that do not have access. Repeated model building and testing could then yield models with similar predictive and

exploratory power suitable for use in a large portion of the world.[25] Future studies should also investigate using machine learning ensemble methods, such as bagged trees and random forests, to compare the important variables that emerge from single decision trees predicting UL capacity to the variable importance measures available from these methods. A second solution would be to provide the software capabilities to medical centers (by state, region, nation, or country) and provide education and training to create a unique version of prediction models of UL capacity that would be realistic to use within each geographical area. This solution would likely produce the most accurate prediction models, however it is unrealistic that these efforts could be implemented on such a large scale. Finally, the rehabilitation field must come together to educate clinicians and researchers of the strengths and limitations of these models, and to offer resources on how to use these algorithms can be used in other countries. Suggestions could include conducting a reassessment at various time points within the first three months post stroke to update predictions and interventions to maximize a person's full potential for recovery of UL capacity post stroke.[14]

In Chapter 3, five categories of UL performance were formed from five UL performance variables in people with and without neurological UL impairments. Future studies could investigate if the same number of input performance variables and the same number of categories emerge when including larger, more heterogeneous data from people with and without other UL problems. Wearable sensors have provided the ability to measure UL performance in daily life, however there has been limited progress to integrate this data into standard rehabilitation research and therapies.[57] One reason is because currently a large number of single variables have been proposed in various research studies in an attempt to capture similar or related constructs of UL movement in daily life.[58] These variables can be mathematically complex and are

challenging to interpret from a clinical perspective. Of these single variables, validation efforts are scarce and even less work has been done to determine if these variables are sensitive to change.[59] Another limitation is that the variables that have been proposed to date tend to align with only one specific patient population, with much of the development occurring in the stroke rehabilitation population. For example, the use ratio quantifies the relative time, in hours, that one UL is active compared to the other. This single variable is most relevant to rehabilitation populations with asymmetric effects,[60, 61] such as stroke, hemiparetic cerebral palsy, and limb amputation/prosthetic use. There are many other clinical populations that present with UL problems that would benefit from the ability to measure UL performance in daily life, but do not have asymmetrical deficits. Given these limitations it seems unrealistic to continue proposing/developing, validating, and determining sensitivity to change of single variables across clinical populations with UL problems. The categories that were formed in Chapter 3 have the potential to streamline implementation of wearable sensor data into routine rehabilitation care. They could be a tool for clinicians to measure UL performance in daily life. UL performance is a complex construct that is likely multi-dimensional [58, 62, 63] and therefore may not be well-represented by a single performance variable.[62, 63] The five categories from Aim 2 were formed from five UL performance variables that measure aspects of the duration, magnitude, variability, and symmetry of UL activity in daily life. As a result, the categories of UL performance capture the multi-dimensional complexities of UL activity in daily life of people with and without neurological UL deficits. UL disability stems from a range of biological causes including neurological (e.g. stroke, multiple sclerosis), musculoskeletal (e.g. UL fractures, adhesive capsulitis), and other medical conditions (e.g. post breast cancer treatment). People with and without UL problems generally need to complete similar UL activities to

function in their daily life. Because of these reasons, it seems realistic that these categories could be validated across multiple rehabilitation populations with health conditions that cause UL disability. Additional studies will be required to determine what constitutes a meaningful change in UL performance, either of multivariate categories or of single variables. There have been some efforts to identify the minimal clinically important difference (MCID) of a single variable from wearable sensors (e.g. activity counts),[64, 65] but much more work is needed. An efficient approach might be to validate the categories first and then determine what a meaningful change in UL performance categories would be. Categories could provide more accessible information to people with health conditions and their health providers compared to single variables. In other biomedical science fields, formation of categories which encompass multi-dimensional measures have facilitated research and clinical decision making for persons with health conditions such as, spinal cord injury,[66] heart failure,[67, 68] and chronic obstructive pulmonary disease (COPD).[69] Categories also are used for general physical activity levels, such as, "sedentary" through "highly active".[70-72] With the physical activity categories, individuals assigned (by a clinician or via consumer-grade wearable sensors) to any one category have a variety of biological conditions, impairments, capacity limitations, and personal and environmental factors, but the category helps set goals to improve physical activity.

Use of validated multi-dimensional UL performance categories have the potential to ease the burden on rehabilitation clinicians that work with clients with UL problems of various causes. Typically rehabilitation professionals do not see clients with UL problems due to a single condition. Instead, a clinician's case load may include people with stroke, spinal cord injury, cardiac problems, and amputation.[73, 74] People with health conditions generally access services when they have difficulty performing activities in daily life.[75, 76] Regardless of the cause of a

person's UL problems, self-identified rehabilitation goals are almost always directed at improving performance of activities in daily life.[77] UL performance categories, therefore, have the potential to measure the outcome of interest for clinicians and people with health conditions that cause UL problems. Once the categories are validated, future studies could involve stakeholder engagement of rehabilitation providers and people with health conditions that impact the ULs to understand the barriers and facilitators to using categories to convey UL performance information. A future option could be to offer a user-friendly, software package to rehabilitation clinicians that would calculate the UL performance variables from data extracted from wrist-worn accelerometers. Based on a person's values across the variables, a category of UL performance could be determined and used to communicate current UL performance and used to set goals for future UL performance. Based on the aspects of movement (duration, magnitude, variability, and symmetry) selected to form the categories, it is possible that these categories could be highly relevant for many clinical conditions affecting UL performance in daily life, not just those with stroke.

In Chapter 4, we used machine learning techniques to predict the categories of UL performance from clinical measures and participant demographic information. We identified some factors important to UL performance after stroke, but there is still a large portion of unexplained variance. In this analysis, we lacked the capability to partition the data into training, validation, and testing sets because of the limited sample size. Our results are a first attempt at exploring the relationship of early clinical measures and participant demographics to the subsequent UL performance categories. Future studies with larger data sets should continue to explore what other factors could influence UL performance in daily life. We suspect that other personal factors will need to be explored, such as employment status (pre- and post- stroke),

marital status, and other social supports. It has also been suggested that UL performance is influenced by both intrinsic and extrinsic sources of motivation.[15, 78, 79] Intrinsic motivation has led to better long-term behavioral outcomes for medication adherence,[80] weight loss maintenance,[81] and persistence and enhanced subjective well-being.[82] A serious challenge is that measuring intrinsic motivation is difficult, and new measures may need to be developed to understand the influence of intrinsic motivation on UL recovery after stroke. Other personal factors that should be explored to understand their impact on UL performance after stroke include self-regulation,[82-84] outcome expectation, [84] and perceived control.[83]

5.4 Conclusion

This dissertation work has conducted an external validation of an existing prediction model, developed categories of UL performance in people with and without neurological UL deficits, and explored how early clinical measures and participant demographic information were associated with subsequent categories of UL performance after stroke. Our findings provide strong support for the importance of measuring recovery of the UL across ICF domains, not just with impairment and capacity level measures. The UL performance categories formed in Chapter 3 offer the future opportunity to transition measurement of UL performance in daily life for person's receiving UL rehabilitation away from the current confines of rehabilitation research labs and into standard of care. In Chapter 4, different machine learning techniques were explored to predict the categories of UL performance from information routinely captured after the onset of a stroke. Collectively this work provides preliminary measurement tools that could eventually be available to rehabilitation clinicians following subsequent validation efforts. Additionally, this work provides a rich exploration into the strengths, weaknesses, and limitations of analytical methods and their impact on validation efforts.

5.5 References

1. Waddell KJ, Strube MJ, Bailey RR, Klaesner JW, Birkenmeier RL, Dromerick AW et al. Does Task-Specific Training Improve Upper Limb Performance in Daily Life Poststroke? Neurorehabil Neural Repair 2017;31(3):290-300.

2. Doman CA, Waddell KJ, Bailey RR, Moore JL, Lang CE. Changes in Upper-Extremity Functional Capacity and Daily Performance During Outpatient Occupational Therapy for People With Stroke. Am J Occup Ther 2016;70(3):7003290040p1-p11.

3. Rand D, Eng JJ. Disparity between functional recovery and daily use of the upper and lower extremities during subacute stroke rehabilitation. Neurorehabil Neural Repair 2012;26(1):76-84.

4. Andrews K, Stewart J. Stroke recovery: he can but does he? Rheumatol Rehabil 1979;18(1):43-8.

5. Hidaka Y, Han CE, Wolf SL, Winstein CJ, Schweighofer N. Use it and improve it or lose it: interactions between arm function and use in humans post-stroke. PLoS Comput Biol 2012;8(2):e1002343.

6. Schweighofer N, Han CE, Wolf SL, Arbib MA, Winstein CJ. A functional threshold for long-term use of hand and arm function can be determined: predictions from a computational model and supporting data from the Extremity Constraint-Induced Therapy Evaluation (EXCITE) Trial. Phys Ther 2009;89(12):1327-36.

7. Lang CE, Waddell KJ, Barth J, Holleran CL, Strube MJ, Bland MD. Upper limb performance in daily life plateaus around three to six weeks post stroke. Neurorehabilitation and Neural Repair 2021;35(10):903-14.

8. Barth J, Lohse KR, Konrad JD, Bland MD, Lang CE. Sensor based categorization of upper limb perforance in daily life of persons with and without neurological upper limb deficits. Frontiers in Rehabilitation Science 2021;2.

9. Stinear CM, Byblow WD, Ackerley SJ, Smith MC, Borges VM, Barber PA. PREP2: A biomarker-based algorithm for predicting upper limb function after stroke. Ann Clin Transl Neurol 2017;4(11):811-20.

10. Smith MC, Ackerley SJ, Barber PA, Byblow WD, Stinear CM. PREP2 algorithm predictions are correct at 2 years poststroke for most patients. Neurorehabilitation and neural repair 2019;33(8):635-42.

11. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2016.

12. Walford I, Rondina JM, Ward N. Patient-specific prediction of long-term outcomes will change stroke rehabilitation for the better. J Neurol Neurosur Ps 2021;92(6):572-.

13. Rosso C, Lamy JC. Prediction of motor recovery after stroke: being pragmatic or innovative? Current Opinion in Neurology 2020;33(4):482-7.

14. Selles R, Andrinopoulou E-R, Nijland R, van der Vliet R, Slaman J, van Wegen E et al. Computerised patient-specific prediction of the recovery profile of upper limb capacity within stroke services: the next step. Journal of Neurology, Neurosurgery & Psychiatry 2021.

15. Hidaka Y, Han CE, Wolf SL, Winstein CJ, Schweighofer N. Use It and Improve It or Lose It: Interactions between Arm Function and Use in Humans Post-stroke. Plos Computational Biology 2012;8(2).

16. Dromerick AW, Geed S, Barth J, Brady K, Giannetti ML, Mitchell A et al. Critical Period After Stroke Study (CPASS): A phase II clinical trial testing an optimal time for motor recovery after stroke in humans. Proc Natl Acad Sci U S A 2021;118(39).

17. Bonkhoff AK, Hope T, Bzdok D, Guggisberg AG, Hawe RL, Dukelow SP et al. Bringing proportional recovery into proportion: Bayesian modelling of post-stroke motor impairment. Brain 2020;143(7):2189-206.

18. Bonkhoff AK, Hope T, Bzdok D, Guggisberg AG, Hawe RL, Dukelow SP et al. Recovery after stroke: the severely impaired are a distinct group. J Neurol Neurosurg Psychiatry 2021.

19. Rand D, Eng JJ. Predicting daily use of the affected upper extremity 1 year after stroke. Journal of Stroke and Cerebrovascular Diseases 2015;24(2):274-83.

20. van der Vliet R, Selles RW, Andrinopoulou ER, Nijland R, Ribbers GM, Frens MA et al. Predicting Upper Limb Motor Impairment Recovery after Stroke: A Mixture Model. Annals of Neurology 2020;87(3):383-93.

21. Winters C, van Wegen EE, Daffertshofer A, Kwakkel G. Generalizability of the proportional recovery model for the upper extremity after an ischemic stroke. Neurorehabilitation and neural repair 2015;29(7):614-22.

22. Lundquist CB, Nielsen JF, Brunner IC. Prediction of Upper Limb use Three Months after Stroke: A Prospective Longitudinal Study. Journal of Stroke and Cerebrovascular Diseases 2021;30(11):106025.

23. Gebruers N, Truijen S, Engelborghs S, De Deyn PP. Predictive value of upper-limb accelerometry in acute stroke with hemiparesis. Journal of Rehabilitation Research and Development 2013;50(8):1099-106.

24. Gebruers N, Truijen S, Engelborghs S, De Deyn PP. Prediction of upper limb recovery, general disability, and rehabilitation status by activity measurements assessed by accelerometers or the Fugl-Meyer score in acute stroke. American Journal of Physical Medicine & Rehabilitation 2014;93(3):245-52.

25. Lundquist CB, Nielsen JF, Arguissain FG, Brunner IC. Accuracy of the Upper Limb Prediction Algorithm PREP2 Applied 2 Weeks Poststroke: A Prospective Longitudinal Study. Neurorehabilitation and Neural Repair 2021;35(1):68-78.

26. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2(3):18-22.

27. Brownlee J. Bagging and Random Forest Ensemble Algorithms for Machine Learning 2016 [cited 2021 11/15/2021]. Available from: URL: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/?utm_source=drip&utm_medium=email&utm_campaign=Machine+Learning+Mastery+Crash+Course&utm_content=Combine+predictions+with+ensemble+methods.

28. Yiu T. Understanding Random Forest: How the algorithm works and why it is so effective 2019 [cited 2022 1/16/2022]. Available from: URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

29. Kim JK, Choo YJ, Chang MC. Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: Development of Practical Models. Journal of Stroke and Cerebrovascular Diseases 2021;30(8):105856.

30. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

31. Lundquist CB, Nielsen JF, Brunner IC. Prediction of Upper Limb use Three Months after Stroke: A Prospective Longitudinal Study. J Stroke Cerebrovasc 2021;30(11).

32. Buxbaum LJ, Varghese R, Stoll H, Winstein CJ. Predictors of arm non-use in chronic stroke: a preliminary investigation. Neurorehabil Neural Repair 2020;34(6):512-22.

33. Sundet K, Finset A, Reinvang I. Neuropsychological Predictors in Stroke Rehabilitation. J Clin Exp Neuropsyc 1988;10(4):363-79.

34. Thilarajah S, Mentiplay BF, Bower KJ, Tan D, Pua YH, Williams G et al. Factors Associated With Post-Stroke Physical Activity: A Systematic Review and Meta-Analysis. Arch Phys Med Rehabil 2018;99(9):1876-89.

35. Miller A, Pohlig RT, Wright T, Kim HE, Reisman DS. Beyond Physical Capacity: Factors Associated With Real-world Walking Activity After Stroke. Arch Phys Med Rehabil 2021;102(10):1880-7 e1.

36. Awad L, Reisman D, Binder-Macleod S. Distance-Induced Changes in Walking Speed After Stroke: Relationship to Community Walking Activity. Journal of neurologic physical therapy : JNPT 2019;43(4):220-3.

37. Fulk GD, He Y, Boyne P, Dunning K. Predicting Home and Community Walking Activity Poststroke. Stroke 2017;48(2):406-11.

38. Fulk GD, Reynolds C, Mondal S, Deutsch JE. Predicting Home and Community Walking Activity in People With Stroke. Archives of Physical Medicine and Rehabilitation 2010;91(10):1582-6.

39. Michael KM, Allen JK, Macko RF. Reduced Ambulatory Activity After Stroke: The Role of Balance, Gait, and Cardiovascular Fitness. Archives of Physical Medicine and Rehabilitation 2005;86(8):1552-6.

40. Ezekiel L, Collett J, Mayo NE, Pang L, Field L, Dawes H. Factors Associated With Participation in Life Situations for Adults With Stroke: A Systematic Review. Archives of Physical Medicine and Rehabilitation 2019;100(5):945-55.

41. Hackett ML, Yapa C, Parag V, Anderson CS. Frequency of depression after stroke: a systematic review of observational studies. Stroke 2005;36(6):1330-40.

42. French MA, Miller A, Pohlig RT, Reisman DS. Depressive Symptoms Moderate the Relationship Among Physical Capacity, Balance Self-Efficacy, and Participation in People After Stroke. Physical Therapy 2021;101(12).

43. Miller A, Pohlig RT, Reisman DS. Social and physical environmental factors in daily stepping activity in those with chronic stroke. Topics in stroke rehabilitation 2021;28(3):161-9.

44. Zhang L, Yan T, You L, Li K. Barriers to Activity and Participation for Stroke Survivors in Rural China. Archives of Physical Medicine and Rehabilitation 2015;96(7):1222-8.

45. Simonsick EM, Guralnik JM, Fried LP. Who Walks? Factors Associated with Walking Behavior in Disabled Older Women With and Without Self-Reported Walking Difficulty. Journal of the American Geriatrics Society 1999;47(6):672-80.

46. French MA, Moore MF, Pohlig R, Reisman D. Self-efficacy Mediates the Relationship between Balance/Walking Performance, Activity, and Participation after Stroke. Topics in stroke rehabilitation 2016;23(2):77-83.

47. Salbach NM, Mayo NE, Robichaud-Ekstrand S, Hanley JA, Richards CL, Wood-Dauphinee S. Balance self-efficacy and its relevance to physical function and perceived health status after stroke. Archives of physical medicine and rehabilitation 2006;87(3):364-70.

48. Bailey RR. Examining daily physical activity in community-dwelling adults with stroke using social cognitive theory: an exploratory, qualitative study. Disability and Rehabilitation 2020;42(18):2631-9.

49. Miller A, Wright T, Wright H, Thompson E, Pohlig RT, Reisman DS. Readiness to Change is Related to Real-World Walking and Depressive Symptoms in Chronic Stroke. Journal of Neurologic Physical Therapy 2021;45(1):28-35.

50. Danks KA, Pohlig RT, Roos M, Wright TR, Reisman DS. Relationship between walking capacity, biopsychosocial factors, self-efficacy, and walking activity in persons poststroke. Journal of Neurologic Physical Therapy 2016;40(4):232-8.

51. Durcan S, Flavin E, Horgan F. Factors associated with community ambulation in chronic stroke. Disability and Rehabilitation 2016;38(3):245-9.

52. Waddell KJ, Tabak RG, Strube MJ, Haire-Joshu D, Lang CE. Belief, confidence, and motivation to use the paretic upper limb in daily life over the first 24 weeks after stroke. Journal of Neurologic Physical Therapy 2019;43(4):197-203.

53. Winstein C, Varghese R. Been there, done that, so what's next for arm and hand rehabilitation in stroke? Neurorehabilitation 2018;43(1):3-18.

54. Hoonhorst MH, Nijland RH, van den Berg PJ, Emmelot CH, Kollen BJ, Kwakkel G. Does transcranial magnetic stimulation have an added value to clinical assessment in predicting upper-limb function very early after severe stroke? Neurorehabilitation and neural repair 2018;32(8):682-90.

55. Stinear CM, Smith M-C, Byblow WD. Prediction Tools for Stroke Rehabilitation. Stroke 2019;50(11):3314-22.

56. Connell LA, Smith MC, Byblow WD, Stinear CM. Implementing biomarkers to predict motor recovery after stroke. NeuroRehabilitation 2018;43:41-50.

57. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field. Sensors 2020;20(20).

58. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of Wearable Sensing Technology for Movement: Pushing Forward into the Routine Physical Rehabilitation Care Field. Sensors (Basel) 2020;20(20).

59. Lang CE, Cade WT. A step toward the future of seamless measurement with wearable sensors in pediatric populations with neuromuscular diseases. Muscle Nerve 2020;61(3):265-7.

60. Uswatte G, Miltner WH, Foo B, Varma M, Moran S, Taub E. Objective measurement of functional upper-extremity movement using accelerometer recordings transformed with a threshold filter. Stroke 2000;31(3):662-7.

61. Lang CE, Bland MD, Bailey RR, Schaefer SY, Birkenmeier RL. Assessment of upper extremity impairment, function, and activity after stroke: foundations for clinical decision making. J Hand Ther 2013;26(2):104-15.

62. Smith BA, Lang CE. Sensor Measures of Symmetry Quantify Upper Limb Movement in the Natural Environment Across the Lifespan. Arch Phys Med Rehabil 2019;100(6):1176-83.

63. David A, Subash T, Varadhan SKM, Melendez-Calderon A, Balasubramanian S. A Framework for Sensor-Based Assessment of Upper-Limb Functioning in Hemiparesis. Front Hum Neurosci 2021;15:667509.

64. Li YC, Liao WW, Hsieh YW, Lin KC, Chen CL. Predictors of Clinically Important Changes in Actual and Perceived Functional Arm Use of the Affected Upper Limb After Rehabilitative Therapy in Chronic Stroke. Archives of Physical Medicine and Rehabilitation 2020;101(3):442-9.

65. Chen HL, Lin KC, Hsieh YW, Wu CY, Liing RJ, Chen CL. A study of predictive validity, responsiveness, and minimal clinically important difference of arm accelerometer in real-world activity of patients with chronic stroke. Clinical Rehabilitation 2018;32(1):75-83.

66. Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A et al. International standards for neurological classification of spinal cord injury (revised 2011). The journal of spinal cord medicine 2011;34(6):535-46.

67. Ovbiagele B, Goldstein LB, Higashida RT, Howard VJ, Johnston SC, Khavjou OA et al. Forecasting the future of stroke in the United States: a policy statement from the American Heart Association and American Stroke Association. Stroke 2013;44(8):2361-75.

68. Schlotter F, Orban M, Rommel K-P, Besler C, von Roeder M, Braun D et al. Aetiology-based clinical scenarios predict outcomes of transcatheter edge-to-edge tricuspid valve repair of functional tricuspid regurgitation. Eur J Heart Fail 2019;21(9):1117-25.

69. Fromer L, Cooper CB. A review of the GOLD guidelines for the diagnosis and treatment of patients with COPD. Int J Clin Pract 2008;62(8):1219-36.

70. Tudor-Locke C, Bassett DR. How many steps / day are enough? Preliminary pedometer indices for public health. . Sports Med 2004;34:1-8.

71. Tudor-Locke C, Hatano Y, Pangrazi RP, Kang M. Revisiting "how many steps are enough?". Med Sci Sports Exerc 2008;40(suppl 7):S537-S43.

72. Schmidt MD, Cleland VJ, Shaw K, Dwyer T, Venn AJ. Cardiometabolic rish in younger and older adults across an index of ambulatory activity. Am J Prev Med 2009;37:278-84.

73. Buchanan JL, Rumpel JD, Hoenig H. Charges for outpatient rehabilitation: Growth and differences in provider types. Archives of Physical Medicine and Rehabilitation 1996;77(4):320-8.

74. Dixon MG, Lutfy C. Outcomes Among Patients Referred to Outpatient Rehabilitation Clinics After COVID-19 diagnosis - United States, January 2020-March 2021 (vol 70, pg 967, 2021). Mmwr-Morbid Mortal W 2021;70(33):1126-.

75. Budtz CR, Mose S, Christiansen DH. Socio-demographic, clinical and psychological predictors of healthcare utilization among patients with musculoskeletal disorders: a prospective cohort study. BMC Health Serv Res 2020;20(1):239.

76. Lentz TA, Beneciuk JM, George SZ. Prediction of healthcare utilization following an episode of physical therapy for musculoskeletal pain. BMC Health Serv Res 2018;18(1):648.

77. Waddell KJ, Birkenmeier RL, Bland MD, Lang CE. An exploratory analysis of the self-reported goals of individuals with chronic upper-extremity paresis following stroke. Disabil Rehabil 2016;38(9):853-7.

78. Chin LF, Hayward KS, Brauer SG. Factors Influencing Paretic Upper Limb Use During First 4 Weeks After Stroke: A Cross-Sectional Accelerometry Study. American Journal of Physical Medicine & Rehabilitation 2021;100(2).

79. Janssen H, Bird ML, Luker J, McCluskey A, Blennerhassett J, Ada L et al. Stroke survivors' perceptions of the factors that influence engagement in activity outside dedicated therapy sessions in a rehabilitation unit: A qualitative study. Clinical Rehabilitation 2022;0(0):1-9.

80. Williams GC, Rodin GC, Ryan RM, Grolnick WS, Deci EL. Autonomous regulation and long-term medication adherence in adult outpatients. Health Psychology 1998;17(3):269-76.

81. Williams GC, Grow VM, Freedman ZR, Ryan RM, Deci EL. Motivational predictors of weight loss and weight-loss maintenance. Journal of Personality and Social Psychology 1996;70(1):115-26.

82. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68-78.

83. Bandura A. Social cognitive theory: An agentic perspective. Annu Rev Psychol 2001;52:1-26.

84. Bandura A. Social Cognitive Theory of Self-Regulation. Organizational Behavior and Human Decision Processes 1991;50(2):248-87.

Appendix A

The table below is the Supplemental table of the loadings of the UL performance variables included in the final solution from Chapter 3. A total of two principal components explained the most variance in the UL performance variables. Interestingly, as more UL performance variables were included less total variance was explained.

Supplementary Table 3.1. Loadings on principal component one and two of the five UL performance variables included in the final solution

Variable Name	PC1 Loadings	PC2 Loadings
Paretic/ND Hrs	0.48	-0.31
Non-paretic/D Hrs	0.34	-0.78
Median acceleration paretic/ND (counts)*	0.45	0.39
Acceleration variability of paretic/ND (counts)*	0.48	0.22
Use Ratio	0.47	0.30
Total Variance Explained	76.4%	17.6%

Appendix B

This study fulfilled the degree requirement for the Masters of Science in Clinical Investigation (MSCI) degree. These results may be of interest to the committee because they highlight that while accelerometry is a tool that, while measuring quantity of movement, can also reflect the use of general compensatory movement patterns of the upper limb in persons with chronic stroke.

This study has been published:

Barth J, Klaesner JW, Lang CE. Relationships between accelerometry and general compensatory movements of the upper limb after stroke. *Journal of NeuroEngineering and Rehabilitation*. October 2020; 17(138): 1-10.

B.1 Abstract

Background: Standardized assessments are used in rehabilitation clinics after stroke to measure restoration versus compensatory movements of the upper limb. Accelerometry is an emerging tool that can bridge the gap between in- and out-of-clinic assessments of the upper limb, but is limited in that it currently does not capture the quality of a person's movement, an important concept to assess compensation versus restoration. The purpose of this analysis was to characterize how accelerometer variables reflect upper limb compensatory movement patterns after stroke.

Methods: This study was a secondary analysis of an existing data set from a Phase II, single-blind, randomized, parallel dose-response trial (NCT0114369). Sources of data utilized were: 1) a compensatory movement score derived from video analysis of the Action Research Arm Test (ARAT), and 2) calculated accelerometer variables (quantifying time, magnitude and variability of upper limb movement) from the same time point during study participation for both in-clinic and out-of-clinic recording periods.

Results: Participants had chronic upper limb paresis of mild to moderate severity. Compensatory movement scores varied across the sample, with a mean of 73.7 $\pm$ 33.6 and range from 11.5 to 188. Moderate correlations were observed between the compensatory movement score and each accelerometer variable. Accelerometer variables measured out-of-clinic had stronger relationships with compensatory movements, compared with accelerometer variables in-clinic. Variables quantifying time, magnitude, and variability of upper limb movement out-of-clinic had relationships to the compensatory movement score.

Conclusions: Accelerometry is a tool that incorporates aspects of both quantity of upper limb movement and general compensatory movement patterns of the upper limb in persons with chronic stroke. Individuals who move their limbs more in daily life with respect to time and variability tend to move with less movement compensations and more typical movement patterns. Likewise, individuals who move their paretic limbs less and their non-paretic limb more in daily life tend to move with more movement compensations at all joints in the paretic limb and less typical movement patterns.

B.2 Introduction

As advances in medicine persist, more people are surviving a stroke. Over 80% of those affected will have persistent hemiparesis of their upper limb.[1] These people will be left with chronic disability when trying to complete their activities of daily living (ADL), and an even larger number will not resume their normal daily activities completed prior to stroke.[2] At this time, physical and occupational therapy is the only option available to improve upper limb use after stroke. The ultimate goal of these therapies is to restore the use of the upper limb to the same level it was used before the stroke. Most individuals, however, only partially regain function of their upper limb requiring compensations of the upper limb to complete daily tasks. The differentiation between restoration of upper limb movement and compensation is an area of high interest in stroke rehabilitation.[3] Compensatory movements can be thought of on multiple levels, including a change in behavior (e.g. completion of an activity by a spouse rather than the individual) and a change in context (e.g. using a built up spoon for self- feeding). For the purposes of this paper, compensatory movements will refer to completion of the same movement but with an alternative movement pattern. Specifically, this level of compensatory movements typically describe accessory movements of the head, trunk and upper limb that an individual

incorporates in order to accomplish tasks. A simple example is that if an individual lacks shoulder flexion, or the ability to raise their arm in front of them, the individual lifts their arm by raising it more to the side and bending forward with the trunk.[4,5] Many in the neurorehabilitation field view compensation and restoration as a dichotomy, where individuals will either be classified as using compensatory movement patterns or restored movement patterns. Return of upper limb function may be better conceptualized as a gradient, with individuals having degrees of compensatory movement patterns.[6]

Currently, many in-clinic standardized assessments have some aspects that measure use of compensatory movement patterns. For example, the Reaching Performance Scale specifically assesses compensatory movements of the upper limb during reaching in people with hemiparesis.[7] The Wolf motor function tests, functional ability scale reduces scores if movement compensations were observed during item completion.[8] The Fugl-Meyer arm motor scale, an impairment scale, focused on movement patterns, takes points off where specific compensatory movements are observed on each item.[9] Additionally the Action Research Arm Test (ARAT) scores individuals completing functional reaching tasks with consideration of the quality of the reach and grasp pattern along with the fluidity or precision of the task.[10,11] Standardized assessments have the ability to measure upper limb functional capacity and compensatory movements of the upper limb after stroke, however these assessments only capture one piece of upper limb recovery after stroke.

The current gold standard in the field to measure quality of movement or compensatory movements is through the use of 3D kinematics.[12] Kinematics provides the most detailed assessment of how an individual moves after stroke. It is not realistic, however, to use kinematics in the clinic for all patients due to cost of equipment, time required to test, and

training of personnel. This leaves standardized assessments to be the alternative and most accessible measure of compensatory movement patterns. This gap in measurement has lead our lab to question how we might utilize our existing accelerometry methodology to capture some of these changes in compensatory movement.

In-clinic assessments are limited in that they measure the individual's ability to use the limb in a standardized, structured setting, leaving the individuals actual activity of the limb during daily life unaccounted for. Over the past five years, methodology has been developed to measure upper limb activity in daily life using wearable sensors (accelerometers).[13,14] Accelerometry can quantify how much and how often a person uses their affected limb during their daily life, bridging the gap between in and out-of-clinic assessment. Current accelerometer metrics quantify time, magnitude and variability of movement of the upper limb.[15-19] A limitation of current accelerometry methods is that they quantify the amount of movement, but do not capture the quality of a person's movement, an important concept to assess compensation versus restoration.

The purpose of this secondary analysis was to characterize how accelerometer variables reflect upper limb compensatory movement patterns after stroke. Relationships between compensatory movement patterns and accelerometer variables were calculated for both in-clinic and out-of-clinic time points. Both time points were included as the in-clinic time includes completion of standardized assessments and participation in an intensive upper limb therapy protocol. Due to the nature of the therapy protocol, we anticipated there may be different relationships because during the in-clinic time participants are intentionally training their affected limb. The out-of-clinic recordings captures the individual in their free-living environment, providing a more realistic picture of how the individual uses their upper limb in

daily life. It is hypothesized that quantitative metrics from accelerometers both in and out-of-clinic will have moderate associations with compensatory movement patterns of the upper limb.

B.3 Methods

This study was a secondary analysis of an existing data set from a Phase II, single-blind, randomized, parallel dose-response trial (NCT0114369).[20] Sources of data utilized were: 1) a compensatory movement score derived from video analysis of the Action Research Arm Test (ARAT), and 2) calculated accelerometer variables from the same time point during study participation.

B.3.1 Participants

Inclusion criteria were (1) ischemic or hemorrhagic stroke as determined by neurologist and consistent with neuroimaging; (2) time since stroke $\geq$ 6 months; (3) cognitive skills to actively participate, as indicated by scores of 0 to 1 on items 1b and 1c of the National Institutes of Health Stroke Scale (NIHSS); (4) unilateral upper limb weakness, as indicated by a score of 1 to 3 on items 5 (arm item) on the NIHSS; and (5) mild-to-moderate functional motor capacity of the paretic upper limb, as indicated by a score of 10 to 48 on the ARAT. [10,11] Exclusion criteria were (1) participant unavailable for 2 month follow-up (2) inability to follow-2-step commands; (3) psychiatric diagnoses; (4) current participation in other UL stroke treatments (ex/Botox); (5) other neurological diagnoses; (6) participants living further than 1 hour away or were unwilling to travel for assessments and treatment sessions; and (7) pregnancy. The clinical trial was approved by the Washington University Human Research Protection Office and all subjects provided informed consent prior to trial participation.

B.3.2 Compensatory movement score

A compensatory movement score was derived from video recordings of available baseline or subsequent ARATs. We first developed a checklist to quantify the degree of movement compensations of the upper limb. Compensatory movement information was synthesized from 9 standardized assessments of the upper limb measuring quality of movement or compensatory movement patterns. [7,8,10,21-26] Descriptions of compensatory movement patterns of the upper limb were extracted from the assessments and organized to generate the list of items on the checklist. The checklist was piloted and refined following feedback from licensed physical and occupational therapists. The Supplemental Table provides the final checklist.

Items selected for the checklist were compensatory behaviors specific to each joint. Compensatory behaviors on the checklist [7,8,10,21-26] included movements at the head, trunk, shoulder, elbow, forearm, wrist, fingers, and fluidity/movement precision. The administration of the ARAT adhered to the standardized instructions recommended by Yozbatiran et al. [10], participants were not provided with instructions regarding how the task should be completed. Compensations were scored as present or absent from the videotaped completion of the ARAT. For example, potential trunk compensations could be: excessive trunk flexion or excessive trunk side bending/rotation. In addition to compensations at each joint, an item labeled fluidity and precision of moment was added to capture jerky or uncoordinated sub-movements and multiple attempts to complete a task.[7,8,10,21-26]

Raters were current physical therapy students and one undergraduate summer intern. Non-licensed individuals were selected to decrease bias. In piloting, we found that licensed therapists tended to rate compensatory movement scores higher due to anticipation of expected movement patterns, whereas students simply rated if a compensatory movement was present (+1

point) or absent (0 points). Raters were trained prior to beginning scoring videos for data collection. Raters were provided with a manual that described the movement compensations. Then, raters scored a video side-by-side with a trainer (JB), where they discussed and highlighted each type of movement compensation. Finally, raters independently scored 3 videos of subjects with varying degrees of movement compensations. When the rater scoring was deemed to be acceptably close to the trainer ($\pm$10 points) they were allowed to score independently. If the score varied by more than $\pm$10 points (+/- 3% error), the rater continued to review videos with the trainer. This process continued until the rater became independent. Once training was complete, each video was scored by 2 raters, if total scores differed by over $\pm$10 points, a third rater scored the video. Scores were averaged for use in the final analysis. Possible scores range from 0 to 261 points, with lower scores indicating fewer observed compensations, or better movement quality.

B.3.3 Accelerometer variables

Data were extracted from bilateral, wrist worn accelerometers (wGT3X+, Actigraph, Pensacola, FL, USA) for 24 hours at the selected time point matching the video that was scored. Sleep was not excluded from the analysis, persons with stroke have irregular sleep patterns which would prove challenging to extract definitive time for sleep from the data.[27,28] Accelerometers are a valid and reliable instrument to capture upper limb movement in daily life in individuals after stroke[16,18,29-32] and non-disabled adults.[13,14,28]

For the selected time point, accelerometers were donned at the beginning of their session, prior to their in-clinic assessments and intensive upper limb therapy, then worn for an entire day afterward. Accelerometers were returned on the next treatment session and the data were downloaded using ActiLife 6 software (Actigraph Corp, Pensacola, FL, US). Accelerometers measure UL movement along 3 axes with activity counts, where 1 count=

0.001664g. Data were sampled at 30Hz. Data were binned into 1-second epochs, and activity counts across each axis were combined creating a single vector magnitude value.[17] Using custom-written software in MATLAB (Mathworks Inc, Natick, MA, USA), ten variables were calculated for in-clinic and out-of-clinic time from the recorded data. Recording time was separated into 1.5 hours of in-clinic time which included upper limb assessments and intensive therapy (targeting repetitions of upper limb movement) and then 22.5 hours of out-of-clinic wear. Variables quantified different aspects of upper limb movement and can be conceptualized into variables measuring movement time, movement magnitude, and movement variability.[13,14,18,33,34] Table B.1 provides a summary of variables. In addition, two newly proposed variables were calculated, the jerk asymmetry index[35] and the spectral arc length.[36,37] These variables were calculated as they have been proposed to measure smoothness of movement, an aspect of quality of movement, by others in the field.

Table B.1. Accelerometer variables

Variable Name	Description
Time	
Isolated Non-Paretic Limb Activity[28]	Time, in hours, that the non-paretic limb is moving, while the paretic limb is still.
Isolated Paretic Limb Activity[28]	Time, in hours, that the paretic limb is moving, while the non-paretic limb is still.
Bilateral Activity[28,38]	Time, in hours, that both upper limbs are moving together.
Use Ratio[16,30,39]	Ratio of hours of paretic limb movement, relative to hours of non-paretic limb movement.
Magnitude	
Paretic Limb Magnitude[40,41]	Magnitude of accelerations of the paretic limb, in activity counts.*
Bilateral Magnitude[28,38]	Intensity, or magnitude of accelerations, of movement across both arms, in activity counts.*
Magnitude Ratio[28,38,41]	Ratio of the magnitude of paretic UL accelerations relative to the magnitude of the non-paretic UL accelerations. This ratio reflects the contribution of each limb to activity, expressed as a natural log.
Variability	
Variability of Paretic Movement [40,41]	Standard deviation of the magnitude of accelerations across the paretic limb, reflecting the variability of paretic limb movement, in activity counts*
Variability of Bilateral Movement[40,41]	Standard deviation of the magnitude of accelerations across both limbs, reflecting the variability of bilateral upper limb movement, in activity counts.*
Variation Ratio[40,41]	Ratio of the variability of paretic limb accelerations relative to the variability of the non-paretic limb accelerations, reflecting the relative variability in the paretic limb.
Smoothness	

Unimanual Jerk Asymmetry Index [35]	Ratio of the average jerk magnitude between the paretic upper limb and the non-paretic upper limb. Higher jerk represents less smooth movement, and an index value of 0 represents similar smoothness of movement in the paretic and non-paretic limbs. Values are bounded between -1 to +1.
Spectral Arc Length[36,37]	A measure of movement smoothness that quantifies movement intermittencies independent of the movement's amplitude and duration. Longer spectral arc lengths are reflective of less smooth or less coordinated movement.

* Activity counts are computed by the Actilife proprietary software such that 1 activity count=0.001664g

B.3.4 Analysis

All data were analyzed in R, an open source statistical computing program. The main analyses evaluated the relationships between the compensatory movement scores and each calculated accelerometer variable. Spearman rank correlations were chosen because relationships between compensatory movement scores and accelerometer variables were not assumed to be linear. Criteria for statistical significance was set at $\alpha < 0.05$. The following criteria were used to interpret correlation coefficients: coefficients of rho $\leq$ 0.25 or below were considered low, coefficients ranging from 0.26 to 0.50 were considered moderate, coefficients from 0.51 to 0.75 were considered good, and those greater than 0.75 were considered excellent.[42] Beyond the individual relationship analysis, an exploratory, step-wise multiple regression evaluated how multiple accelerometer variables might collectively explain the variance in compensatory movement scores.

B.4 Results

Participants

Demographics of the participants are provided in Table B.2 and have been reported elsewhere.[20] Overall, the sample had chronic upper limb paresis post stroke of mild to moderate severity. Compensatory movement scores were highly variable across the sample, with a mean of 73.7 $\pm$ 33.6, and a range from 11.5 to 188. This range indicates that none of the subjects were free from compensatory movements, and no subject used the maximum amount of

compensations defined by the checklist. The majority of movement compensations were observed at the shoulder (28%). The second highest observed compensations were at the trunk (22%), followed by the fingers (21%), fluidity and movement precision (14%), elbow (5%), wrist (4%), head (3%), and finally forearm (2%).

Table B.2: Characteristics of Sample, Values are means ± SD (range) or % of total sample unless otherwise specified.

Descriptors (n= 78)	
Age (Years)	61.9 ± 10.5 (32, 85)
Gender	35% Female 65% Male
Type of Stroke	72% Ischemic 13% Hemorrhagic 15% Unknown
Ethnicity	99% Non-Hispanic/Latino 1% Hispanic/Latino
Months post stroke (median, min/max)	12 , 5/221
Affected Limb	46% Left 54% Right
% Concordance*	51%
% Independent with ADL	79%
Baseline ARAT score	32.4 ± 11.2 (10 - 48)
Compensatory Movement Score	73,7 ± 33.6 (11.5 - 188)
Baseline Use Ratio	0.66 ± 0.23 (0.22 - 1.32)

*Concordance is the percent of individuals whose paretic UL was their dominant UL

Relationships of variables to compensatory movement

Overall, moderate correlations were observed between the compensatory movement scores and each accelerometer variable. Figure 1 shows the correlation coefficients and their 95% confidence intervals for each accelerometer variable, calculated from both in-clinic and out-of-clinic time. For most of the accelerometer variables, higher scores are better, making most of the correlation coefficients negative.

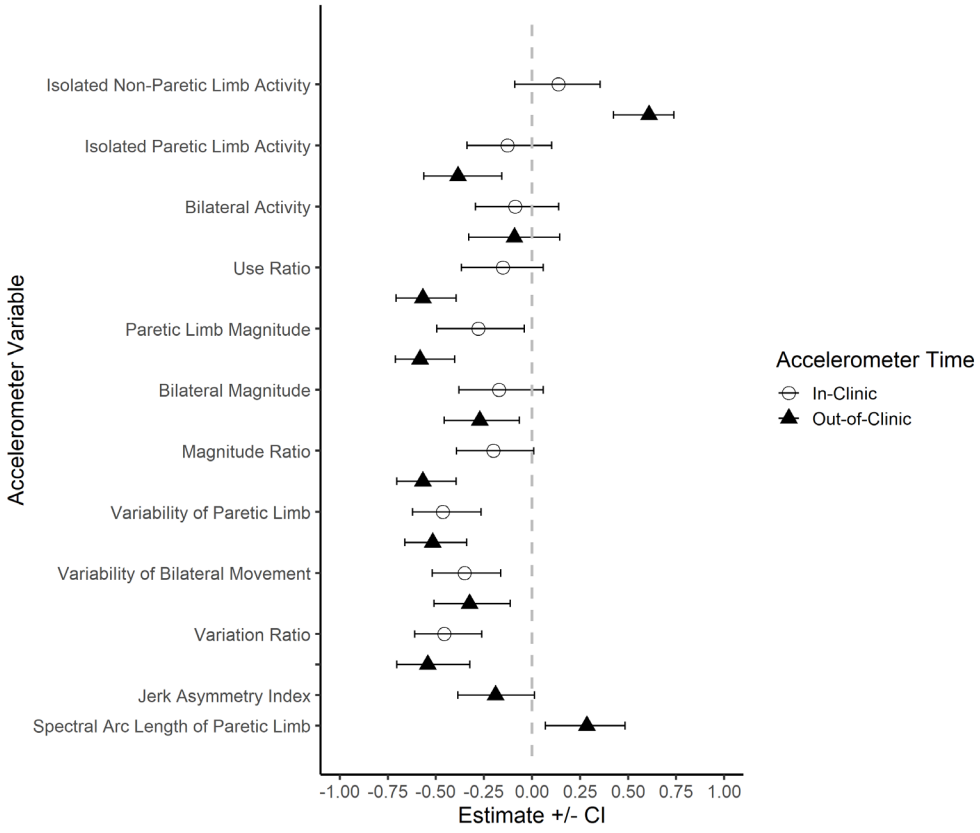


Figure B.1. Relationships (x-axis) of compensatory movement scores to accelerometer variables (y-axis). Open symbols are in-clinic calculations, and closed symbols are out-of-clinic calculations. Error bars are 95% confidence intervals for each correlation coefficient. Lack of statistical significance occurs when error bars cross the vertical dashed line at 0.

More than half of the accelerometer variables had similar relationships with compensatory movement scores when calculated from both in-clinic and out-of-clinic time. Figure 2 is a scatterplot of one such variable, variability of bilateral movement, where Figure 2A illustrates its relationship to the compensatory movement score in-clinic (rho = -0.35, $p < 0.001$), and Figure 2B its relationship out-of-clinic (rho = -0.32, $p < 0.01$). This moderate relationship indicates that individuals with more movement compensations tended to have less movement variability of the upper limbs, regardless of in which environment they were moving.

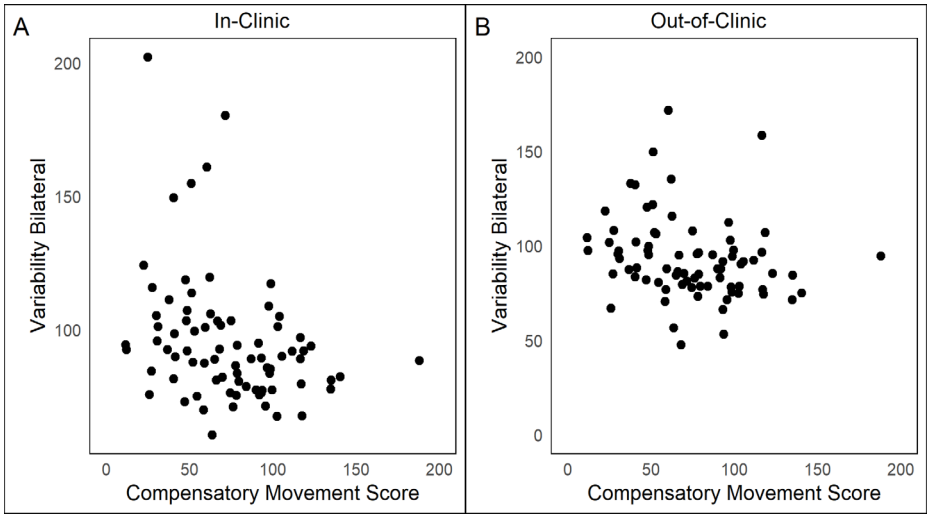


Figure B.2. Relationship of bilateral movement during in-clinic time (**a.** rho= -0.32, p<0,001) (**b.** rho= -0.35, p<0.01). This accelerometer variable has a similar moderate relationship both in and out-of-clinic.

Other accelerometer variables had a stronger relationship with compensatory movement scores, when calculated from time out-of-clinic versus in-clinic. Figure 3 shows scatterplots of two variables, isolated non-paretic limb activity and use ratio plotted relative to the compensatory movement score. Figure 3A illustrates the relationship of isolated non-paretic limb activity to compensatory movement score in-clinic (rho = 0.14, p = 0.23), and Figure 3B its relationship out-of-clinic (rho = 0.61, p <_0.0001). The stronger positive relationship out-of-clinic indicates that individuals with more compensatory movements moved their non-paretic limb only more while out-of-clinic. The use ratio also had a stronger negative relationship with the compensatory movement score out-of-clinic. Figure 3C illustrates the use ratio in-clinic to the compensatory movement score (rho = -0.15 p = 0.18), and Figure 3D its relationship out-of-clinic (rho = -0.57, p < 0.0001). The strong relationship out-of-clinic indicates that, at home, individuals with more compensatory movements had a lower use ratio, indicating less relative paretic limb activity. None of the accelerometer variables had a stronger relationship during in-clinic time versus out-of-clinic time.

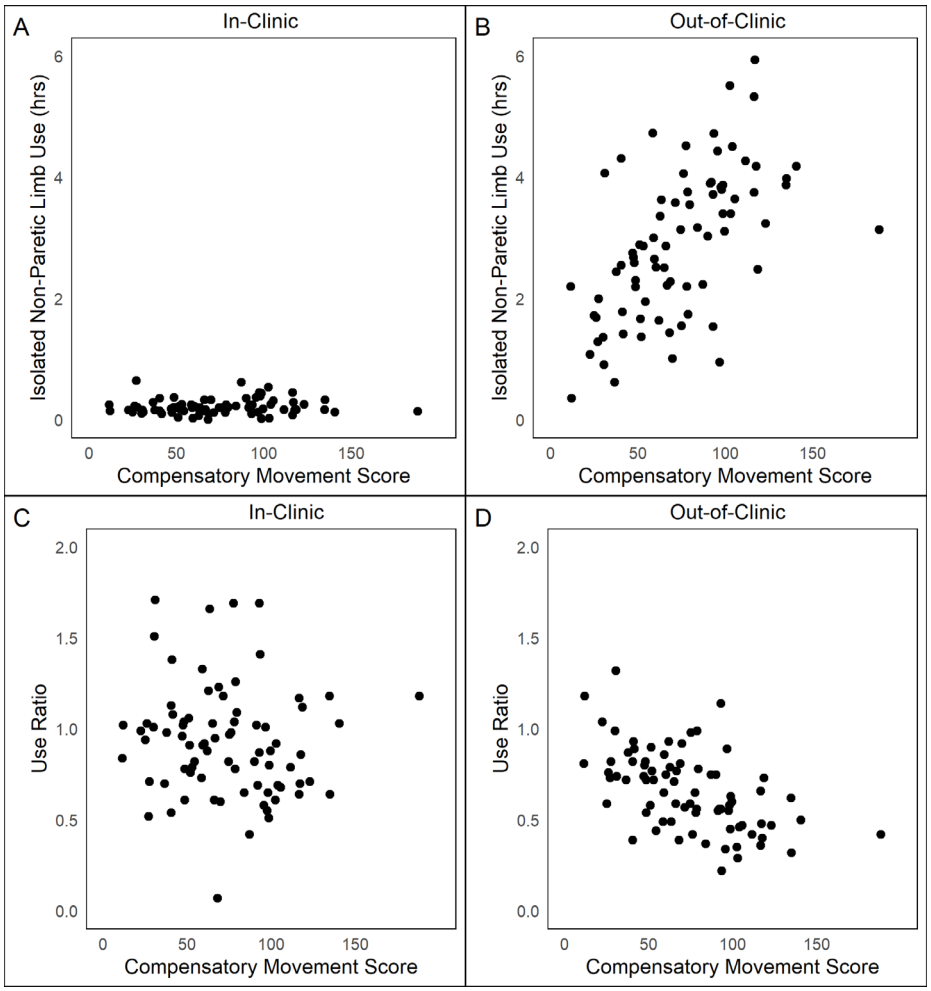


Figure B.3. Relationship of isolated use of the non-paretic limb to compensatory movement score, both in-clinic (3A, rho = 0.14, p = 0.23) and out-of-clinic (3B, rho = 0.61, p < 0.0001). Relationship of the use ratio to the compensatory movement score in-clinic (3C, rho = -0.15, p = 0.18) and out-of-clinic (3D, rho = -0.57, p = 0.18) These variables both had a little to no relationships in-clinic, yet good relationships out-of-clinic.

Two variables have been proposed to reflect movement smoothness as an aspect of quality of movement.[35-37] Figure 4 shows the relationship of the compensatory movement score to the jerk asymmetry index (Figure 4A, rho = -0.19, p = 0.09) and to the spectral arc length of the paretic limb (Figure 4B, rho = 0.29, p < 0.01). Both variables had low relationships with the compensatory movement score.

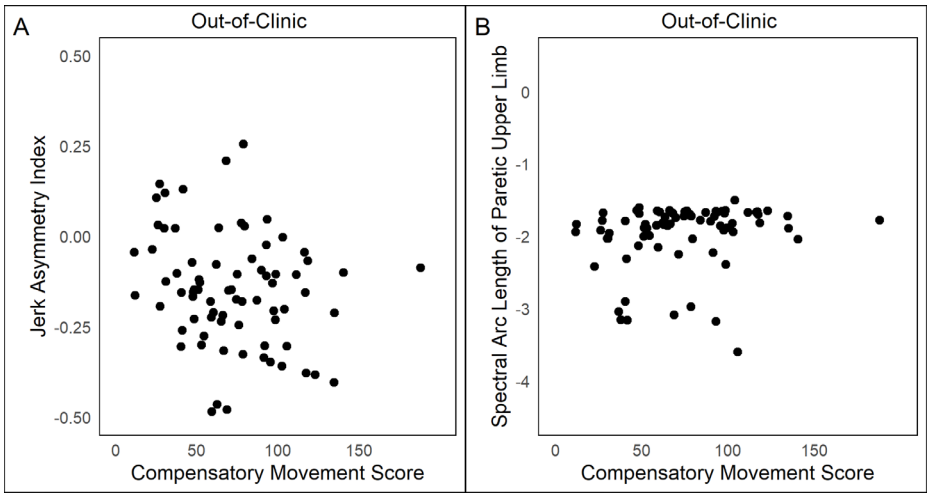


Figure B.4: Relationship of two newly proposed metrics that quantify quality of upper limb movement. 4A: Relationship of the Jerk Asymmetry Index to compensatory movement scores (rho = -0.19, p = 0.09). 4B: Relationship of the spectral arc length of the paretic limb to compensatory movement scores (rho = 0.29, $p < 0.01$). In 4B, one outlier with a spectral arc length of > -6 has been omitted from the plot. Both variables are from out-of-clinic time and had a low relationship with the compensatory movement score.

Last, an exploratory multiple regression evaluated which combination of accelerometer variables explained the most variance in the compensatory movement score. Using a stepwise approach to select variables, two time-based variables explained the most variance. The use ratio out-of-clinic and the hours of isolated non-paretic limb use out-of-clinic together explained 37% of the variance in the compensatory movement score ($R^2 = 0.37$, $p \leq 0.0001$).

B.5 Discussion

This study was a secondary analysis of an existing dataset that explored the relationships between accelerometer variables and compensatory upper limb movements in individuals with chronic hemiparesis. Individuals in the sample had a range of compensatory movements observed during the video scoring. Most accelerometer variables had a moderate relationship with the degree of compensatory movements of the upper limb for both in and out-of-clinic time points. This study used a novel approach to quantify compensatory movement patterns of the limb at a single time point, during completion of a standardized assessment. The developed

checklist is proposed as an idea for a new scale that would need further work in terms of examining its psychometric properties. These results indicate that accelerometry incorporates aspects of both quantity of upper limb use and general compensatory movement patterns of the upper limb.

Most accelerometer variables had moderate relationships with compensatory movement scores. The variables more strongly associated with compensatory upper limb movements quantified time, magnitude, and variability of upper limb movement while participants engaged in activity out-of-clinic. For example, the strong relationship of movement compensations to isolated use of the non-paretic limb aligns with clinical expectations[3,38,43] that individuals who have more movement compensations of the paretic upper limb, frequently use their non-paretic limb to complete daily tasks at home. Likewise, individuals who use more movement compensations have less variability in both paretic and bilateral limb movements. In general, reduced movement variability is considered to align with "an unhealthy pathological state or an absence of skill." [44] Individuals who use more compensatory movements have fewer options for movement available.[45,46]

Some accelerometer variables tended to be have stronger relationships with compensatory movement scores when quantified from out-of-clinic recordings vs. in-clinic recordings. This is illustrated visually in Figure 1, where more closed triangles are further from the zero line than open circles are. The in-clinic recordings here are from participation in an intensive, progressive, upper limb trial, where individuals are trained to use their affected paretic limb for functional activities.[20] Weaker relationships of some variables in-clinic confirms that therapy sessions were promoting activity of the affected upper limb. We note that the intent of the training protocol was to improve upper limb functional capacity, not to reduce movement

compensations.[20] During in-clinic recordings, the accelerometer variables measure what an individual does during the training protocol. The out-of-clinic time measures how an individual moves their upper limbs during daily life.[31,47] Based on the moderate or strong relationships, out-of-clinic, accelerometer variables reflect not just quantity of upper limb movement, but also collective use (more vs. less) of compensatory movements of the upper limb.

Limitations

Several limitations should be considered when interpreting these data. First, video recordings of a standardized assessment were used to quantify movement compensations as a proxy for compensatory movements that would occur throughout the recording period. Given that research and therapy participants often try to do their best on tests in front of an assessor, [28,48-50] using these videos to quantify compensatory movements may be an under-estimate of the compensatory movements participants engage in throughout the day. Second, video-recording of the assessment was chosen to quantify compensatory movements over the video-recording of the therapy session. This decision was made because the assessment was the same for all, while the therapy sessions involved individualized therapeutic activities of different amounts, i.e. making it hard to compare across subjects. While the ARAT standardized assessment captures most upper limb movement components,[51] one cannot rule out the possibility that alternative compensations might be observed within the therapy session or at home. Collectively these two limitations mean that we may have underestimated upper limb compensatory movements, and perhaps also underestimated the strength of the relationships of the accelerometer variables to the compensatory movement score. A third limitation is the use of coding from videos instead of using kinematic analysis of movement compensations. [12] Kinematic data from this sample does not exist. It is anticipated that using a kinematic analysis would not diminish the relationships of

accelerometer variables to movement compensations of the upper limb, rather future studies using kinematics could be used to validate the relationships found here. Additionally, kinematic analysis could expand upon those relationships by indicating the specific movement compensations an individual is using with their upper limb, not just the general quantification used here. The round table currently suggests kinematics as the gold standard, in the future we envision that accelerometry may be a way to measure specific movement compensations of the upper limb if sensors were cheaper and smaller.

B.6 Conclusion

This study quantified movement compensations of the upper limb and determined their relationship to accelerometer variables. Individuals who move their limbs more in daily life with respect to time and variability tend to move with less movement compensations and more normal movement patterns. Likewise, individuals who move their paretic limbs less and their non-paretic limb more in daily life tend to move with more movement compensations at all joints in the paretic limb and less normal movement patterns. These results suggest that, for people with upper limb paresis due to chronic stroke (> 6 months), movement quality is not be an independent construct from movement quantity. While accelerometers as a tool can provide some information on movement quality, more work is needed to improve the methodology.

B.7 Supplemental Table

Table B.1. Compensatory movement scoring checklist

Side Impaired	GRASP						GRIP				PINCH						GROSS MOTOR			TOTAL
	1. 4" block	2. 1" block	3. 2" block	4. 3" block	5. 3" ball	6. Stone	1. Pour water glass to glass	2. 1" tube	3. 1 cm tub	4. washer over bolt	1. ball bearing R and T	2. Marble I and T	3. BB L and T	4. BB I and T	5. M R and T	6. M L and T	1. hand behind head	2. hand top head	3. hand mouth	
HEAD																				0
Inappropriate head Flexion																				0
Inappropriate head rotation either with or without head flexion																				0
TRUNK																				0
Excessive trunk flexion																				0
Excessive trunk side bending/ rotation																				0
SHOULDER																				
Inadequate abduction and shoulder external rotation when it is REQUIRED for the task																				0
Lack of Shoulder Flexion resulting in one or more of the following : shoulder hiking/ shoulder abduciton/ unable to elevate arm																				0
Unable to stablize limb at beginning of task																				0
Unable to stabilize limb at end range																				0
ELBOW																				
Excessive elbow flexion or inadequate elbow extension																				0
FOREARM																				
Excessive shoulder abduction instead of forearm pronation when pouring, different from unimpaired side																				0
Excessive supination in reaching or grasping																				0
WRIST																				
Excessive wrist flexion during grasp																				0
Excessive wrist flexion in release of object																				0
FINGERS																				0
Use of compensatory grasp pattern																				0
Inadequate hand opening to grasp the item																				0
Inadequate hand opening to release the item																				0
Lack of opposition of digits																				0
FLUIDITY and PRECISION of MOVEMENT																				
Jerky/ Multiple Sub-movements during arm translation																				0
More than 2 attempts to grasp object																				0
TOTAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Items to create this checklist were synthesized from the following assessments, Reaching Performacne Scale [7], Upper Extremity Fugl- Meyer [52], Wolf Motor Function Test, [8] Action Research Arm Test, [51] Chedoke McMasters, [23] Stroke Rehabiliation Assessment of Movement (STREAM), [22] Motor Evaluation Scale for Upper Extremity in Stroke Patients (MESUPES), [25] Motor Assessment Scale (MAS), [24]and Quantative assessment of upper extremity function.[26] Compensatory behaviors on the checklist included movements at the head, trunk, shoulder,elbow, forearm, wrist, fingers and fluidity/movement. Compensations were scored as present or absent from the videotaped completion of the ARAT.

B.8 References

1. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association. *Circulation*. Mar 20 2018;137(12):e67-e492. doi:10.1161/cir.0000000000000558

2. Dobkin BH. Rehabilitation after stroke. *New England Journal of Medicine*. 2005;352(16):1677-1684.

3. Levin MF, Kleim JA, Wolf SL. What do motor "recovery" and "compensation" mean in patients following stroke? *Neurorehabilitation and neural repair*. 2009;23(4):313-319.

4. Lum PS, Mulroy S, Amdur RL, Requejo P, Prilutsky BI, Dromerick AW. Gains in upper extremity function after stroke via recovery or compensation: Potential differential effects on amount of real-world limb use. *Top Stroke Rehabil*. Jul-Aug 2009;16(4):237-53. doi:10.1310/tsr1604-237

5. Roby-Brami A, Feydy A, Combeaud M, Biryukova E, Bussel B, Levin MJAns. Motor compensation and recovery for reaching in stroke patients. 2003;107(5):369-381.

6. DeJong SL, Birkenmeier RL, Lang CE. Person-specific changes in motor performance accompany upper extremity functional gains after stroke. *Journal of applied biomechanics*. 2012;28(3):304-316.

7. Levin MF, Desrosiers J, Beauchemin D, Bergeron N, Rochette A. Development and validation of a scale for rating motor compensations used for reaching in patients with hemiparesis: the reaching performance scale. *Physical therapy*. 2004;84(1):8-22.

8. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf motor function test as outcome measure for research in patients after stroke. *Stroke*. 2001;32(7):1635-1639.

9. See J, Dodakian L, Chou C, et al. A standardized approach to the Fugl-Meyer assessment and its implications for clinical trials. *Neurorehabilitation and neural repair*. 2013;27(8):732-741.

10. Yozbatiran N, Der-Yeghiaian L, Cramer SC. A Standardized Approach to Performing the Action Research Arm Test. *Neurorehabilitation and Neural Repair*. 2008;22(1):78-90. doi:10.1177/1545968307305353

11. Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra-and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. *Archives of Physical Medicine and Rehabilitation*. 2001;82(1):14-19.

12. Kwakkel G, Wegen EV, Burridge J, et al. Standardized measurement of quality of upper limb movement after stroke: Consensus-based core recommendations from the Second Stroke Recovery and Rehabilitation Roundtable. *International Journal of Stroke*. 2019:1747493019873519.

13. Bailey RR, Klaesner JW, Lang CE. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. 2014;9(7):e103135.

14. Bailey RR, Lang CE. Upper extremity activity in adults: referent values using accelerometry. *Journal of rehabilitation research and development*. 2014;50(9):1213.

15. Uswatte G, Miltner WH, Foo B, Varma M, Moran S, Taub E. Objective measurement of functional upper-extremity movement using accelerometer recordings transformed with a threshold filter. *Stroke*. 2000;31(3):662-667.

16. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of Accelerometry for Monitoring Real-World Arm Activity in Patients With Subacute Stroke: Evidence From the Extremity Constraint-Induced Therapy Evaluation Trial. *Archives of Physical Medicine and Rehabilitation*. 2006/10/01/ 2006;87(10):1340-1345. doi:https://doi.org/10.1016/j.apmr.2006.06.006

17. Bailey RR, Lang CE. Upper-limb activity in adults: referent values using accelerometry. *Journal of Rehabilitation Research and Development*. 2013;50(9):1213-1222. doi:10.1682/JRRD.2012.12.0222

18. Bailey RR, Birkenmeier RL, Lang CE. Real-world affected upper limb activity in chronic stroke: an examination of potential modifying factors. *Topics in stroke rehabilitation*. 2015;22(1):26-33.

19. Doman CA, Waddell KJ, Bailey RR, Moore JL, Lang CE. Changes in upper-extremity functional capacity and daily performance during outpatient occupational therapy for people with stroke. *American Journal of Occupational Therapy*. 2016;70(3):7003290040p1-7003290040p11.

20. Lang CE, Strube MJ, Bland MD, et al. Dose response of task-specific upper limb training in people at least 6 months poststroke: A phase II, single-blind, randomized, controlled trial. *J Annals of neurology*. 2016;80(3):342-354.

21. Fugl-Meyer AR JL, Leyman I, Olsson S, Steglind S. FUGL-MEYER ASSESSMENT UPPER EXTREMITY (FMA-UE) Assessment of sensorimotor function. 1975;

22. Daley K, Mayo N, Wood-Dauphinée S. Reliability of scores on the Stroke Rehabilitation Assessment of Movement (STREAM) measure. *Physical therapy*. 1999;79(1):8-23.

23. Gowland C, Stratford P, Ward M, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. *Stroke*. 1993;24(1):58-63.

24. Carr JH, Shepherd RB, Nordholm L, Lynne D. Investigation of a new motor assessment scale for stroke patients. *Physical therapy*. 1985;65(2):175-180.

25. Van de Winckel A, Feys H, van der Knaap S, et al. Can quality of movement be measured? Rasch analysis and inter-rater reliability of the Motor Evaluation Scale for Upper Extremity in Stroke Patients (MESUPES). *Clinical rehabilitation*. 2006;20(10):871-884.

26. Carroll D. A quantitative test of upper extremity function. *Journal of Chronic Diseases*. 1965;18(5):479-491.

27. Sadeh A, Acebo C. The role of actigraphy in sleep medicine. *Sleep medicine reviews*. 2002;6(2):113-124.

28. Bailey RR, Klaesner JW, Lang CE. Quantifying real-world upper-limb activity in nondisabled adults and adults with chronic stroke. *Neurorehabilitation and Neural Repair*. 2015;29(10):969-978. doi:10.1177/1545968315583720

29. Dobkin BH, Martinez C. Wearable Sensors to Monitor, Enable Feedback, and Measure Outcomes of Activity and Practice. *Curr Neurol Neurosci Rep*. Oct 6 2018;18(12):87. doi:10.1007/s11910-018-0896-5

30. Uswatte G, Foo WL, Olmstead H, Lopez K, Holand A, Simms LB. Ambulatory Monitoring of Arm Movement Using Accelerometry: An Objective Measure of Upper-Extremity Rehabilitation in Persons With Chronic Stroke. *Archives of Physical Medicine and Rehabilitation*. 2005/07/01/ 2005;86(7):1498-1501. doi:https://doi.org/10.1016/j.apmr.2005.01.010

31. Waddell KJ, Strube MJ, Bailey RR, et al. Does task-specific training improve upper limb performance in daily life poststroke? *J Neurorehabilitation*. 2017;31(3):290-300.

32. Waddell KJ, Strube MJ, Tabak RG, Haire-Joshu D, Lang CE. Upper Limb Performance in Daily Life Improves Over the First 12 Weeks Poststroke. *Neurorehabilitation and neural repair*. 2019;33(10):836-847.

33. Bailey RR, Klaesner JW, Lang CEJN, repair n. Quantifying real-world upper-limb activity in nondisabled adults and adults with chronic stroke. 2015;29(10):969-978.

34. Urbin M, Bailey RR, Lang CEJJonptJ. Validity of body-worn sensor acceleration metrics to index upper extremity function in hemiparetic stroke. 2015;39(2):111.

35. de Lucena DS, Stoller O, Rowe JB, Chan V, Reinkensmeyer DJ. Wearable sensing for rehabilitation after stroke: Bimanual jerk asymmetry encodes unique information about the variability of upper extremity recovery. IEEE; 2017:1603-1608.

36. Balasubramanian S, Melendez-Calderon A, Roby-Brami A, Burdet E. On the analysis of movement smoothness. *Journal of Neuroengineering*. 2015;12(1):112.

37. Balasubramanian S, Melendez-Calderon A, Burdet E. A robust and sensitive metric for quantifying movement smoothness. *IEEE transactions on biomedical engineering*. 2012;59(8):2126-2136.

38. Bailey RR, Klaesner JW, Lang CE. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. *PLOS ONE*. 2014;9(7):e103135. doi:10.1371/journal.pone.0103135

39. Lang CE, Waddell KJ, Klaesner JW, Bland MD. A method for quantifying upper limb performance in daily life using accelerometers. *JoVE (Journal of Visualized Experiments)*. 2017;(122):e55673.

40. Urbin MA, Waddell KJ, Lang CE. Acceleration metrics are responsive to change in upper extremity function of stroke survivors. *Archives of Physical Medicine and Rehabilitation*. 2015/05/01/ 2015;96(5):854-861. doi:https://doi.org/10.1016/j.apmr.2014.11.018

41. Urbin M, Bailey RR, Lang CE. Validity of body-worn sensor acceleration metrics to index upper extremity function in hemiparetic stroke. *Journal of neurologic physical therapy: JNPT*. 2015;39(2):111.

42. Portney LG. *Foundations of Clinical Research: Applications to Evidence-Based Practice*. FA Davis; 2020.

43. Krakauer JW. Arm function after stroke: from physiology to recovery. Citeseer; 2005:384-395.

44. Stergiou N, Decker LM. Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Human movement science*. 2011;30(5):869-888.

45. Krabben T, Prange GB, Molier BI, Rietman JS, Buurke JH. Objective measurement of synergistic movement patterns of the upper extremity following stroke: An explorative study. 2011:1-5.

46. Sethi A, Patterson T, McGuirk T, Patten C, Richards LG, Stergiou N. Temporal structure of variability decreases in upper extremity movements post stroke. *Clinical Biomechanics*. 2013/02/01/ 2013;28(2):134-139. doi:https://doi.org/10.1016/j.clinbiomech.2012.11.006

47. Andrews K, Steward JJR. Sroke recovery: he can but does he? 1979;18(1):43-48.

48. Kunkel A, Kopp B, Müller G, et al. Constraint-induced movement therapy for motor recovery in chronic stroke patients. *Archives of Physical Medicine and Rehabilitation*. 1999/06/01/ 1999;80(6):624-628. doi:https://doi.org/10.1016/S0003-9993(99)90163-6

49. André JM. "Functional motor amnesia" in stroke (1904) and "learned non-use phenomenon"(1966). *Journal of rehabilitation medicine*. 2004 36(3):138-140. doi:10.1080/16501970410026107

50. Johnson M. Quantifying learned non-use after stroke using unilateral and bilateral steering tasks. 2011:1-7.

51. Lyle. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *International journal of rehabilitation research*. 1981;4(4)

52. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J Rehabil Med*. 1975 1975;7(1):13-31.