Adaptive Algorithms for Personalized Health Monitoring

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

by

Matthew Morrow Engelhard

December 2016

APPROVAL SHEET

This Dissertation
is submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Author Signature: _____

This Dissertation has been read and approved by the examining committee:

Advisor: Stephen D Patek

Committee Member: Laura E Barnes

Committee Member: Peter A Beling

Committee Member: Myla D Goldman

Committee Member: John C Lach

Committee Member: _____

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

December 2016

Adaptive Algorithms for Personalized Health Monitoring

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Doctor of Philosophy (Systems and Information Engineering)

by

Matthew M. Engelhard

December 2016

Abstract

Health monitoring has entered the era of precision medicine. With support from President Obama's Precision Medicine Initiative, the National Institutes of Health have renewed their focus on prevention, management, and treatment strategies that are tailored to individual patients. The initiative relies heavily on genomics and bioinformatics, but clinical informatics and mobile health technologies have been consistently emphasized, including the real-time monitoring of physiologic data. A patient's heart rate can now be measured with a wristband and transmitted wirelessly to their clinician; skin impedance can be sampled every second to develop a personalized stress profile. As illustrated by these examples, the most distinctive feature of monitoring is the repeated, sequential collection of physiologic measurements by a computing platform. Consequently, monitoring systems have capabilities not found elsewhere in precision medicine: they can interact with the patient and adapt in real time.

In this work, we develop monitoring algorithms uniquely suited for health care applications. They are designed for different monitoring scenarios with distinct data types, but each one adapts its outputs or decisions to a growing history of observations. Our central hypothesis is that health monitoring benefits immensely from an adaptive approach – more so than other monitoring applications – due to fundamental differences between engineered and biological systems. These differences pertain not only to variability between persons, but also to our knowledge of physiology, our access to salient system parameters, and the importance of subjective experiences in health care.

Many of our objectives have been motivated by a target application, walking ability in

multiple sclerosis (MS). MS is a disease of the central nervous system which can produce almost any neurological sign or symptom, making adaptation and personalization all the more critical. After presenting a case study in MS, we formulate adaptive algorithms for three common health monitoring scenarios. The first is a physiologic signal monitoring algorithm designed for signals that vary substantially between persons or over time. The second, adaptive symptom reporting, personalizes its queries to accurately track disability while reducing the burden placed on the patient. The third, active health event identification, learns to classify events from the patient's perspective by requesting event labels at opportune times. Each algorithm is validated with data from persons with MS.

Approval Sheet

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Systems and Information Engineering)

Matthew M. Engelhard

Matthew M. Engelhard

This dissertation has been read and approved by the Examining Committee:

Stephen D. Patek

Stephen D. Patek, Advisor

Peter A. Beling

Peter A. Beling, Committee Chair

Laura E. Barnes

Laura E. Barnes

John C. Lach

John C. Lach

Myla D. Goldman

Myla D. Goldman

Accepted for the School of Engineering and Applied Science:

Craig H. Benson

Craig H. Benson, Dean, School of Engineering and Applied Science

December 2016

To Karen and Josie

Acknowledgments

Each member of my committee made an essential contribution to this work. Dr. Stephen Patek, my advisor, shaped me as an academic researcher and pointed me in the right direction whenever I was lost. This research is equal parts his vision and mine. Dr. Myla Goldman, my clinical research mentor, drove our work forward and pushed me to be my best. She taught me everything I know about clinical research and inspired me as a role model and friend. Dr. John Lach brought me out of PhD seclusion to join the INERTIA team, giving me a research community. He challenged me to see the big picture and taught me how to bridge the gap between engineering and clinical impact. My chair, Dr. Peter Beling, gave me technical guidance at critical points, leading to our work in active learning. Crafting lectures and exam problems with him as a teaching fellow was perhaps the most memorable experience of graduate school. Dr. Laura Barnes has been a continual source of wisdom and support. As the expert in machine learning and health informatics, she repeatedly lead us past technical and conceptual obstacles. I am so thankful to have had the privilege of working with each of them.

I'd also like to thank Dr. Gerard Learmonth, who encouraged me to apply to the PhD program, advised me in my first year, and supported me during a difficult time. I owe special thanks to Karen Schmidt, Kristina Sheridan, Josh Inouye, Stephen Adams, INERTIA, and the Goldman Lab, who supported this research with ideas, comments, and hard work.

And of course I'd like to thank my family, who smiled and nodded for 8 years as I talked through one "big idea" after another. Will I ever get a real job?

Contents

List of Tables

List of Figures

Frequently Used Abbreviations

6MW	Six-Minute Walk
AEI	Active Event Identification
ASR	Adaptive Symptom Reporting
CRC	Category Response Curve
DIF	Differential Item Functioning
DTW	Dynamic Time Warping
EDSS	Expanded Disability Severity Scale
GRM	Graded Response Model
HMM	Hidden Markov Model
IRT	Item Response Theory
MCID	Minimal Clinically Important Difference
MFIS	Modified Fatigue Impact Scale
MLE	Maximum Likelihood Estimation
MS	Multiple Sclerosis
MSWS-12	Multiple Sclerosis Walking Scale
PRO	Patient-Reported Outcome
RSOI-DTW	Rotation, Scale, and Offset Invariant DTW
SDT	Standard Deviation Threshold

Chapter 1

Introduction

Health monitoring has entered the era of precision medicine. With support from President Obama's Precision Medicine Initiative[2], the National Institutes of Health (NIH) have renewed their focus on "prevention and treatment strategies that take individual variability into account"[3]. The initiative is broad, with genomics and bioinformatics at the forefront, but clinical informatics and mobile health technologies have also been emphasized, including "real-time monitoring of glucose, blood pressure, and cardiac rhythm"[3]. Individualized care will be realized through a "new taxonomy of disease" based on novel physiologic measurements and molecular markers – genes, proteins, and cellular metabolites – in addition to the usual clinical symptoms and signs. Once in place, treatment can be tailored to specific cellular phenotypes within the new taxonomy[4].

Like genomics and other "-omics", monitoring draws upon new kinds of measurements as technology progresses. Its distinctive feature, however, is sequential, real-time measurement collection via a computing platform. A patient's heart rate can be measured with a wristband and transmitted wirelessly to their clinician; skin impedance can be sampled every second to develop a personalized stress profile. This data can be immediately processed and acted upon, giving monitoring systems a capability not found elsewhere in precision medicine: real-time *adaptation*. Monitoring algorithms can adapt to the information they observe to improve

their outputs or behaviors. Moreover, many monitoring platforms have opportunities to directly interact with patients, which itself is a form of decision-making that can benefit from adaptation. Patient interactions are a rich source of information not adequately tapped by most monitoring systems.

This dissertation is focused on algorithms designed for health monitoring – though they can be applied in other domains – that adjust to the individual being monitored. Some adjustments account for heterogeneity among persons and disease processes; others bridge the gap between objective measurements and subjective experiences; and still others streamline interactions with patients to make monitoring less burdensome. The common thread is personalization and adaptation: an *adaptive* algorithm draws upon an evolving history of observations to guide its current behavior. This notion is formalized in Chapter 2.

Our central hypothesis is that health monitoring benefits immensely from an adaptive approach. Indeed, it is our belief that adaptation is a *necessity* in many health monitoring scenarios, much more so than monitoring in other domains, due to fundamental differences between engineered and biological systems. These differences pertain to our level of understanding of the system itself, our access to salient system parameters, and the importance of subjective experiences in health care. Engineered systems (of the same model) are more or less identical, whereas every biological system is different. Monitoring techniques designed for engineered systems have been applied to health care, sometimes with excellent results, but other times the aforementioned factors limit their effectiveness.

This work has been guided in many respects by our target application, walking ability in multiple sclerosis (MS). MS is a disease of the central nervous system (CNS) which can produce almost any neurological sign or symptom. It is therefore the quintessential example of a heterogeneous disease; even walking pathologies with similar neurologic correlates tend to have distinctive idiosyncracies from person to person. In our experience, a one size fits all approach to monitoring performs poorly in MS. At the very least, an adequate training dataset should span the full range of MS heterogeneity, which would require massive, unprecedented

data collection. On the other hand, our personalized approach to monitoring has achieved promising results, as we will show. While MS is an extreme case, we argue in Chapter 2 that this characteristic is quite general in health care.

The specifics of a health monitoring algorithm depend heavily on the type of data involved. In Chapters 4 and 6, we monitor physiologic time series recorded by wearable sensors on different scales and with different statistical properties. In these scenarios, variability between persons and over time is perhaps the foremost challenge for a monitoring platform. In contrast, Chapter 5 – and to some extent, Chapter 6 – deals with patient-reported symptoms and patient-defined events, respectively, which can be subjective in nature. Subjective information is important in health care but not present in the monitoring of engineered systems. It is therefore not surprising that common monitoring algorithms are poorly equipped to handle it. In these Chapters, we show how adaptation and patient interaction can be used to learn about the patient experience.

In fact, the relationship between subjective and objective measurements, and the extent to which it varies between persons, is itself a topic of considerable interest. Patient perceptions are at the center of the growing patient-centered care movement, which is believed to promote appropriate health care utilization, reduce costs, and perhaps even improve outcomes. Patient-centered care and improved patient engagement have improved self-reported health status by helping patients and providers find common ground[5], and when care providers use a patient-centered approach, patients utilize health care services less often[6] with reduced associated cost[7, 8]. Remote health monitoring has played an important role in this movement; indeed, over 500 studies have assessed mHealth interventions, with remote monitoring of chronic conditions being one of the most common and consistent targets[9]. On average, these interventions have had a small but significant positive effect on targeted behaviors[10]. Adaptive algorithms can learn a correspondence between subjective and objective measurements on the fly through patient interaction. This has the potential to revolutionize our understanding of the patient experience, and it is the focus of Chapter 6.

As wearables and other smart devices continue to proliferate, real-time health monitoring will become ubiquitous. Personal and professional-grade monitoring systems will collect physiologic measurements with unprecedented frequency and precision, process them in real time, and immediately present them to care providers to facilitate clinical decision-making. However, the usefulness of data they gather is determined largely by the algorithms which interpret it. In this work, we show how adaptation can be used to augment physiologic data with rich contextual information, including unique patient physiology, experiences, and preferences.

1.1 Outline

The six remaining chapters of this dissertation are organized as follows. Chapter 2 discusses the predominant monitoring paradigm for mechanical, electrical, and structural systems, contrasting it with an *adaptive* approach. Medical and health applications have adopted this paradigm despite the profound differences between biologic and engineered systems. Although health monitoring has been successful in many scenarios under the current approach, we argue that adaptive, personalized monitoring is often more natural.

Chapter 3 introduces our primary application, remote monitoring of walking ability in multiple sclerosis (MS). After a brief MS overview, we discuss clinical walking assessment to orient the reader to the goals of remote monitoring and the types of data involved. We then revisit the argument for adaptive monitoring in this context.

Chapters 4 - 6 present adaptive algorithms for three distinct health monitoring scenarios. Chapter 4 presents a personalized approach to physiologic signal monitoring designed for signals which vary significantly between persons. Chapter 5 concerns adaptive symptom reporting (ASR), an algorithm which tracks patient-reported health. ASR draws upon continually acquired patient data to personalize symptom reporting content and timing. Chapter 6 develops active health event identification, which learns to map the monitoring

system's internal states to events experienced by the patient in real time. We begin with a hidden Markov model (HMM) and extend it to online, active state identification.

Algorithm-specific background is discussed at the beginning of these three chapters, and future directions are discussed at the end. We close with a summary and some thoughts on the future of health monitoring in Chapter 7.

1.2 Scope And Contributions

Algorithm design is but one component of health monitoring. Monitoring ecosystems include the patient on one end, a care provider on the other, and many elements in between. These include sensor and hardware development, wireless communications, data processing, integration with EHR, data presentation, and human factors, just to name a few. Data processing alone consists of several steps, including preprocessing, feature selection, and algorithm training and design. This work focuses specifically on algorithms which process an established set of features. Aspects of preprocessing and feature extraction are discussed in a few cases, but they are secondary to algorithm design.

A more appropriate title might have been "A Few Adaptive Algorithms for Personalized Health Monitoring". We make no attempt to exhaustively explore adaptive approaches to the myriad monitoring problems in health care. Rather, we present three personalized algorithms appropriate for specific monitoring scenarios. These algorithms are quite general, with many possible applications, and we discuss a number of planned extensions. Still, there are many monitoring scenarios to which none of our algorithms apply.

On the other hand, Chapters 4 - 6 survey a diverse set of monitoring scenarios, and we cover several distinct data types commonly encountered in real-time health monitoring. Imaging findings, lab results, and clinical notes are not discussed, but these are not commonly monitored in real time. Importantly, the potential applications of our algorithms are broad

enough to support our argument that adaptation is essential in health monitoring, which is intended to be a significant contribution of this work.

Chapters 3 - 6 all present clinical data from persons with MS, and all three algorithms involve some form of clinical validation. In some cases, one or more additional validation studies have also been presented. These results prove the algorithms work as intended, and they explore a number of design considerations. However, this is only the first step of many needed to pave the way for routine clinical use. Reaching this goal would require study in multiple populations culminating in large-scale clinical trials.

Remote monitoring and mobile health have been criticized for having an inadequate evidence base[11]. We make a modest contribution to this literature through several clinical studies. Knowing the obstacles involved, however, the objective of this work is not to bring these algorithms directly into clinical use. Instead, this work highlights the unique advantages and capabilities of adaptive approaches to health monitoring in order to encourage further exploration and algorithm development.

Chapter 2

Motivating an Adaptive Approach

The most common approach to system monitoring involves a cycle of data collection and analysis followed by the deployment of a customized but unchanging monitoring platform. This approach has been successfully applied to a staggering number and variety of problems, including many in health care, often with excellent results. In many health monitoring applications, however, the system being monitored has characteristics that are fundamentally at odds with a traditional, static approach to monitoring. Such characteristics include variability between persons, the importance of subjective experiences, and limitations in our understanding of human physiology. Consequently, an *adaptive* approach, in which the monitoring platform learns about the monitored system over time without human intervention, is often more appropriate in health monitoring.

This chapter makes the case that adaptive algorithms offer tremendous advantages in the monitoring of physiologic systems. After discussing the traditional approach to monitoring, we formally define an *adaptive* algorithm, provide examples and common applications, and contrast the adaptive and static approaches to monitoring. The largest portion of the chapter highlights properties of medicine and human physiology that call for system adaptation. Lastly, we highlight recent work in remote interventions to underscore the potential benefits of an accurate, personalized monitoring system.

In later chapters, we present adaptive algorithms for three distinct health monitoring scenarios. The arguments in this chapter apply to all three scenarios, revealing a common thread behind our research program, but they are also intended as a self-contained contribution of this work.

2.1 The Tested Approach to Monitoring

Networks of sensors are routinely deployed to monitor the "health" of a wide variety of systems, including engines and other mechanical systems, electrical circuits and devices, and civil infrastructure. The assessments derived from monitoring systems are often called *prognostics*, and are usually intended to provide an early warning of system failure. Sensor-based monitoring is a cost-effective alternative to routine inspection and maintenance, so much so that in some applications – for example, the monitoring of aircraft engines – it has made maintenance schedules obsolete.

System development and deployment is an iterative cycle requiring years of data collection, testing, and the knowledge of many experts. In aircraft monitoring, "companies work together over many years to develop parameters for known defects"[12]. This cycle leads to the development, refinement, and standardization of a set of measurements – we will call them *features* – sufficient to generate the prognostics needed to assess system health. Each feature is specified in terms of both sensors and preliminary data processing. In structural monitoring, for example, an acoustic sensor might be used to determine a structure's vibrational response, which is processed via Fourier analysis to generate features[13].

System prognostics and alarms are then derived from the features. In electronics monitoring, for instance, "it is necessary to identify the precursor variables for monitoring, and then develop a reasoning algorithm to correlate the change in the precursor variable with the impending failure"[14]. The "reasoning algorithm" might be a simple parameter threshold which triggers maintenance when exceeded; or, toward the other extreme, a large, many-layered

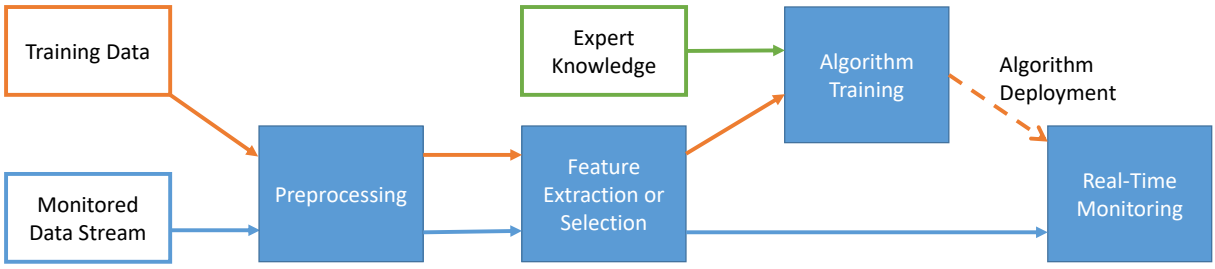


Figure 2.1: Traditional Model Deployment (Banaee et al. [1])

neural network [12, 15]. Herein we define a *monitoring algorithm* as any method, whether simple or complex, that transforms features into prognostics or alarms for system assessment. Much like parameter development, "years of accumulation of knowledge is typically necessary to establish all the necessary rules" comprising a monitoring algorithm[12].

We will refer this model of system deployment as the *traditional* model or *static* model to contrast it with the adaptive approach. A block diagram of this model, which was inspired by a similar diagram created by Banaee et al.[1], is shown in Figure 2.1. As the diagram shows, data and expert knowledge are used to train the monitoring algorithm, which is then deployed. After deployment, the algorithm does not change, except (possibly) through additional cycles of training and deployment.

Not surprisingly, this development model is also the most common approach to health monitoring. A simple example may be found in any intensive care unit: whenever a patient's heart rate, blood pressure, or oxygen saturation falls outside of a pre-defined acceptable range, an alarm goes off. These physiologic measurements are simple features that serve as inputs to a thresholding algorithm. In this case, algorithm training consists in setting the alarm threshold based on empirical data and expert knowledge.

More sophisticated systems are also common, particularly in remote and mobile health monitoring. Many utilize state of the art supervised and unsupervised learning[16], and fantastic successes have been achieved. A compelling example is the HeRO system developed by Moorman et al., which used heart rate prognostics to reduce mortality from sepsis in very low birth weight infants[17]. Another is the APACHE prognostic, which estimates mortality

risk by monitoring physiologic and clinical variables[18]. These systems are much more complicated and data-driven than a heart rate threshold, but they still fit the traditional development model as we have described it: data collection, feature engineering, and algorithm training are followed by deployment of a fine-tuned but static monitoring system.

The traditional model has several shortcomings when it comes to health monitoring. First, it is not capable of context awareness, a recognized need in remote monitoring and preventive medicine[1]. Moreover, model accuracy is limited by the fact that "biomedical and healthcare factors that affect diseases are not fully known"[19]. The importance of personalization is widely recognized[3], as discussed in the previous chapter, but developing personalized systems under the traditional model requires data collection and analysis at unprecedented scale. Separate models must be developed for each branch of a patient taxonomy so that "person-specific" models – which would more accurately be labeled as *subpopulation* models – can be applied to each patient. Many experts propose this approach[20] despite the almost insurmountable practical difficulties it entails. Instead, these shortcomings can be addressed more easily and elegantly with adaptive algorithms.

2.2 Adaptive Algorithms

An adaptive algorithm might be described as one which utilizes prior experiences to adjust to circumstances at the time of execution. Possible adjustments include dynamically allocating resources or memory, optimizing decision-making, or improving predictions. In this work, we are concerned specifically with algorithms whose outputs and/or actions depend on prior observations. Letting $o_{1,t} = (o_1, ..., o_t)$ be a sequence of observations at times 1 to t, where $o_i \in O$ for all $i \in \{1, ..., t\}$, we formally define an algorithm A to be *adaptive* if either the mapping f from observations to outputs $y \in Y$ or the mapping g from observations to actions $u \in U$ depends on $o_{1,t}$. To indicate this dependence, we write $f_{o_{1,t}} : O \to Y$ and $g_{o_{1,t}} : O \to U$, respectively, which may be abbreviated as f_t and g_t.

As a simple example, consider a binary classifier operating on a data stream. At each time t, the classifier f utilizes the observation o_t to make a class prediction $f(o_t) \in \{0, 1\}$. Typically such a classifier is static: it has been trained with observations from a previous deployment, but does not evolve as new data is obtained. As a result, $f(o_{t_1}) = f(o_{t_2})$ whenever $o_{t_1} = o_{t_2}$. In contrast, suppose instead that the classifier draws on recent observations to improve its predictions. The classifier is now written f_t to indicate its dependence on $o_{1,t}$, and $o_{t_1} = o_{t_2}$ no longer implies that $f_{t_1}(o_{t_1}) = f_{t_2}(o_{t_2})$ due to the new information obtained between times t_1 and t_2. The second algorithm is adaptive, because f depends on t, while the first is not.

Rosenblatt's perceptron algorithm fits this description. Supposing that observations are points in $\mathbb{R}^n$, the perceptron labels each observation o_t as 1 or -1 based on the sign of $v * o_t$, where v is a prediction vector. Adaptation occurs through updates to v, which take place whenever the prediction is incorrect. Specifically, v is updated as $v + y * o_t$, where y is the correct label for o_t[21]. Minor variations of this algorithm have been used in a variety of applications, a notable example being adaptive spam filtering[22]. It is ideal for classifiers which must learn in real-time, because computation and memory requirements are constant. Perceptron-like algorithms and other adaptive classifiers could be used in health monitoring to identify clinical events whose observations are different from person to person or change over time. The algorithms developed in Chapters 4 and 6 utilize this kind of adaptation.

Alternatively, g may depend on t. For instance, consider a system that tracks its location subject to strict energy requirements. It controls a high-powered sensor that provides accurate location and velocity estimates, but quickly drains the battery when activated. It also has several other sensors which are always on. At each time t, the system must decide whether to activate its high-powered sensor, so that $g_t(o_t) \in \{$ 'activate', 'don't activate' $\}$. This decision depends on the system's estimate of the current circumstances, which in turn depends on previous observations. As before, $o_{t_1} = o_{t_2}$ does not imply that $g_{t_1}(o_{t_1}) = g_{t_2}(o_{t_2})$, because the function g is not time-invariant.

Our adaptive symptom reporting algorithm, developed in Chapter 5, is an example of

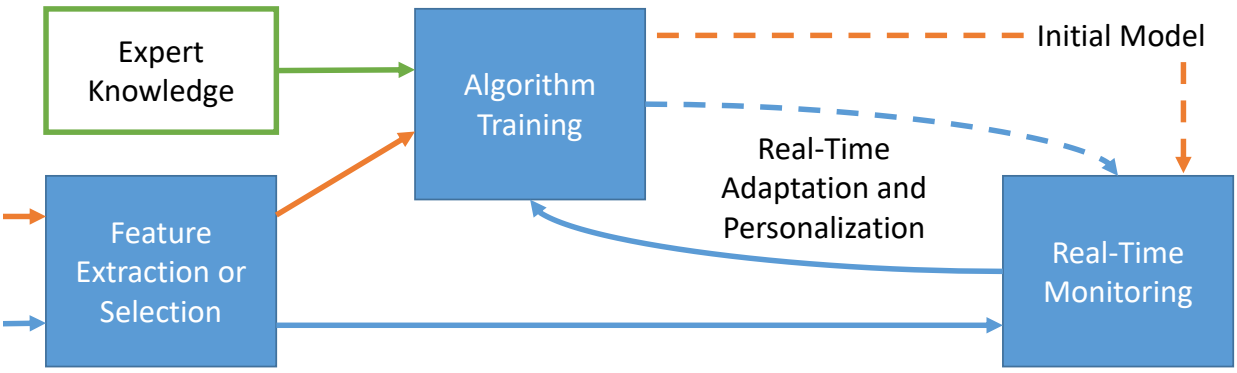


Figure 2.2: Adaptive Model Deployment

this kind of adaptation. The questions selected by the system depend on the system's information about subject disability, which in turn depends on prior observations. Chapter 6 also involves adapting decisions based on the observation history. In fact, any Partially Observed Markov Decision Process (POMDPs) satisfies the definition even when solved offline, because actions depend on the belief state, which in turn depends on $o_{1,t}$. Model adaptation (e.g. Bayes-adaptive POMDP) is sufficient but not necessary to be adaptive in our sense.

Adaptive model deployment is depicted in Figure 2.2. With respect to Figure 2.1, we have zoomed in on the right side of the diagram to emphasize algorithm deployment. Unlike the traditional approach, initial model deployment is followed by a continual cycle of real-time adaptation, which includes personalization. In some cases, the initial model is trivial, incorporating no acquired knowledge or data. The perceptron algorithm is an example of this, as the prediction vector v is typically initialized to the zero-vector. Another is the incremental learning of a Hidden Markov Model (HMM) starting with randomly initialized parameters and/or non-informative priors. When a trivial model is chosen for initial deployment, all system knowledge is acquired on the fly. More commonly, the initial model has been fitted to population-level data, and real-time adaptation serves to tailor this model to the individual being monitored.

Although our definition of adaptation is quite broad, it is important to note that the

standard machine learning techniques commonly used in health monitoring do *not* satisfy it, nor do the "person-specific" models commonly prescribed in the precision medicine literature, which are fitted to subpopulations with shared characteristics[20]. In the following sections we argue that in many health monitoring settings, adaptation is critical.

2.3 The Need for Adaptive Health Monitoring

Adaptive algorithms can offer advantages in the monitoring of *any* system, including the engineered systems for which non-adaptive algorithms have an established track record. For instance, BAPOMDPs have been proposed for monitoring progressive damage to machining tools used in manufacturing[23]. However, we argue in this section that adaptation is particularly important – sometimes *essential* – in health monitoring due unique characteristics of physiologic systems and, to a lesser extent, health care. Variability between persons is perhaps the most important, but several additional, less obvious characteristics can also limit the effectiveness of a static monitoring approach. In this section, we discuss these characteristics in turn.

2.3.1 Variability Between Persons

When monitoring engineered systems, specifications are known and consistent. Within a particular model, each object or system produced is approximately identical to the last, by design. Imprecisions in manufacturing or construction may introduce small differences between objects of the same model, but they are typically well characterized, and each monitored parameter has a normal operating range that quantifies acceptable variability, which is consistent from object to object. Different models have distinct characteristics, but they are known and planned, and the number of models is finite.

In health, the opposite is true: every person is different. "Normal" values of physiologic variables depend on health status and phenotype, which is comprised of effectively infinite

variables, most of them unknown. EEG and other physiologic time series have person-specific signatures as distinctive as fingerprints[24], and the physiologic measurements associated with the same event can vary substantially between persons.

As an example of the latter, consider panic attacks, which are defined as episodes of intense fear coupled with somatic symptoms such as increased heart rate (HR), sweating, shaking, and nausea. There are 11 of these symptoms in total, but a patient may experience as few as four of them[25]. As this definition implies, the types of symptoms associated with panic attacks are quite variable between persons, as are the physiologic measurements associated with them. One of the most consistent symptoms is heart rate elevation, yet it accompanies only a slight majority of attacks (58%), and when present, it is highly variable between persons[26]. In particular, heart rate elevation is more pronounced in patients who have had repeated attacks[27]. Galvanic skin response also exhibits variable responsiveness, with more severe changes reported in patients who have had repeated attacks[28]. This variability makes it quite challenging to accurately detect panic attacks based on population-level training data. On the other hand, symptoms are consistent for a given patient, which has made it possible to develop panic disorder subtypes[29]. This suggests that adaptive, personalized monitoring might be more accurate than traditional, population-based methods.

Another example may be found in diabetes care, where a central problem is the detection of meals, insulin boluses, and activities based on glucose levels[30]. This problem is complicated by the fact that event-related changes in glucose are highly person-specific. Indeed, the glycemic response to foods has been described as a "personal attribute", with up to five-fold variability in the area under the time/glucose curve for specific foods[31]. Responses to insulin also vary; in fact, regular exercise is encouraged in type II diabetes partly because it increases patients' sensitivity to insulin[32]. Again, an adaptive algorithm could improve detection by adjusting to person-specific factors.

More fundamentally, there is no practical, commonly used equivalent to the normal operating range for a physiologic variable. The nearest analogue is the *reference range*, but

on closer inspection the two concepts are quite different. Reference ranges for vital signs, lab values, and other clinical variables are defined empirically in terms of observed population values, often simply as 2 standard deviations about the mean. Therefore, the reference range describes variability *between* persons, not within an individual. The unresolved debate over the definition of tachycardia, meaning an unusually fast heart rate, highlights this distinction. Although a resting rate below 100 beats per minute is defined as normal, rates in the 90s have been associated with increased risk of mortality[33, 34]. In other words, that which is healthy in a small percentage of the population is pathologic in the majority. This phenomenon occurs again and again when dealing with physiologic measurements. In contrast to engineered systems, the normal operating range differs from person to person.

As we have described, the traditional approach to variability between persons is to characterize physiology for a taxonomy of subpopulations. This approach has also been applied to reference ranges. Returning to our heart rate example, population variability can be modeled as a mixture of Gaussians, with each mixture component corresponding to a particular subpopulation[34]. In general, once subpopulations have been defined, variability can be characterized for each of them, leading to "personalized" reference ranges. Because they pertain to a homogeneous subpopulation, it is tempting to suppose these ranges reflect *intra*-person variability, but in our view, this is a mistake: the size of populations has nothing to do with the distinction between inter- and intra-person variability. References ranges always pertain to the latter, whereas the normal operating range pertains to the former.

In addition to physiologic variability, there are profound behavioral differences between subjects that affect the interpretation of physiologic parameters. As a simple example, a person's heart rate depends on their sleep-wake status, which in turn depends on sleep patterns. Additional sensor measurements provide context in some cases, but may not be available in others. More critically, it is often impossible to quantify, predict, and mitigate behavioral factors prior to deployment of a monitoring system. Consider the measurement of galvanic skin response, which is affected by skin health and the use of skin care products[35].

It would be impossible to account for all such factors in advance, but an adaptive system can adjust to them after deployment.

Patients' use of monitoring devices themselves is an important source of behavioral variability that can directly affect signal features. Each device has its own unique combination of sensors, placement location(s) on the body, and attachment method(s), all of which lead to variability between patients. In our experience in gait monitoring, device orientation and attachment vary significantly even when trials are supervised by trained clinical staff. Monitoring devices for engineered systems can be precisely installed, whereas every patient wears their device differently.

Adaptive monitoring offers an elegant, unified solution to many problems of variability. If there is some degree of consistency within an individual, an adaptive algorithm can – in theory – adjust to that individual's physiology and behaviors given an appropriate framework, sufficient observation history, and sufficient processing time. Instead of relying on a reference range, intra-person variability can be characterized over time and used to detect anomalies. Admittedly this is a simplified and optimistic outlook, and adaptive algorithms certainly face many non-trivial obstacles, both practical and theoretical. However, we feel that traditional monitoring is not an option in many health applications due to the impossibility of capturing the full range of inter-person variability in any training set.

2.3.2 System Drift

Physiologic systems tend to *drift*, or change over time. System drift can be caused by behavioral changes, disease progression, aging, and environmental factors. Returning to our previous examples, a gradual change in a person's glycemic response to foods is system drift; it could be caused by exercise, weight change, or other physiologic factors. If a panic attack sufferer begins taking a beta blocker, the heart rate elevation they experience during panic attacks might decrease substantially. Drift is more permanent than intra-person variability,

which pertains to transient system fluctuations. To account for drift, a monitoring system must be able to change its behavior over time.

The concept of drift is more similar to that of inter-person variability, but pertains to the same person at two different points in time. It implies that the monitoring system's mapping from observations to outputs or decisions cannot be static. If drift occurs between times t and $t+1$, then f_t and g_t should not be equal to f_{t+1} and g_{t+1}. In other words, adaptation is necessary by definition.

While an adaptive algorithm ultimately handles inter-person variability and drift through the same mechanism, there is an important practical difference. Personalization is the adaptation of a population-level model to the individual, whereas drift requires ongoing adjustments to the personalized model over time. An adaptive algorithm begins with an "initial model", as shown in Figures 2.2. Through a cycle of real-time training updates, the model becomes personalized. When drift occurs, this same update cycle keeps the model personalized and current. If no drift were present, adaptation could cease after converging to a correct or best model, but in the presence of drift, the adaptation process must continue indefinitely.

Inter-person variability and drift have important implications for the *learning rate* of a monitoring system. Informally, the learning rate is a parameter of an adaptive algorithm that determines how much weight is placed on new information compared to previously acquired knowledge. When the learning rate is high, the system will adapt quickly, but important prior knowledge can be lost. In a traditional monitoring system, the learning rate is zero, meaning that the system does not learn from new information. Early in algorithm deployment, the learning rate should be high to allow personalization to take place. Over time, it should be decreased to make the system more resilient to transient fluctuations. When physiologic drift is present, however, the learning rate should always be greater than zero.

In the presence of drift, model adaptation can serve a double purpose. Its primary purpose is to improve algorithm outputs or behaviors, but as a secondary purpose, drift itself can also

be quantified. Since drift may reflect underlying disease progression, behavioral changes, or other clinically meaningful properties, this information may be useful to care providers. When monitoring panic attacks, for instance, adaptation might be employed first and foremost to improve panic attack detection. As added value, however, adaptation itself can serve as a measure of physiologic change. We do not explore this possibility in detail, leaving it as a potential direction of future work.

2.3.3 Subjectivity and Ground Truth

A foundational difference between health monitoring and other monitoring applications lies in the importance of subjective patient experiences. Ultimately the goal of health care is to improve patient well-being, which is distinct from the objective measurements traditionally used in health care. Growing emphasis on the patient perspective has led to the patient-centered care movement and increased use of patient-reported outcomes (PROs), which are clinically validated symptom questionnaires. Both practices are supported by an expanding body of evidence. Patient-centered care has helped patients and providers find common ground[5], and when care providers use a patient-centered approach, patients utilize health care services less often[6] with reduced associated cost[7, 8]. The NIH recognizes the importance of subjective experiences, as demonstrated by its PROMIS initiative, which is designed modernize PRO development and application[36]. Electronic PRO monitoring has improved health-related quality of life[37], and remote monitoring can support the patient narrative by giving patients unprecedented data access and control. In fact, remote monitoring of chronic conditions has been one of the most common and consistent targets for mobile interventions[9].

Subjective experiences can be incorporated in health monitoring several ways. Experiences can be the outcome of monitoring – the prognostics of the system, so to speak – which requires the system to learn a correspondence between subjective experiences and objective measurements. Subjective reports can also be inputs to the system, serving as information

about the outside world. To be consistent with our previous terminology, these inputs are the system's observations. Both cases are explored in this work and involve distinct challenges. In Chapter 5, PRO responses are used as observations, but in Chapter 6, occurrences of patient-defined events are also predicted by the algorithm. the algorithm learns to define events in the patient's own terms. By nature, the relationship between subjective experiences and objective measurements can differ dramatically from patient to patient, making inter-patient variability all the more pronounced. An extreme example is pain, which can range from absent to extreme for the same medical event depending on the patient.

The concept of "ground truth" is essential in monitoring and other machine learning applications. When monitoring engineered systems, the ground truth might be structural integrity or system wear; it is the true system state – a gold standard – that prognostics are designed to assess. Typically there is some method, albeit costly, to obtain ground truth or approximate it to high precision, and this information is critical in system training and validation. Unfortunately, ground truth is often elusive or absent in health monitoring, because it is rare to find labels that can be directly, reliably measured. As an extreme example, the outcome of interest might be a subjective experience, in which case the term "ground truth" is somewhat of a misnomer and the aforementioned considerations apply. More commonly, ground truth is a hidden physiologic parameter or the presence or absence of a disease. In these cases, the ground truth is objective in theory, but often deceptively difficult or impossible to obtain. Diagnoses themselves can be evolving definitions rather than definitive labels, particularly for diseases which lack a well-established pathogenesis. This includes our target application, multiple sclerosis[38].

While inter-subject variability makes traditional monitoring *difficult*, these concerns can make it *infeasible*. The traditional approach assumes that prognostics can be developed and understood on a population level, an assumption that does not hold when monitoring targets subjective experiences, or when ground truth is poorly understood. Consequently, population-level training is not possible in many health monitoring applications, making

personalization the *only* reasonable option. Of course, the traditional development cycle can be used to develop a person-specific algorithm, but it must be repeated for each subject. Under these conditions, an adaptive approach is much more attractive.

2.3.4 Patient Burden

Unlike engineered systems, health monitoring ultimately depends on the patient's cooperation. Sensors for engineered systems are often built into the system itself, whereas health monitoring systems must be worn or used – correctly – by the patient on a regular basis. Consequently, they must be designed to minimize patient inconvenience, or burden.

Monitoring might be burdensome in a number of ways. Sensors and devices can be large, heavy, uncomfortable, or unsightly. Some require regular charging, syncing, or updates. Many systems require patient input to document symptoms, contextualize measurements, or complete an assessment task. In our monitoring trial, discussed in Chapter 3, patients were required to wear two devices and complete five different surveys every month. The surveys took up to an hour to complete, particularly for subjects with impaired coordination or vision. This was trial designed specifically to be convenient for subjects, yet it involved several burdensome components nonetheless.

Results from mobile health trials suggest that patient burden is an important limiting factor. Adherence to mobile health interventions is often low, and it declines over time. In diabetes care, for instance, 50% of participants dropped out of a phone-based management intervention[39], and an interactive voice response service saw a decrease in call completion over a three to six month period[40]. Dietary self-monitoring adherence declined from roughly 70% to less than 20% in a mobile-enabled weight loss intervention[41], and meal entries declined by 25% between weeks one and two of a mobile intervention for irritable bowel syndrome[42]. As we report in Chapter 3, adherence also declined in our remote monitoring trial.

The importance of patient burden to system design may be under-recognized because of selection bias found in monitoring trials, including our own. As reported by the American Heart Association, the subjects who enroll in mobile health trials tend to be unusually motivated and compliant[43]. They may volunteer to benefit research or society rather than themselves, and in many cases receive monetary compensation for participating. Consequently, adherence to monitoring trials may be inflated compared to real-world deployment. With this in mind, it is all the more important to design monitoring systems to minimize patient burden. These systems must do more with less, shifting the cost/benefit ratio in patients' favor.

Adaptive algorithms can reduce burden via several mechanisms. Actively reducing sensor usage can increase the time between battery charges, and lowering memory consumption can limit syncing. The current work concerns adaptive algorithms that improve decisions or classifications by drawing upon an evolving history of observations. Chapters 5 and 6 both utilize adaptive decision-making to get the most out of each patient interaction. In Chapter 5, the system minimizes the number of patient responses needed to actively track disability status. Past responses and prior information about disability progression are used to select the most informative questions from a question bank. In Chapter 6, an active learning framework is used to optimally elicit contextual information from patients. Over time, the system learns to identify health events in patient-centric terms.

2.4 Related Work: Remote Health Interventions

Remote health monitoring can transform medical care in a number of ways. Passive monitoring systems, which simply report health information at a distance, can give care providers early warnings, decrease costs by reducing in-person visits, and extend care to persons who would not otherwise have access. Recently, however, there has been growing interest in mobile health *interventions*, which act on health information to modify behaviors. Remote interventions can

encourage healthy behaviors or manage care; in effect, the system can automate elements of routine care to further reduce costs and boost access and prevention. Even more than passive monitoring, remote interventions rely on personalization to inform their decision-making, making effective adaptation a critical system component. As we will see, the algorithms presented in Chapters 5 and 6 are designed for active monitoring systems. This last portion of the chapter summarizes recent work in remote interventions, emphasizing some commonly identified challenges. Similar challenges have been identified in our target population, as presented in Chapter 3.

A large number of studies have provided continual feedback to subjects based on physical activity monitoring. Personalized feedback from a handheld personal digital assistant (PDA) increased weekly physical activity in underactive adults[44], and personalized advice seems to be better received, though not more effective, than generic advice as a behavioral intervention[45]. Messages sent by text or email have been used to promote walking and physical activity with mixed results[46, 47, 48]. When patients with coronary heart disease were given a personalized activity program, a pedometer, and regular encouragement using tutorials and email contact, they were more physically active at follow-up compared to controls [49], and a tablet-based exercise coach improved daily step counts over 30 days in sedentary, older adults[50]. Very recently, text messages were shown to increase moderate to vigorous physical activity, as measured by a Fitbit One, though this increase did not seem to be sustained[51]. This result is typical: most studies report modest benefit, but there have been few controlled trials in unbiased samples.

Similar paradigms has been used in diabetes care to encourage patients to eat healthy and stay active. Regular, personalized text messages have lowered A1C levels in several small studies[52, 53, 54]. A larger study of 123 subjects over 3 months found subjects given a cellphone-based reminder, recommendation, and encouragement system improved their A1C and blood pressure more than matched subjects not given this system[55]. Remote monitoring of blood glucose coupled with web-based physician feedback was also associated

with lower A1C levels compared to a control group[56], and a randomized clinical trial of a several-tiered mobile health intervention in 163 patients with type 2 diabetes found that the top tier of patients improved A1C levels more than controls[57]. Compared to other mobile interventions, these diabetes care systems have been unusually effective.

Remote interventions have also been used to treatment of anxiety and panic disorder. Cognitive behavioral group treatment assisted by held-held computer was as effective as traditional treatment in reducing social phobia symptoms and behaviors[58]. Computer-assisted cognitive behavioral treatment was also effective in a randomized trial of 18 patients with panic disorder (PD)[59], a finding that was successfully replicated in 186 patients with PD[60]. A personal digital assistant has been used to improve treatment adherence in bipolar disorder[61], and several recent reviews highlight opportunities for remote monitoring in mental health care[62][63]. Recently, researchers used a smartphone-based platform to assist in treatment of child anxiety[64].

The results of these trials, which span several distinct medical specialties, have been mixed to positive. The effects of smartphone-based physical activity interventions have been "modest at best"[65]. Results in BG management have been stronger, particularly in the well-studied population of type 2 diabetics. One review reports "strong evidence" that mobile health interventions improve BG control[66], while a second review found small benefit from computer-based interventions[67]. Mobile technology is believed to be a promising platform for psychosocial intervention[68], but continued research is needed to measure its effectiveness.

Importantly, while all of the studies cited here draw upon mobile platforms, and many have incorporated some method of personalization, *none* have utilized an adaptive approach like the ones we propose. Common symptom reporting platforms ask repeated, redundant questions, and physical activity measurements fail to incorporate contextual information. The effectiveness of active monitoring has been limited in large part by the inflexible, static character of most monitoring algorithms. Results from our remote monitoring trial (Chapter 3) help to support this claim.

2.5 Summary

In this chapter, we have discussed the traditional approach to monitoring and contrasted it with an *adaptive* approach. Adaptive algorithms can be useful in a number of monitoring problems, but they are particularly well-suited to health monitoring due to its many unique characteristics. The most straightforward of these is variability between persons, which has physiologic, behavioral, and environmental components. Human physiology also changes over time, which is another obstacle for traditional algorithms easily overcome by the adaptive approach. The algorithms we propose are inherently personalized, allowing them to learn about patient experiences without relying on an objectively-defined ground truth. Adapting to patients makes it possible to minimize their effort without compromising information quality, which appears to be an important factor limiting adherence to many mobile interventions.

Adapting to patients on the fly offers an additional, critical advantage not yet emphasized: it does not rely on unprecedented data collection or the painstaking development of a patient taxonomy. Instead of learning about every possible type of patient ahead of time, an adaptive algorithm simply learns the *current* patient as it goes. While there are many technical milestones to pass, they are quite attainable compared to the massive collection, processing, and coordination of patient data required to achieve personalization using the traditional approach.

In the next chapter, we revisit these issues in the context of our target application, walking ability in multiple sclerosis. Our efforts to develop adaptive algorithms for three distinct health monitoring scenarios are presented in Chapters 4 - 6.

Chapter 3

Case Study: Walking Ability in Multiple Sclerosis

Before launching into the technical contributions of this work, we provide a case study in multiple sclerosis (MS), a chronic disease of the brain and spinal cord. While the algorithms presented herein apply broadly to health monitoring and beyond, there are several deep-rooted connections to MS. First and foremost, MS exemplifies the characteristics discussed in the previous chapter. Its symptoms and signs are both diverse and unpredictable in their type, location, and rate of progression. An adaptive approach is therefore ideal in MS, because the need for personalization and continual adjustment is greater than in other monitoring scenarios. Second, this case study is the primary source of data used to develop and validate the algorithms is Chapters 4 – 6.

More specifically, our study concerns *walking impairment* in MS. Some degree of walking impairment is experienced by almost all persons with MS, and assessing walking ability is one of the cornerstones of MS care, with a broad range of associated outcome measures. In general, remote monitoring can help to evaluate treatments and/or care, facilitate behavior modification, and empower patients. Walking impairment is therefore an ideal case study for remote monitoring, because it encompasses all of these facets. Moreover, walking assessment

involves several different types of data, including both subjective and objective observations. The three primary data types collected in our trial require different approaches to monitoring, leading to the distinctive methodologies and algorithms presented in later chapters.

We begin the chapter with a clinical background section, then describe the results of our remote monitoring study in MS. The discussion here includes insights that have inspired the algorithms in subsequent chapters, but readers interested only in technical content may skip directly to the algorithms in Chapters 4, 5, or 6.

3.1 Clinical Background

In this section, we provide clinical background to orient the reader to terminology and concepts needed to understand the target application. Clinical information appears repeatedly in the validation studies found in Chapters 4 – 6, for which this section can be used as a clinical reference. Additionally, we emphasize the general characteristics of health monitoring discussed in Chapter 2 throughout this chapter in the context of MS, further strengthening the arguments in favor of adaptive monitoring.

3.1.1 Overview of MS

MS is a disease of the central nervous system (CNS) in which the body's own immune system attacks myelin, an acellular white substance which surrounds and insulates neuronal axons. These attacks can disrupt neuronal communication through several mechanisms, including inflammation, reduced axonal conduction velocity, and destruction of the axons themselves. The resulting scars, or *lesions*, can occur anywhere in the CNS, including the spinal cord. Common symptoms include changes in vision, changes in sensation, muscle weakness, loss of coordination or balance, fatigue, incontinence, and cognitive impairment[38].

Lesions vary not only in location, but also in severity and rate of appearance: MS can progress quickly or remain stable for many years. The disease course can be categorized as

relapsing, characterized by the sudden onset of new symptoms followed by complete or partial resolution, or *progressive*, characterized by gradual appearance or worsening of symptoms. Persons with MS can transition from a relapsing course to a progressive one or experience aspects of both, prompting clinicians to define several clinical subtypes[69].

The symptoms experienced by patients depend on lesion location and severity. For example, damage to the motor cortex can lead to muscle weakness, whereas damage to the optic nerve can produce a visual deficit. Since lesions vary widely between persons, so too do their manifestations. Symptoms can be mild or severe, localized or widespread, and temporary or permanent. There are characteristic patterns, to be sure – some areas of the CNS are more commonly affected than others – but as a disease process, MS is marked by heterogeneity.

3.1.2 Walking Pathologies

Although no symptom is universal, walking impairment is a central feature of MS. Walking ability tends to degrade over time, and the typical MS patient needs a cane to walk 28 years after diagnosis [70]. The widely used Expanded Disability Status Scale (EDSS) depends heavily on walking; for example, an EDSS of 6.0 is defined by the use of a cane or crutch[71]. When walking ability is compromised, persons with MS become less independent, leading to lowered quality of life[72]. Consequently, walking impairment is reported to be the most challenging aspect of disease[73]. For these reasons, walking assessment is a core component of routine care and a common research outcome. There is even a medication designed specificaly to improve walking in MS – the "walking pill", dalfampridine – which has completed phase 3 trials[74, 75].

Like MS itself, MS-related walking impairment is quite heterogeneous, and distinct varieties of impairment are associated with particular regions of the brain. Sensory, motor, and balance deficits can all cause motor impairment: there is the unsteady, wide-based gait caused by cerebellar damage; the swinging, "circumducting" gait associated with lesions to

the contralateral motor cortex; and numerous other varieties and combinations, so that each person's walking-related symptoms are unique.

Due to the clinical importance of walking ability and relative ease of collecting activity data with commercial technologies, walking impairment is a natural target for remote monitoring. As we will see, however, transforming this data into clinically meaningful assessment is deceptively difficult. Much of this difficulty is due to the profound differences between walking pathologies and disability progression found in MS.

3.1.3 Walking Outcomes

In clinical care and research, validated disability assessments are called *outcomes*. Clinical outcomes are needed to quantify disability for treatment evaluation, care evaluation, prognosis, and clinical research. In a chronic disease setting, health monitoring consists in ongoing outcome assessment and reporting, thus understanding outcomes is essential when developing a health monitoring platform.

There are three varieties of outcomes related to walking in MS: clinician-reported neurological findings, objective assessments of performance, and patient-reported outcomes. All three appear repeatedly in this work, so we describe them in turn.

Neurological Examination

The neurological exam is a physical exam focused on neurological findings. It is administered and scored by a physician or other trained care provider. The exam includes cognitive, sensory, and motor testing, balance and coordination assessment, and qualitative gait evaluation.

The Expanded Disability Status Scale (EDSS) is a clinician-reported outcome that depends primarily on the neurological exam. It can be broken down into seven functional system scores (FSS): vision, brainstem, pyramidal, cerebellar, sensory, bowel and bladder, and cognitive. The brainstem FSS pertains to cranial nerve function, pyramidal pertains to motor function, and cerebellar pertains to coordination and balance[71]. While the limitations of the EDSS are

widely recognized, the scale remains one of the most universal measures of disease progression in MS. Is the primary exam-based MS outcome, and is useful in a variety of clinical contexts.

Timed Walks

Three timed walk tests are routinely used in MS care and research. All of them measure average gait speed, but over varying distance and duration. The timed 25-foot walk (T25FW) is the briefest of the three. A care provider leads subjects to a start line, instructs them to walk as quickly as possible to a finish line 25 feet away, and records the time elapsed between the two lines. The result can be directly reported or used to calculate speed. The T25FW is commonly used in both routine care and clinical research[76].

The six-minute walk (6MW) is a much longer walk intended to measure both walking capacity and dynamic motor fatigue[77]. Subjects are instructed to walk back and forth in a 75-foot corridor for six minutes. Instructions are given according to a script which emphasizes walking as quickly as possible while remaining safe. The 6MW is reported as the distance covered. It is commonly used in clinical research, and has been shown to induce physiologic changes related to fatigue in MS subjects[78, 79].

Lastly, the two-minute walk (2MW) was developed to be a less time-consuming alternative to the 6MW. Aside from the duration, it is identical to the 6MW. While the 2MW and 6MW are highly correlated[80], the latter is believed to be more sensitive to motor fatigue[77].

The timed walks are all measures of walking *capacity*, which might be defined as the highest speed or step rate a person is able to achieve.

Habitual Physical Activity

Habitual physical activity (HPA) measurement has attracted attention as a walking outcome because it captures the real-world impact of walking disability. Mechanical pedometers have been used to count daily steps for many years, but wearable electronic devices now make recording and reporting steps and activity data much more convenient. HPA has added

importance in MS care, because routine physical activity improves quality of life[81], and it may even limit disability progression to some degree[82].

As a newer outcome, HPA measurement is an active area of study. Daily step counts – the most commonly reported HPA statistic – are correlated to other walking outcomes[83], but also affected by behavioral factors[84]. As an alternative, HPA has been classified as moderate or vigorous physical activity based on cut-points. This approach has instrinsic limitations; for example, it does not take advantage of the sequential nature of HPA data[85]. Further, the step rates associated with physical activity vary with age, body type[86], and disability status[87], therefore a personalized approach is required to achieve accurate classification. This is part of the motivation for active event identification (Chapter 6).

3.1.4 Patient Reported Outcomes

In clinical care and research, symptoms are quantified from the patient perspective with rigorously validated patient-reported outcome measures (PROs) covering domains such as fatigue, pain, and walking ability. Use of PROs is expanding due to support from the NIH[36] and increasing emphasis on patient-centered care, which can improve outcomes[5]. PROs are used to evaluate treatments, aid in clinical decision-making, evaluate care, and more recently to inform and empower the patients themselves. As discussed in Section 2.3.3, PROs are designed to incorporate the patient perspective in health care and research, recognizing that oftentimes there is a disconnect between objective measurements and patient experiences. Indeed, many symptoms important to patients, such as pain or mild cognitive impairment, are poorly captured by traditional measures.

A patient-reported outcome (PRO) is symptom-related questionnaire that has been clinically validated by collecting many responses from the population of interest, then evaluating their psychometric properties, including reliability, validity, and dimensionality. Questions often focus on functional impairment to limit subjectivity, improving the reliability of assessment. For example, a PRO to assess strength might ask subjects to rate their

difficulty turning a door handle. In MS alone, PROs are used to assess walking, fatigue, depression, pain, quality of life, and many other aspects of the disease.

This work relies heavily on the MS Walking Scale and Modified Fatigue Impact Scale, which are described below. In addition, we utilize the Godin Leisure-Time Exercise Questionnaire (GLTEQ), in which patients self-report their physical activity[88], and the MS Performance Scales (PS), which assess 11 different symptom domains using single-question rating scales[89]. PROs are the focus of Chapter 5, which develops an adaptive approach to remote PRO collection.

The MS Walking Scale

The 12-item MS Walking Scale (MSWS-12) was developed in 2003 as a comprehensive measure of mobility incorporating the patient perspective[90]. Since then, the MSWS-12 has become the primary patient-reported outcome (PRO) measure of walking ability in MS, with 205 citations of the primary manuscript in Web of Science. It was the only PRO for MS-related walking impairment identified by a 2012 review[91], and it was used to validate response criteria in phase 2 and 3 trials of dalfampridine, the "walking pill"[92, 93].

The MSWS-12 asks patients to rate the severity of their walking disability using a five-point rating scale. Each item asks "In the past 2 weeks, how much has your MS...", followed by a specific symptom or functional change. For example, the first item asks how much MS has "Limited your ability to walk". All items utilize the following rating scale: 1 (Not at all), 2 (A little), 3 (Moderately), 4 (Quite a bit), and 5 (Extremely)[90]. Because of its importance to this work, the full MSWS-12 is depicted in Figure 3.1.

To score the MSWS-12, items are summed together and linearly normalized to fall between 0 and 100. This approach has several recognized shortcomings[94]. All items are weighted equally, implying that they are equally important indicators of disability, which does not appear to be the case, and the scale's minimal clinically important difference (MCID) is high. Recently, item response theory (IRT) has been applied to the MSWS-12 to address these

Study # _____
Subject # _____

Multiple Sclerosis Walking Scale (MSWS-12)

- These questions ask about ***limitations to your walking*** **due to MS during the past two weeks.**
- For each statement, **please *circle* the *one* number that best describes your degree of limitations.**
- Please answer ***all*** **questions even if some seem rather similar to others, or seem irrelevant to you.**
- ***If you cannot walk at all*****, please tick this box** ☐

In the past two weeks, how much has your MS…	Not at all	A little	Moderately	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking indoors? (e.g. holding on to furniture, using a stick, etc.)	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors? (e.g. using a stick, a frame, etc.)	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Figure 3.1: The MS Walking Scale

concerns[95, 96], as discussed in Chapter 5, but summed scoring remains the most common approach.

The Modified Fatigue Impact Scale

The 21-item Modified Fatigue Impact Scale (MFIS) is one of the components of the MS Quality of Life Inventory[97]. Like the MSWS-12, it uses a five-point rating scale to assess fatigue from the patient perspective. MFIS items ask about the frequency of fatigue-related

symptoms, scored from 0 to 4: 0 (Never), 1 (Rarely), 2 (Sometimes), 3 (Often), and 4 (Almost Always).

Unlike the MSWS-12, the MFIS can be broken into three subscales pertaining to physical fatigue, cognitive fatigue, and psychosocial fatigue, and psychometric analysis has verified the scale's multidimensional nature[98]. The MFIS is scored by summing all items without normalization. The three subscales are scored by summing all items on that scale.

3.1.5 Minimal Clinically Important Difference

The concept of a minimal clinically important difference (MCID) is important in outcomes research, and therefore in health monitoring. Intuitively, MCID is the smallest change in an outcome measure that has clinical significance, where significance can be defined in many different ways. Most often clinical significance is defined as patient-perceived benefit[99], though it has also been defined in terms of other outcomes measuring a similar construct[84]. In this work, the MCID for the MSWS-12 is defined as the smallest change consistently associated with changes in objective walking outcomes.

The MCID is used in Chapter 5 to contextualize the measurement error of adaptive symptom reporting. We consider errors larger than the MCID to be unacceptable, whereas errors much smaller than the MCID are inconsequential. The concept of a MCID goes hand in hand with health monitoring: when monitoring systems are used to estimate or predict clinically relevant values, the MCID should guide the interpretation of error magnitude.

3.2 Pilot Study: Remote Monitoring in MS

To learn about the real-world challenges encountered in health monitoring and explore potential benefits, we conducted a field study in 31 participants with MS. This study is the primary motivation for the work proposed here, and also a rich source of data needed to validate the methods we describe.

MS is a promising target for remote monitoring because of the progressive nature of the disease, the unpredictability of relapses, and the importance of ongoing assessment. Over 80% of persons with MS use the internet on a weekly basis and 90% can navigate an electronic health record[10], making internet-based interventions technically feasible in MS. Consequently, a growing number of mobile interventions have focused on the MS population. An informational website for MS patients and their families received positive feedback[100], and MSDialog, a mobile and web-based patient-reported outcome (PRO) platform, was well-received by patients and care providers[101]. For these reasons, technology-based monitoring is part of a consensus vision for the future of MS care[8].

While adherence to remote monitoring and interventions had been studied in other populations, comparatively little was known about remote monitoring in MS. The aforementioned studies were limited both in scope and in their presentation of results, which focused primarily on positive experiences reported by patients. The current study explores remote monitoring in MS more thoroughly by evaluating our monitoring platform in terms of feasibility, reliability, adherence, and subject-perceived benefits.

3.2.1 Study Procedures

This study was approved by the University of Virginia (UVA) Institutional Review Board for Health Sciences Research. Subjects were recruited from the UVA Department of Neurology patient population, and written consent was obtained. Recruited subjects had (1) clinically definite MS[102]; (2) the ability to walk, possibly with an assistive device (e.g. cane, walker); (3) high-speed internet access via computer or tablet in the home; and (4) self-assessed ability to complete web-based questionnaires.

Subjects participated for approximately six months (24 weeks), with in-person assessment at baseline and six months. Baseline and six-month assessments included collection of demographic information; neurologic exam by Neurostatus-certified staff for EDSS staging; all three timed walk tests (6MW, 2MW, T25FW); and completion of several PROs, including

the MSWS-12, MFIS, Godin Leisure Time Exercise Questionnaire (GLTEQ)[88], Patient-Determined Disease Steps (PDDS)[103], and Performance Scales (PS) covering 11 distinct symptom domains such as mobility and vision[89].

Two accelerometer-based devices, Fitbit and ActiGraph, were used to record daily physical activity for one week per month. ActiGraph was also used to collect inertial data during the 6MW and 2MW. Habitual physical activity (HPA) data collected via ActiGraph are used to validate the algorithms presented in Chapter 6. Inertial data from the 6MW is used to validate the work presented in Chapter 4.

Web-Based Patient-Reported Outcome (wbPRO) Collection

A UVA-hosted web portal allowed subjects to report symptoms from home and view their symptom history. The web portal was created specifically for this study. Figure 3.2 shows the navigation page, which features a global "How are you feeling today?" (HAYFT) question, scored from 1 to 10, and links to the following four questionnaires: MSWS-12, MFIS, GLTEQ, and PS. Figure 3.3 depicts the web-based MFIS linked from the navigation page. The 11 PS were adapted for the portal and labeled as the "Symptom Tracker", as seen in Figure 3.2. The history of responses to HAYFT, MSWS-12, MFIS, GLTEQ, and each PS could be viewed as a graph or a table. A dedicated "Symptom Tracker" page allowed subjects to compare severity between symptoms and view recent trends. Subjects were oriented to the web portal at baseline visit with a 15-minute, face-to-face tour and tutorial.

Subjects were required to complete each of the five questionnaires at least once per month. Additional use of the web portal was encouraged but not required. Subjects rated the utility of the web portal and provided free-response feedback at the six month visit.

3.2.2 Participant Demographics

Thirty-one subjects completed all study requirements. By design, recruited subjects were evenly dispersed across the disability spectrum up to an EDSS of 6.5. Nine had mild disability

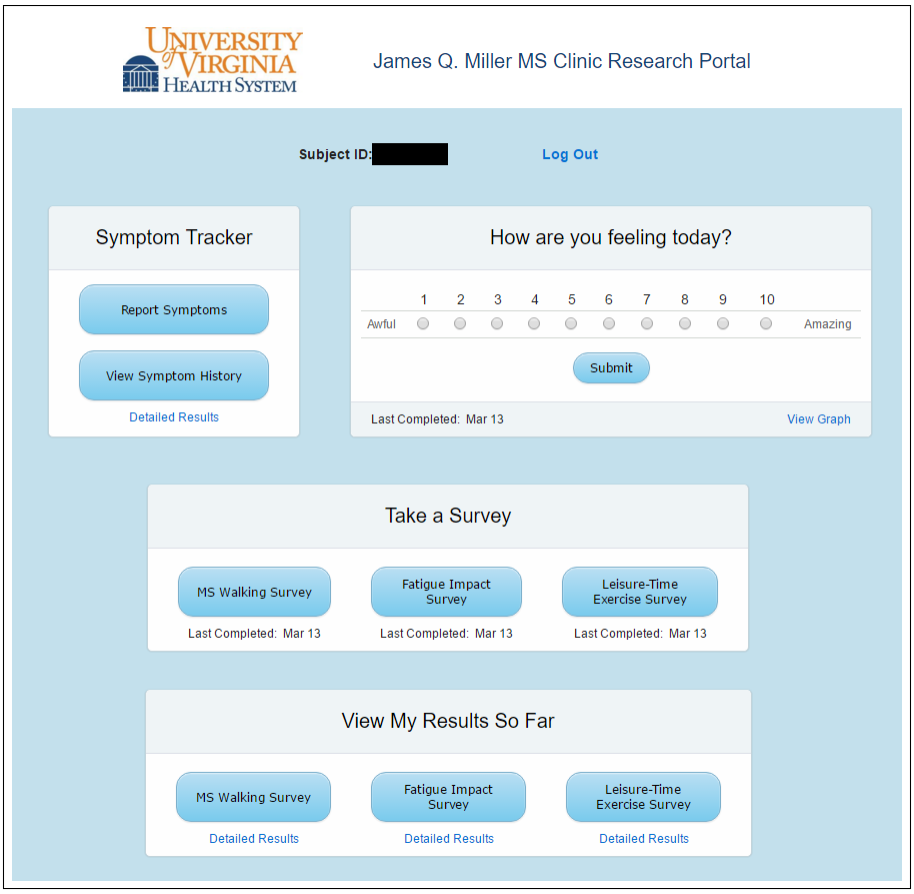


Figure 3.2: Web Portal Landing Page

(EDSS 0 to 2.5), 11 had moderate disability (EDSS 3 to 4.5), and 10 had severe disability (EDSS 5 to 6.5). All subjects were ambulatory, but 13 used assistive devices inside the home (7 cane; 4 walker; 2 hand bars), and three additional subjects used a cane outside of the home. Demographics and outcome measures at baseline visit are summarized in Table 3.1. Subjects were 93.5% female (29/31), with a median age of 48 years (range: 27 years to 61 years). No subjects experienced a MS relapse during the study, and the largest EDSS progression was 1.5 (median = 0, range = -2 to 1.5, IQR = -0.5 to 0.5).

3.2.3 Feasibility and Reliability

Only 12.9% of subjects had technical difficulty with the web portal log-in process. In each case, this was solved by phone call with the study coordinator. Most subjects (87.1%)

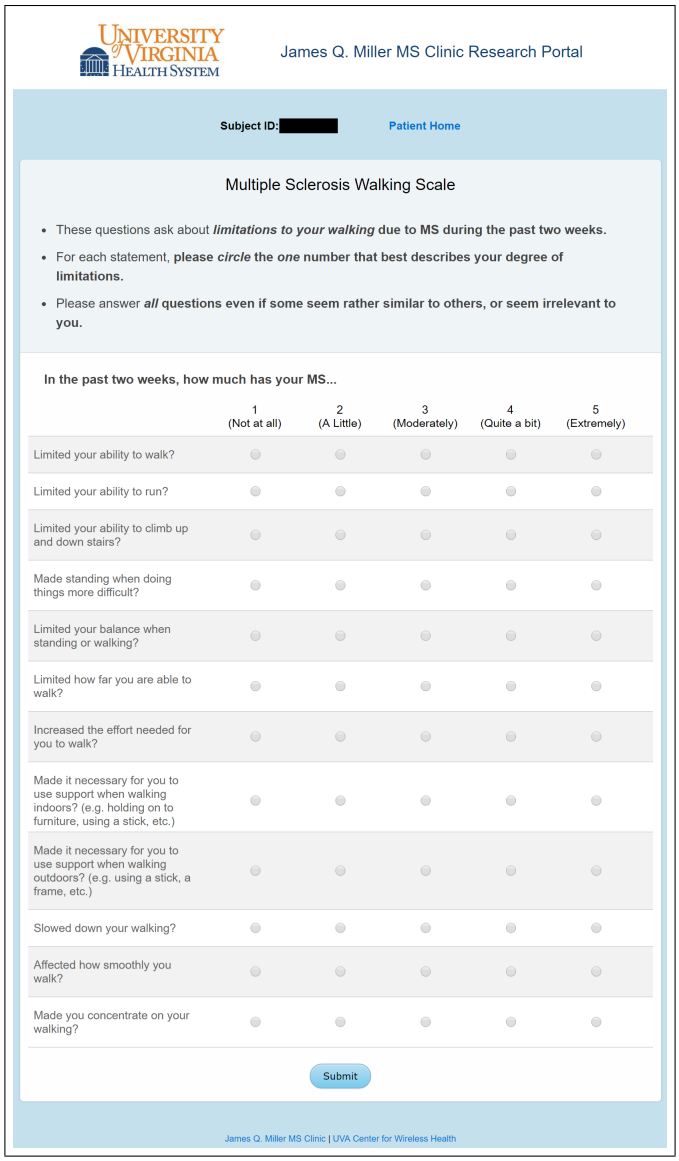


Figure 3.3: Web-based MS Walking Scale (MSWS-12)

completed all questionnaires the first month per study requirements. In total, 77.4% had no difficulty with the web portal, and 22.6% had some difficulty. These groups were examined for shared characteristics such as age, disease severity, disease subtype, employment status and education, but contrary to expectations, none were identified.

Altogether, technical difficulties occurred at a rates consistent with other, similar studies. For example, Haase et al. found that 83% of patients can quickly adjust to new software and 87% report weekly internet use[104], similar to the 87% of our cohort who did not require

Table 3.1: Subject Demographics and Disability Outcomes

	Median	IQR	Range
Age	48	44 - 56	27 - 61
EDSS	3.5	2.5 - 6	2 - 6.5
PDDS	3	1 - 4.5	0 - 7
IADL Score	2	0 - 4	0 - 9
MSWS-12 Score	31.3	8.9 - 79.7	0 - 100
MFIS Total	36	23 - 49	0 - 67
T25FW	5.0	3.9 - 12.7	3.3 - 67.4

EDSS = Expanded Disability Status Scale; PDDS = Patient Determined Disease Steps; IADL = Instrumental Activities of Daily Living; MSWS-12 = MS Walking Scale; MFIS = Modified Fatigue Impact Scale; T25FW = Timed 25-foot Walk

Table 3.2: Reliability of wbPRO collection based on correlation to standard PROs

	Correlation between PROs (r)		
Encounter	MSWS-12	MFIS	Symptom Tracker
Baseline (web-based/standard)	0.973	0.944	0.922
Six Months (web-based/standard)	0.973	0.912	0.959
Baseline vs Six Months (both standard)	0.933	0.780	0.886
Baseline vs Six Months (both web-based)	0.957	0.794	0.941
Baseline (standard) vs Six Months (web-based)	0.936	0.816	0.801
Baseline (web-based) vs Six Months (standard)	0.944	0.789	0.934

technical assistance.

Table 3.2 quantifies the reliability of wbPROs, measured as the correlation between wbPROs and standard PROs. First, standard PROs from baseline and six month visits were compared to wbPROs from the first and last months, respectively. These completions were up to two weeks apart, so there is some longitudinal component to the comparison. Standard and web-based MSWS-12 scores were most highly correlated ($r = 0.973$ at both the baseline and six month visits). MFIS totals were least highly correlated on average, though correlation was still strong ($r = 0.944$ and $r = 0.912$ at baseline visit and six month visit, respectively). For comparison, the baseline PROs and six month PROs are less strongly correlated; in fact, these values are lower than their counterparts when comparing month 1 wbPROs with month 6 wbPROs. No subjects reported difficulty with the web-based questionnaires themselves.

While not surprising, this result reinforces a foundational assumption of remote symptom reporting, namely that remote symptom reports may be trusted. Our wbPROs faithfully reproduced the standard PROs, whereas a less strict web-based interpretation might have produced different results. No subjects reported difficulty with the questionnaires themselves. Some variability between standard PROs and web-based PROs was expected given the subjective nature of the questionnaires and the time delay between completions (up to 2 weeks). These cross-sectional correlations were stronger than the corresponding longitudinal correlations, suggesting that variability was within expected limits. Moreover, longitudinal correlations were similar regardless of the modality (paper/web-based). In the MS population – and more than likely other clinical populations as well – it is reasonable to utilize symptom reports as inputs to a monitoring system.

3.2.4 Adherence

Subject retention was high, with only two participants dropping out – initially there were 33 subjects – and the majority of the subjects' monthly requirements were met. Subjects wore Fitbit *much* more often than required, with 17 of the 31 subjects wearing it for at least 14 days per month, double the requirement. However, completion rates (including Fitbit wear) declined over the course of the study.

Figure 3.4 shows average monthly completions for each questionnaire and days worn for Fitbit among all subjects in the first half of the study (months 1 – 3) compared to the second half (months 4 – 6). A majority of subjects met requirements in the first half. In the second half, a majority met HAYFT and Fitbit requirements but failed to meet MSWS-12, MFIS, PS, and GLTEQ requirements. The median number of completions for the MSWS-12 and MFIS was 2, one below the minimum requirement, and the decline in total completions was significant by paired t-test for all questionnaires (HAYFT $p = 0.001$; MSWS-12 $p = 0.031$; MFIS $p = 0.008$; PS $p = 0.011$; GLTEQ $p = 0.004$).

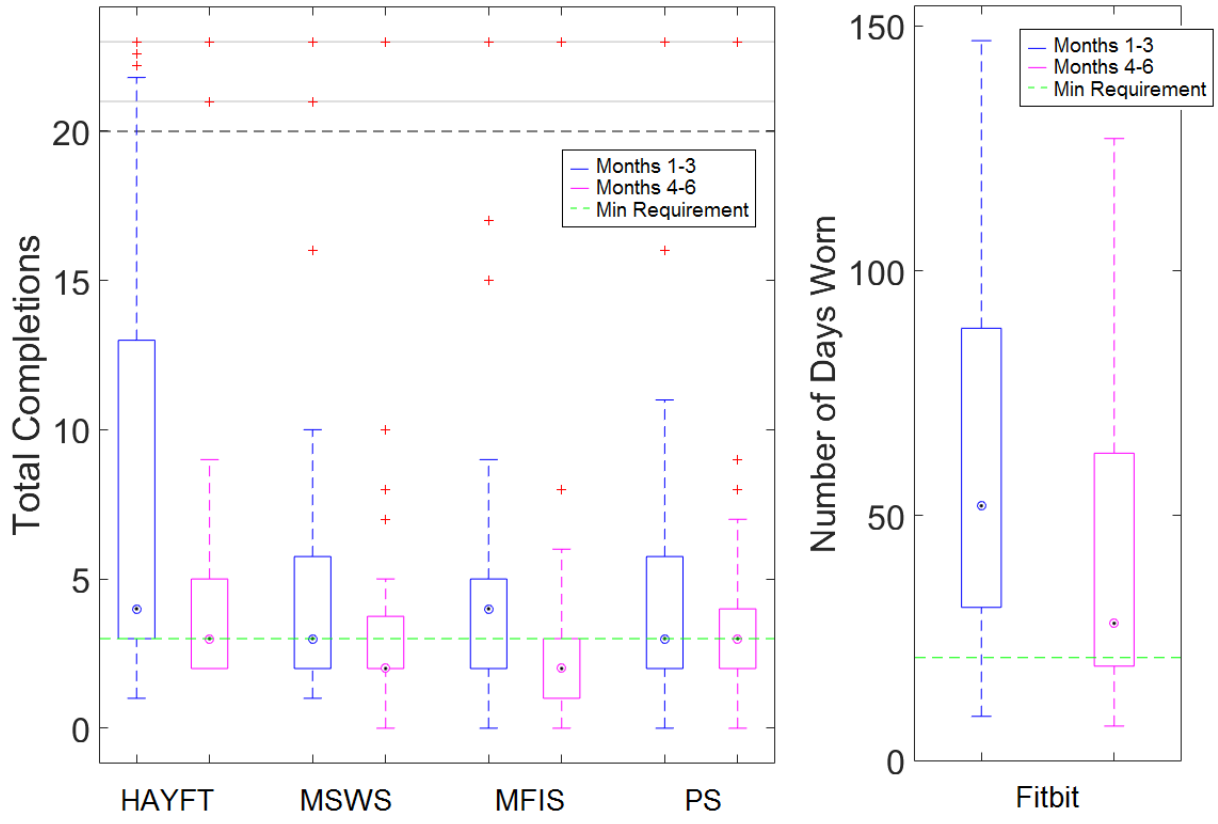


Figure 3.4: Participation Declined Over Time

There were several notable outliers in terms of completion rates: for example, one subject completed HAYFT every day (185 completions). Low household income (< $10,000) was the only demographic or disease-related factor significantly associated ($p < 0.001$) with "super-completer" status (> 100 total completions).

Subjects who exceeded requirements (> 1 completion per month per survey, on average) had higher EDSS ($p = 0.026$) than others by Mann-Whitney rank sum test. Figure 3.5 shows differences in median FSS and EDSS between these groups. Subjects who exceeded requirements also had higher FSS, with significant differences found in vision ($p = 0.049$), cerebellar ($p = 0.006$), sensory ($p = 0.032$), and bowel/bladder ($p = 0.023$) scores by Mann-Whitney rank sum test. The trend was also present in the remaining FSS, but it did not reach statistical significance.

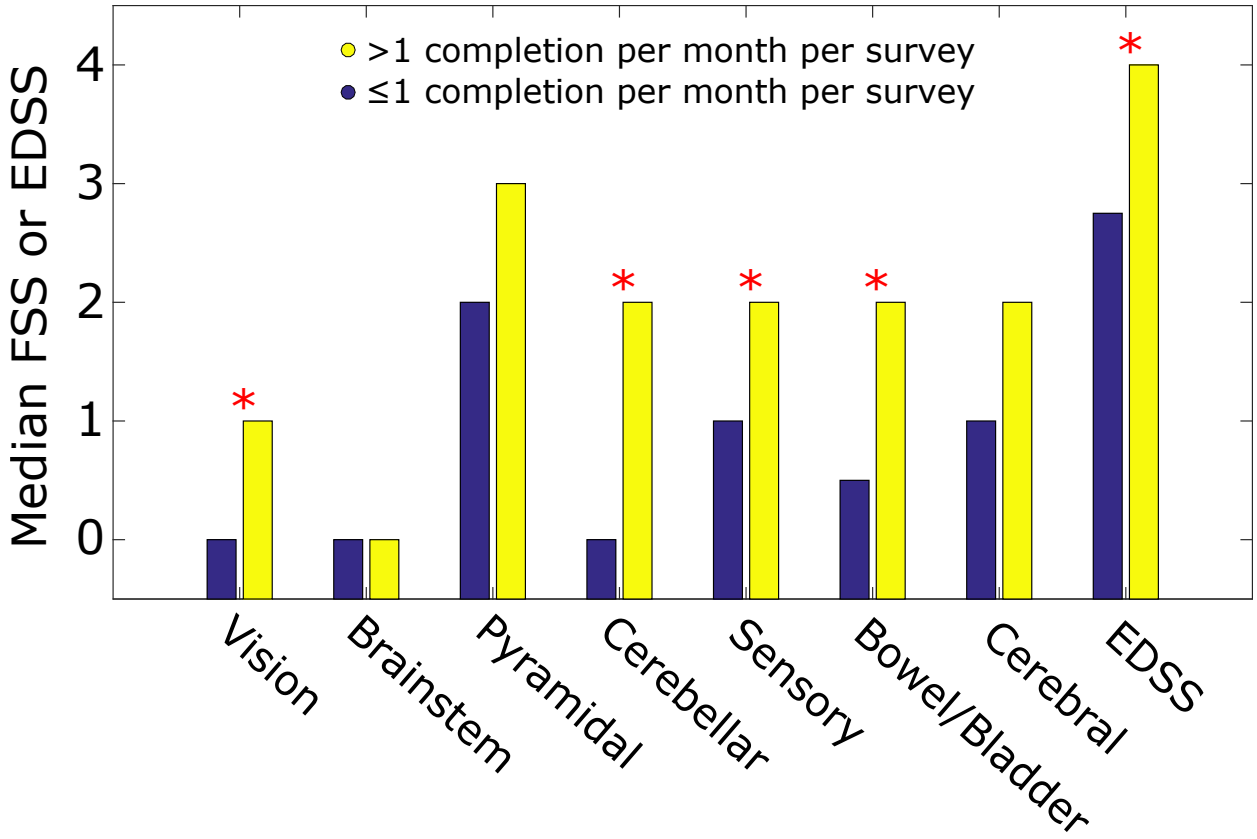


Figure 3.5: Patients with Higher Disability Completed More Surveys

Declining completion rates are a major obstacle to mHealth adoption that future interventions must address. Much like other remote interventions[105, 39, 40, 41], our subjects were initially engaged with high completion rates, but engagement dropped as subjects became more realistic about intervention benefits. Combined with the feasibility and reliability results, this strongly suggests that patients find monitoring burdensome. Even though our patients were *able* to comply with study requirements, many chose not to, particularly in the second three months. This conclusion is supported by free response feedback, as discussed in the next section.

The association between disease severity and completion rates is promising, as subjects with more severe symptoms tended to report them more often. Nevertheless, low completion rates point to a need for more innovative, individualized approaches to monitoring. Patients are excited about these technologies at first, but appear to conclude over time that sustained

effort is not worthwhile. As the evidence both here and in the literature demonstrates, subject interest tends to fade over weeks to months. Thus any technology demanding consistent, prolonged effort from the patient may be doomed to fail as patients resume their old habits.

Many studies have combated low compliance by opting for short form PROs, reducing patient burden. This strategy was tested in our study through the "Symptom Tracker", which allowed patients to rate individual symptoms using clinically validated scales. Ultimately, however, there is a trade-off between PRO length and score validity. Instead of reducing the number of questions, we favor intelligent question selection based on prior information about the patient, as developed in Chapter 5.

3.2.5 Subject Feedback

Approximately 60% of subjects viewed all results monthly. This was the minimum requirement for completion, meaning that almost 40% did not meet the requirement. Very few subjects viewed results weekly or daily; for example, 12.9% of subjects viewed the Symptom Tracker on a weekly basis or more. The GLTEQ was viewed least often. Most subjects felt that most of the questionnaires were "Moderately" useful in helping them understand their MS, though GLTEQ was more often seen as "Not at all" useful. The Symptom Tracker was perceived as the most useful, with 25.8% reporting it as "Quite a bit" useful. One subject reported the MSWS-12 as "Extremely" useful, but this was an exception to the rule. Subjects also rated the usefulness of the surveys and Fitbit when communicating with their care provider, with results shown in Figure 3.6. Several subjects found the Fitbit "Extremely" useful, but more of them did not. Not surprisingly, the frequency of survey completion and Fitbit wear correlated with perceived utility ($r = 0.653$ for Fitbit). In the words of one participant, "The Fitbit was useful in challenging myself to walk more and a great way to share the info with doctor". The Symptom Tracker, MSWS-12, and MFIS were perceived as "Moderately" useful, whereas HAYFT and GLTEQ were "Not at all" useful. In general, fewer subjects saw the questionnaires as useful when communicating compared to understanding their MS. One

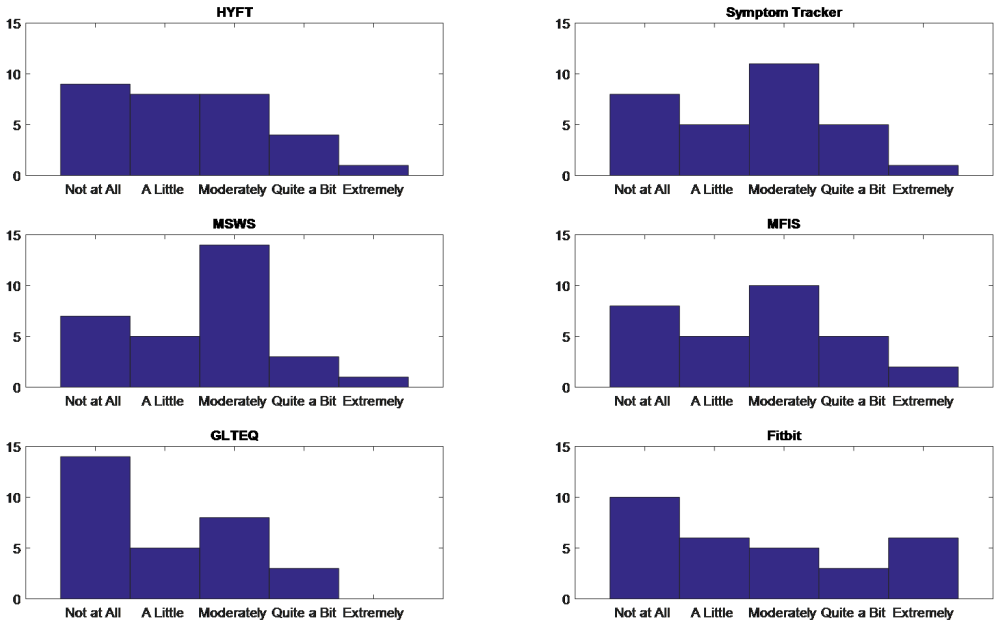


Figure 3.6: Utility when Communicating with Care Provider

notable exception was the MFIS, which 16.7% of subjects found "Quite a bit" useful when communicating with their provider.

Frequency of viewing was positively correlated with perceived utility on both questions: when subjects viewed results more often, they were more likely to report them as useful. This association reached statistical significance for HAYFT, Symptom Tracker, MFIS, and GLTEQ, but not MSWS-12. There were three clear themes identified in the free response feedback. First, a majority of subjects (51.6%) commented that monitoring helped them understand aspects of their disease. For example, one subject wrote "I found it very useful because I could see what days were good and why and what days were bad and why". A smaller but substantial portion (16.1%) said the symptom history would be useful when communicating with a care provider, for instance: "I think it would be more useful to the doctor to have an overview of the symptoms between office visits". Lastly, subjects commented on the lack of patient-specific content and/or timing in symptom questionnaires (16.1%). One said "Because my walking is so bad I felt a lot did not pertain to me", while another said "I probably would

not answer the questions within periods where my symptoms remained largely unchanged."

These results point to several causes of poor adherence. First, most subjects simply did not find symptom reporting to be very useful. Adherence to GLTEQ was poor because it was "Not at All" useful, whereas the Fitbit, which many subjects found "Extremely" useful, had better adherence than any of the questionnaires. Importantly, PROs were perceived as more useful than HAYFT, suggesting that subjects appreciate the clinical value and precision of these tools. On closer inspection, we see that perceived utility varies between subjects. When disease is more severe, symptom reporting is more relevant, so subjects complete PROs more often. Further, the subjects themselves explain in their free responses that when questions are appropriate and timely, they are more useful.

Taken together, our observations suggest that subjects make reasonable choices about the effort they put in. They are compliant when symptom reporting has value, but less so when it is inappropriate or uninformative. Our symptom reporting system could not distinguish between these scenarios, because it did not adapt to subjects. One observation drives home the point: HAYFT was perceived as less useful than other surveys, yet it was completed more often. Since this one-question survey was easy to complete, its cost/benefit ratio was low even though it was not very useful. The implication for monitoring is that an appropriate cost/benefit ratio must be maintained. An intelligent monitoring system should elicit responses only when they are likely to be informative, otherwise patients will not comply. This is the motivation behind adaptive symptom reporting, which is presented in Chapter 5.

3.2.6 Walking Capacity and Activity Behaviors

To better understand the relationship between HPA and walking capacity, daily step counts were compared to other measures of walking, including the six minute walk (6MW). Correlation could only be assessed on a group level, because walking ability did not change significantly in any subject. Figure 3.7 contains boxplots of daily Fitbit step counts for each subject.

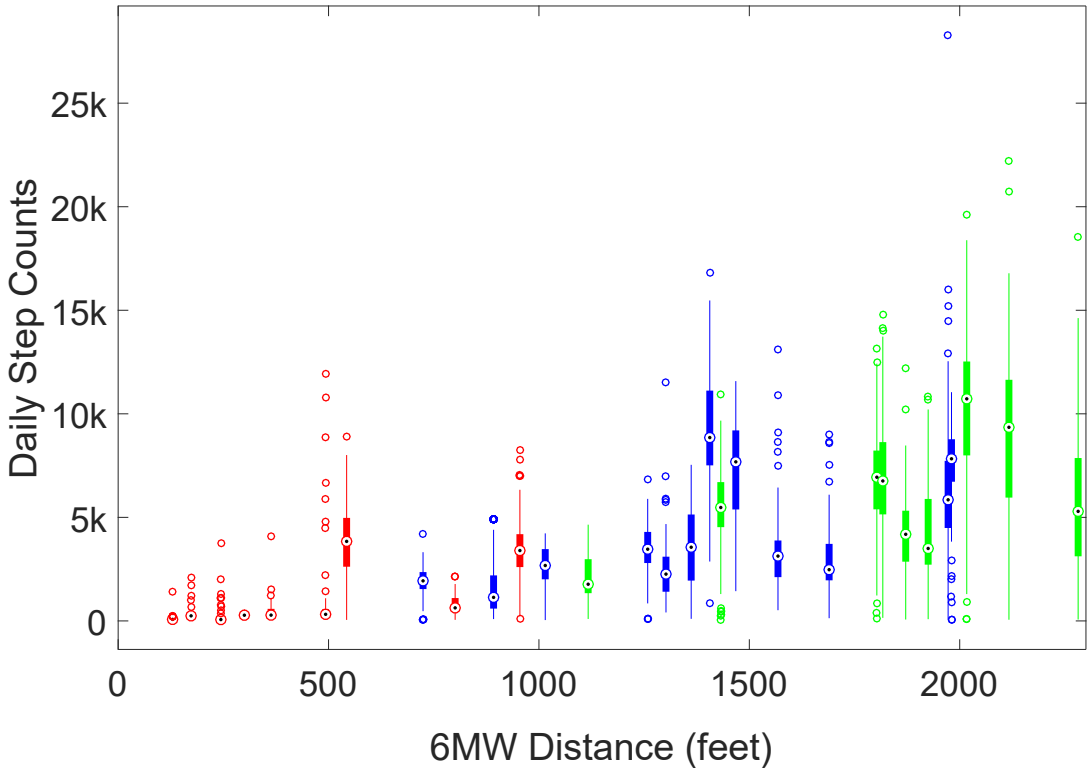


Figure 3.7: Boxplots of daily step counts positioned by the subject's 6MW distance

Each boxplot is colored according to that subject's their disability level: the red boxplots (severe subjects) are toward the left (low 6MW distance) and they are not very high (low daily step counts), while the green box plots (mild subjects) are toward the right with higher counts. A relationship is present, but it is weakened by individual behaviors: some subjects who are able to walk choose not to, and others who are less able choose to walk as much as they can. Activity levels should be contextualized with information about walking capacity, but person-specific patterns not related to disability are also important.

As Figure 3.7 shows, daily step counts are an imprecise measure of walking capacity. Indeed, daily counts do not reliably change when patient-reported walking ability changes[84], and they explain less than half of the variance in objective walking outcomes[106, 107]. This is perhaps not surprising, as total daily activity is affected by a multitude of personal, environmental, and social factors in healthy adults[108, 109]. Unfortunately, they are not

a reliable measure of behavior, either: daily counts are not significantly correlated with self-reported physical activity in MS[110]. In MS, both disability and behavioral factors have an influence on daily activity. Thus steps counts are a coarse measure of both walking capacity and walking behavior, but a precise measure of neither.

In Chapter 6, we explore this phenomenon in more detail. Active event identification is used to accurately determine when subjects are walking or running, helping us to uncouple walking capacity from walking behaviors. Similarly, event identification is needed in many other health monitoring applications to contextualize data from a variety of sensors. The algorithms for event identification presented in Chapter 6 were motivated by HPA in MS and validated using our ActiGraph dataset.

3.3 Summary

Our pilot study contributes to the growing evidence base for remote monitoring in MS care by evaluating web-based PRO collection in terms of feasibility, reliability, adherence over a six month period, and subject-perceived benefits. Persons with MS are able to participate technology-based monitoring, but their willingness to do so could be improved through personalization and adaptation. More specifically, an adaptive approach can be used to optimize patient interactions and minimize unnecessary burden. While remote monitoring has potential to facilitate patient-centered care and improve engagement, reaching it will require innovative solutions to combat poor compliance and adherence.

Each of the arguments for adaptive health monitoring from Chapter 2 can be applied to the monitoring of walking ability in MS. Indeed, MS inspired many of these arguments, and it epitomizes the characteristics we have described. This chapter has used an extended example to illustrate differences between health monitoring and other monitoring scenarios. In this application, adaptive algorithms are the most – arguably the *only* – reasonable approach to monitoring given the virtually unlimited heterogeneity of MS-associated walking impairment.

The other arguments for adaptation presented in Chapter 2, including system drift and the problem of subjectivity, are equally potent in MS.

With this background in place, we begin the technical content of this work: algorithm development. Chapter 4 tackles the problems of heterogeneity and patient burden through a personalized signal monitoring algorithm validated using inertial data. Our algorithms naturally incorporate personalization by evaluating the current signal with respect to previously encountered templates. Burden is minimized by incorporating rotation, scale, and offset invariance, which allows us to relax constraints on device attachment and patient clothing and footwear. Chapter 5 tackles burden and the problem of subjectivity through an algorithm designed specifically for PRO-based monitoring. Prior information about disability progression and PRO responses is used to optimize question selection in real-time, minimizing the questions patients must answer to maintain accurate disability estimates. Chapter 6, the culmination of this work, tackles all of the issues presented in Chapter 2 in the context of health event detection and identification. Our algorithm learns to characterize and identify medically relevant events in patient-centered terms, bridging the gap between objective measurements and subjective patient experiences. The approach is validated using HPA data from persons with MS. In each case, adaptation provides important advantages compared to a more traditional approach.

Chapter 4

Physiologic Signal Monitoring

This chapter presents our first algorithm, a personalized approach to physiologic signal monitoring based on the popular dynamic time warping (DTW) algorithm. DTW uses dynamic programming to optimally align two sequences. In intuitive terms, it measures the similarity between a sequence of interest and a second, template sequence. Our approach is designed to overcome the problems of heterogeneity and drift, as described in Chapter 2, by using each subject as his or her own control. This is accomplished by choosing baseline measurements from the same subject as the template sequences. In this way, DTW-based monitoring is inherently personalized: rather than tracking specific, pre-defined signal features, it directly tracks accumulated changes to the signal over time. We believe this strategy is superior to traditional, feature-based monitoring for tracking changes to heterogeneous or poorly characterized signals, including MS gait pathology, because it makes very few assumptions about the signal itself or the nature of changes which may appear.

This chapter is more application-specific than subsequent chapters. Our DTW modifications can be applied to a wide variety of health monitoring scenarios, but they were developed with inertial data in mind. Nevertheless, this work and its potential applications and extensions are excellent examples of adaptive and/or personalized monitoring.

In adaptive event identification, presented in Chapter 6, the system's internal model can

drift to some degree, but the algorithm's goal is to identify states, not track model drift. In this chapter, on the other hand, the goal is to quantify the drift itself. As a result, it should be applied to physiologic signals that can change either transiently or progressively as the result of underlying pathology. From a statistical standpoint, Chapter 6 assumes that samples are independent and identically distributed given an underlying state, whereas the algorithms discussed in this chapter are most appropriate when successive samples are highly correlated.

We begin with background on the DTW algorithm and a brief review of applications of DTW relevant to this work. Alternative approaches to qualitative gait monitoring are also discussed. With this background in place, we develop DTW as the core routine of a physiologic signal monitoring algorithm via three distinct modifications. The first is a method for applying DTW to signals that occur infrequently in a time series, requiring identification and extraction. The second is the Warp Score, an alternative DTW statistic that has proven useful in our target application. We suspect its usefulness reflects underlying differences between health monitoring and more typical DTW applications, and we believe the Warp Score may be equally useful in other health applications. The third and most substantial modification is rotation, scale, and offset invariant DTW (RSOI-DTW), which is designed for real-world monitoring of inertial data. Rotation invariance allows the algorithm to match gait cycles regardless of device orientation, which is a necessity in many non-supervised (i.e. not supervised by a care provider) health monitoring scenarios due to the possibility of patient error in device placement. Scale and offset invariance help to mitigate false alarms caused by changes in speed, surface, footwear, and other variables.

The methodology and mathematics behind all three modifications are presented Section 4.2, and corresponding validation studies are presented in Section 4.3. In these studies, we judiciously select specific modifications to illustrate their value.

4.1 Background

DTW, the core subroutine of our approach, may be viewed as a distance measure: it directly compares the current signal to a previously observed template sequence. Basing monitoring on a distance measure has several advantages over traditional feature engineering in many applications. It requires no prior knowledge about relevant features of interest, and is not specific to a particular signal or pathology. When signal features are used to monitor walking quality, they must be chosen carefully to minimize the loss of important information. This can be difficult or impossible, as salient features vary between diseases, pathologies, and persons; what is abnormal in one person might be typical in another, and different pathologic processes – say, hemiparesis and Parkinsonian gait – often have little in common. Consequently, extensive data collection, expert knowledge, and iterative refinement are required to develop a feature-based monitoring algorithm.

Further, the distance measure approach uses each subject as his or her own control, eliminating inaccuracy due to differences in sensors, device placement, and physical characteristics. Each device is different, with its own unique combination of sensor types, placement location(s) on the body, and attachment method(s). In the face of these multiple sources of heterogeneity, any broadly applicable technique must be capable of "tuning" itself to the situation at hand. Using DTW, which is differential by nature, is a simple, elegant solution.

Note that this is a specific case of the argument for adaptive algorithms presented in Chapter 2, because a distance measure is naturally adaptive by our definition: it depends on previously stored observations, as explained in the next section. In Section 4.2.4, we speculate about truly adaptive extensions of DTW which continually update a set of template sequences in a principled manner.

The significance of specific patterns observed by our monitoring algorithm depends on the application. In some scenarios, including gait monitoring, similarity to baseline sequences should return if normal physiology can be restored. In chronic, progressive conditions, on the other hand, the distance to baseline will increase on an ongoing basis, and new baselines

should be established to more precisely detect further degradation. A heart arrhythmia would have high distance to normal, sinus rhythm, for example, but it would quickly decrease when normal rhythm was restored.

4.1.1 Dynamic Time Warping (DTW) Basics

DTW is a common algorithm for curve registration, in which sequences are aligned to minimize the Euclidean distance between them. Here we offer a brief, formal description of the DTW algorithm. For a more comprehensive treatment, we recommend the influential work of Keogh and Ratanamahatana[111].

The DTW algorithm takes two sequences $X_{1,m}$ and $Y_{1,n}$ as inputs and returns a measure of similarity d_{DTW}, often called the DTW distance, between them. In this work, $X_{1,m}$ is a previously observed template sequence – we often refer to it as a baseline measurement – and $Y_{1,n}$ are the current observations which are being evaluated. The x_i and y_i are three dimensional acceleration vectors in our validation studies.

Our DTW implementation also returns warped sequences X^W and Y^W derived from X and Y by (possibly) repeating terms to improve alignment. More precisely, $X^W = ((x_1)^{a_1}, ..., (x_m)^{a_m})$ and $Y^W = ((y_1)^{b_1}, ..., (y_n)^{b_n})$, where $(\cdot)^k$ denotes k repetitions of a term, and the a_j and b_j are positive integers found by the algorithm. Using this notation, the DTW distance d_{DTW} is the squared Euclidean distance between X^W and Y^W, defined as follows:

Definition 1. *Given sequences $A_{1,N}$ and $B_{1,N}$, the squared Euclidean distance between A and B is:*

$$d(A, B) = \sum_{i=1}^{N} \|a_i - b_i\|^2 \tag{4.1}$$

where $\|\cdot\|$ is the usual Euclidean norm.

To compute d_{DTW}, X^W, and Y^W, we first construct an $(m \times n)$ matrix D, where $D_{(i,j)} = \|x_i - y_j\|^2$. Intuitively, we then find the minimum cost path through D from $D_{(1,1)}$ to $D_{(m,n)}$ subject to a path constraint. Letting w_k be the k^{th} element of a warping path W –

a possible path through D – we constrain W to allow only three moves: repeat the current point in X, repeat the current point in Y, or move to the next point in both. Formally, if $w_k = (i_k, j_k)$, then $w_{k+1} \in \{(i_k + 1, j_k), (i_k + 1, j_k + 1), (i_k, j_k + 1)\}$. The optimal path from (i, j) to (m, n) and its cost $C_{(i,j)}$ are computed using dynamic programming, where $C_{(m,n)} = D_{(m,n)}$, and the remaining $C_{(i,j)}$ are given by the following recursion:

$$C_{(i,j)} = D_{(i,j)} + min\{C_{(i+1,j)}, C_{(i+1,j+1)}, C_{(i,j+1)}\} \tag{4.2}$$

This process may be carried out row-wise or column-wise, with $C_{(i,j)} = \infty$ for $i > m$ or $j > n$. The final DTW distance is $C_{(1,1)}$, and the warping path W along with the warped sequences X^W and Y^W may be recovered from C.

Although we refer to d_{DTW} as a distance measure, it is not a distance metric in the mathematical sense. The DTW distance is non-negative ($d_{DTW}(A, B) \geq 0$) and symmetric ($d_{DTW}(A, B) = d_{DTW}(B, A)$) for all sequences A and B, but it does not satisfy the triangle inequality. Further, while $d_{DTW}(A, A) = 0$ for any sequence A, it is also possible to have $d_{DTW}(A, B) = 0$ for distinct sequences A and B.

Unless otherwise stated, we have resampled the input sequences to have the same length and limited the warping path to the Sakoe-Chiba band [112] to reduce computation, so that $C_{(i,j)} = \infty$ whenever $|j - i|$ is greater than one fourth the length of the inputs.

Figure 4.1 illustrates DTW in a study subject with mild MS disability (Section 4.3.6). The asterisks show the same peak before and after phase correction. Warps may be identified by the horizontal segments, which indicate repeated data points. Two such segments have been circled.

4.1.2 DTW as Adaptive

The physiologic signal monitoring algorithms presented in this chapter are *personalized*, but unlike the algorithms presented in Chapters 5 and 6, they are not truly adaptive.

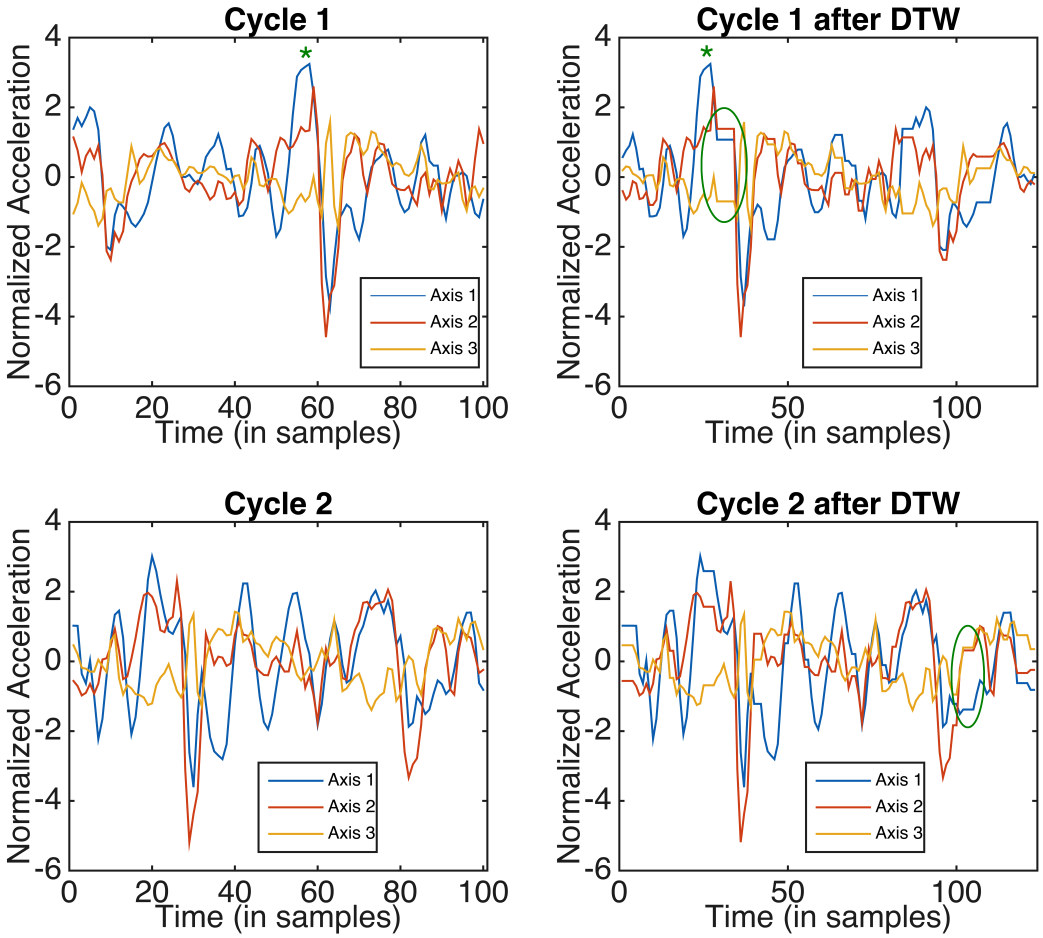


Figure 4.1: DTW aligns gait cycles in a subject with mild MS disability.

Personalization is achieved via person-specific baseline measurements: the algorithms assess subjects with respect to their own baseline. As a result, they are not confused by differences between persons or types of pathology. Recall, however, that we defined an *adaptive* algorithm to have a mapping from observations to either outputs or actions that depends on previous observations. DTW does not involve a set of actions, so the question is whether its outputs – the DTW distances d_{DTW} – depend on previous observations. While d_{DTW} does depend on subject-specific baseline measurements, the implementations of DTW in this chapter are not adaptive, because baseline measurements do not change.

A truly adaptive signal monitoring algorithm could be very similar to ours, with DTW as

its core subroutine, but it would need to manage an evolving set of template sequences instead of using a static baseline set. At a minimum, this requires criteria to add an observation sequence. If constant memory is desired, the system must also choose template sequences to discard. In Section 4.2.4 we speculate on these requirements. Importantly, our algorithms hold to the spirit of the adaptive approach by performing well despite many sources of heterogeneity, and this work paves the way for truly adaptive DTW-based monitoring.

4.1.3 DTW Applications and Variants

DTW is a versatile technique for signal alignment and time series data mining, and as such, it has been used in a broad range of applications, including speech recognition[113], shape and object identification[114], and gesture recognition[115], just to name a few. DTW has seen numerous applications in gait analysis alone, including activity recognition [116] and biometric gait recognition[117][118], and it has been used to monitor cardiac filling cycles[119]. To our knowledge, however, this work has been the first application of DTW to the detection and monitoring of gait pathology[120].

Due to the algorithm's generality and popularity, many DTW variants have been proposed. A prominent example is derivative DTW, which replaces the signal with its estimated derivative[121], and efficient algorithms for real-time data stream monitoring have been demonstrated[122]. Efficient DTW implementation would be critical in a real-time monitoring scenario.

Algorithms for combinations of rotation, scale, and/or offset invariance have also been proposed, though our iterative algorithm for three dimensional inertial data appears to be new. Affine DTW achieves scale and offset invariance[123], and a rotation invariant implementation has been used to analyze two-dimensional handwriting samples [124]. Another rotation invariant algorithm based upon principal component analysis has been used for gait recognition[125].

4.1.4 Algorithms for Monitoring Gait Pathology

Here we highlight several existing monitoring algorithms for gait pathology. With a few notable exceptions – including those discussed in the previous section – the vast majority of gait monitoring algorithms follow the traditional approach (Figure 2.1), which has been used successfully in the activity recognition literature.

Inertial devices have been used to monitor walking ability in several different contexts. The ISway test uses an accelerometer to measure postural sway resulting from neurological impairment [126]. The iTug system uses inertial sensors to partly automate the Timed Up and Go test, a clinical measure of balance and mobility [127]. Several other accelerometer-derived features have shown promise is indicators of disability severity [128], including the recently-developed causality index [129]. Many studies have used daily step counts derived from inertial sensors as an outcome measure, but comparatively few have assessed features intrinsic to individual gait cycles.

Additionally, a number of gait analysis algorithms not based on DTW have incorporated rotation invariance. An algorithm for model-based classification of upper limb movements and walking activities incorporated sensor orientation correction[130], and a linear dynamical model based method for activity classification was robust to sensor mounting errors, including rotation [131].

4.2 Dynamic Time Warping for Physiologic Signals

This section presents the methodology behind three novel DTW modifications useful in health monitoring developed in this work. The results which follow (Section 4.3) demonstrate their value. All of our results required at least one of these modifications.

4.2.1 Cyclic DTW

DTW is designed to align similar sequence regions by introducing *warps* – which are nothing more than point-wise repetitions – in either input sequence. DTW does not alter the order of regions in the input sequences, and in particular, it cannot circularly shift sequences by moving regions from the end to the beginning, or vice-versa. In other words, DTW is not equipped to correct phase differences in periodic signals.

This is a significant limitation when working with physiologic signals. Consider using DTW to monitor an ECG signal supposing that a baseline, previously observed ECG sequence were known. To measure the similarity of the current signal to baseline, we must choose a starting point for the test sequence. In some cases, easily identified sequence regions make this easy. For example, we might choose a point just before atrial depolarization – which is easily observed as the "P wave" – as the starting point for all of our sequences. However, oftentimes establishing this kind of landmark is not feasible, and the landmark itself may not be reliable. In atrial flutter, for example, there are multiple P waves per cycle (as defined by the QRS complex).

The same problem arises in inertial data from persons with severe MS. In healthy adults, there are several easily identified gait cycle regions, including the heel strike and toe-off. However, in MS these features can be irregular or absent. Consequently, extracting appropriate input sequences for DTW is itself a difficult problem.

Cyclic DTW is our solution to this dilemma. It packages together several well-known approaches to facilitate the DTW-based processing of physiologic signals. The following description provides numeric values appropriate for gait cycles, but extending the approach to other signals is a simple matter of parameter adjustment. We note that the "starting point" problem is a common one which is even more difficult in an online setting. Many DTW adaptations have been proposed to efficiently test all starting points for a possible distance-based match[122].

First, the cycle rate was identified via fast Fourier transform (FFT) as the highest

peak within the 0.2 – 1.4 Hz band, which corresponds to 24 – 168 steps per minute. Non-overlapping segments of corresponding length were extracted as cycles. Since this method does not guarantee that cycles are in phase, a phase-invariant implementation of DTW was used to correct this difference. In other words, all possible circular shifts had to be checked. Given cycles $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_n)$, DTW was run between $A_i = (a_i, \ldots, a_n, a_1, \ldots, a_{i-1})$ and B for all $i \in \{1, \ldots, n\}$.

To reduce computation, early DTW termination and Keogh lower bounding and were used to pre-emptively abandon starting points that could not be optimal[114]. Early DTW termination is the more straightforward of the two. While cycling through all starting points, the lowest DTW distance d_{min} observed so far is stored. As the DTW algorithm moves through the columns (or rows) of the cost matrix C, it compares d_{min} to the minimum possible total cost $C_{min} = \min_i C_{(i,j)}$ (or $C_{min} = \min_j C_{(i,j)}$) for the current starting point. If $C_{min} \geq d_{min}$, then the current starting point cannot achieve a total cost better than d_{min}, so the algorithm immediately terminates its dynamic programming loop and moves to the next starting point.

Keogh lower bounding places a lower bound on the DTW distance between two sequences. If $\text{LB\_Keogh}(A_i, B) \geq d_{min}$, then starting position i need not be checked, since $d_{DTW}(A_i, B) \geq \text{LB\_Keogh}(A_i, B) \geq d_{min}$. Calculation of this lower bound is beyond the current scope, but relatively straightforward[114].

Figure 4.1 illustrates the cyclic DTW procedure. Two resampled, normalized cycles from the same subject (left) have been shifted and warped to align similar features (right). The DTW distance is the Euclidean distance between the warped cycles on the right. The warping length is the number of warps inserted.

4.2.2 Distance Score and Warp Score

DTW operates on pairs of sequences, returning a distance and warping path for each pair. Since monitoring typically involves a *set* of baseline sequences and many test sequences, it

requires a method to summarize many measurements into a single, clinical statistic. This is particularly important in clinical applications, since a single, interpretable value must be provided. Given a set of baseline sequences $\mathbb{A}$ and test sequences $\mathbb{B}$, where $|\mathbb{A}| = M$ and $|\mathbb{B}| = N$, we define *Distance Score* as follows:

$$\mathrm{DS} = \frac{1}{2}\left(\underset{A\in\mathbb{A}}{\mathrm{median}}\left(\min_{B\in\mathbb{B}}(d_{DTW}(A,B))\right) + \underset{B\in\mathbb{B}}{\mathrm{median}}\left(\min_{A\in\mathbb{A}}(d_{DTW}(A,B))\right)\right) \tag{4.3}$$

In words, we first find the best match for each sequence among all sequences of the other set. For each set, we then take the median best match; the median is preferred to make the process robust to outlying, abnormal sequences. Finally, we average the two medians. This process has been used previously by Boulgouris et al. to summarize DTW distances[132].

Unlike other applications, we have also provided a statistic based on the warping path: the Warp Score. To calculate the Warp Score, warping paths were first summarized in terms of the warping length, defined as the difference in length between input cycles and their DTW-aligned counterparts. This length measures how much cycles were "stretched" – by repeating samples – during alignment. Warping lengths were then reduced to a single Warp Score by the same process used to reduce DTW distances to the Distance Score.

Our clinical results suggest that the Warp Score is useful in gait monitoring. All results were stronger using Warp Score as a measure of gait deterioration rather than the Distance Score. For example, the Warp Score showed stronger Spearman correlation to the MSWS-12 compared to the Distance Score ($r_s = 0.429$ vs $r_s = 0.263$). In Section 4.3.6, all clinical insights are presented in terms of the Warp Score.

This finding was surprising, as DTW distance is more commonly reported than warping length; in fact, the Warp Score is an novel contribution of this work not found elsewhere in the DTW literature to our knowledge. This may reflect a fundamental difference between the current work and other applications of DTW. Typically, DTW is used to overcome signal distortions to determine whether two signals match, but here it's the distortions themselves – the "stretch" captured by the Warp Score – that are manifestations of gait pathology. This

is true in other applications as well. Generally speaking, health monitoring is concerned with the progression or appearance of an underlying pathological process that actively distorts a signal. The goal is not to determine whether the signals are the same when distortion is removed, but rather to quantify the *degree* of distortion. Further work is needed to clarify the difference in meaning between these measures and to study the Warp Scores value in other health monitoring applications.

4.2.3 Rotation, Scale, and Offset Invariant DTW

Scale and offset invariant DTW, in which one sequence can be scaled or shifted to improve similarity, has been developed by several authors, notably Chen et al., who evaluated an iterative algorithm similar to ours on a number of time series data sets [123]. When analyzing gait, scale and offset invariance may mitigate variability due to walking surface, shoe type, attachment method, and moderate changes in speed.

Rotation, scale, and offset-invariant DTW (RSOI-DTW) takes the next step by incorporating rotation invariance. This property applies to three-dimensional, vector valued signals, including inertial data. An algorithm is defined to be *rotation invariant* if arbitrarily rotating the signal prior to processing has no effect on algorithm output. In practice, this means that physically rotating the device does not disrupt or confuse the monitoring system.

RSOI-DTW is an iterative algorithm that alternates between optimizing the rotation, scaling, and offset of the sequence Y, and optimizing the warping path using DTW. The former is an instance of the Procrustes problem, which may be solved in closed form using singular value decomposition. The details of this problem are beyond the current scope, but may be found in [133].

Formally we define rotations as elements of $SO(3)$, the (3×3) orthogonal matrices of determinant 1. $SO(3)$ are the rigid rotations in $\mathbb{R}^3$, excluding reflection; they correspond with the rotations possible for a rigid physical object. These matrices form a group under

multiplication: in particular, they are invertible, and the inverses and products of rigid rotations are also rigid rotations.

This section first defines the transformations allowed in RSOI-DTW – the RSO transformations – then provides the RSOI-DTW algorithm. Finally, it proves that RSOI-DTW is rotation, scale, and offset invariant under typical circumstances, and the algorithm is guaranteed to terminate.

Definition 2. *An RSO transformation is an affine transformation* $f : \mathbb{R}^3 \to \mathbb{R}^3$ *of the form* $f(x) = sRx + b$*, where* $s \in \mathbb{R}^+, R \in SO(3)$*, and* $b \in \mathbb{R}^3$*.*

Proposition 1. *The RSO transformations are closed under composition and inverse, thus forming a subgroup of the affine group.*

Proof. Let $f_\alpha(x) = s_\alpha R_\alpha x + b_\alpha$, and $f_\beta(x) = s_\beta R_\beta x + b_\beta$. The inverse of f_α is given by $(f_\alpha)^{-1}(x) = \frac{1}{s_\alpha}(R_\alpha)^{-1}x - \frac{1}{s_\alpha}(R_\alpha)^{-1}b_\alpha$. By inspection, $\frac{1}{s_\alpha} \in \mathbb{R}^+$ and $-\frac{1}{s_\alpha}(R_\alpha)^{-1}b_\alpha \in \mathbb{R}^3$, and $(R_\alpha)^{-1} \in SO(3)$ because $SO(3)$ is closed under inverse. So, the RSO transformations are closed under inverse.

The composition of f_α and f_β is given by $(f_\alpha \circ f_\beta)(x) = (s_\alpha s_\beta)(R_\alpha R_\beta)x + (s_\alpha R_\alpha b_\beta + b_\alpha)$. As before, we note that $s_\alpha s_\beta \in \mathbb{R}^+$ and $(s_\alpha R_\alpha b_\beta + b_\alpha) \in \mathbb{R}^3$; and the closure of $SO(3)$ under composition guarantees that $R_\alpha R_\beta \in SO(3)$. Thus the RSO transformations are closed under composition. □

Definition 3. *Given an RSO transformation* f *and a sequence* $X = (x_1, ..., x_n)$*, where* $x_i \in \mathbb{R}^3\ \forall i$*, the sequence* $f(X)$ *is defined to be* $(f(x_1), ..., f(x_n))$*.*

With this background in place, we now present the RSOI-DTW algorithm:

This algorithm can be restricted for a particular use case by limiting f to a subset of the RSO transformations. If rotation is not a concern, R may be held to I, the identity matrix, reducing RSOI-DTW to scale and offset invariant DTW, as developed in [123]. Similarly, if scaling is not a concern, s may be held to 1.

Rotation, Scale, and Offset Invariant DTW

procedure RSOI-DTW(X, Y)
  $X^W \leftarrow X$
  $Y^W \leftarrow Y$
  $d \leftarrow \infty$
  repeat
    $d_{old} \leftarrow d$
    $f^* \leftarrow \operatorname{argmin} d(X^W, f(Y^W))$
    ▷ where f is an RSO transformation
    $(d, X^W, Y^W) \leftarrow DTW(X, f^*(Y))$
  until $(d_{old} - d)$¡ϵ
  return $d, X^W, f(Y^W)$
end procedure

Proposition 2. *RSOI-DTW is rotation, scale, and offset invariant. More precisely, let X, Y, and Z be sequences in $\mathbb{R}^3$ of equal length, where $Z = f_z(Y)$ for some RSO transformation f_z. If $d(X^W, f(Z^W))$ has a unique minimizer f^* at each iteration of the RSOI-DTW algorithm, then RSOI-DTW(X, Y) = RSOI-DTW(X, Z).*

Proof. Let f_z be the RSO transformation taking Y to Z, so that $f_z(Y) = Z$, and suppose $f_z(Y^W) = Z^W$ at the beginning of the i^{th} iteration of RSOI-DTW(X, Y) and RSOI-DTW(X, Z). We proceed by induction; note that the $i = 1$ case holds, since we initialize $Y^W \leftarrow Y$ and $Z^W \leftarrow Z$.

Given a unique RSO transformation f^* that minimizes $d(X^W, f(Z^W))$ in iteration i, the RSO transformation $(f^* \circ f_z)$ must be the unique minimizer of $d(X^W, f(Y^W))$. To see this, suppose there were some other RSO transformation g such that $d(X^W, g(Y^W)) \leq d(X^W, (f^* \circ f_z)(Y^W))$. Since $Y^W = f_z^{-1}(Z^W)$, where f_z^{-1} is the inverse of f_z, we have:

$$\begin{aligned} d(X^W, (g \circ f_z^{-1})(Z^W)) &= d(X^W, (g(f_z^{-1}(Z^W))) \\ &= d(X^W, g(Y^W)) \\ &\leq d(X^W, (f^* \circ f_z)(Y^W)) \\ &= d(X^W, (f^*(f_z(Y^W))) \\ &= d(X^W, f^*(Z^W)), \end{aligned}$$

violating our assumption that f^* is the unique minimizer.

Knowing that f^* and $(f^* \circ f_z)$ are the unique minimizers found in line 7 of the i^{th} iteration of RSOI-DTW(X, Z) and RSOI-DTW(X, Y), respectively, we conclude that the input to the DTW subroutine is $f^*(Z^W)$ in either case. Because of this, DTW returns the same distance d and warping path W in both cases, guaranteeing $f_z(Y^W) = Z^W$ at the beginning of the $(i+1)^{th}$ iteration, completing our inductive proof. □

Intuitively, since the first step in RSOI-DTW is to optimally rotate, scale, and shift the input, the sequences Y^W and Z^W are both transformed to $f^*(Z^W)$ in the first iteration of the algorithm, and subsequent processing is identical.

Proposition 3. *RSOI-DTW terminates.*

Proof. First, notice that $f(X^W) = f(X)^W$, because applying the warping path W is a repetition of terms, and the transformation f is applied once to each term in either case.

Let d_i, W_i, and f_i be the distance, warping path, and optimal RSO transformation found in iteration i, so that $d_i = d(X^{W_i}, f_i(Y)^{W_i})$. In iteration $(i+1)$, f_{i+1} is chosen to minimize $d(X^{W_i}, f(Y)^{W_i})$, thus $d(X^{W_i}, f_{i+1}(Y)^{W_i}) \leq d_i$. And DTW finds the warping path

W_{i+1} minimizing $d(X, f_{i+1}(Y))$. Together, we have:

$$d_{i+1} = d(X^{W_{i+1}}, f_{i+1}(Y)^{W_{i+1}}) \leq d(X^{W_i}, f_{i+1}(Y)^{W_i}) \leq d_i.$$

Since this holds for all i, the sequence $\{d_i\}_{i\in\mathbb{N}}$ is monotonically decreasing. Further, since the d_i are squared Euclidean distances, they are bounded below by zero. Therefore this sequence converges by the monotone convergence theorem, guaranteeing termination of the RSOI-DTW algorithm. □

In our view, the RSO transformation is the most general affine transformation reasonable for inertial data. Nevertheless it preserves the cycle features needed to distinguish persons and walking styles, as we will show in the next section.

4.3 Validation Studies

The validation process proceeded in three phases. Each phase had distinct objectives and tested distinct combinations of DTW modifications. We began with a very small, simple study designed to confirm the rotation invariance of RSOI-DTW. This was followed by a larger exploration of pathology detection and fast walking in healthy volunteers. Both trials succeeded, giving us every reason to believe the approach would detect real pathology in an MS cohort. Our efforts culminated in a clinical study in which inertial data from the 6MW was analyzed using DTW. A total of 115 subjects – 86 with MS and 29 healthy controls – participated in this trial.

This study did *not* require rotation invariance, as it was conducted in a highly controlled clinical environment. RSOI-DTW makes DTW robust to real-world variability, but it is not necessary for a single, supervised trial conducted in a hospital corridor. However, our two other modifications – the Warp Score and Cyclic DTW – proved critical to our clinical results.

In all three cases, applying our general monitoring approach to the specific use case led to the development of a novel methodology.

4.3.1 Study Procedures and Participants

In each validation study, subjects walked in a corridor while wearing a single ActiGraph GT3X accelerometer on their left hip, secured using an elastic belt. Specific procedures differed between studies, as described below. In each case, the data was downloaded using ActiLife software, divided by person and trial, and segmented into gait cycles, defined as the data between consecutive left heel strikes. The data must be segmented before applying DTW or RSOI-DTW, which take two gait cycles – a template and a test cycle – as inputs.

Before running DTW, cycles were resampled to be 100 samples in length. Cycles were also normalized by subtracting the mean along each axis, then dividing all samples by the standard deviation of the vector magnitude. These steps were designed to mitigate the effects of cycle length and gait speed on the results. This was an application-specific step which may not be appropriate in other monitoring scenarios.

Sensor Rotation

To verify that RSOI-DTW is invariant under a real rotation – that is, a mis-orientation of the sensor – three participants completed a walking trial with the sensor in four different orientations: no rotation, a 90° rotation about the medial-lateral axis, a 180° rotation about the medial-lateral axis, and a 180° rotation about the vertical axis. Subjects walked with their normal, casual gait each time. The 125 rotated signals were compared to the non-rotated signals using DTW and RSOI-DTW, and the resulting distances were used to recognize subjects. Gait cycles were manually segmented by identifying the heel strike, which creates a characteristic spike in acceleration.

Investigating Walking Speed and Simulated Pathology

As a preliminary investigation of walking speed and pathology detection, 21 healthy volunteers participated in an initial walking trial. Each subject was asked to walk down a long corridor four times to demonstrate four different styles of gait: casual walking, fast walking, ataxic walking, and right leg circumduction. Ataxic walking is seen in persons with balance difficulties, characterized by a wide base and lateral swaying. Circumduction is the outward, circular swinging of one leg in swing phase; it occurs when the leg is rigid or spastic at the knee and/or ankle joint.

Subjects walked with each style in one direction for 40 steps, then turned, paused five seconds, and walked back with the next style. Each style was demonstrated before the trial by a clinically trained research assistant, and subjects were given an opportunity to practice until comfortable.

Multiple Sclerosis Cohort

To validate the approach in a real, clinical monitoring scenario, 86 subjects with clinically definite MS[134] and 29 healthy controls were recruited for a walking study. All study procedures were approved by the University of Virginia (UVa) Institutional Review Board for Health Sciences Research, and written consent was obtained from all participants. Subjects completed a 6MW in a hospital corridor while wearing the ActiGraph accelerometer. Controls with any disability and MS subjects with non-MS causes of walking impairment were excluded. Prior to the walk, subjects underwent Expanded Disability Status Scale (EDSS) assessment by Neurostatus-certified staff. The MFIS and MSWS-12 were also collected.

Gait cycles were highly irregular in many subjects, particularly those with more advanced disability. This made it impossible to segment cycles by identifying heel strikes, as in the previous two trials. Instead, the approach outlined in Section 4.2.3 – which requires cyclic DTW – was used to select and compare cycles. Cycles from minute two of the 6MW were used as the baseline set for DTW. Minute two was the preferred baseline due to evidence

Table 4.1: Demographics and Outcome Measures in MS Subjects and Controls

	MS	Control
N (Female/Male)	86 (73/13)	29 (20/9)
Age, median (IQR)	46 (38 – 52)	40 (32 – 48)
EDSS, median (IQR)	2.5 (2 – 3.5)	NA
MSWS-12 Score, median (IQR)	14.6 (0 – 45.8)	NA
MFIS Score, median (IQR)	29 (10.5 – 45.5)	9 (2.5 – 25)
6MW Distance*, median (IQR)	1574 (1352 – 1873)	2009 (1793 – 2166)

*distance reported in feet; IQR: Inter-Quartile Range; EDSS: Expanded Disability Status Scale; MSWS-12: MS Walking Scale; MFIS: Modified Fatigue Impact Scale; 6MW: Six-Minute Walk

that subjects' slow down dramatically during minute one[77]. This result was confirmed in our cohort. Anecdotally, we have observed that subjects adjust their pace and rhythm in the first minute.

Participant demographics are summarized in Table 4.1. EDSS and MSWS-12 were not assessed in the control subjects, so they are reported as not applicable. In statistical calculations, however, control subjects were assigned a zero on the EDSS and MSWS-12. Though the EDSS can be nonzero outside of MS, we have excluded controls with any form of disability, making this a fair assumption. MFIS results were collected in both MS and control subjects, as symptoms of fatigue can be present in controls and the MFIS items are not worded to be MS-specific.

4.3.2 Empirical Rotation Invariance

Our theoretical results show that RSOI-DTW is rotation invariant from a mathematical standpoint. Based primarily on our first study, we demonstrate its rotation invariance in practice by showing that real mis-orientations of the sensor have minimal effect on output. In particular, RSOI-DTW successfully distinguishes subjects (i.e. gait recognition) despite changes in sensor orientation.

First, we illustrate RSOI-DTW by applying it to arbitrarily chosen cycles from a healthy volunteer (Figure 4.2). The raw cycles C_1 and C_2 are shown in the top and middle left plots,

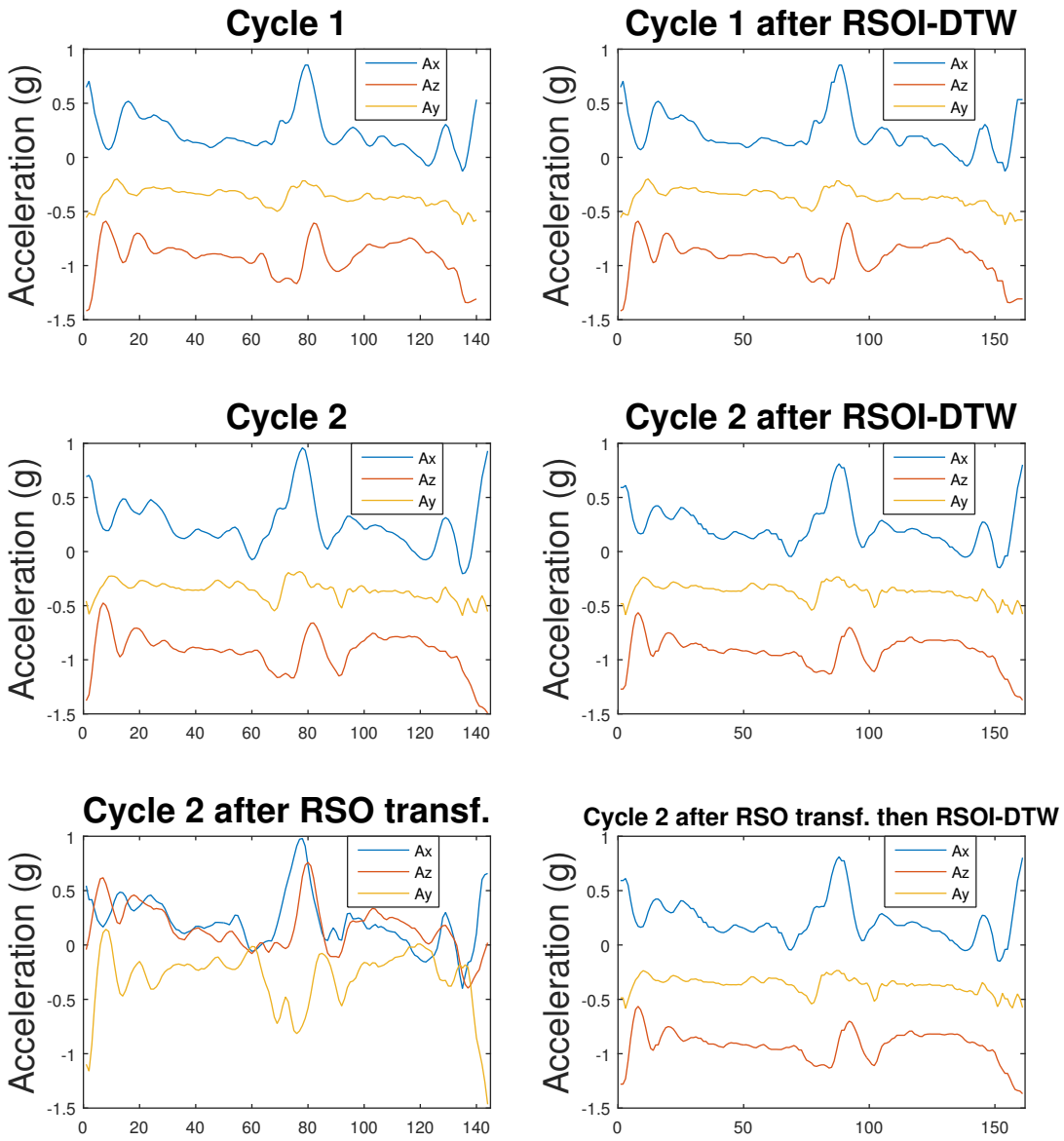


Figure 4.2: Example of RSOI-DTW Rotation Invariance

respectively. The bottom left plot shows C_2 after a randomly chosen RSO transformation f is applied. The right panel shows how RSOI-DTW alters the cycles. When running RSOI-DTW(C_1, C_2), a warped version of C_1 (top right) and a warped, transformed version of C_2 (middle right) are returned. Here the rotation, scale, and offset are small, because sensor

Table 4.2: 1NN Cycle Recognition in the Presence of Sensor Mis-Orientation

	DTW	RSOI
# Correct	24	125
(%)	19.2	100.0

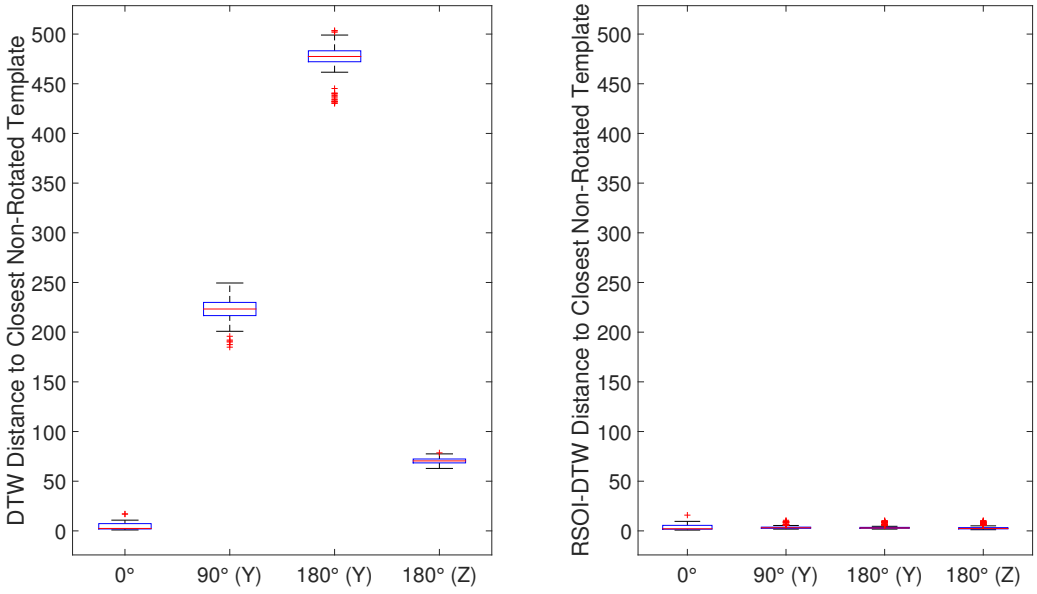


Figure 4.3: DTW and RSOI-DTW Distances between Correctly Oriented Cycles and Incorrectly Oriented Cycles from the Same Subject

alignment was consistent and no rotation was applied. When running RSOI-DTW$(C_1, f(C_2))$, the same two plots are returned. The warped version of C_1 (not shown) is identical to the top right plot, and the warped, rotated version of $f(C_2)$ (bottom right) is identical to the plot above it, because RSOI-DTW is invariant under f.

Using data from our initial verification study, 1-nearest neighbor (1NN) recognition was perfect under RSOI-DTW but worse than random under DTW (Table 4.2), and RSOI-DTW distances were similar for all four orientations (Figure 4.3). Thus RSOI-DTW was rotation invariant in our initial testing.

As additional support, randomly chosen RSO transformations were applied to each of our 21 healthy participants' 12 casual gait cycles. RSOI-DTW was used to compare the

transformed cycles to the original, non-transformed cycles, for a total of $\binom{252}{2}$, or 31,626, comparisons. These distances were compared to the corresponding distances obtained without first applying a transformation. In all 31,626 cases, the RSOI-DTW distances were identical up to rounding error, never differing by more than 10^{-13}. Together with the experimental result, this strongly suggests that RSOI-DTW is RSO invariant in practice when used on inertial time series data.

4.3.3 Empirical Convergence

Figure 4.4 shows histograms of the RSOI-DTW convergence rate when comparing (a) pairs of casual walking cycles, and (b) fast walking cycles to casual walking cycles. In over 60,000 runs per plot, the algorithm most often required 7 iterations, rarely over 20, and never over 30. To ensure local optimality, we insisted that $d = d_{old}$ for convergence, meaning the warping path was stable. In our data, using a less strict (e.g. 10^{-5}) convergence criterion typically reduces the number of iterations by one, and never more than two.

As shown in the RSOI-DTW algorithm, each iteration calls DTW once and the Procrustes algorithm once. DTW involves $O(N^2)$ computations, where N is the length of the inputs: computation is proportional to the number of pairings between points in X and points in Y. However, RSOI-DTW must recover the warping path W in addition to the distance d, requiring a second trip through the cost matrix and adding a multiple of N^2 computations. Run time is increased, but not by more than a factor of two. The Procrustes problem requires only $O(N)$ computations [133], and in our data set, DTW occupies the vast majority of run time. A conservative run time estimate may be obtained by multiplying the DTW run time by $2I$, where I is the number of iterations required.

4.3.4 Detecting Simulated Pathology

In the pathology detection problem, the algorithm must decide (yes/no) whether an unknown cycle represents normal gait or possible pathology. This is done by setting a threshold on

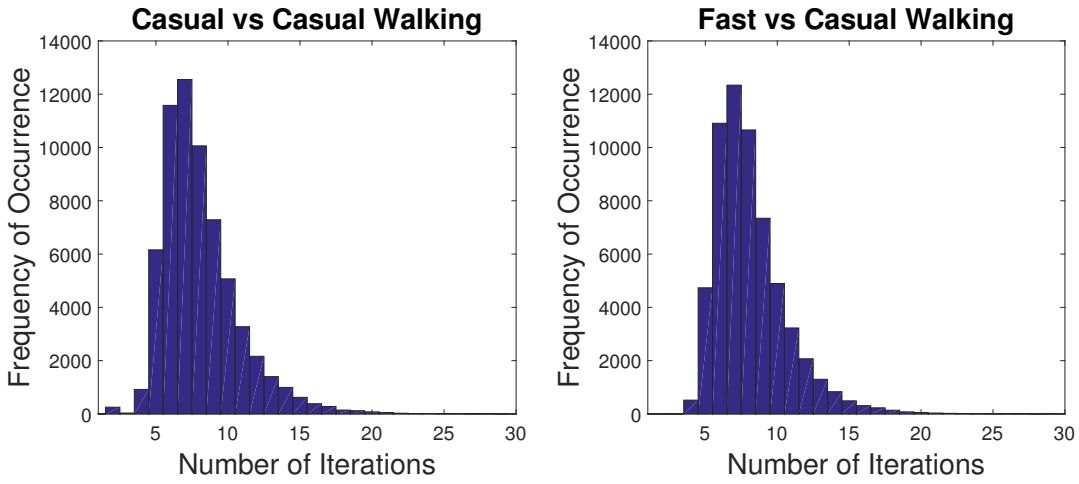


Figure 4.4: Histograms of the Number of RSOI-DTW Iterations

Table 4.3: EER for Detection of Simulated Pathology using DTW and RSOI-DTW

	DTW EER (%)	RSOI EER (%)
Mean	0.0	0.4
Median	0.0	0.0
Min	0.0	0.0
Max	0.0	8.3

the best match between the tested cycle and the known subject's template cycles. Table 4.3 shows that the two simulated pathologies were easily recognized in all subjects under both DTW and RSOI-DTW distances. The equal error rate (EER) was nonzero in only one subject under RSOI-DTW. As we have described, only RSOI-DTW is capable of this accuracy despite the variability possible in a real-world use case.

4.3.5 Effects of Walking Speed

A gait monitoring algorithm shouldn't classify normal walking patterns as pathology, regardless of speed. In particular, fast walking should *not* register as pathology. The same is true in gait recognition: ideally, fast walking and slow walking would both be recognized. Scale and offset invariant DTW may help to match fast walking to normal walking templates by warping and scaling cycles. Unfortunately, Table 4.4 shows that in practice, both DTW and RSOI-DTW struggle to distinguish pathology from fast walking: the equal error rate (EER)

Table 4.4: EER when Distinguishing Normal Gait (Casual or Fast) from Simulated Pathology

	DTW EER (%)	RSOI EER (%)
Mean	43.8	31.3
Median	50.0	39.6
Min	0.0	0.0
Max	52.3	50.0

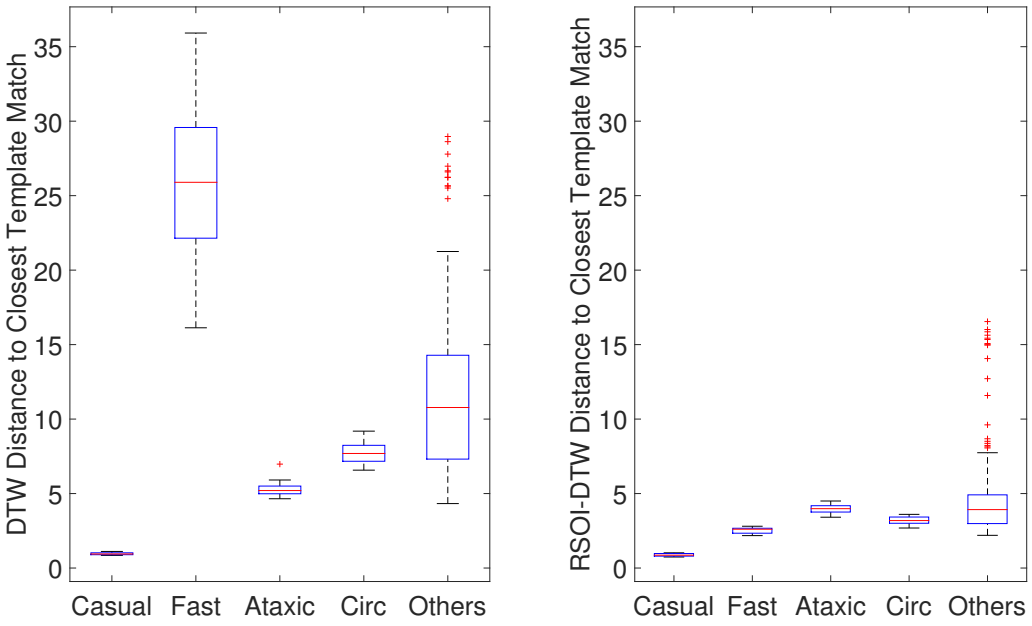


Figure 4.5: Distances to the Closest Casual Gait Template for Several Groups of Cycles using DTW (left) and RSOI-DTW (right) in an Example Subject

is very high with both approaches. As before, test cycles were compared to the subject's casual walking cycles, but this time the test cycles included both simulated pathology *and* fast walking. RSOI-DTW slightly improved performance compared to DTW, but overall performance was poor with both approaches.

Figure 4.6 illustrates this difficulty in a single subject: while RSOI-DTW dramatically lowered the distances between fast cycles and template cycles compared to DTW, it was not enough for reliable pathology detection or gait recognition.

In light of this difficulty, either (1) walking speed must be consistent between cycles, or (2) the template set must include cycles at many speeds. To test the latter, fast walking

Table 4.5: Improved EER after including Fast Cycles in the Template Set

	DTW EER (%)	RSOI EER (%)
Mean	4.9	8.0
Median	0.0	1.1
Min	0.0	0.0
Max	31.3	33.3

cycles were added to the template set in each subject, with results shown in Table 4.5. With this modification, simulated pathology could again be distinguished from normal walking (casual or fast). Compared to the near perfect results in the previous section, however, the EER is high in several subjects. RSOI-DTW improves the EER in two subjects, but worsens it in others.

4.4 Clinical Insight: Motor Fatigue

As discussed in Section 3.1.3, the six-minute walk (6MW) is a commonly used, objective walking assessment primarily used in research settings. Subjects walk "as far and as fast as possible" for six minutes without resting, and the distance walked is measured at the end of each minute[77]. Due to its longer duration, the 6MW is believed to be more sensitive to dynamic motor fatigue than other objective measures. This is particularly important early in the disease course, when signs of walking impairment may be difficult to elicit. Physiologic changes occur in MS subjects during the walk; for example, oxygen consumption increases over the first three minutes[78]. Unfortunately, direct evidence of motor fatigue has been elusive due to the inherent difficulty of measuring qualitative changes in walking. Some practitioners believe the two-minute walk (2MW) is sufficient for clinical assessment, because the two tests are highly correlated[80]. Based on clinical experience, however, we believe that dynamic motor fatigue only emerges after several minutes of walking in many mildly disabled subjects. If so, the 2MW is inadequate as a measure of motor fatigue.

To improve our understanding of dynamic motor fatigue in MS, we used cyclic DTW

to identify progressive changes to gait cycles occurring during the 6MW in MS subjects. We refer to these changes as "gait deterioration", noting that they are less prominent in healthy volunteers (as shown in the sections to follow). Our goal was to demonstrate clinically relevant information emerges beyond minute two, supporting the distinct clinical utility of the longer walk test, and to connect this information to motor fatigue. First, we show that gait deterioration is present, and it is more prominent in MS subjects than in healthy volunteers. We then connect gait deterioration to motor fatigue using patient-reported outcomes measuring fatigue (MFIS) and walking ability (MSWS-12) (Section 3.1.4).

4.4.1 Progressive Gait Deterioration

Figure 4.6 shows the mean Warp Score (left) and variance among Warp Scores (right) among subjects grouped by EDSS in minutes 3 to 6. Warp Scores were much higher in severe subjects (EDSS > 4.5) than in all other groups ($p < 0.001$). Variance decreased over time in the severe group, but increased in other groups. Warp Scores were higher in moderate subjects (EDSS 3 – 4.5) compared to controls even in minute 3 (Cohen's $d = 0.706$, $p = 0.01$), with steadily increasing variance. Moderate subjects had significantly higher warp scores than mild subjects (EDSS 0 – 2.5) in minute 3 (Cohen's $d = 0.700$, $p = 0.006$), but this gap narrowed by minute 6 (Cohen's $d = 0.202$, $p = 0.41$). Mild subjects and controls had similar Warp Scores in minutes 3 and 6 (Cohen's $d = 0.024$, $p = 0.92$ and Cohen's $d = 0.186$, $p = 0.44$, respectively). Variance among mild subjects increased sharply between minutes 5 and 6 from just over 4 to almost 10. Warp Scores increased between minutes 3 and 6 in mild (Cohen's $d = 0.786$, $p < 0.001$) and moderate subjects (Cohen's $d = 0.374$, $p < 0.001$).

Mild subjects Warp Scores were similar to controls on average, but the sharp increase in variance in minute 6 is important. Compared to controls, mild subjects Warp Scores were less consistent: some increased toward the end of the walk, while others were unchanged. The regression model (Figure 4.7) shows that these differences are meaningful, as they improve prediction of MSWS-12 scores. In contrast, moderate subjects had higher Warp Scores

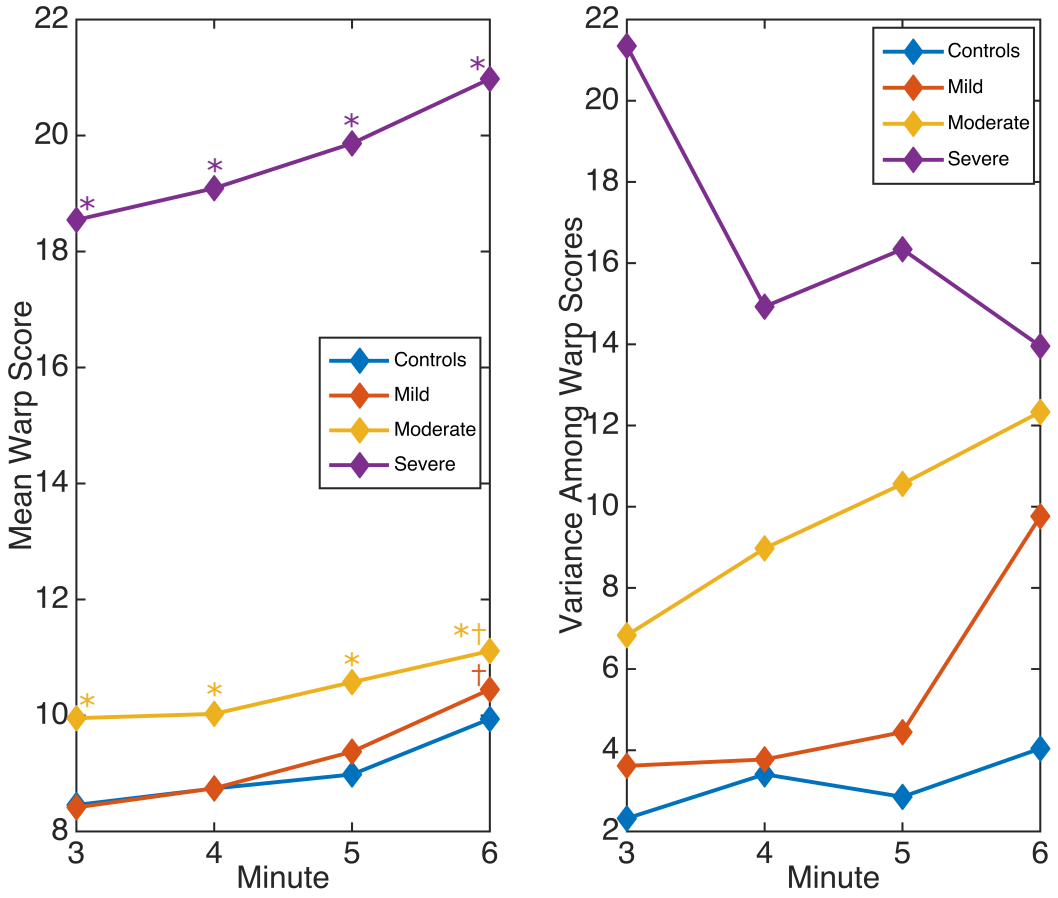


Figure 4.6: Warp Score Progression in Subjects Grouped by Disability Level

throughout the walk, with steadily increasing variance. Evidently gait deterioration appeared earlier in these subjects, becoming more common and extreme over time. Severe subjects scores are much higher on average, and unlike the other groups, variance decreases. This occurs because some subjects Warp Scores are high in minute 3, but most are high by minute 6, increasing consistency within the group.

4.4.2 Linking Gait Deterioration to Motor Fatigue

Correlations between gait deterioration and clinical outcomes may be found in Table 4.6. Several of these outcomes are not discussed in this chapter, but they are presented in Section

Table 4.6: Spearman Correlations and Partial Spearman Correlations between Clinical Outcomes and Warp Scores in All Subjects

	Spearman Correlation		Partial Spearman Correlation	
	r_s	p	pr_s	p
MSWS-12	0.429*	<0.0001	0.363*	<0.0001
MFIS	0.273*	<0.0001	0.136	0.0050
$MFIS_{phy}$	0.293*	<0.0001	0.161*	0.0008
EDSS	0.367*	<0.0001	0.257*	<0.0001
Cerebellar FSS	0.464*	<0.0001	0.258*	0.0001
Sensory FSS	-0.052	0.4251	-0.258*	0.0001
Pyramidal FSS	0.471*	<0.0001	0.241*	0.0002
Bowel and Bladder FSS	0.458*	<0.0001	0.170	0.0091
Vision FSS	0.227*	0.0004	0.098	0.1347
Cerebral FSS	0.164	0.0108	-0.063	0.3330
Brainstem FSS	0.254	0.0001	0.003	0.9666

*statistically significant ($p < 0.001$); MSWS-12: MS Walking Scale; MFIS: Modified Fatigue Impact Scale; $MFIS_{phy}$: MFIS Physical Subscale; EDSS: Expanded Disability Status Scale; FSS: Functional System Score

3.1. All MSWS-12 items had statistically significant corrected correlation to the Warp Score ($p < 0.001$) except Item 5 (Limited your balance when standing or walking?) ($p = 0.011$). Corrected correlation was strongest in Item 7 (Increased the effort needed for you to walk?) ($pr_s = 0.383$) and Item 6 (Limited how far you are able to walk?) ($pr_s = 0.331$). Corrected correlations were also significant for the following MFIS items: 4, 6, 9, 10, 17, 18, 20. Five of these items are on the MFIS physical subscale, with Item 20 (I have limited my physical activities) being strongest ($pr_s = 0.261$). Two items not on the physical subscale also reach significance: Item 9 (I have been limited in my ability to do things away from home) ($pr_s = 0.161$) and Item 18 (My thinking has been slowed down) ($pr_s = 0.163$).

The MSWS-12 regression model with lowest Bayesian Information Criterion (BIC) (Figure 4.7) included only walking speed and the Warp Score as predictors, both of which easily reached statistical significance ($p < 10^{-10}$). This model explained almost 74% of total MSWS-12 variance ($r^2 = 0.739$, adjusted $r^2 = 0.715$). In contrast, simple linear regression on walking speed or Warp Scores alone explained less of the total variance ($r^2 = 0.683$ and $r^2 = 0.480$, respectively).

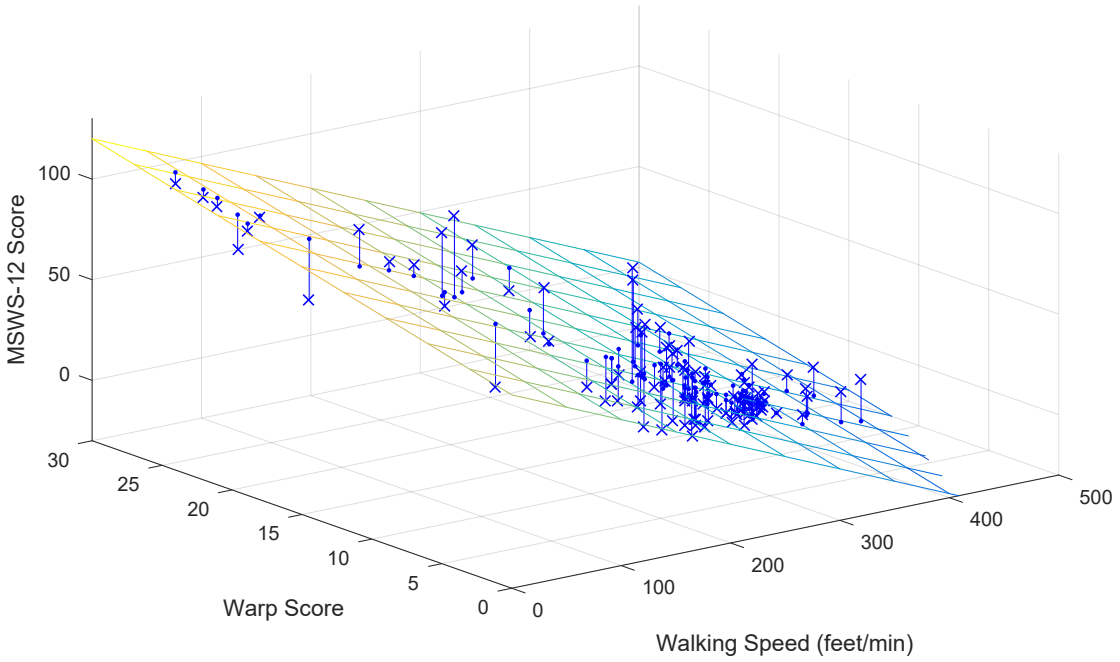


Figure 4.7: Final MSWS-12 Regression Model

Warp Scores and MSWS-12 scores were more strongly correlated by minute 6 ($pr_s = 0.322$, $p < 0.001$) compared to minute 3 ($pr_s = 0.215$, $p = 0.025$). Consequently, the MSWS-12 model performed better in minute 6 ($r^2 = 0.753$, adjusted $r^2 = 0.748$) than in minute 3 ($r^2 = 0.720$, adjusted $r^2 = 0.715$).

In one notable mild subject (EDSS = 2.0), the Warp Score rose from 15.5 in minute 5 to 23.5 in minute 6. Simple regression on walking speed estimated this subjects MSWS-12 as 12.4, higher than the average score among mild subjects ($\mu = 6.4$). However, their true MSWS-12 score was 27.1, resulting in a large model error (14.7). Interestingly, they also had an unusually high MFIS response of 3 on two items: Item 6 (I have had to pace myself in my physical activities) and Item 21 (I have needed to rest more often or for longer periods). In contrast, the multiple regression model accurately estimated the MSWS-12 score (estimate = 26.9, error = 0.2).

4.4.3 Summary of Clinical Results

As hypothesized, Warp Scores progressed over the course of the walk, particularly in subjects with mild disability. Progression was highly variable between subjects: gait deterioration was substantial in some, and negligible in others. Importantly, Warp Scores correlate with subjectively reported walking difficulty (MSWS-12) and physical fatigue (MFIS), showing that gait deterioration is clinically meaningful. Among individual items, the corrected correlation was strongest for MSWS-12 Item 7, which asks how much MS has "Increased the effort needed for you to walk". On the MFIS, corrected correlations were strongest for items related to physical fatigue.

These findings can be partly explained by disease severity; but importantly, correlations persist even after controlling for walking speed. Thus the Warp Score tells a different story compared to other objective measures, one confirmed by subjective report. This novel information is drawn out over the walk, as Warp Scores rise in mild and moderate subjects between minutes 3 and 6. Moreover, there is an increase in correlation to the MSWS-12 associated with this rise, suggesting that symptoms reported by subjects emerge with continued walking. As case in point, one subjects Warp Score rose only in minute six. Their speed was steady throughout the walk, yet they had a high score on the MFIS physical subscale compared to other subjects with similar disability. Walking speed was not sensitive to this subjects fatigue, but the Warp Score detected it.

Subsequent analysis has revealed that the Distance Score primarily measures physical fatigue, the Warp Score measures both physical fatigue and balance, and walking speed – the conventionally reported statistic – is affected by a wide variety of symptoms, including cognitive fatigue[135]. Additional details and results may be found in the original publication[79]. This result is the culmination of our work in DTW. By carefully extending the core DTW routine, we have constructed a novel, personalized monitoring algorithm with demonstrated clinical value.

4.5 Summary and Future Work

This chapter has developed a general approach to the monitoring of physiologic signals based on a distance measure – namely, dynamic time warping – rather than extracted signal features. While DTW can be viewed as a signal feature, the DTW distance and warping length for a given observation sequence are computed with respect to a second baseline sequence. As such, DTW is a *personalized* signal feature, making it well suited to be the core subroutine for adaptive signal monitoring.

Three modifications of DTW have been developed to facilitate its use in physiologic signal monitoring. The first of these combines several established methods into a single methodology appropriate for physiologic signals. The second is a novel summary statistic, the Warp Score, which may be more appropriate in a monitoring setting than the DTW distance. Indeed, it provided superior results in our clinical cohort. The last is an iterative algorithm for achieving rotation, scale, and offset invariance, which we believe is required in order to obtained meaningful results from a non-supervised, real-world algorithm deployment. RSOI-DTW was originally published in [120]. Each modification has been supported through one or more validation studies.

Compared to many other approaches, DTW is intuitive and interpretable. This is particularly true if progression can be characterized using representative sequences which evolve over time. In our experience, DTW can be explained from first principles to clinical collaborators, which is an important, underestimated advantage over black box approaches.

Currently, the approach is powerful because (1) it makes very few assumptions about signal characteristics, and (2) it is inherently personalized, with each subject serving as his or her own control. In principle, it is also powerful because (3) it can adapt to subjects in real time, and (4) it maintains a library of examples which illustrate signal progression over time.

Long-term, real-world monitoring of physiologic signals faces many challenges not encountered in this work. RSOI-DTW is robust to many real-world variables – mis-oriented sensors, changing speeds, variable surfaces, etc – but it has been tested in a controlled setting. A

prospective trial is needed to validate the algorithm's ability to detect pathology "in the wild", but there are a number of challenges to overcome, including development of an online, efficient implementation.

As discussed in Section 4.1.2, a DTW-based approach to signal monitoring is truly adaptive only if new baseline sequences are added as observations are encountered. The solution is not simply to add sequences that match the baseline, which would result in expanding memory and computational requirements. More importantly, such a strategy would never "forget", which runs counter to the spirit of adaptive monitoring. Ideally we'd like to capture representative sequences over time to characterize signal progression in detail. Finding representative sequences among a finite set of observations can be accomplished with clustering algorithms for arbitrary distance metrics, such as the k-medoids algorithm or affinity propagation. These algorithms may be an appropriate starting point for an algorithm that maintains an evolving library of representative sequences. Based on these results, we anticipate that a similar approach could be successfully applied much more broadly in health monitoring.

Most importantly, future work should focus on continued application of personalized and adaptive algorithms in a variety of health monitoring scenarios. Prospective trials will be explored in both clinical and mobile health settings.

Chapter 5

Adaptive Symptom Reporting

Adaptive Symptom Reporting (ASR) is a novel approach to longitudinal PRO assessment which uses personalization, domain-specific knowledge, and Bayesian inference to substantially reduce the number of questions required for accurate trait estimation. The foundation of ASR is Item Response Theory (IRT), which formalizes the relationship between PRO items and the underlying trait they are intended to measure. In this work, we repeatedly refer to PROs and clinical applications, yet ASR can be applied to any questionnaire which satisfies the assumptions inherent in IRT modeling.

PROs are becoming increasingly important. Health care is now viewed as a service, and the patient-centered care movement has placed greater emphasis on patient experiences than ever before. PROs are now used for treatment evaluation[74], care evaluation, and health system evaluation, and the NIH supports PRO development through its PROMIS initiative[136]. At the same time, PROs are burdensome to patients, who find themselves providing the same answers over and over again. Recent work in item response theory (IRT) has improved PROs from a psychometric standpoint[137], but it has not removed this redundancy or alleviated patient burden.

ASR leverages IRT to create a statistical model of symptom progression. This model is used to drive a modern symptom reporting framework which dynamically selects PRO items

based on evolving knowledge of subject disability status. The algorithm maintains a running estimate of subject disability based on the history of past PRO responses. It then uses this estimate to optimize subject interactions by selecting PRO items most useful in pinpointing that particular subject's disability. In the end, patient burden is significantly reduced, in terms of the number of PRO responses, with minimal loss of information.

The chapter begins with brief background on IRT and the specific IRT model used in this work, the graded response model (GRM). We review relevant literature in two areas: longitudinal IRT and IRT applied to PROs. ASR is then presented, beginning with IRT model development and culminating with the ASR algorithm itself. The discussion of model development highlights considerations most relevant to ASR. Finally, the overall approach is validated via our target application, walking ability in MS. Most notably, the ASR algorithm is evaluated retrospectively with responses to the MSWS-12, which we described in Section 3.1.4. The chapter summary includes limitations and plans for continued ASR development.

Chapters 4, 5, and 6 each focus on a different variety of adaptation. ASR adapts by tailoring question selections – its *actions* – to the patient based on their response history. Referring back to Chapter 2, ASR addresses the problems of patient burden and subjectivity by modernizing and streamlining PRO collection. Variability between persons can also arise in PRO development, where it is called differential item functioning (DIF), as discussed in Section 5.2.1. However, an individualized model of PRO responses would not be useful due to the subjective nature of PRO measurements. Here we adjust to heterogeneity by prompting subjects only with relevant, timely questions.

5.1 Background

5.1.1 Item Response Theory (IRT)

IRT – the foundation of ASR – is a family of models and techniques which formalize the relationship between latent trait and sets of items designed to measure them. IRT is commonly

used in educational testing to measure test-taking ability, and in psychology to measure personality traits and other psychological constructs. In both cases, the quantity being measured is hidden and continuous-valued. In general, IRT can be used whenever one may assume that item responses are influenced by one or more such constructs, which are most often called *traits*.

The trait-item relationship is commonly modeled as a logistic function; in fact, this is the well-known Rasch model[138]. IRT is much like logistic regression, except the value of the continuous, predictor variable is unknown. To draw an analogy, IRT is to logistic regression as factor analysis is to linear regression. IRT and factor analysis build a model to estimate a hidden trait, whereas regression classifies or predicts from a known predictor variable.

A number of IRT models have been developed to fit different assumptions and item formats. In this work, we focus primarily on the graded response model (GRM) and other models appropriate for the rating scales found in most PROs. In addition to the GRM, this includes the polytomous Rasch model (PRM)[138] and the generalized partial credit model (GPCM) [139], but we present mathematical details only for the GRM. For discussion of other models, we recommend the comprehensive book by de Ayala[140].

Fitted rating scale models are often presented using category response curves (CRCs), which specify the probability of observing a particular response, or category, as a function of the latent trait. These probabilities are derived from multiple logistic functions, as we show in equations 5.1 - 5.4. The total number of model parameters – which govern the CRCs – depends on the form of the model and the number of item categories. Figure 5.1 shows the fitted CRCs for Item 1 ("How much has your MS limited your ability to walk?") of the GRM for the MSWS-12, which we discuss later in the chapter. As walking disability increases, the subject becomes more likely to choose higher values on the rating scale.

The GRM specifies a relationship between a trait θ and responses to item i via five model parameters: a discrimination parameter α_i and four category threshold parameters β_{ij}, $j \in \{2, 3, 4, 5\}$. As shown below, $j = 1$ is a special case. The probability $P(C_i = j \mid \theta)$ of

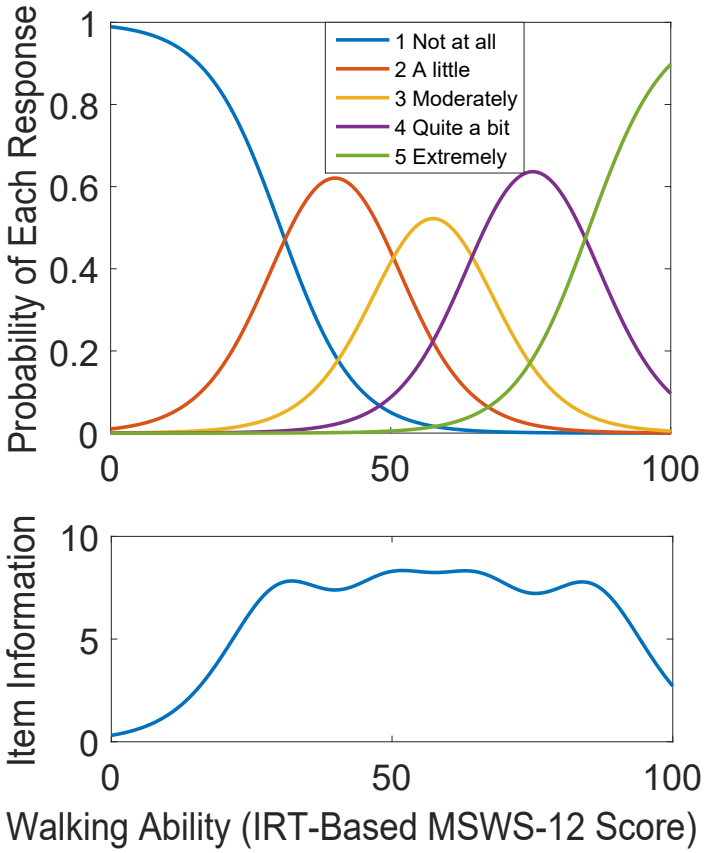


Figure 5.1: Category Response Curves and Item Information for MSWS-12 Item 1

response category j on item i for a person with walking ability θ is given by the following equations. Note that these have been adapted to suit 1-5 scale indexing, as found on the MSWS-12.

$$P(C_i \geq 1 \mid \theta) = 1 \tag{5.1}$$

$$P(C_i \geq 6 \mid \theta) = 0 \tag{5.2}$$

$$P(C_i \geq j \mid \theta) = \frac{e^{\alpha_i * (\theta - \beta_{ij})}}{1 + e^{\alpha_i * (\theta - \beta_{ij})}}, j \in \{2, 3, 4, 5\} \tag{5.3}$$

$$P(C_i = j \mid \theta) = P(C_i \geq j \mid \theta) - P(C_i \geq (j+1) \mid \theta) \tag{5.4}$$

The CRCs are the functions $f(\theta) = P(C_i = j \mid \theta)$. The bottom panel of Figure 5.1 shows the item information, which is based on the Fisher information about θ carried by the item i. The item information for item i is given by the following[140]:

$$I_{ij}(\theta) = \frac{\left(\frac{dP(C_i=j|\theta)}{d\theta}\right)^2}{P(C_i = j \mid \theta)} \tag{5.5}$$

$$I_i(\theta) = \sum_{j=1}^{5} I_{ij}(\theta) \tag{5.6}$$

Given the fitted GRM, walking ability may be estimated as the value of θ maximizing the likelihood of observed MSWS-12 responses (maximum likelihood estimation [MLE]). A prior distribution over θ may also be used (maximum *a posteriori* [MAP] or expected *a posteriori* [EAP]). This method avoids the unbounded estimates returned by MLE when all responses have maximum value (e.g. all MSWS-12 responses = 5) or minimum value (e.g. all MSWS-12 responses = 1). In this work, trait value is estimated as the expected value of the posterior using a uniform prior over a suitable range of θ. Herein we refer to latent trait estimation as IRT-based scoring.

To fit an IRT model, a large set of response data must be collected. The model is usually fitted to responses by expectation maximization (EM) with subject traits as the hidden variables. In the E-step, a point estimate of model parameters is used to generate a distribution over trait values. Then, in the M-step, model parameters are optimized (MLE) with respect to this distribution. A number of alternative estimation procedures can also be used; for details, see the following sources[140, 141, 142].

5.1.2 Longitudinal IRT

Longitudinal IRT refers to the use of IRT to interpret item responses collected over time, in sequence, from the same person. The goal of a monitoring platform is to infer system

health using longitudinal measurements. IRT has been applied to longitudinal data from a number of domains with a variety of goals in mind. In political science, it has been used to find "change points" in the preferences of Supreme Court justices based on their voting history[143]. In educational testing, longitudinal methods were used to estimate students' test-taking ability while accounting for possible learning between measurements[144]. Item parameters can be estimated directly from longitudinal data or fixed beforehand, as in the current work; Andrade and Tavares have published a method for tracking IRT-based ability when item parameters are fixed[145].

However, our work differs from the majority of longitudinal IRT analyses, which have focused on accurate parameter and/or ability estimation after the data has been collected. In contrast, our algorithm is designed to reduce subject burden by optimizing item selection in real time. To our knowledge, ours is the first longitudinal, IRT-based algorithm designed for real-time subject interaction and question selection. Moreover, it is the first algorithm of its kind for mobile PRO assessment.

5.1.3 IRT in the PRO Literature

IRT offers important benefits to clinical practice and research. Items with poor performance can be identified by their parameters and replaced to reduce measurement error, potentially lowering the MCID (see Section 3.1.5). IRT uses a standard error of estimate (SEE) that varies with trait level (instead of a single standard error of measurement), making it possible to quantify PRO precision from patient to patient. Changes in score can be meaningfully compared across the spectrum of disability, improving measurement of treatment effectiveness in diverse samples. Unlikely response patterns can be detected, so that if a patient misunderstands a question or the examiner incorrectly transcribes a response, the error can be corrected. Finally, IRT facilitates development of computerized adaptive PRO measures, which can decrease length without increasing error, reducing burden to patients[141, 137]. In

Section 5.3.3, we describe our effort to facilitate clinical adoption of the IRT-based MSWS-12 scoring.

Because of these benefits, IRT analyses have become common in the PRO literature. Indeed, Hobart et al. recommended secondary IRT or Rasch analysis of the MSWS-12 in their initial publication[90]. IRT has been used to evaluate the Disabilities of the Arm, Shoulder, and Hand in persons with MS[146], and the Modified Fatigue Impact Scale (MFIS) has been fitted to a Rasch model[98]. Velozo et al. used Rasch analysis to create a computerized adaptive measure of walking, climbing, and running[147]. Both IRT and classical test theory were used to develop a computerized adaptive testing (CAT) item bank in multiple languages for quality of life assessment in MS[148]. Recently, a computerized adaptive version of the MS International Quality of Life Questionnaire (MusiQoL) was created using a multidimensional graded response model (GRM) similar to the one-dimensional GRM in the current work[149].

As discussed in Section 3.1.4, the NIH has recognized the need to improve and standardize PRO measures, founding the Patient-Reported Outcomes Measurement Information System (PROMIS) in 2004. The goal of the PROMIS initiative is to create a consensus-based framework for self-reported health using IRT methods[136].

The most common application of IRT in the PRO literature is the development of "short forms" which reduce PRO length by selecting questions with high information. Figure 5.1 illustrates the short form development process. Short forms have been developed for pain, fatigue, sleep disturbances, and many other symptoms[150]. Emphasis on short forms underscores the problem of patient burden. Unfortunately, item information depends on trait level, so a short form – or any other static PRO – cannot be as effective as an adaptive PRO on a per question basis.

CATs like the one developed by Michel et al. improve on short forms by adapting to the patient as more questions are answered. However, they are not equipped to draw upon previous responses. Our ASR algorithm functions much like a CAT, but it incorporates a model of disability progression over time. Thus it can reason about patient disability on an

Figure 5.2: PRO Short Form Development

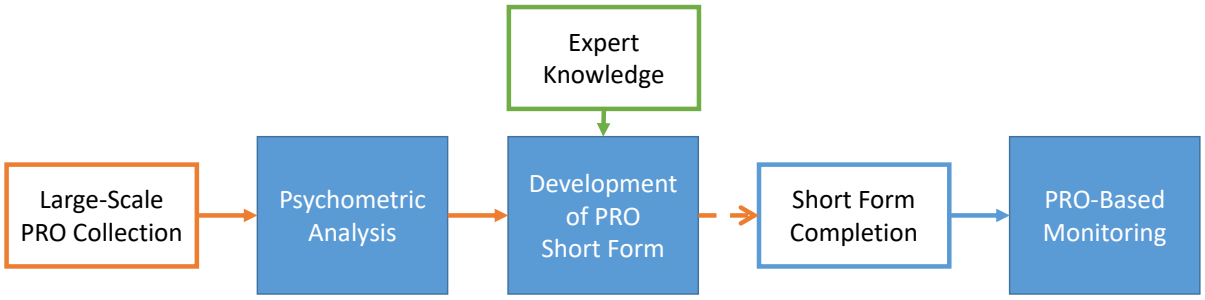
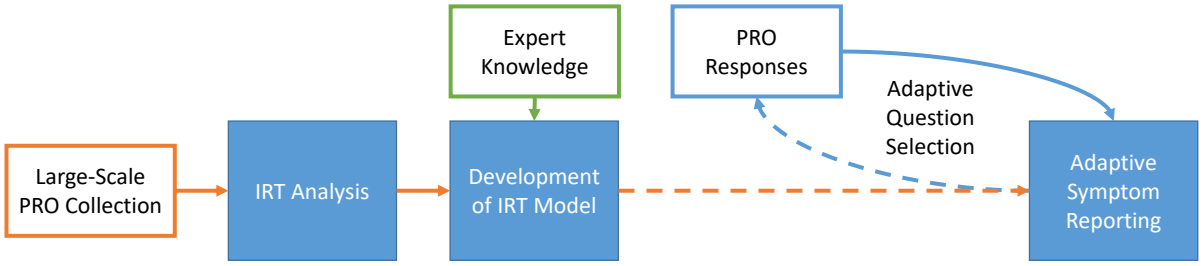


Figure 5.3: Deployment of ASR



ongoing basis to dramatically reduce the number of patient responses required for accurate disability estimation.

5.2 Adaptive Symptom Reporting (ASR)

ASR deployment for a particular PRO proceeds in two phases through the process depicted in Figure 5.3. First, an IRT model of the PRO must be developed. A large number of responses – ideally more than 500[151] – are collected and analyzed with IRT to develop an IRT model. Expert knowledge is needed to guide the IRT analysis, particularly in the model selection phase. Note that initial PRO development also requires expert knowledge and psychometric analysis to select and test potential items, respectively.

The IRT model specifies $P(R \mid \theta)$, the likelihood of trait values given an observed response vector. A second distribution, $P(\theta_t \mid \theta_{t-1})$, must also be fixed. In this work, t and $(t-1)$ are successive observations, where the time between observations may vary, therefore we refer to

the time elapsed between them as $\delta(t-1,t)$, or simply δ_t. The distribution $P(\theta_t \mid \theta_{t-1}) = f(\delta_t)$ is a function of δ_t governing trait progression over time.

In our validation studies, $P(\theta_t \mid \theta_{t-1})$ has been fixed based on literature review and expert knowledge, as described in Section 5.3.4. When possible, this choice should be informed directly by data collection and updated as more observations are encountered.

The second phase of ASR deployment is the implementation as a mobile app followed by distribution to patients. In this work, we have simulated this phase by validating ASR using previously collected longitudinal PRO data from MS subjects. Our plans for prospective deployment are described in Section 5.4.2.

5.2.1 Building the IRT Model

In this section, we overview the first phase of ASR deployment: development of the IRT model. The fitting of IRT models goes far beyond this brief treatment, of course; the books by de Ayala [140] and Embretson and Reise [141] are good starting points. These comments are intended to briefly orient the reader to model fitting, emphasize the measurement properties required for ASR, and highlight ASR- and PRO-specific considerations.

Assessing Scale Dimensionality

Like all applications of IRT, ASR depends on the assumption that item responses are influenced by an underlying trait. This trait may have more than one dimension, though we focus on the one dimensional case in this work. Understanding the dimensionality of the latent trait is one of the earliest steps in IRT model development and thus ASR as well.

Many packages are available for dimensionality assessment. We have used the *psych* package for our preliminary psychometric assessment[152]. One, two, and three-factor latent trait solutions may be extracted via maximum likelihood using the polychoric correlation matrix, which is appropriate when working with polytomous responses[153].

The theory behind IRT and ASR extends to the multidimensional case, and extending the ASR algorithm is straightforward. However, our implementation performs a numeric integration each time it runs, so expanding model dimensionality would substantially increase computation. The tradeoff between model accuracy and computation should be considered carefully when working with multidimensional traits. In the multidimensional case, a more elegant, less computationally intensive approach may be needed to keep run time within reasonable limits.

Choosing a Model

The choice of IRT models is constrained by the format of items, but typically there are several reasonable choices. In this work, the polytomous Rasch model [138], graded response model (GRM) [139], and generalized partial credit model (GPCM)[154] were all suitable for the rating scale format found on the MSWS-12. Models may always be compared using the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC), and likelihood ratio chi-square tests are appropriate when considering the addition of specific parameters. If dimensionality assessment is inconclusive, models of varying dimension may be directly compared in terms of BIC, AIC, and model fit.

We have fitted our models to MSWS-12 responses via maximum likelihood estimation using the *mirt* package for R[155]. Other R packages are available (e.g. *ltm* and *eRm*), and there are several standalone programs as well. MLE most commonly proceeds via expectation maximization, though some packages operate directly on the liklihood function using gradient-based methods.

The inclusion or omission of discrimination parameters has added significance when developing a model for ASR. Working from equations (5.3), (5.4), and (5.5), we may conclude that item information is proportional to the square of the discrimination parameter α_i for item i. Consequently, models which incorporate item-specific discrimination parameters – which includes the GRM and GPCM but not the polytomous Rasch model – have higher

variability in information between items. As we will see, item selection in ASR is driven by item information, therefore items with higher discrimination parameters are more likely to be chosen. We believe this behavior is beneficial. If model estimation is accurate, items with higher discrimination parameters are more closely related to the latent trait, making them more informative. Still, it must be noted that the choice of model, and in particular the choice to incorporate discrimination parameters, has important consequences for ASR that have not been explored.

Model Fitting and Assessment

Goodness of fit assessment is an active area of study in the IRT literature[141]. Simpler models such as the Rasch model have an established set of fit indices, but the issue is not settled for the GRM and other polytomous models. In this work, we assess model fit with the Z_h and S-X2 statistics[156, 157] and person fit with the Z_h statistic[156], both calculated in *mirt*. Model-based CRCs and item information plots were generated with mirt and reflect the fitted model parameters. Empirical CRCs were plotted by dividing subjects into ten groups according to model-based log-odds score, then calculating category response frequencies for each group. Similarity between model-based CRCs and empirical CRCs serves as a preliminary assurance that model assumptions are reasonable for a given dataset.

Fit assessment is critical to ASR, just as it is to IRT, because trait estimates are only as good as the model that generated them. This model is fixed throughout ASR deployment, so the effects of poor fit are far-reaching. ASR adapts by choosing questions based on previous observations, but it is not *model-adaptive*: its internal statistical model of the world does not change. This distinction is discussed further in Chapter 6, as active event identification *is* model-adaptive, and in our discussion of future directions for ASR.

Differential Item Functioning (DIF) analysis is an important element of fit assessment which tests for trait-item relationships that vary between subpopulations. DIF would be present, for example, if older subjects were more likely to answer a test item correctly than

younger subjects with the same test-taking ability. A common method to test a specific item for DIF proceeds as follows. First, the population is divided into the subpopulations in which DIF is suspected, and distinct models are fitted for each subpopulation. In these models, only the parameters of the item being tested are allowed to vary between subpopulations. The models are then compared to a baseline model with a likelihood ratio chi-square test, which measures the statistical significance of the improved likelihood[158].

There are a number of alternative procedures to detect DIF[140], but we've employed the aforementioned procedure in the this work. Two primary DIF analyses were conducted, one based on age and the other based on sex. In the age-based DIF analysis, subjects were divided into two subpopulations of equal size by using the median age (48) as the cut-off to define groups. All MSWS-12 items were individually tested. These analyses were conducted to determine whether age or sex affected response patterns despite controlling for the IRT-based score.

DIF analysis is important for the same reasons as fit assessment. If items function differently based on demographic traits, then the estimates generated by ASR may be less reliable in certain subpopulations. If necessary, this can be solved by "tuning" the IRT model to a particular subpopulation. Personalized model adaptation should also be explored, but may not be reasonable when tracking latent constructs unless anchored to objective findings.

5.2.2 The ASR Algorithm

In adaptive symptom reporting (ASR), we leverage an IRT model for a PRO – in this case, the GRM for the MSWS-12 – to maintain a running estimate of disability. Instead of requiring *all* PRO responses in each assessment session, the ASR system sequentially chooses items to keep uncertainty about PRO-defined disability below a specified threshold. More formally, the ASR procedure estimates $P(\theta_t|R_{1,\dots,t})$, the distribution over disability at time step t given the complete history of item responses $R_{1,\dots,t}$ observed so far, the IRT model, and a

probability distribution $P(\theta_t|\theta_{t-1}) = f(\theta_t, \theta_{t-1}, \delta_t)$ governing possible changes in disability as a function of δ_t, the time elapsed between $[t-1]$ and t.

The algorithm begins by estimating $P(\theta_0|R_0)$ based on the full set of responses R_0 using traditional IRT, as described in the previous section. In this work, R_0 are the paper-based MSWS-12 responses collected during the initial visit. In subsequent opportunities to interact with the subject, the algorithm begins by updating its distribution over θ based on δ_t, the time elapsed since the previous iteration. This update depends on $P(\theta_t|\theta_{t-1})$ as follows:

$$P(\theta_t|R_{1,\ldots,t-1}) = \int P(\theta_t, \theta_{t-1}|R_{1,\ldots,t-1})d\theta_{t-1} \tag{5.7}$$

$$= \int P(\theta_t|\theta_{t-1}, R_{1,\ldots,t-1})P(\theta_{t-1}|R_{1,\ldots,t-1})d\theta_{t-1} \tag{5.8}$$

$$= \int P(\theta_t|\theta_{t-1})P(\theta_{t-1}|R_{1,\ldots,t-1})d\theta_{t-1} \tag{5.9}$$

Note that by simplifying (5.8) to (5.9), we assume that θ_t is conditionally independent of prior responses given θ_{t-1}. The expected value of this distribution is then used to estimate the information which could be obtained from each item i: $I_i[t] = I_i(\mathbf{E}(P(\theta_t|R_{1,\ldots,t-1})))$. A single item is selected at random with probability proportional to the square of its information $(I_i[t])^2$, and the subject is prompted to respond to this item. Because of the retrospective nature of the current work, we have simulated this action by using only the response(s) to the selected item(s) in subsequent calculations.

Once the new response is obtained, the ASR system again updates its distribution over θ to incorporate the new information:

$$P(\theta_t|R_{1,\ldots,t}) = \frac{P(R_t|\theta_t, R_{1,\ldots,t-1})P(\theta_t|R_{1,\ldots,t-1})}{\int P(R_t|\theta_t, R_{1,\ldots,t})P(\theta_t|R_{1,\ldots,t-1})d\theta_t} \tag{5.10}$$

$$P(\theta_t|R_{1,\ldots,t}) = \frac{P(R_t|\theta_t)P(\theta_t|R_{1,\ldots,t-1})}{\int P(R_t|\theta_t)P(\theta_t|R_{1,\ldots,t-1})d\theta_t} \tag{5.11}$$

$$\propto P(R_t|\theta_t)P(\theta_t|R_{1,\ldots,t-1}) \tag{5.12}$$

Similar to before, the simplification between (5.10) and (5.11) follows because the probability of an item response is conditionally independent of previous responses given the current trait level. Note that $P(R_t|\theta_t)$ is the term specified by the IRT model.

Finally, the algorithm checks to see whether the uncertainty in the new estimate is acceptable by calculating the standard deviation of $P(\theta_t|R_{1,\ldots,t})$ and comparing it to the threshold ϵ, a tunable parameter of the ASR algorithm. If $P(\theta_t|R_{1,\ldots,t}) < \epsilon$, the algorithm ends the current session; otherwise, it selects a *different* item – one not yet obtained in the current session – using the same procedure as before. This continues until either (1) the desired threshold ϵ is reached, or (2) all items have been selected.

Adaptive Symptom Reporting Algorithm

procedure ASR($P(\theta_{t-1}|R_{1,\ldots,t-1}), \delta_t, \epsilon$)
  $P(\theta_t|\theta_{t-1}) \leftarrow f(\theta_t, \theta_{t-1}, \delta_t)$
  $P(\theta_t|R_{1,\ldots,t-1}) \leftarrow \int P(\theta_t|\theta_{t-1})P(\theta_{t-1}|R_{1,\ldots,t-1})d\theta_{t-1}$
  $I_i[t] \leftarrow I_i(\mathbf{E}(\theta_t|R_{1,\ldots,t-1}))$ for all items i
  repeat
    select an item i with probability $\propto (I_i[t])^2$
    prompt subject with item i
    $P(\theta_t|R_{1,\ldots,t}) \leftarrow P(R_t|\theta_t)P(\theta_t|R_{1,\ldots,t-1})$
    $I_i[t] \leftarrow 0$
  until $stdev(P(\theta_t|R_{1,\ldots,t})) < \epsilon$
  return $P(\theta_t|R_{1,\ldots,t})$
end procedure

In this study, MSWS-12 responses had already been collected, so the algorithm was run at pre-defined intervals. In contrast, an online implementation of ASR could prompt subjects at

any time. We suggest running the algorithm daily, which would result in prompting whenever uncertainty about θ falls below ϵ. No modification to the algorithm would be needed.

Our item selection procedure was designed to favor higher-information items while avoiding repeated selection of the same limited subset. Due to the discrimination parameters present in the model, some items have much higher information than others: in fact, some items do not have highest information for *any* value of θ. Though higher-information items are heavily weighted by the ASR procedure, lower-information items always have a chance of being selected. This choice may improve the user experience, and it avoids potential pitfalls due to item bias, imperfect estimation of IRT parameters, and other real-world deviations from IRT assumptions.

5.3 Validating ASR: Walking Ability in MS

In this section, we use ASR to track walking ability in MS. It contains a number of application-specific insights, but the majority of results are needed to show that ASR functions as intended. The highlight of the section is perhaps the discussion of tradeoffs between estimation accuracy and the number of questions selected by the algorithm. Reducing the number of questions necessarily results in a loss of information – nothing is free – but our results show tradeoffs that are favorable in typical applications.

We begin with an overview of the two studies required for ASR validation. After presenting our IRT model of the MSWS-12, we describe our efforts to facilitate its use in clinical care and research. Finally, we present validation results for the ASR algorithm. ASR accurately tracks walking disability in our cohort while reducing patient burden – in terms of the number of questions – by 42-75%.

5.3.1 Study Design

Two separate studies were need to validate ASR corresponding to the phases we have described: building the IRT model and validating ASR itself. These phases lent themselves to very different study designs. Building the model required a single MSWS-12 collection from a large cohort of subjects, whereas ASR validation required repeated collection from a much smaller cohort. Both studies were conducted through the MS clinic at the UVA Department of Neurology.

Large-Scale MSWS-12 Collection

The first study was a large-scale collection of MSWS-12 responses by mail. Here we describe the study itself and the demographic characteristics of responders and non-responders, who were similar in age, sex, disease subtype, and disease duration. We have no data on the disability status of non-responders, so it's impossible to know whether disability severity is similar between groups. However, IRT does not require the sample to be unbiased in terms of the trait being modeled[140].

Results of the model fitting process are presented in the next section.

Procedures Study procedures were approved by the University of Virginia (UVa) Institutional Review Board for Health Sciences Research. A four-page survey packet containing the MSWS-12, MFIS, and a dietary survey was mailed to 604 UVa Department of Neurology MS clinic patients with clinically definite MS[102]. Packets were keyed and coded to protect patient confidentiality. A chart review was conducted on all 604 subjects, including those who did not respond. Age, sex, MS subtype, and disease duration were obtained for all subjects. Responders timed 25-foot walk (T25FW) and neurological exam findings were obtained when available from clinic visits within the previous year.

To assess sample representativeness, demographic characteristics were compared between responders and non-responders using t-test and chi-square as appropriate for continuous and

Table 5.1: Demographics and summary statistics

	Responders	Non-Responders	p-value
Number of subjects	293	311	
Gender (female)	74.4%	76.5%	0.12
Age, median (range)	48 (19 – 80)	49 (14 – 81)	0.14
MS subtype			
Relapsing remitting MS	71.5%	75.6%	
Primary progressive MS	8.5%	6.2%	0.67
Secondary progressive MS	12.5%	10.8%	
Progressive relapsing MS	7.5%	7.4%	
Disease duration, mean (range)			
<1 year	1.0%	4.3%	
1-5 years	26.0%	20.1%	0.17
5-10 years	29.5%	28.4%	
10-15 years	16.0%	20.3%	
>15 years	27.5%	26.9%	
MSWS-12 score, mean (range)	40.4 (0 – 100)	NA	
T25FW speed, mean (range)	4.85 (0.37 – 9.71)	NA	

*speed reported in feet per second; MSWS-12: MS Walking Scale; T25FW: Timed 25-foot Walk

ordinal variables, respectively. Disease duration was categorized (<1 year, 1-5 years, 5-10 years, 10-15 years, >15 years) and analyzed as an ordinal variable.

Responders The response rate (293/604) was similar to the rate reported by Hobart et al. in their initial psychometric evaluation (766/1530)[90]. Demographics of responders and non-responders are shown in Table 5.1. Responders were 74.4% female, with a median age of 48 (range: 19 – 80). There were no statistically significant differences between responders and non-responders in terms of age ($p = 0.12$), sex ($p = 0.14$), MS subtype ($p = 0.67$), or disease duration ($p = 0.17$). The mean MSWS-12 score among responders was 40.4 with a standard deviation of 34.1, and the mean T25FW speed was 4.85 ft/sec with a standard deviation of 1.77 ft/sec. Nine of the 293 MSWS-12 records were incomplete, leaving 284 available for IRT analysis. Our IRT model of the MSWS-12 from this analysis was originally present in [159].

Retrospective Validation

With the IRT model of the MSWS-12 in place, ASR was validated retrospectively using MSWS-12 data from the pilot study described in Chapter 3. A complete description of study procedures and demographic characteristics of the study cohort may be found in Section 3.2. Briefly, 31 subjects with MS completed the MSWS-12 using a web portal over a six month period. MSWS-12 responses were also collected in person at the beginning and end of the study.

MSWS-12 Scores at initial visit spanned the full range of values (0 - 100), with 12, 5, 6, and 8 subjects in the first through fourth quartiles, respectively. Subjects completed the MSWS-12 an average of 8 times, two more than the minimum study requirement. In total there were 246 sets of MSWS-12 responses (sessions) available for analysis. Complete sets of MSWS-12 responses were available for each session, but only the responses requested by ASR were provided to it, simulating a prospective trial.

The complete set of responses from the initial visit were used to initialize ASR. Using our established notation, R_1 are the initial visit responses which determine $P(\theta_1|R_1)$.

5.3.2 IRT Model of the MSWS-12

This section presents our IRT model of the MSWS-12, which was fitted to the responses to our mail survey. Results are organized to parallel the structure of the previous section in order to illustrate the methodological considerations discussed therein. ASR is valid only insofar as this model is valid, so it is important to carefully assess dimensionality and model fit. Moreover, the item information results have important implications for ASR algorithm design.

Dimensionality and Model Selection

A single latent factor was able to explain approximately 87% of the total variance among responses. Adding a second factor increased the total variance explained by only 2% to

Table 5.2: Information criteria for all exploratory models.

Model	BIC	AIC
1D Rasch	6112.5	5933.7
1D GRM	5977.7	5758.8
1D GPCM	6021.5	5802.5
2D GRM	5919.7	5660.7
2D GPCM	5951.3	5692.2
3D GRM	5972.7	5677.2
3D GPCM	6009.1	5713.6

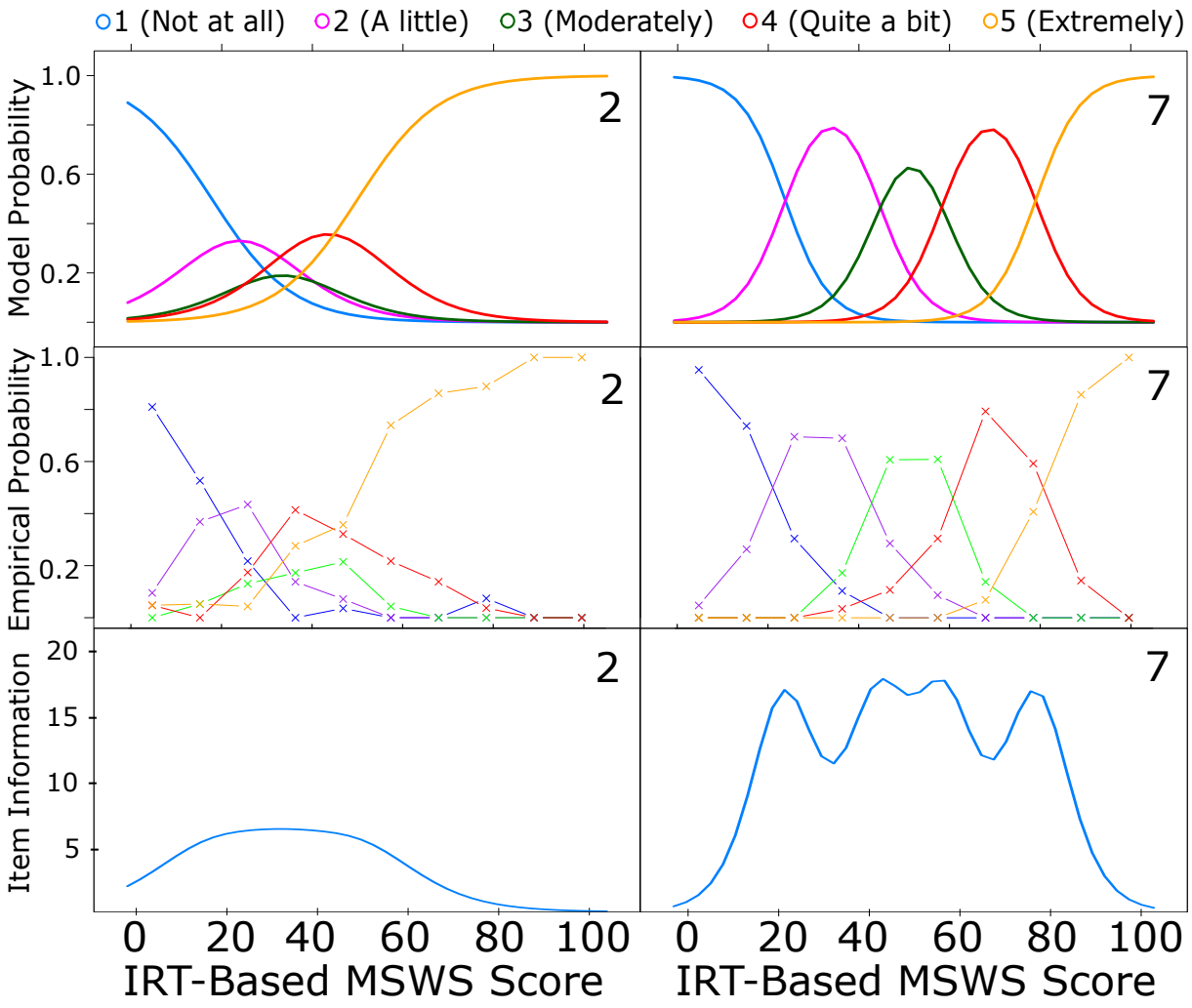
BIC: Bayesian Information Criterion; AIC: Akaike Information Criterion; 1D, 2D, 3D: One, Two, and Three-Dimensional; Rasch: Polytomous Rasch Model; GRM: Graded Response Model; GPCM: Generalized Partial Credit Model

approximately 89%, and adding a third factor increased it by another 2% to 91%. All factor loadings were greater than 0.90 in the one-factor solution. In the two-factor solution, only Item 7 loaded on the second factor with magnitude greater than 0.20 (0.42 for Item 7). In the three-factor solution, the largest loading on the third factor was 0.23 (Item 12). Factor loadings for the second and third factors did not fit any a priori hypotheses or clinically relevant patterns.

Model comparison results (BIC and AIC) are shown in Table 2. The polytomous Rasch model had the highest BIC and AIC among all models, eliminating it from further consideration. The BIC and AIC favored the one, two, and three-dimensional GRMs over their GPCM counterparts. The two-dimensional GRM outperformed the one-dimensional GRM in the chi-square test ($\chi^2 = 120.2$, $df = 11$, $p < 0.01$) as well as the BIC and AIC (Table 2), but the three-dimensional GRM did not yield further improvement ($\chi^2 = 3.5$, $df = 10$, $p = 0.97$).

The one-dimensional GRM was selected based on the factor analysis results and existing research supporting a one-factor solution[90, 160]. Additional factors were present to some degree, but did not seem to be clinically relevant. While model comparison did favor the two-dimensional GRM, the second dimension added marginal statistical value and no clinical value.

Figure 5.4: CRCs, Empirical CRCs, and Item Information for MSWS-12 Items 2 and 7



Goodness of Fit

Figure 5.4 shows the model-based CRCs, empirical CRCs, and item information for Items 2 and 7. The two sets of curves are well matched on inspection. Discrimination parameters range from 0.112 (Item 4) to 0.217 (Item 10), as shown in Table 5.4. Along with Item 4, Items 2 and 5 have smallest discrimination parameters (0.114 and 0.120). Category thresholds are adequately spaced with the exception of Items 2 and 9, to which response 3 (Moderately) is never most likely.

Goodness of fit was assessed using two statistics appropriate for polytomous items, the Z_h and S-X2 indices[156, 158, 161]. Both fit indices are presented in Table 5.4. The Z_h

Table 5.3: MSWS-12 Discrimination parameters (DP), goodness of fit statistics, and differential item functioning (DIF) statistics

Item	DP	Goodness of Fit				Age DIF		Sex DIF	
	α	Z_h	S-X2	S-X2 df	S-X2 p	χ^2 ($df = 5$)	p	χ^2 ($df = 5$)	p
1	0.139	0.74	20.14	26	0.78	5.02	0.41	1.70	0.89
2	0.114	-0.03	26.28	36	0.88	19.02	0.002**	2.50	0.78
3	0.136	0.58	28.52	27	0.38	8.97	0.11	3.65	0.60
4	0.112	0.32	36.13	35	0.42	0.58	0.99	5.98	0.31
5	0.120	0.57	21.40	26	0.72	5.28	0.38	7.29	0.20
6	0.163	0.78	24.17	24	0.45	2.89	0.72	0.97	0.96
7	0.209	1.20	30.10	22	0.12	9.89	0.08	4.15	0.53
8	0.139	0.63	39.01	34	0.25	4.49	0.48	1.49	0.91
9	0.144	0.58	33.09	31	0.37	7.19	0.21	4.86	0.43
10	0.217	1.26	21.51	25	0.66	4.80	0.44	5.87	0.32
11	0.185	0.94	22.58	20	0.31	5.20	0.39	13.76	0.02*
12	0.139	0.54	51.08	29	0.01**	6.66	0.25	5.23	0.39

index is a normalized measure of the likelihood of observed responses to an item[156]. It approximates a standard normal distribution, so a majority of values should fall between -1 and 1, and absolute values above 2 suggest poor fit. All Z_h values fell within the acceptable range, with a maximum absolute value of 1.26 for Item 10. The S-X2 statistic is a generalized chi-square fit index showing improved performance compared to the traditional chi-square index[157, 161]. An S-X2 p-value less than 0.05 suggests poor fit. Item 12 fit poorly according to the S-X2 index but not the Z_h index, as seen in Table 5.4, and all other items had S-X2 p-values greater than 0.05. These results show that observed responses were consistent with predictions, validating our choice of model.

Item and Test Information

Item information for Items 2 and 7 may be found in the bottom panels of Figure 5.4. The SEE and its squared reciprocal, the total information, may be found in Figure 5.5. Since item information peaks near the response category thresholds, information is greatest and error lowest at intermediate scores[140]. As a result of the discrimination parameters, Item 4 has lowest peak information followed by Items 5 and 2, whereas Item 10 has highest

Figure 5.5: MSWS-12 Total Information and Standard Error of Estimate

peak information followed by Items 7 and 11. Item 2 has lowest average information across the disability spectrum. Items 8 and 9 also have low average information because of low discrimination parameters (0.139 and 0.144) and poor separation between lowest and highest category thresholds (36.3 points and 31.8 points). Items 7, 10, 11, and 6 have highest average information, respectively, and Items 1, 3, and 12 have intermediate average information.

While all items fit the model, item information varies substantially. Information for Items 2 and 7 is shown in Figure 5.4. These results are critical to ASR, as item information is the criterion used to select items. Items with high information, such as Item 7, are more useful when assessing walking ability, as the information quantifies the degree to which responses reduce measurement uncertainty. Items with low information, such as Item 2, are less useful. In fact, Item 7 has higher information than Item 2 at almost all values of disability, the only exceptions being IRT-based scores less than 10.

Occurrences like this one were a concern when designing the ASR algorithm. Large

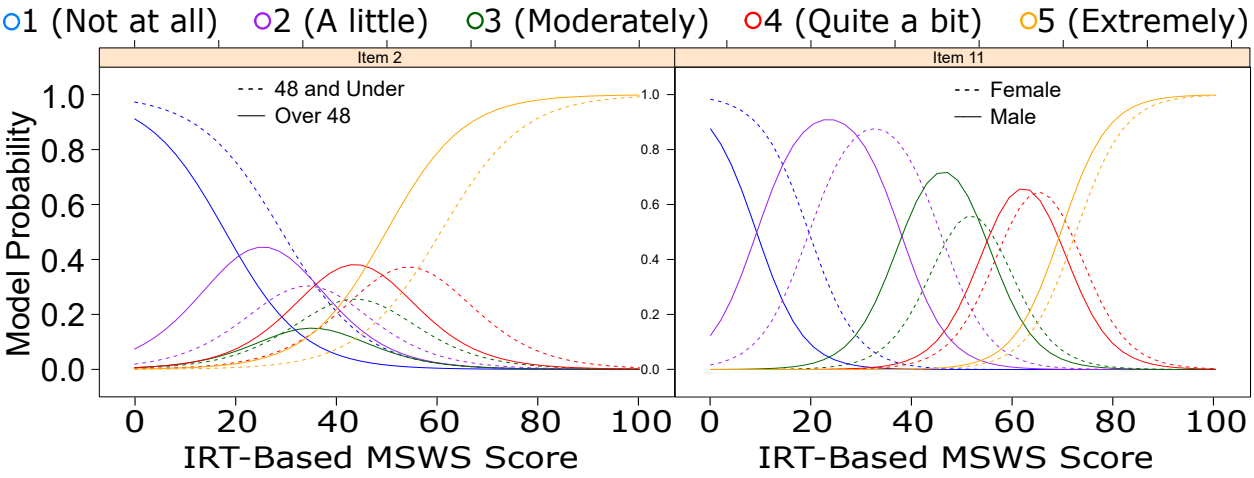
discrepancies in the area under the information curve made some items better than others (in terms of item information) on a consistent basis. Specifically, we noticed that a deterministic item selection procedure would repeatedly choose from a subset of MSWS-12 items and never choose others. Variety in item selection is desirable not only because it is more engaging for subjects, but more importantly because particular items can be biased in individuals or subgroups. Therefore we give all items a chance of being selected by introducing a degree of randomness in the selection procedure. Items are selected randomly with probability proportional to the square of their estimated item information, as previously described in Section 5.2.2.

Differential Item Functioning

DIF was measured by likelihood χ^2 test between subpopulations based on age (> 48 or ≤ 48) and sex (F or M). Results from the DIF analyses are presented in Table 5.4. Item 2 (How much has your MS limited your ability to run?) showed strong evidence of DIF in the age-based analysis ($\chi^2 = 19.02, df = 5, p < 0.01$). Item 11 (How much has your MS affected how smoothly you walk?) showed evidence of DIF in the sex-based analysis ($\chi^2 = 13.76, df = 5, p = 0.02$). In other words, when these items parameters were allowed to vary between subgroups, meaningful improvement was observed via likelihood ratio chi-square testing. Group-specific CRCs for Items 2 and 7 are shown in Figure 5.6. Older subjects were more likely to report running difficulty compared to younger subjects with the same IRT-based score. Men were more likely to report difficulty on Item 11 compared to women.

DIF was also detected for Item 11 (How much has your MS affected how smoothly you walk?), though its performance was otherwise good. DIF is undesirable because it makes scores inconsistent between groups, adding unwanted noise to PRO measurement. This effect can be mitigated by using the IRT-based score.

Figure 5.6: DIF Based on Age (left) and Sex (right) in MSWS-12 Item 2



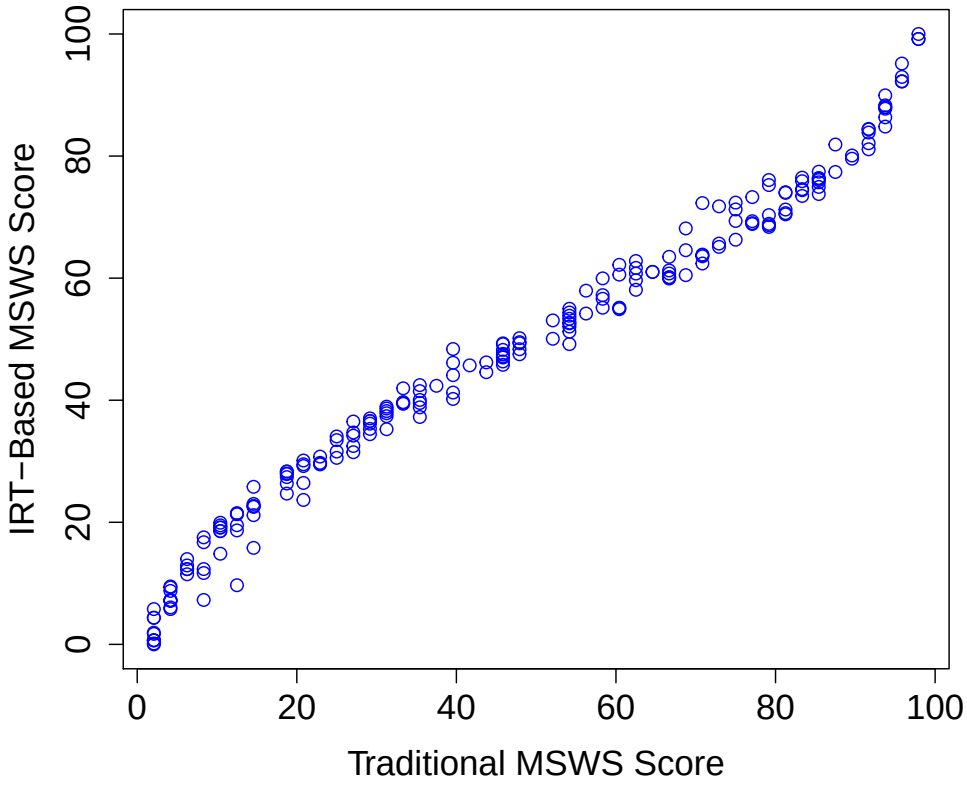
5.3.3 Promoting IRT Adoption

There are a number of barriers to the clinical adoption of IRT methods, including inertia and lack of familiarity with IRT methods. Most importantly, standard PROs can be scored with pen and paper, whereas IRT-based PROs must be scored with a lookup table or a computer. In our view, the widespread adoption of mobile health tools is needed to push IRT methods into mainstream care. Algorithms like ASR may help to encourage this transition. Nevertheless, we have taken two steps to support the adoption of our MSWS-12 model by care providers.

The first involves IRT trait estimates θ, which fall on a log-odds scale, where $\theta \in (-\infty, \infty)$. Unbounded estimates come up in practice when MLE is used. As an example, consider an aptitude test with dichotomous responses. If the subject gives all correct answers – or all incorrect answers – then their aptitude is unbounded: we have no way to distinguish between high aptitude and very high aptitude. This comes up in practice when using the MSWS-12. For example, a person with severe walking disability will have the same responses as another person who can't walk at all, namely all 5's. While this makes sense mathematically, unbounded estimates are not "clinician friendly". Neither are negative trait values or the fact that the scaling and offset of θ are arbitrary[140].

To facilitate clinical adoption, we have placed our IRT-based MSWS-12 on a 0-100 scale,

Figure 5.7: IRT-based MSWS-12 Scoring Compared to Original Scoring



the same as traditional MSWS-12 scoring. Log-odds scores were normalized to fall *between* 0 and 100, and unbounded estimates from MLE were assigned values of 0 or 100 as appropriate. All of the presented results use this scale, though unbounded estimates are not encountered by the ASR algorithm due to the incorporation of a prior distribution over θ. Figure 5.7 shows the correspondence between traditional MSWS-12 scores and the IRT-based scores from our population. The two are highly correlated ($r = 0.989$), and the Spearman correlation is even stronger ($r_s = 0.997$) due to the non-linear but monotonic relationship between them. Scores are not in one-to-one correspondence, because the IRT model takes response patterns into account.

While this scoring system is a bit more intuitive, it still requires complex computations, potentially limiting clinical adoption. In fact, a lookup table works for Rasch models, but it

would be prohibitively large for our model due to the presence of discrimination parameters. To make our scoring easily available to care providers and researchers, we have published a web application using the Shiny platform for R. When MSWS-12 responses are entered, it calculates the traditional score, IRT-based score, expected T25FW, and person fit (Z_h) value for error checking. The tool is intended to provide the advantages of IRT-based scoring to a broad audience, and to help set a precedent for subsequent PRO analyses.

The webapp is available at https://ms-irt.shinyapps.io/e-MSWS-12.

5.3.4 ASR Algorithm Evaluation

ASR was implemented in Matlab R2015b and tested using MSWS-12 responses from our 31 subject cohort. As discussed, walking ability estimates were linearly normalized to a 100-point scale to maintain consistency with traditional MSWS-12 scoring[90]. Values of θ were discretized ($N = 1000$), and integrals were estimated numerically using the trapezoidal rule. The distribution $P(\theta_t|\theta_{t-1})$ was calibrated based on expert opinion and a literature review of MSWS-12 variation during MS relapses[162][163] and over time[164][165][166][167][168].

Testing was designed to (1) ensure that ASR functioned as intended, (2) evaluate its accuracy, and (3) quantify the effect of the standard deviation threshold ϵ on algorithm accuracy and the number of selected questions. A total of 26 thresholds ranging from 1.6 to 9.6 were tested on all subjects. Thresholds were chosen to be less than 10.4, the lowest of several recently reported minimal clinical important difference (MCID) values for the MSWS-12[163], to ensure that clinically meaningful changes in MSWS-12 would not go undetected. Thresholds lower than 1.6 were unnecessary, as a threshold of 1.6 results in selection of all items at all time points.

Because ASR item selection is not deterministic – items are selected with probability proportional to the square of their item information – each subject-threshold combination was repeated five times for a total of $26 * 5 = 130$ rounds of data processing per subject.

Processing the complete dataset took approximately 4.2 hours running in Matlab 2015b on a single core of a typical desktop computer (Intel Core i3-2120 processor running at 3.30 GHz). The 246 data sessions available for analysis were each processed 130 times (26 SDTs $\times$ 5 repetitions), resulting in $130 \times 246 = 31980$ calls of the ASR algorithm. Therefore the average ASR run-time was approximately 0.47 seconds. Moreover, since a single run of ASR involves up to 12 iterations of question selection, the run-time between successive prompts is a fraction of this length. Detailed analysis of algorithm run-time and complexity is outside of the current scope, but unlikely to be a limiting factor in our target application.

Tracking Disability Progression

Results of ASR in three representative subjects are shown in Figure 5.8. An estimated MSWS-12 score has been calculated for each response session using three different versions of the ASR algorithm. Estimates based on full-information ASR are shown in blue: all twelve MSWS-12 responses from each session were used to calculate these estimates. Estimates obtained with a standard deviation threshold (SDT) of 4.8 are shown in red, and those obtained with a SDT of 6.4 are in yellow. The first session (Day 0) corresponds to $P(\theta_0|R_0)$ from the initial visit, which utilizes all 12 questions, so these estimates are the same for all three methods.

Both of the partial information estimates follow the full-information estimate, but average error is lower with the more conservative (lower) SDT. Confidence intervals are also lower for the SDT of 4.8 compared to the SDT of 6.4, as expected, and lowest with the full-information estimate. The green dotted line shows the minimal clinically important difference (MCID) of 10.4 in each plot. Plots have been scaled and shifted for clarity, so the height of the MCID is not the same between plots. An error greater than the MCID would by definition change the clinical interpretation of results, so it is important to keep error below this threshold.

The number of questions selected by ASR in each session are shown in Table 5.4. The list is longer for Subject B (middle), who completed the MSWS-12 more often than the other two.

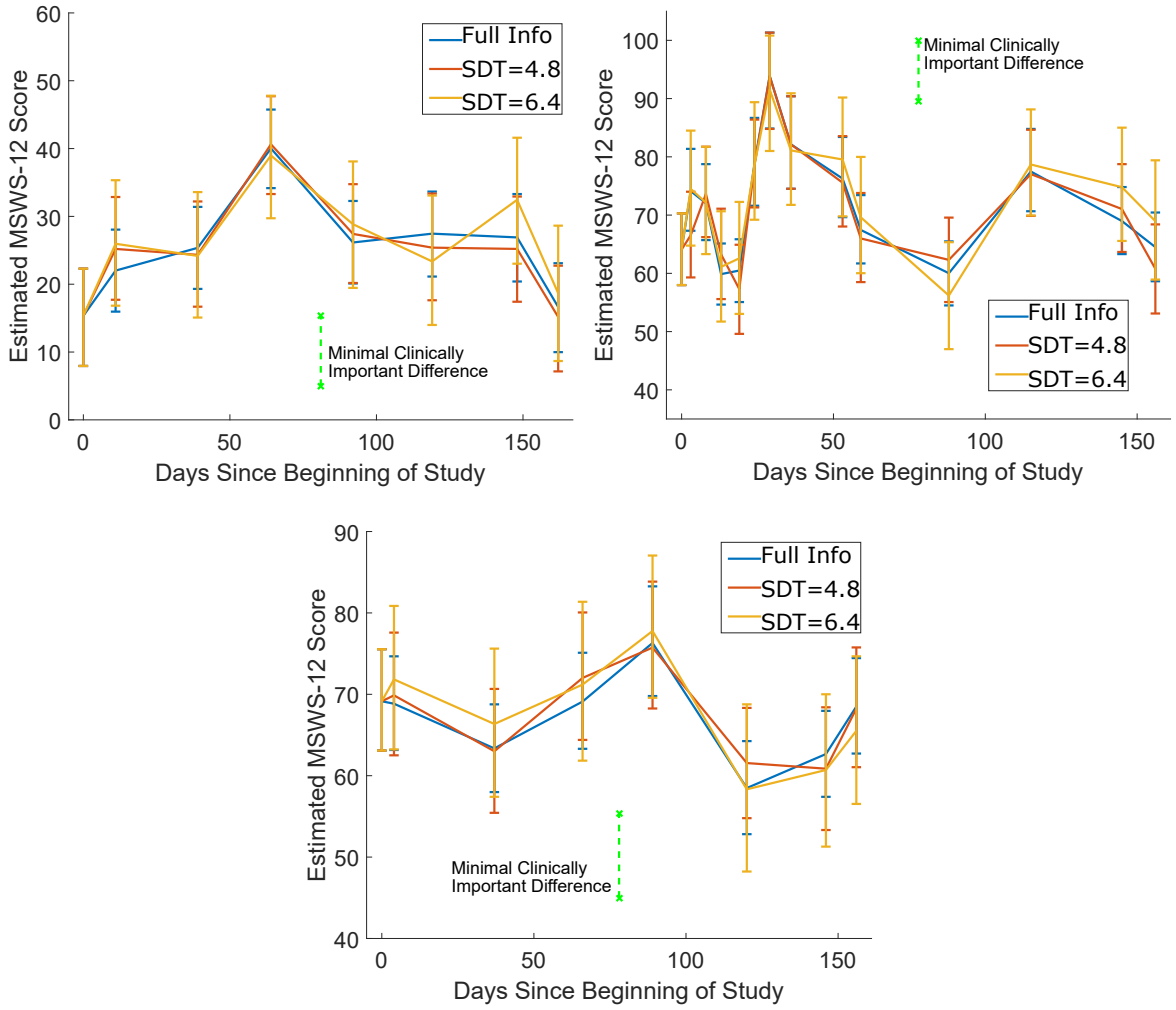


Figure 5.8: Results of ASR in Three Subjects with Three Different SDT Values

Since Day 0 is a special case it is not shown in the table, so the number of plotted points is one more than the number of sessions. Note that ASR tends to require more responses when disability changes. For example, ASR required all 12 responses to accurately estimate the sharp peak in Subject B's walking ability (session 6).

Subject A (top) has lowest disability as measured by the MSWS-12, with most estimates falling between 20 and 30, whereas Subjects B (middle) and C (bottom) have higher disability, with most estimates falling between 60 and 80.

Table 5.4: Number of Questions for Each of the Adaptive Symptom Reporting Examples from Figure 5.8

	Subject A (top)			Subject B (middle)			Subject C (bottom)		
Session	FI	T1	T2	FI	T1	T2	FI	T1	T2
1	12	6	3	12	9	3	12	5	4
2	12	5	3	12	10	5	12	5	3
3	12	6	4	12	4	3	12	7	5
4	12	7	4	12	5	3	12	8	3
5	12	6	3	12	10	8	12	7	3
6	12	8	4	12	12	5	12	4	2
7	12	7	5	12	12	6	12	7	5
8				12	7	6			
9				12	6	4			
10				12	5	2			
11				12	7	3			
12				12	6	4			
13				12	5	3			

FI = Full Information Adaptive Symptom Reporting; T1 = Standard Deviation Threshold 1 (= 4.8); T2 = Standard Deviation Threshold 2 (= 6.4)

Tradeoffs between Estimation Accuracy and Number of Questions

The accuracy of ASR and the number of questions it selects both of which depend on the SDT. As the SDT is decreased, the accuracy improves but typically more responses are required. Accuracy is quantified in two different ways: first, as the difference between partial-information ASR estimates and full-information ASR estimates; and second, as the difference between the partial-information estimates and traditional IRT estimates. It is important to note that there is no "true" value of MSWS-12-defined walking ability, as it is observed only indirectly via MSWS-12 responses.

Figure 5.9 shows how the SDT affected the number of questions (black) and estimation error (red). The plotted points are median values among all subjects, and the error bars show the inter-quartile range (IQR). The MCID of 10.4 is plotted as a green line; it is only relevant to the error (red) portion of the plot. The top plot shows error with respect to full-information ASR, while the bottom plot shows error with respect to traditional IRT estimates. The number of questions is the same in both plots.

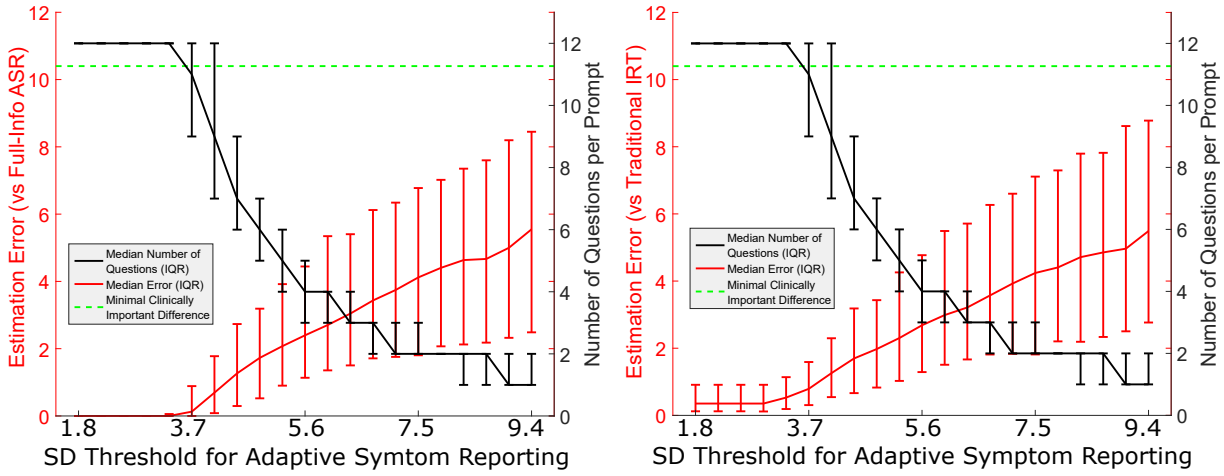


Figure 5.9: Error and Number of Questions vs SDT. Error is Based on Full-Information ASR (left) and Traditional IRT (right).

These plots illustrate the trade-off between the amount of information obtained by ASR and the resulting estimation error: as the number of questions decreased, the error increased. At the three lowest values of SDT (1.6, 2.0, and 2.4) ASR always required all 12 questions (full-information), so error is zero compared to full-information ASR. While the IQR falls below the MCID at all values of the SDT, there were outlying values exceeding this threshold at higher SDT values, as shown in Table 5.4.

Figure 5.10 shows algorithm performance among all subjects for the two SDTs shown in Figure 5.8 (4.8 and 6.4). Histograms show the number of questions (left) and the estimation error compared to full-information ASR (right). The number of questions is much lower at the higher SDT (median values of 3 and 6). However, error almost never exceeds the MCID at an SDT of 4.8 (0.16% of sessions) but occasionally does so at an SDT of 6.4 (4.45% of sessions).

Descriptive statistics for all SDTs may be found in Table 5.4. At an SDT of up to 5.2, less than 1% of sessions result in error exceeding the MCID. This decreases to zero at SDTs of 4.4 or less. Note that at an SDT of 4.4, the median number of questions is 7, a 42% reduction compared to traditional IRT.

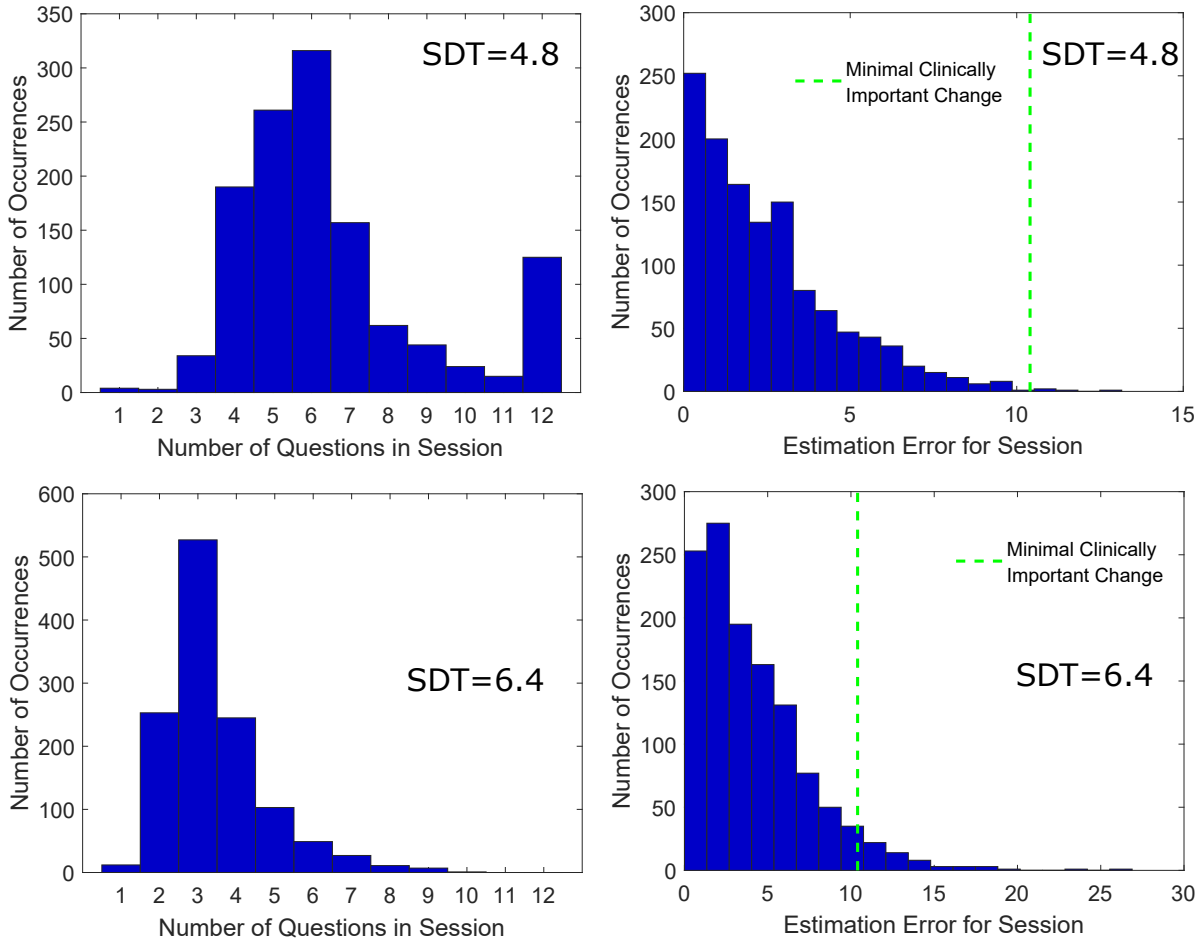


Figure 5.10: Histograms of Number of Questions and Estimation Error at Two Different SDT Values

5.4 Summary and Future Work

Adaptive symptom reporting (ASR) is a new paradigm for subjective disability assessment, combining the accuracy of patient-reported outcomes (PROs) with the convenience of simplified symptom reporting apps. The ASR algorithm was designed for mobile disability assessment: subjects respond to prompts as needed to keep uncertainty about disability within a specified threshold. Mobile implementation is essential to ASR because it requires regular opportunities for patient interaction.

PROs are the accepted clinical standard for subjective disability assessment, with superior measurement properties compared to the shorter question sets commonly found in symptom

Table 5.5: Questions and Estimation Error for Response Sessions Depend on the Standard Deviation Threshold (SDT).

	Number of Questions			Estimation Error (vs Traditional IRT)						Estimation Error (vs Full Info ASR)					
SDT	Mdn	Range	IQR	Mean±SD	Range	Mdn	IQR	99%	95%	Mean±SD	Range	Mdn	IQR	99%	95%
1.6	12	12-12	12-12	0.7±1.0	0-8.2	0.4	0.1-0.9	5.5	2.3	0.0±0.0	0-0.0	0.0	0.0-0.0	0.0	0.0
2.0	12	12-12	12-12	0.7±1.0	0-8.2	0.4	0.1-0.9	5.5	2.3	0.0±0.0	0-0.0	0.0	0.0-0.0	0.0	0.0
2.4	12	12-12	12-12	0.7±1.0	0-8.2	0.4	0.1-0.9	5.5	2.3	0.0±0.0	0-0.0	0.0	0.0-0.0	0.0	0.0
2.8	12	7-12	12-12	0.7±1.0	0-8.2	0.4	0.1-0.9	5.5	2.3	0.0±0.1	0-1.5	0.0	0.0-0.0	0.1	0.0
3.2	12	3-12	12-12	0.8±1.0	0-8.2	0.5	0.2-1.1	5.5	2.4	0.1±0.4	0-2.8	0.0	0.0-0.1	1.7	1.0
3.6	11	2-12	9-12	1.1±1.1	0-8.1	0.8	0.3-1.6	5.5	3.2	0.6±0.8	0-5.2	0.1	0.0-0.9	3.2	2.3
4.0	9	1-12	7-12	1.6±1.4	0-10.1	1.3	0.5-2.3	6.4	4.4	1.1±1.2	0-8.9	0.7	0.1-1.8	4.9	3.6
4.4	7	1-12	6-9	2.1±1.8	0-10.4	1.7	0.7-3.2	7.5	5.6	1.7±1.6	0-8.3	1.3	0.3-2.7	6.7	4.9
4.8	6	1-12	5-7	2.5±2.1	0-13.2	2.0	0.8-3.4	9.3	6.6	2.2±2.0	0-13.5	1.7	0.5-3.2	8.4	6.2
5.2	5	1-12	4-6	3.0±2.4	0-16.6	2.3	1.0-4.3	9.7	7.5	2.7±2.3	0-15.6	2.1	0.9-3.9	9.5	7.1
5.6	4	1-12	3-5	3.4±2.8	0-26.9	2.7	1.3-4.8	11.7	9.0	3.1±2.7	0-25.5	2.4	1.1-4.4	11.4	8.5
6.0	4	1-12	3-4	3.9±3.2	0-27.0	3.0	1.5-5.5	14.2	10.1	3.7±3.1	0-25.6	2.7	1.4-5.3	13.3	9.6
6.4	3	1-10	3-4	4.1±3.3	0-26.9	3.2	1.7-5.7	14.7	10.4	3.8±3.2	0-26.7	3.0	1.5-5.4	14.0	9.9
6.8	3	1-8	2-3	4.4±3.6	0-30.4	3.6	1.8-6.3	16.0	11.1	4.2±3.4	0-27.4	3.4	1.7-6.1	14.8	10.6
7.2	2	1-7	2-3	4.9±4.1	0-26.8	3.9	1.8-6.6	18.8	12.7	4.6±3.8	0-26.6	3.7	1.8-6.3	18.2	11.9
7.6	2	1-6	2-3	5.1±4.2	0-28.3	4.2	1.8-7.1	19.7	13.1	4.8±3.9	0-25.3	4.1	1.8-6.8	18.5	12.3
8.0	2	1-6	2-2	5.4±4.3	0-27.0	4.4	2.2-7.3	18.4	13.7	5.1±4.0	0-25.6	4.4	2.1-7.0	17.3	13.2
8.4	2	1-4	1-2	5.6±4.4	0-30.4	4.7	2.2-7.8	20.1	13.9	5.3±4.1	0-27.4	4.6	2.1-7.4	18.7	13.0
8.8	2	1-4	1-2	5.8±4.6	0-30.4	4.9	2.3-7.8	20.3	15.1	5.5±4.3	0-27.4	4.7	2.2-7.6	19.1	14.4
9.2	1	1-4	1-2	6.2±4.8	0-25.8	5.0	2.5-8.6	21.6	15.9	5.9±4.5	0-25.6	5.0	2.3-8.2	20.0	14.6
9.6	1	1-4	1-2	6.5±4.9	0-25.2	5.5	2.8-8.8	20.8	16.3	6.2±4.6	0-25.1	5.5	2.5-8.4	19.8	15.6

SDT = Standard Deviation Threshold; Mdn = Median; IQR = Inter-Quartile Range; 99% = 99th Percentile of Values; 95% = 95th Percentile of Values

reporting apps. However, mobile PRO collection is uncommon because it is burdensome for patients. The NIH has begun to improve PROs through its PROMIS, which incorporates IRT and computer adaptive testing[136]. ASR takes this a step further by drawing upon IRT methods to track disability – or any other latent trait – as a moving target. By taking full advantage of previous observations, ASR substantially reduces the number of questions required for accurate trait estimation.

Deployment of ASR must be preceded by the development of an IRT model. The model relates the trait of interest – in this case, disability – to each item in a bank of questions designed to assess it. The IRT model is the foundation of ASR, so informed model selection and accurate estimation are critical. In our validation study, ASR was validated with data from the MSWS-12, a PRO for walking ability in MS. The MSWS-12 fits the unidimensional GRM, a common IRT model for polytomous items, though DIF was observed in Items 2 and 11. To facilitate clinical adoption of our model, we have made it available through a web app: https://ms-irt.shinyapps.io/e-MSWS-12. The app allows clinicians and researchers to score the MSWS-12 using IRT and check responses for probable errors.

The ASR algorithm itself was validated in the MSWS-12 dataset presented in Chapter 3. Progression of walking disability was retrospectively estimated in 31 subjects with MS over a

six month period. The ASR algorithm incorporates a threshold parameter, the SDT, which governs the trade-off between accuracy and patient burden. Specifically, the system continues to prompt patients with new questions until uncertainty about disability falls below the threshold (or all questions are used). Here the appropriate choice depends on the application and possibly even the individual subject. When patient burden is a limiting factor, a less accurate estimate may be better than no estimate at all.

Descriptive statistics have been provided to characterize accuracy and question selection across a range of SDTs. The ASR algorithm achieved a 42% reduction in the median number of patient prompts (12 to 7) while *always* keeping error below the minimal clinically important difference (MCID). A 50% reduction (12 to 6) was achieved while allowing only 0.16% of errors to exceed the MCID, and at 75% reduction (12 to 3), 4.45% of errors exceeded the MCID.

Additional testing should cover different PROs in a variety of clinical populations. ASR should also be applied to questionnaires assessing psychological or educational constructs. Indeed, ASR can be applied to any questionnaire suitable for IRT analysis.

5.4.1 Limitations

The fitting of IRT models requires large samples, and our population was modest in size (N = 284). In support of this work, a sample size of 250 was sufficient to accurately recover model parameters for a 15 item, two-parameter model[169], and there is precedent for model fitting and DIF assessment with a similarly sized sample[170]. However, dimensionality assessment for polytomous items can be unreliable using samples smaller than 500[151], and in general, a sample size of 1000 or more is preferred for more complex IRT models[171, 172]. Thus the current item parameters are not definitive.

Further, the model comparison results suggest that MSWS-12 dimensionality may warrant further study despite previous support for a single scale dimension[160, 173]. Item 5 and other balance-related items have comparatively low discrimination parameters, implying

that balance may not be adequately captured by the unidimensional solution. A larger IRT analysis may be needed to conclusively estimate parameters and assess dimensionality.

In a prospective study, ASR question selection would look quite different. Due to the retrospective nature of our ASR validation, the algorithm could not prompt patients at arbitrary times. By the time new information became available, several questions were typically needed to bring uncertainty back within acceptable limits. In contrast, an online implementation of ASR could prompt subjects whenever its uncertainty threshold was exceeded. Rather than asking many questions at infrequent intervals, it could ask a single question every few days. More generally, the frequency of ASR initiation is a second parameter which should be adjusted to suit patient and provider preferences.

Any analysis of PROs is limited by the subjectivity inherent in subject-reported assessment. Subjectivity is a primary benefit of PROs, which are designed to assess disability from the patient perspective. However, subjectivity also complicates validation: there is no "true" value of MSWS-12 defined walking ability, because it is a latent trait not directly observed. Moreover, patients frequently make errors during PRO completion. In this study, we have evaluated the accuracy of partial-information ASR in terms of (1) full-information ASR, and (2) traditional IRT estimates. While this is an appropriate method of validation, it is important to note that both (1) and (2) are themselves imperfect estimates of a latent trait. Like PROs, comprehensive evaluation of ASR would require a number of validation studies incorporating multiple clinical outcomes.

5.4.2 Prospective Validation

As a retrospective analysis of ASR, this work was intended to establish the feasibility and accuracy of the algorithm, setting the stage for a prospective clinical trial. Such a trial is needed as further validation, and to compare ASR and other symptom reporting methods in terms of subject compliance, subject feedback, and the clinical utility of resulting disability

estimates. The current study shows that ASR *works*, but continued evaluation is needed to demonstrate its benefit in a clinical population.

Recently, we have established a link between bladder dysfunction, chronic hydration status, and MS-associated fatigue which warrants further study[174]. This is an excellent use case for ASR, because (1) fatigue can be measured by the unidimensional Fatigue Severity Scale[175], (2) water consumption can be reported by the subject, and (3) hydration status may be self-assessed via urine color. We hypothesize that promoting hydration will mitigate fatigue severity, and that a mobile tool will promote hydration more effectively than conventional behavior modification. A study to test these hypotheses is currently in development. This study will utilize ASR to assess fatigue, providing an opportunity for prospective algorithm validation.

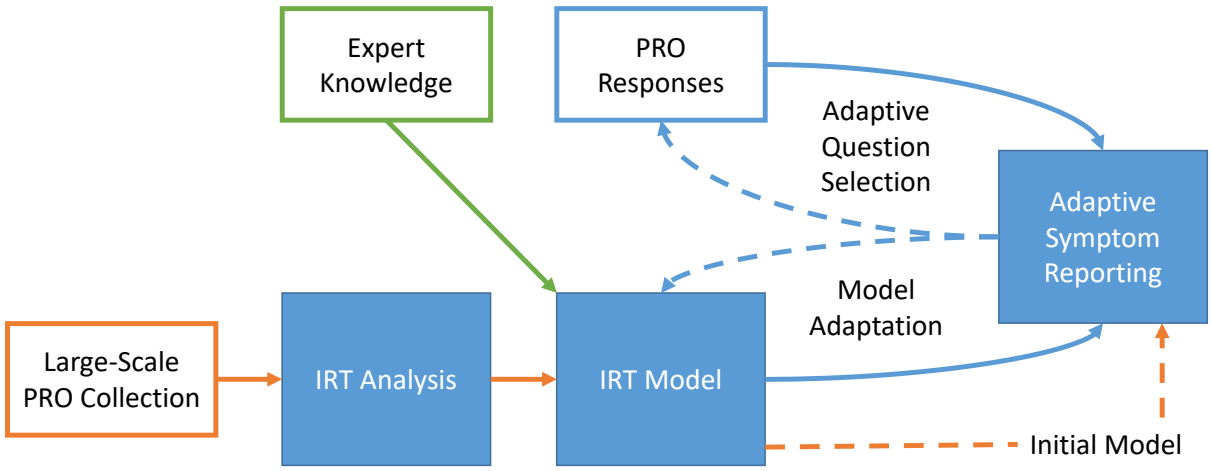
5.4.3 Active Model Learning

ASR adapts to the patient, but its internal model does not adapt. The algorithm tackles the problem of patient burden by reducing the number of responses, and it partly addresses the problem of subjectivity by defining disability on the patient's terms. However, the current implementation of ASR does not address variability between persons or potential model drift.

We envision a future version of ASR which is model-adaptive, meaning that its internal model will adjust over time. This model consists in two probability distributions: $P(R|\theta)$ and $P(\theta_t|\theta_{t-1}) = f(\delta_t)$. This system can be described as self-improving; a diagram of its deployment is depicted in Figure 5.11.

Unlike other model-adaptive systems, however, model-adaptive ASR should not adapt to an individual. Since trait estimation is relative by nature – the trait scale itself is arbitrary – allowing parameters to drift between persons would cause the trait scale itself to drift, making comparisons between persons meaningless. Instead, adaptation should occur on a population level. As observations are collected, the IRT and trait progression models can be refined, improving the system as a whole.

Figure 5.11: Possible Self-Improving ASR System



The modernization of health monitoring depends on self-improving systems like this one. The overhead required for iterative algorithm development is prohibitive; such systems might be deployed in research settings, but not for routine care. Widespread deployment of intelligent monitoring systems will be achieved only when these systems can improve themselves on the fly.

Chapter 6

Active Health Event Identification

Our last algorithm, active event identification (AEI), improves device-based event classification by asking patients to label events as they occur. The primary problem is one of *active* learning, which we later define in more detail. In simple terms, the active learning problem consists in choosing the most opportune, valuable times to request information, where value is defined as improvement to the system's internal model. Patient burden remains one of our foremost concerns, so AEI is designed to be judicious in its requests.

AEI is designed for monitoring systems which utilize sensor measurements on an ongoing basis to decide whether one or more medically-relevant events has occurred. Sensor measurements might include step counts, heart rate, galvanic skin response, and a wide range of common to obscure physiologic measurements and/or signal features. Importantly, our framework assumes the state is *known* to the patient; in other words, the he or she can tell whether an event is occurring and provide that information to the system. This assumption is fair in many domains, including meal detection, physical activity detection, and anxiety/panic disorders.

As a first step toward AEI, we present Hidden Markov Models (HMMs) as an appropriate probabilistic structure for the classification of health events. Like many other algorithms designed to classify or act upon sequential data, AEI supposes that an underlying Markov

chain governs transitions between events. Even without AEI, the use of HMMs addresses one of the major obstacles faced by monitoring systems: heterogeneity between persons and the need for personalization. In the first portion of this chapter, we show that applying HMMs to physical activity data from persons with MS clarifies its clinical interpretation compared to more traditional approaches. Step count data from an ActiGraph accelerometer is fitted to a HMM with negative binomial state to observation probabilities, leading to several novel clinical insights. Specifically, we show that the HMM can help to distinguish changes in walking capacity from changes in behaviors when interpreting physical activity data.

We then move to the technical contributions of the chapter, including our presentation of the AEI algorithm. AEI is a strongly online algorithm that incorporates recent advances in active learning and the incremental learning of HMMs. By quantifying query value in terms of the expected reduction in entropy over a population of models, AEI is able to select informative queries to quickly learn about health events. The AEI algorithm is first validated using synthetic data, then with a physical activity testbed developed based on the fitted models described in the first part of the chapter. In both cases, we show that active query selection improves classification accuracy much more quickly than random query selection. We close with a summary of results and some closing comments on the advantages of AEI, future directions, and related research.

6.1 Problem Features

This chapter concerns health events whose physiologic manifestations satisfy several statistical assumptions. While HMMs and AEI have broad applications, it's important to specify circumstances under which they do and don't apply. In this section, we state assumptions made throughout this chapter and contrast them with Chapters 4 and 5.

We begin by supposing that a set of discrete physiologic or behavioral states can explain our physiologic measurements. These are the states S of our model; we often refer to them

as *underlying* or *hidden*. In our target application, the states are daily physical activities, including walking, running, and inactivity. In diabetes care, a state might be a meal, run, or any other activity that affects blood glucose. In panic disorder, the states might relate to heightened stress or lack thereof. States are observed only indirectly via one or more signals of interest such as heart rate, respiratory rate, or galvanic skin response. Exercise increases step counts and heart rate, whereas meals increase blood glucose. The underlying state is never known with certainty, yet it is reflected in the data. Consequently, our *beliefs* about the state depend on the data in a Bayesian sense.

We also assume that observations are *memoryless*, meaning that they depend only on the state, not on past observations. Given the current physical activity, for instance, the subject's current step counts is independent of earlier counts. Similarly, carbohydrate intake depends on the current meal, not on its value a few moment ago. While these statements are approximations – history always matters to some degree – there is a sharp contrast with the physiologic signals discussed in Chapter 4. In physiologic signal monitoring, we were interested in periodic signals with high correlation between successive samples, including inertial data. Such data should not be fitted directly to a HMM.

The assumption of memorylessness is often more valid when observations are less frequent. Physiologic measurements do not change instantaneously, so the degree of correlation often depends on measurement frequency. Step counts can be viewed as independent (given the current activity) when binned every minute, for instance, but not when binned every five seconds.

A central assumption of AEI, though not HMMs in general, is that the subject has privileged information about their activities not directly available to the monitoring system. The system continually records physiologic measurements, but the user knows what's happening at the time. To justify this assumption, however, events must be defined in patient-centric terms. The subject doesn't know when they're "moderately active", as defined by energy expenditure, but they know when they're taking a walk. In most cases the distinction seems

unimportant: subjects know when they're walking, eating, or having a panic attack. On closer inspection, however, even "simple" events involve some degree of subjectivity: how many bites make a meal? In AEI, we must avoid such ambiguities and remember that activities are defined *from the patient's perspective.*

6.2 Hidden Markov Model (HMM) Basics

In this section, we briefly formulate the event classification model as a standard HMM, which will be augmented in later sections. Let $\mathcal{M} = \{S, O, \pi, \phi, \lambda\}$ be a HMM with states in S, observations in O, initial state distribution π, state transition probabilities ϕ, and state to observation probabilities λ. Taken together, the HMM parameters $\{\pi, \phi, \lambda\}$ are denoted as θ. Note that this use of θ is not related to its use in Chapter 5. We now discuss each element of $\mathcal{M}$ in turn.

Our observations O are the physiologic measurements obtained by the monitoring system, which satisfy the properties described in the previous section. In our target application, observations are minute-wise step counts derived from an individual patient's accelerometer data.

The elements of S are distinct physiologic or behavioral states which help to explain the observations. In our target application, walking ability in MS, we identify them with walking or running, though in general the relationship to patient-defined activities may not be one-to-one. The number of states $|S|$, denoted N, is a fixed feature of the model, although models with different numbers of states can certainly be compared (as in Section 6.3.2). Labels for the states are arbitrary, but each state has a characteristic state to observation probability distribution learned in the model fitting process.

The state transition probabilities define the probability of moving from one state to another between successive observations. These probabilities form an $(N \times N)$ matrix ϕ. The $(i, j)^{th}$ entry of ϕ is ϕ_{ij}, the probability of transitioning from state i to state j, and in

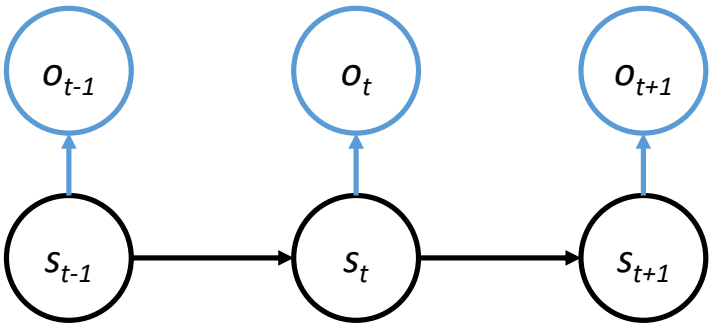


Figure 6.1: Probabilistic Graphical Model of HMM

particular, ϕ_{ii} describes the probability of remaining in state i (i.e. *not* transitioning). By giving the probability of transitioning to state j only in terms of state i, we've implied the Markov property:

$$P(S_k = s_k | S_{(k-1)} = s_{(k-1)}, \ldots, S_1 = s_1) = P(S_k = s_k | S_{(k-1)} = s_{(k-1)}) \tag{6.1}$$

Additionally, the transition probabilities are time-invariant, meaning that ϕ does not depend on the stage (or time). In fact *all* of the HMM parameters are time-invariant throughout this work.

Figure 6.1 depicts a graphical model of a HMM, where the arrows indicate conditional dependence. Note that the state depends only on the previous state, and observations depend only on the current state. This model formalizes the assumptions discussed in the previous section.

6.3 Clinical Insight: Physical Activities in MS

The first contribution of this chapter is to apply a standard statistical technique – the HMM – to the quantification of habitual physical activity (HPA) in MS, leading to several new clinical insights. Based on this method and subsequent analysis, we have identified precise, independent measures of walking capacity and walking behavior, which can be

used as outcomes in MS care. To our knowledge, this is the first application of HMMs to HPA in a clinical patient population. From a methodological perspective, these results illustrate the HMM as a versatile method of unsupervised event classification, supporting is use in subsequent sections as the foundation of AEI. The strongest evidence for the general applicability of HMMs to real-world problems, however, is their widespread use in decision theory, dynamic programming, and reinforcement learning. In this section, we use the HMM to classify distinct physical activities such as walking and running, but this approach is capable of classifying any health-related events that satisfy the problem features described earlier in the chapter.

Measurement of HPA has emerged as a walking outcome in multiple sclerosis (MS)[176] and Parkinsons disease[177], with broad applications in health care[178, 179]. Unlike other objective walking outcomes, HPA addresses the impact of disease on real-world functioning. HPA is particularly important in MS care, because physical activity appears to mitigate disability progression. Specifically, exercise training can lead to modest improvements in walking ability[180, 181, 182, 183], and reduced physical activity is associated with worsening symptoms independent of disease course[183]. Consequently, several studies have evaluated HPA in MS. Daily step counts are significantly correlated with walking outcomes[106, 83], and MS subjects are active less often then controls[184, 185].

Currently, HPA is commonly reported as (1) daily step counts or movement counts; or (2) time spent in moderate physical activity (MPA) and vigorous physical activity (VPA). These statistics are easy to calculate and interpret, but they perform poorly as outcome measures. Daily counts do not reliably change when patient-reported walking ability changes[84] and explain less than half of the variance in objective walking outcomes[83, 106], making them a poor measure of walking capacity. Nor are they a reliable measure of behavior: daily counts are not significantly correlated with self-reported physical activity in MS[110]. Indeed, total daily activity is affected not only by walking capacity, but also by the multitude of personal, environmental, and social factors that influence physical activity behaviors in

healthy adults[108, 109]. Daily counts are a coarse measure of both walking capacity and walking behavior, but a precise measure of neither.

The second approach to HPA interpretation involves classifying activities as MPA, VPA, or neither by placing pre-specified cut-points on the step rate[184, 185]. Each step count is labeled independently: if it falls between the i^{th} cutpoint and the $(i+1)^{th}$ cutpoint, it is classified as event i. The U.S. Department of Health and Human Services recommends 150 minutes of MPA or 75 minutes of VPA per week in adults with disabilities[186]. However, these guidelines define MPA and VPA in terms of energy expenditure, not step rates, and the relationship between the two depends on age[86], body weight[187], the length and weight distribution of the legs[86], and disease-related movement limitations. Cut-points have been adjusted by disability status[188, 189, 87], partly addressing this concern, but truly accurate MPA and VPA classification requires person-specific calibration[190]. Consequently, the difference in MPA and VPA between MS subjects and controls[184] is difficult to interpret. A person with limited walking capacity might not reach MPA step rates even when their energy expenditure is high, causing a difference in capacity to be misinterpreted as a difference in behavior.

Further, the cut-point approach has several additional limitations easily overcome by the use of HMMs. First, classification is overly sensitive to brief fluctuations in the data. Whenever a threshold is crossed, the classification changes, resulting in many brief activity periods. As a result, the cut-point approach tends to overestimate the number of activity periods and underestimate their length. More importantly, cut-points are affected by the typical drawbacks of static monitoring discussed in Chapter 2. Since they are set ahead of time based on population-level data, cut-point based activity classification cannot adapt to particular patients or circumstances. Returning to our previous terminology, the cut-point approach is not equipped to deal with the problems of heterogeneity or drift. We'd like to know when the patient walked, ran, or biked; how intense these activities tended to be; and whether their frequency and intensity has changed over time. Instead of reporting time

spent in static categories – "mild activity", "moderate activity", etc – activity should be summarized in terms the patient and physician can understand and discuss. Use of an HMM partly achieves this goal through personalized, context-aware classification. AEI fully realizes it by learning to identify activities in patient-defined terms, providing a patient-centered narrative that is annotated with physiologic measurements.

In this section, we describe the application of HMMs to physical activity data from 88 persons with MS and 38 healthy controls. Our primary clinical hypothesis was that this approach would be able to identify subject-specific activity characteristics and distinguish walking capacity from behavior, thereby giving clinicians precise behavior and capacity outcomes in MS. A secondary hypothesis was that walking and running often fall outside of literature-based MPA and VPA cut-points. These hypotheses were validated in a medium sized MS cohort through correlational analysis to other walking outcomes. We believe this work is the most detailed statistical analysis of HPA to date in MS.

6.3.1 Study Design

Recruitment and Study Procedures

All study procedures were approved by the University of Virginia (UVa) Institutional Review Board for Health Sciences Research, and written consent was obtained from all participants. Subjects with clinically definite MS[102] were recruited from the UVa Neurology Department outpatient clinic population. Successfully recruited subjects were asked to bring someone with similar demographic characteristics but without MS to their first study visit to serve as a control. All recruited subject were age 18-64 years and able to ambulate for six minutes, possibly with an assistive device. Subjects with neurological impairment from other diagnoses (e.g. stroke), orthopedic limitations of the lower extremity, morbid obesity, or known cardiac or respiratory disease were excluded. Subjects were instructed to discontinue fatigue-related medications 48 hours prior to their appointment.

Baseline demographics for all subjects included age, sex, race, employment status, and level of education. Current MS therapies and MS-related history were reviewed and documented. A neurologic examination was conducted by Neurostatus-certified staff for Expanded Disability Status Scale (EDSS) assessment. MS subjects were classified as mild (EDSS 0-2.5), moderate (EDSS 3.0-4.0), or severe (EDSS$\geq$ 4.5) walking impairment. The timed 25-foot walk (T25FW) was assessed prior to 6MW testing, and additional patient-reported outcomes (PROs) were collected: the MS Walking Scale (MSWS-12) and Modified Fatigue Impact Scale (MFIS). MSWS-12 was not collected in control subjects, as it assesses mobility limitations attributed to MS, however MFIS was collected.

All subjects completed a six-minute walk (6MW) while wearing an ActiGraph GT3X accelerometer on their non-dominant hip secured with an elastic belt. Subjects also wore a Polar S610i heart rate monitor. The 6MW took place in a 75-foot hospital corridor using the instructions and script developed by Goldman et al[77]. Distance was manually recorded in 1-minute epochs for all 6 minutes. Heart rate was measured at baseline and at the end of each minute. 6MW step rates were recorded by ActiGraph (total steps $\div$ six minutes).

Subject Demographics

A total of 88 subjects with MS[102] and 38 healthy controls were recruited. Demographics and clinical outcomes for this population are presented in Table 6.1. Among MS subjects, 52.3% had mild disability, 35.2% moderate, and 12.5% severe. A higher proportion of subjects were female among MS subjects (83.0%) compared to controls (71.1%), though this difference was not significant. Education was similar between groups, but age, employment status, years since symptom onset and diagnosis, and clinical outcomes showed statistically significant differences. Disease subtype was also significantly different between groups, with 95.7% of mild MS subjects having relapsing-remitting disease compared to 77.4% of moderate MS and 45.5% of severe MS. MSWS-12 responses spanned the full range of possible scores (0-100).

Table 6.1: Demographics and Clinical Outcomes

Variable	Control	Mild	Moderate	Severe	p-value
Number of Subjects	38 (30.2%)	46 (36.5%)	31 (24.6%)	11 (8.7%)	
Sex, F:M (% Female)	27:11 (71.1%)	38:8 (82.6%)	25:6 (80.7%)	10:1 (90.9%)	0.4194
Age, mean ± SD	35.05 ± 12.38	41.43 ± 9.94	47.19 ± 7.85	46.00 ± 8.66	< 0.0001*
6MW Distance, mean ± SD	2079.9 ± 258.9	1826.5 ± 279.2	1619.7 ± 306.6	537.2 ± 327.3	< 0.0001*
6MW Step Rate, mean ± SD	123.40 ± 10.11	117.83 ± 15.54	109.99 ± 18.79	35.86 ± 27.76	< 0.0001*
T25FW, mean ± SD	3.59 ± 0.50	4.17 ± 1.02	4.49 ± 0.80	22.36 ± 17.96	< 0.0001*
MSWS Score, mean ± SD	NA	14.35 ± 19.73	30.10 ± 24.81	84.85 ± 17.46	< 0.0001*
MFIS Total, mean ± SD	17.22 ± 9.96	24.28 ± 17.46	39.03 ± 16.35	48.73 ± 15.23	< 0.0001*

*$p < 0.01$; 6MW: 6-Minute Walk; T25FW: Timed 25-Foot Walk; MSWS: MS Walking Scale; MFIS: Modified Fatigue Impact Scale

Our population spanned the full range of ambulatory MS-related disability by EDSS (range 1-6.5) and MSWS-12 (range 0-100), with approximately equal numbers of control subjects, mild MS subjects, and moderate to severe MS subjects. Beyond this range (EDSS> 6.5), HPA measurement is not appropriate.

Habitual Physical Activity Collection

All subjects wore an ActiGraph device on their non-dominant hip daily over a 7-day period beginning the day after their appointment. Each subject was given the same device used for their 6MW and instructed to wear it during waking hours except while bathing or swimming. Subjects completed an activity log to verify the days and times the device was worn, and both the device and log were returned by mail at the end of the 7-day period.

Steps were detected and aggregated into minute-wise step rates (steps/minute) using the ActiLife 6 software program, then exported to Matlab R2015b for subsequent processing. Wear days were declared valid only if steps were detected for at least 10 hours[83], and subjects with fewer than 6 valid days were excluded from the analysis.

For the conventional statistical analysis, step rates were classified as moderate physical activity (MPA), vigorous physical activity (VPA), or neither using disability-specific cut-points, which vary by MSWS-12 Score[87]. Moderate-to-vigorous physical activity (MVPA) was defined as either MPA or VPA.

Statistical Analysis

Demographics, outcomes, and activity statistics were compared between groups (control, mild MS, moderate MS, and severe MS) by ANOVA and chi-square test as appropriate for numeric and categorical variables, respectively. Linear (Pearson) correlations have been used to quantify relationships between activity statistics and clinical outcomes, all of which are continuous or approximately continuous (MFIS and MSWS-12). T25FW measurements have been converted to speed (feet/sec) prior to correlational analysis to avoid non-linear relationships to other walking outcomes, which are measured in distance or step rate.

Correlation coefficients were interpreted as strong (> 0.6), moderate (0.3-0.6), or weak (< 0.3). Significance levels have been adjusted based on the number of comparisons per analysis. A significance level of $\alpha = 0.004$ was used for demographics and outcomes ($m = 12$). Significance levels of $\alpha = 0.001$ and $\alpha = 0.0001$ were used when comparing activity statistics and identifying significant correlations, respectively, due to the larger number of values tested ($m = 28$ and $m = 128$, respectively). Correlations with $p < 10^{-8}$ are also reported.

Stepwise regression was used to quantify the contribution of clinical outcomes to each activity statistic. The possible predictors were 6MW distance, 6MW step rate, T25FW speed, MSWS-12, MFIS, MFIS physical subscale, MFIS cognitive subscale, and MFIS psychosocial subscale, along with all interaction terms between them. Predictors were added or removed based on the p-value of an F-test of the change in sum of squared errors. These results were replicated with a different selection criterion (BIC). A p-value < 0.05 cut-point was used when selecting predictors, but only p-values less than 0.0001 were interpreted as significant.

Fitting Subject-Specific HMMs

Each subjects minute-wise step rates were fitted to a HMM with a "not-worn" state (with step rate = 0) and 2-6 additional states with negative binomial state to observation probabilities using the Baum-Welch algorithm[191]. States were initialized with means evenly spaced between 0 and the subjects maximum step rate (MSR), with variance set to MSR *div* N, the

number of states. The model with lowest BIC was selected as the final model. Step rates were then fitted to this model using the Viterbi algorithm[192].

The negative binomial (NB) distribution, which was used to model all states other than the "not worn" state, is a discrete distribution with parameters ρ and p. The probability $\lambda_s(o)$ of observing o from state s is given by $\lambda_s(o) = P(O = o \mid S = s) \sim \text{NB}(\rho; \upsilon)$. The NB is described by the following probability mass function (PMF):

$$P(O = o \mid S = s) = \binom{o + \rho - 1}{o} \upsilon^o (1 - \upsilon)^\rho \tag{6.2}$$

This PMF is defined only for non-negative values of o, as expected. The NB distribution may be viewed as a mixture of Poisson distributions where the rate parameter is gamma-distributed. In other words, if the step rate follows a gamma distribution, then step counts are NB-distributed. The NB distribution has proven superior to the Poisson and others when modeling step counts [193], which are over-dispersed relative to the Poisson.

Figure 6.3 illustrates the fitting process in a single subject. There are five activity states, one for each curve in the left panels. These curves are the NB PMFs for the various states, which give the probability of particular step counts (on the x axis) while in the given state.

Active states (e.g. walking, running) were identified by their expected value (>MSR/2) and coefficient of variance ($\leq$ 8). If only one active state was present, it was identified as walking unless the expected step rate exceeded 130. If two active states were present, the state with higher expected step rate was identified as running, and the other state was identified as walking. No subjects had more than two active states, thus further identification procedures were not needed. Since activities were classified by step rates alone, additional activities (e.g. stair climbing) were not specifically identified. The habitual walking step rate (HWSR) and habitual running step rate (HRSR) were defined as the expected step rates during walking and running, respectively.

6.3.2 Clinical Significance of HPA Summary Statistics

In this section, we present standard statistical results not based on the HMMs. They are presented to illustrate the current approach to HPA in clinical research. This serves as a constrast to our HMM-based results and rounds out the clinical picture for this clinical insight section.

We begin with conventional statistics, which include average daily steps and moderate-to-vigorous physical activity. These are the statistics reported by most studies of HPA, whose limitations we have highlighted, and our findings are consistent with other results from the clinical literature. More interestingly, we present compelling results for the MSR, which we position as the best available HPA-based measure of walking capacity, and show that relationships between conventional statistics and clinical outcomes are dramatically affected by disability level.

Conventional Statistics

Average daily steps, MPA, and MVPA were significantly different between controls, mild MS, moderate MS, and severe MS (p$<$ 0.0001). On average, controls had approximately twice as much MVPA as mild MS subjects (196.8%) and over four times that of moderate MS subjects (464.5%). MPA, VPA, and MVPA were zero for all severe MS subjects (Table 6.2).

In MS subjects, average daily steps were strongly and significantly correlated with the 6MW step rate ($r = 0.676$,$p < 10^{-11}$) and 6MW distance ($r = 0.676$, $p < 10^{-12}$) (Fig 6.2, bottom panels). Average daily steps were also significantly correlated with T25FW speed ($r = 0.675$, $p < 10^{-12}$), MSWS-12 ($r = -0.627$, $p < 10^{-10}$), and MFIS ($r = -0.490$, $p < 10^{-5}$). Total MVPA was moderately and significantly correlated with 6MW distance, T25FW speed, MFIS, MFIS Phy., and MFIS Psych. ($p < 10^{-4}$). Among the demographic characteristics, total MVPA was significantly associated only with age ($r = -0.437$, $p < 10^{-6}$). Additional correlations may be found in Table 6.3. None of these correlations were statistically significant in control subjects.

Table 6.2: HPA Statistics by Group

Variable	Controls	Mild MS	Moderate MS	Severe MS	p-value
Conventional Summary Statistics					
Average Daily Steps	7952.03 ± 3466.61	6347.18 ± 2961.43	5270.82 ± 2217.68	1703.39 ± 1009.68	< 0.0001*
Total MPA Time	113.55 ± 125.35	58.57 ± 77.01	27.90 ± 37.87	0.00 ± 0.00	0.0001*
Total VPA Time	22.97 ± 51.51	10.83 ± 28.92	1.48 ± 7.36	0.00 ± 0.00	0.0410
Total MVPA Time	136.53 ± 147.36	69.39 ± 98.89	29.39 ± 40.92	0.00 ± 0.00	< 0.0001*
HMM-Based Statistics					
Max Step Rate	141.92 ± 24.66	123.43 ± 25.90	113.06 ± 23.51	52.09 ± 30.65	< 0.0001*
Walking Step Rate	104.00 ± 8.70	102.23 ± 11.82	97.52 ± 11.85	45.39 ± 40.69	< 0.0001*
Running Step Rate	151.05 ± 17.67	146.80 ± 20.32	NA	NA	0.6151
Total Walking Time	133.03 ± 143.22	70.11 ± 85.50	30.58 ± 52.38	13.45 ± 38.46	0.0001*
Total Running Time	30.66 ± 57.74	16.09 ± 38.17	0.00 ± 0.00	0.00 ± 0.00	0.0088
Total Active Time	163.68 ± 164.56	86.20 ± 110.43	30.58 ± 52.38	13.45 ± 38.46	< 0.0001*
Number of Walks	28.47 ± 33.14	13.61 ± 19.42	5.06 ± 8.23	6.09 ± 17.71	0.0002*
Number of Runs	1.95 ± 4.23	1.48 ± 4.59	0.00 ± 0.00	0.00 ± 0.00	0.0990
Num of Active Periods	29.45 ± 33.50	12.89 ± 18.61	5.06 ± 8.23	6.09 ± 17.71	0.0001*
Longest Walk	16.58 ± 14.99	13.39 ± 15.63	6.52 ± 9.80	1.45 ± 3.24	0.0014
Longest Run	10.26 ± 16.39	5.46 ± 12.68	0.00 ± 0.00	0.00 ± 0.00	0.0026
Longest Active Period	20.53 ± 16.62	17.85 ± 20.08	6.52 ± 9.80	1.45 ± 3.24	0.0001*

*$p < 10^{-3}$; PA: Physical Activity; HPA: Habitual PA; MPA: Moderate PA; VPA: Vigorous PA; MVPA: Moderate-to-Vigorous PA

Table 6.3: Correlations to Clinical Outcomes

	6MW Distance	6MW Step Rate	T25FW Speed	MSWS Score	MFIS Total	MFIS Phy.	MFIS Cog.	MFIS Psych.
Conventional Summary Statistics								
Maximum Step Rate	0.801**	0.863**	0.755**	-0.756**	-0.504*	-0.569 **	-0.383	-0.564*
Avg Daily Steps	0.676**	0.676**	0.675**	-0.627**	-0.490*	-0.537*	-0.391	-0.493*
MPA Time	0.429*	0.407	0.442*	-0.411*	-0.430*	-0.464*	-0.338	-0.427*
VPA Time	0.264	0.261	0.304	-0.251	-0.254	-0.303	-0.165	-0.239
MVPA Time	0.415*	0.396	0.436*	-0.397	-0.413*	-0.453*	-0.315	-0.406*
HMM-Based Statistics								
HWSR	0.701*	0.815**	0.670*	-0.717*	-0.237	-0.279	-0.189	-0.121
HRSR	0.553	0.676	0.721	-0.703	-0.331	-0.460	-0.187	-0.378
Walking Time	0.355	0.302	0.377	-0.377	-0.440*	-0.490*	-0.324	-0.454*
Running Time	0.247	0.244	0.232	-0.211	-0.225	-0.278	-0.143	-0.189
Active Time	0.363	0.319	0.376	-0.370	-0.424*	-0.481*	-0.305	-0.425*
Longest Walk	0.338	0.322	0.322	-0.409*	-0.412*	-0.436*	-0.331	-0.429*
Longest Run	0.217	0.237	0.219	-0.215	-0.193	-0.257	-0.105	-0.153
Longest Active Period	0.369	0.375	0.391	-0.415*	-0.375*	-0.451*	-0.240	-0.384

*$p < 10^{-4}$; **$p < 10^{-8}$; PA: Physical Activity; HPA: Habitual PA; MPA: Moderate PA; VPA: Vigorous PA; MVPA: Moderate-to-Vigorous PA; HWSR: Habitual Walking Step Rate; HRSR: Habitual Running Step Rate

In the stepwise regression models, both the 6MW step rate and MFIS Phy. were significant predictors of average daily steps, explaining 51.6% of the total variance. T25FW speed and the MFIS Phy. were significant predictors of MVPA, explaining 32.8% of the total variance. T25FW was the most significant predictor of MPA, VPA, and MVPA (Table 6.4).

There were no significant relationships between HPA and clinical walking outcomes in control subjects, suggesting that in the absence of disability, HPA is purely behavioral. From a statistical standpoint, the low variability in capacity among controls makes it difficult to identify a correlation.

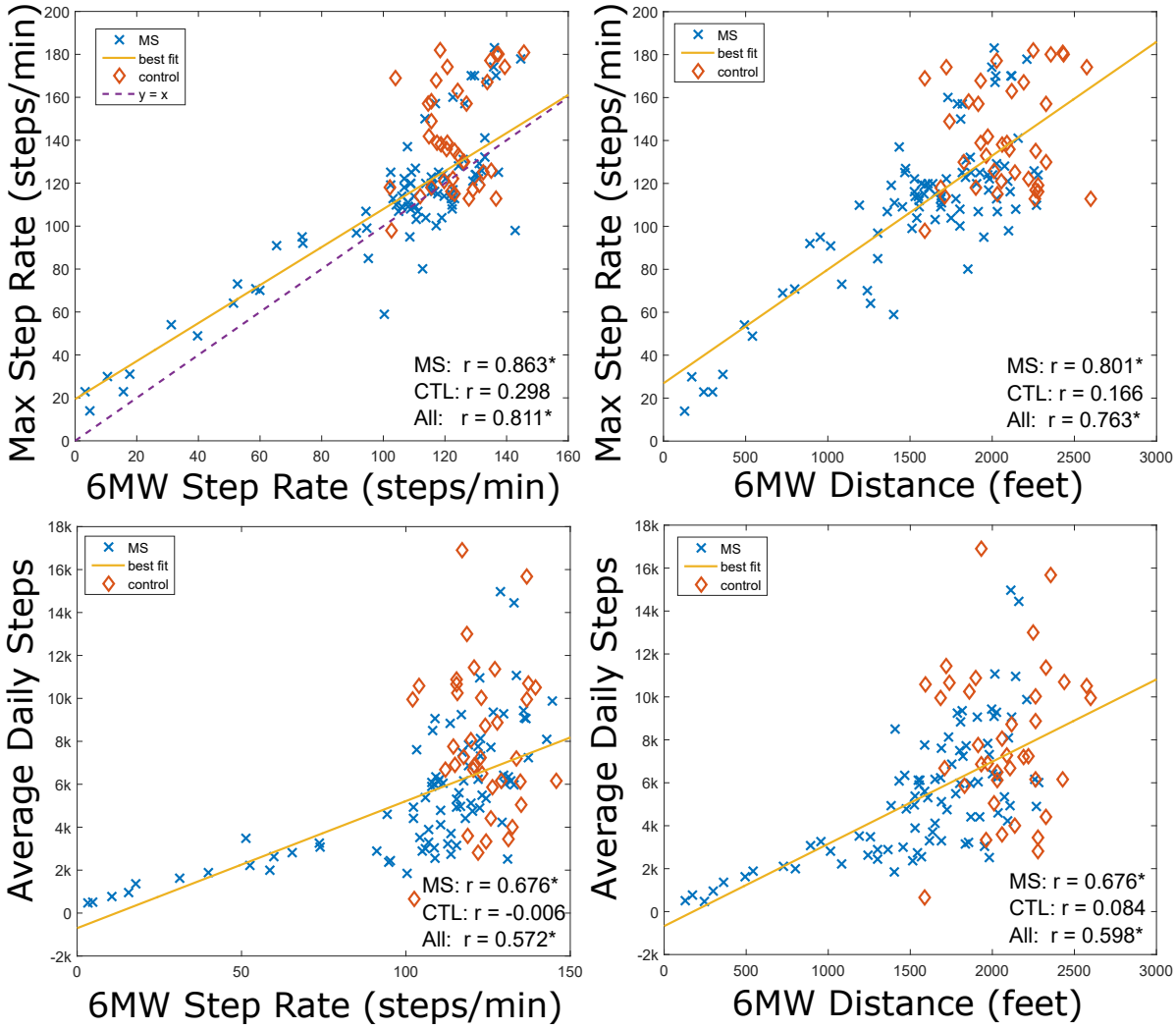


Figure 6.2: Correlations between the 6MW and Two Competing Measures of Capacity

Maximum Step Rate

Among the conventional statistics, the most significant clinical finding involved the maximum step rate (MSR). One might argue from first principles that the MSR is just as valid as the 6MW or T25FW as a direct measure of capacity. The MSR is the highest step rate achieved over a week-long period, whereas the 6MW and T25FW measure speed achieved during in-clinic testing. Each has its limitations as a measure of capacity, which might be defined as the highest speed or step rate the subject is *able* to achieve. All three depend on subject effort in some way, but the 6MW and T25FW are also affected by circumstances and

symptoms at the time of testing. More practically, these measures differ in the duration of effort: the MSR approximates the maximum rate subjects can sustain for a minute, compared to the longer 6MW or shorter T25FW. The results presented in this section show that the MSR is an excellent measure of walking capacity. It is certainly the best HPA-based measure of capacity currently known. We believe the 6MW is the optimal measure of capacity due to its longer duration and the wealth of supporting results, while the MSR is its real-world compliment.

To begin, the MSR was most strongly correlated with clinical walking outcomes among all of the HPA statistics. Specifically, the MSR was strongly correlated with 6MW step rate ($r = 0.863$, $p < 10^{-25}$), 6MW distance ($r = 0.801$, $p < 10^{-20}$), T25FW speed ($r = 0.755$, $p < 10^{-16}$), MSWS-12 ($r = -0.756$, $p < 10^{-16}$), and MFIS ($r = 0.504$, $p < 10^{-25}$) in MS subjects (Table 6.3). Figure 6.2 shows its superior performance over average daily steps when estimating 6MW step rate and distance. It is not surprising that the MSR correlates more strongly with 6MW step rate than 6MW distance; this difference can be explained by variability in stride length, which determines the ratio of step rate to distance traveled.

These correlations support the MSR as a valid measure of walking capacity. Indeed, they are stronger than corresponding correlations for any of the other HPA-based statistics, showing that the MSR is the best HPA-based measure of capacity among those we have tested. These results also verify a central assumption of this work, namely that average daily steps – the most commonly used HPA statistic – are influenced by behavioral factors in addition to walking capacity. The correlation observed between 6MW distance and average daily steps ($r = 0.676$) is consistent with correlations reported by other studies ($r = 0.519$ and $r = 0.630$)[106, 83].

The MSR was faster than the 6MW step rate in 68.3% of subjects, resulting in a statistically significant difference between the two measures ($\mu = 10.4$, $p < 10^{-7}$) (Fig 1, top left panel). Stepwise regression retained the 6MW step rate and MFIS as significant predictors of the MSR, with 75.0% of total variance explained (Table 6.4).

Table 6.4: Stepwise Regression Results

Activity Statistic	Significant Predictors	β	p-value	Var. Explained
Conventional Summary Statistics				
Average Daily Steps	6MW Step Rate	93.627*	3.46E-05	51.6%
	MFIS Phy.	132.280	1.63E-01	
	6MW Step Rate:MFIS Phy.	-1.792	2.76E-02	
Total MPA Time	T25FW Speed	27.300	7.50E-04	31.8%
	MFIS Phy.	3.097	1.27E-01	
	T25FW Speed:MFIS Phy.	-0.895	7.67E-03	
Total VPA Time	T25FW Speed	3.678	5.00E-03	9.2%
Total MVPA Time	T25FW Speed	37.482	2.29E-04	32.8%
	MFIS Phy.	4.851	5.61E-02	
	T25FW Speed:MFIS Phy.	-1.276	2.46E-03	
HMM-Based Statistics				
Maximum Step Rate	6MW Step Rate	0.853**	2.51E-20	75.0%
	MFIS	-0.229	4.64E-02	
HWSR	6MW Step Rate	0.516*	1.44E-05	73.4%
	MSWS-12	-0.732	4.26E-03	
	MFIS	0.229	2.47E-02	
HRSR	T25FW Speed	13.690	2.82E-02	52.0%
Total Walking Time	MFIS Phy.	-3.646*	2.89E-06	23.5%
Total Running Time	MFIS Phy.	-0.811	1.14E-02	7.6%
Total Active Time	MFIS Phy.	-4.457*	4.57E-06	22.7%

*$p < 10^{-4}$; **$p < 10^{-8}$; PA: Physical Activity; HPA: Habitual PA; MPA: Moderate PA; VPA: Vigorous PA; MVPA: Moderate-to-Vigorous PA; HWSR: Habitual Walking Step Rate; HRSR: Habitual Running Step Rate; 6MW: 6-Minute Walk; MFIS: Modified Fatigue Impact Scale; MSWS: MS Walking Scale

Differential Correlation by Disability Level

In mild MS, the MSR and average daily steps are moderately to strongly correlated with 6MW step rate ($r = 0.657$, $p < 10^{-6}$; and $r = 0.601$, $p < 10^{-4}$, respectively) and 6MW distance ($r = 0.467$, $p = 0.001$; and $r = 0.548$, $p < 10^{-4}$, respectively). Similarly, T25FW speed is moderately correlated with MSR ($r = 0.398$, $p = 0.006$) and average daily steps ($r = 0.536$, $p < 0.001$) in mild MS. The MFIS, on the other hand, is moderately correlated with the MSR and average daily steps in mild subjects ($r = -0.412$, $p = 0.004$; and $r = -0.358$, $p = 0.014$, respectively) and moderate subjects ($r = -0.424$, $p = 0.018$; and $r = -0.443$, $p = 0.013$, respectively).

As disability progresses, correlations to walking outcomes increase while correlations to

the MFIS decrease. In severe MS, the MSR and average daily steps are almost perfectly correlated with 6MW step rate ($r = 0.982$, $p < 10^{-7}$; and $r = 0.976$, $p < 10^{-6}$, respectively) and 6MW distance ($r = 0.979$, $p < 10^{-6}$; and $r = 0.962$, $p < 10^{-5}$, respectively). T25FW speed also correlates very strongly with MSR ($r = 0.919$, $p < 10^{-5}$) and average daily steps ($r = 0.944$, $p < 10^{-5}$), whereas the MFIS is not significantly correlated with either activity statistic ($p > 0.8$) in severe MS.

In summary, correlations between HPA and clinical outcomes vary substantially across the MS-disability spectrum. Average daily steps are an excellent measure of walking capacity in severe subjects, but they are significantly influenced by fatigue and behavioral factors in moderate and mild MS, precluding them from accurately measuring capacity. Indeed, the correlation between average daily steps and MFIS physical subscale is significant in mild and moderate MS, but not severe MS.

6.3.3 Clinical Significance of HMM-Based HPA Statistics

Having presented the clinical significance of several standard HPA summary statistics, we now illustrate the added clinical value of HMM-based classification. From a clinical perspective, the HMMs uncouple walking capacity from walking behavior, two factors that are tangled together in the HPA statistics in common use. Average daily steps, which are intended primarily to measure capacity, are confounded by behaviors because subjects choose how often they walk; thus, daily steps are not proportional to walking ability. Cut-point based physical activity, which is intended primarily to measure behavior, is confounded by capacity, as not all subjects are equally able to reach the pre-specified, population-level cut-points. The HMM untangles these factors by learning about subjects' habitual activity states. Capacity can be characterized in terms of the intensity of these states, whereas behavior can be characterized by their frequency and duration. The stepwise regression results are a highlight of this section, because they demonstrate that capacity has in fact been uncoupled from our proposed behavioral outcomes, including walking time and running time.

The final results presented in this section pertain to our secondary hypothesis regarding the relationship between physical activity and the best HPA cut-points available. As we show, walking and running fall outside of these cut-points in a large portion of our subjects, further emphasizing the importance of personalized activity classification.

Summary of HMMs

Fewer total model states were identified in MS subjects (median 4, range 3-6) compared to controls (median 5, range 2-6) ($p < 0.001$). Walking was identified in 76.3% of controls, 60.1% of mild MS, 41.9% of moderate MS, and 18.2% of severe MS ($\chi^2 = 15.7$, $p = 0.001$). Running was identified in 31.6% of controls, 19.6% of mild MS, and no subjects with moderate or severe MS ($\chi^2 = 14.7$, $p = 0.002$). In all subjects, the number of active states was significantly associated with younger age ($p < 0.001$), but not sex ($p = 0.095$), level of education ($p = 0.333$), or employment status ($p = 0.586$). MS subjects with more education had more active states ($p = 0.026$).

Figure 6.3 shows HMM classification in a control subject (top) and a subject with mild MS (bottom) during periods of physical activity. Step rates in each minute (right) are identified as running, walking, or other states based on fitted negative binomial distributions (left) and an underlying stochastic model. Note that classification depends on adjacent step rates as well as the current rate. In the mild MS subject, for example, minute 30 is classified as running, yet the step rate is lower than minute 35, which is classified as walking.

The HMMs identified walking or running in a majority of subjects (59.5%), but it is equally important to note that walking was *not* recognized in approximately 40% of subjects. These subjects almost certainly walked at some point during the week, but not enough to justify adding a "walking" state to the model in terms of improvement to the BIC. This result underscores the profound difference between capacity and behavior: a full 40% of subjects were not active often enough for the model to detect it. This rate differed between controls

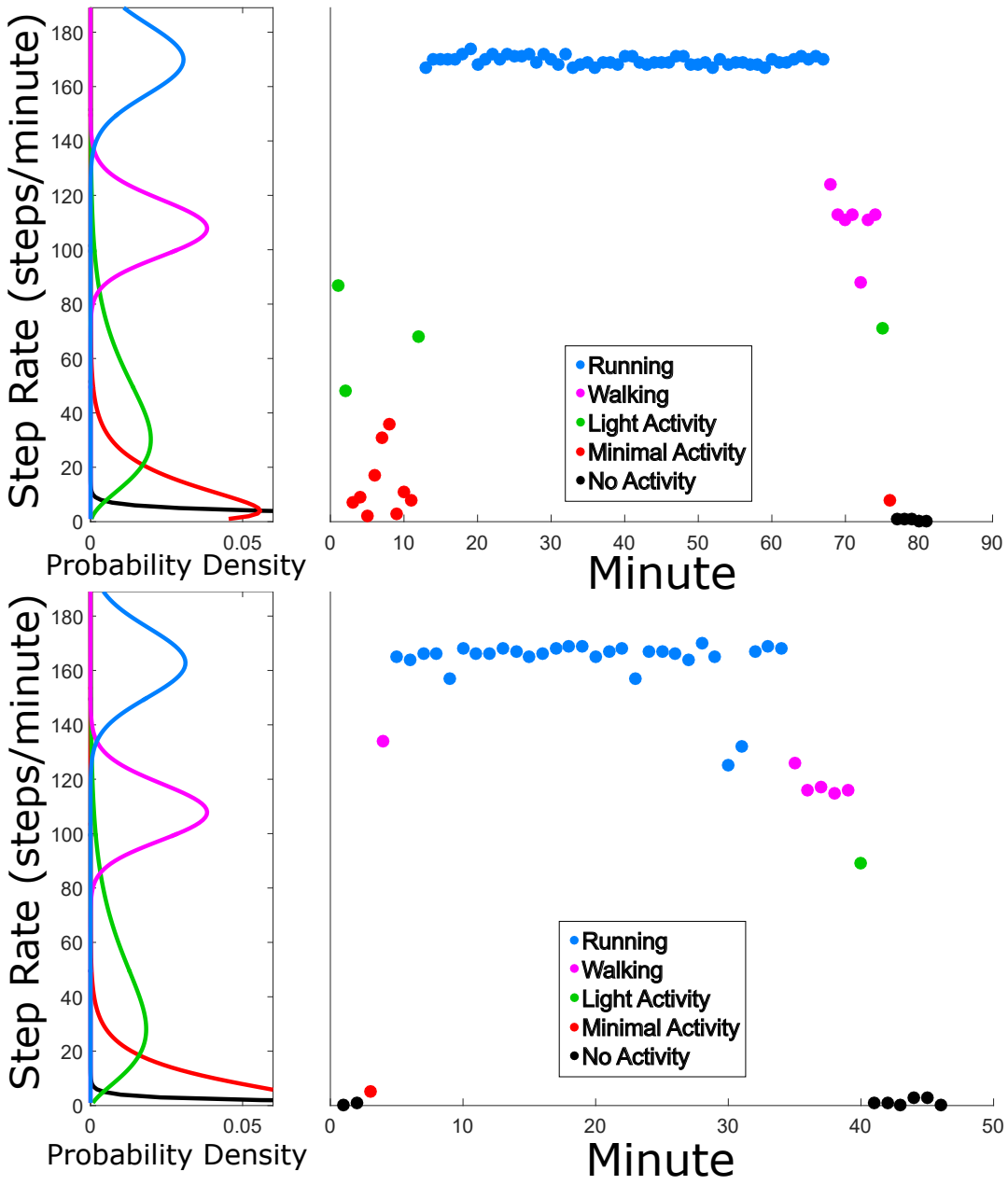


Figure 6.3: Activity Classification by HMM in Two Subjects

(23.7%) and mild to moderate MS (45.5%), supporting the finding that MS subjects are less active than controls[184, 185].

Statistics from HMMs

The habitual walking step rate (HWSR) was similar between controls, mild MS, and moderate MS: the decreases between groups was not statistically significant ($p = 0.196$). However, severe subjects had much lower HWSR ($p < 0.001$). The habitual running step rate (HRSR) was similar in controls and MS subjects ($p = 0.615$). Total walking time and total active time were significantly different between disability groups ($p < 0.001$), with controls having almost twice as much active time compared to mild MS (189.9%), and over five times that of moderate MS (535.3%). Active time was identified in 18.2% of subjects with severe MS; this was in contrast to the conventional analysis, which did not find MVPA in any of the severe subjects. The length of the longest active period differed between groups ($p < 0.001$), with a median of 17.5 minutes in controls (range 0-59), 10 minutes in mild MS (range 0-70), 0 minutes in moderate MS (range 0-30), and 0 minutes in severe MS (range 0-0).

HWSR was strongly correlated with 6MW step rate ($r = 0.815$, $p < 10^{-11}$) and 6MW distance ($r = 0.701$, $p < 10^{-7}$) in MS, but not in controls ($p > 0.4$) (Fig 3). Total active time was moderately and significantly correlated with the MFIS, including the MFIS Phy. and MFIS Psych. ($p < 10^{-4}$), but not with clinical walking outcomes ($p > 10^{-4}$). The longest active period was moderately and significantly correlated with the MSWS-12, MFIS, and MFIS Phy. ($p < 10^{-4}$) (Table 3). Among the demographic characteristics, total active time was significantly associated only with age ($r = -0.470$, $p < 10^{-7}$), much like total MVPA.

In contrast to the MSR, the HWSR was slower than the 6MW step rate in 94.4% of subjects. This difference was statistically significant ($\mu = -18.2$, $p < 10^{-19}$). Moreover, the difference between 6MW step rate and HWSR was moderately correlated with heart rate elevation during the 6MW in all subjects ($r = 0.529$, $p < 10^{-4}$) (Fig 6.4). Step rates during running were significantly higher than the 6MW step rate ($\mu = 21.7$, $p < 10^{-5}$).

While the MSR is the best single measure of walking capacity, the MSR and HWSR can be used together even more effectively. Indeed, their sum has higher correlation to the 6MW step rate ($r = 0.884$, $p < 10^{-14}$) than either factor alone. The MSR tends to over-estimate

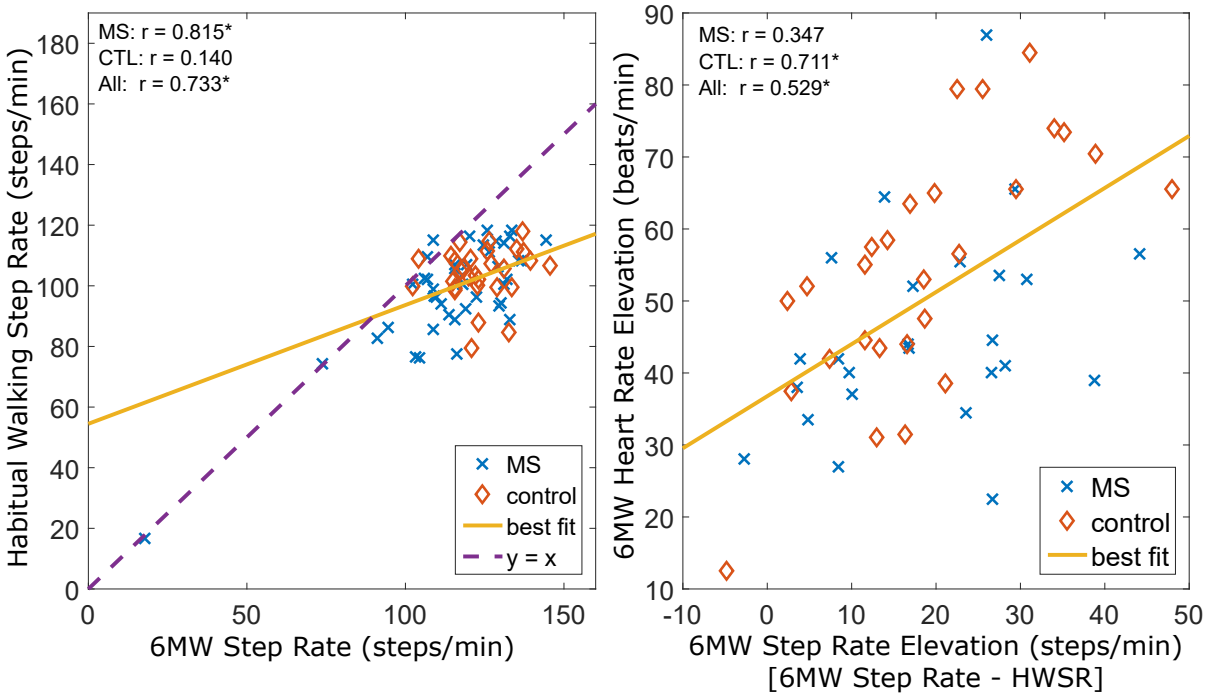


Figure 6.4: Difference between 6MW Step Rate and HWSR Predicts Heart Rate Elevation

6MW step rates, whereas the HWSR tends to under-estimate them. In other words, 6MW step rates are faster than subjects walk on a regular basis, but slower than their maximum rate. Therefore, any HPA-based estimate of the 6MW step rate should be placed between the HWSR and MSR. This result confirms that subjects walk quickly during the 6MW, as instructed, exceeding their habitual pace. In fact, the MSR was lower than the 6MW step rate in almost a third of subjects (31.7%), underscoring the genuine effort they made during the 6MW.

In the stepwise regression models, the 6MW step rate, MSWS-12, and MFIS were all retained as significant predictors of the HWSR, jointly explaining 73.4% of HWSR variance. Only T25FW speed was selected in the HRSR model, explaining 52.0% of HRSR variance. In contrast to the conventional statistics, total active time and the longest active period depended most on physical fatigue, not objective walking outcomes: MFIS physical subscale alone was retained in the models for total walking time, total running time, total active time, longest walk, longest run, and longest active period (Table 6.4). However, the MFIS physical

subscale explained less than 25.0% of the variance in these statistics.

The primary contribution of stepwise regression is to show that, in contrast to MVPA, the HMM activity statistics are not affected by capacity. Given that all subjects in our cohort are able to be physically active, the number and length of activities should not be affected by walking ability (though the associated step rates might be). The conventional statistics are influenced not only physical fatigue but also the T25FW, violating this property. The explanation is straightforward: many subjects with lower capacity dont reach the MPA cutoff while active. In contrast, physical fatigue is the only significant predictor of the HMM statistics. This factor explains less than 25% of variance in HMM statistics, leaving over 75% to unquantified behavioral factors. A behavioral outcome should not be influenced by capacity, but instead evaluate subjects activity in proper context. The HMM statistics achieve this, while the conventional statistics do not.

Activity Cut-Points

Step rates during habitual walking (HWSR) were higher in controls than in MS, and they decreased with increasing disease severity. Median HWSRs fell just above the MPA cut-point in controls, mild MS, and moderate MS (Figure 6.5); cut-points for severe MS have not been developed[87]. The steep drop-off in HWSR in severe subjects may be related to use of assistive devices, which substantially increase energy expenditure[86]. We note that the step rate cut-points are based on MSWS-12 scores, not EDSS, therefore the cut-points shown for moderate MS were applied to two of our mild MS subjects. This difference did not affect any of the results presented.

The HWSR and HRSR varied substantially not only between disability groups, but also between subjects with similar disability (Figure 6.5). The median HWSR was 105.5 in controls, 102.0 in mild MS, 96.7 in moderate MS, and 45.4 in severe MS. HWSRs fell below subject-appropriate MPA cut-points[87] in 30.1% of controls, 35.7% of mild MS subjects, 38.5% of

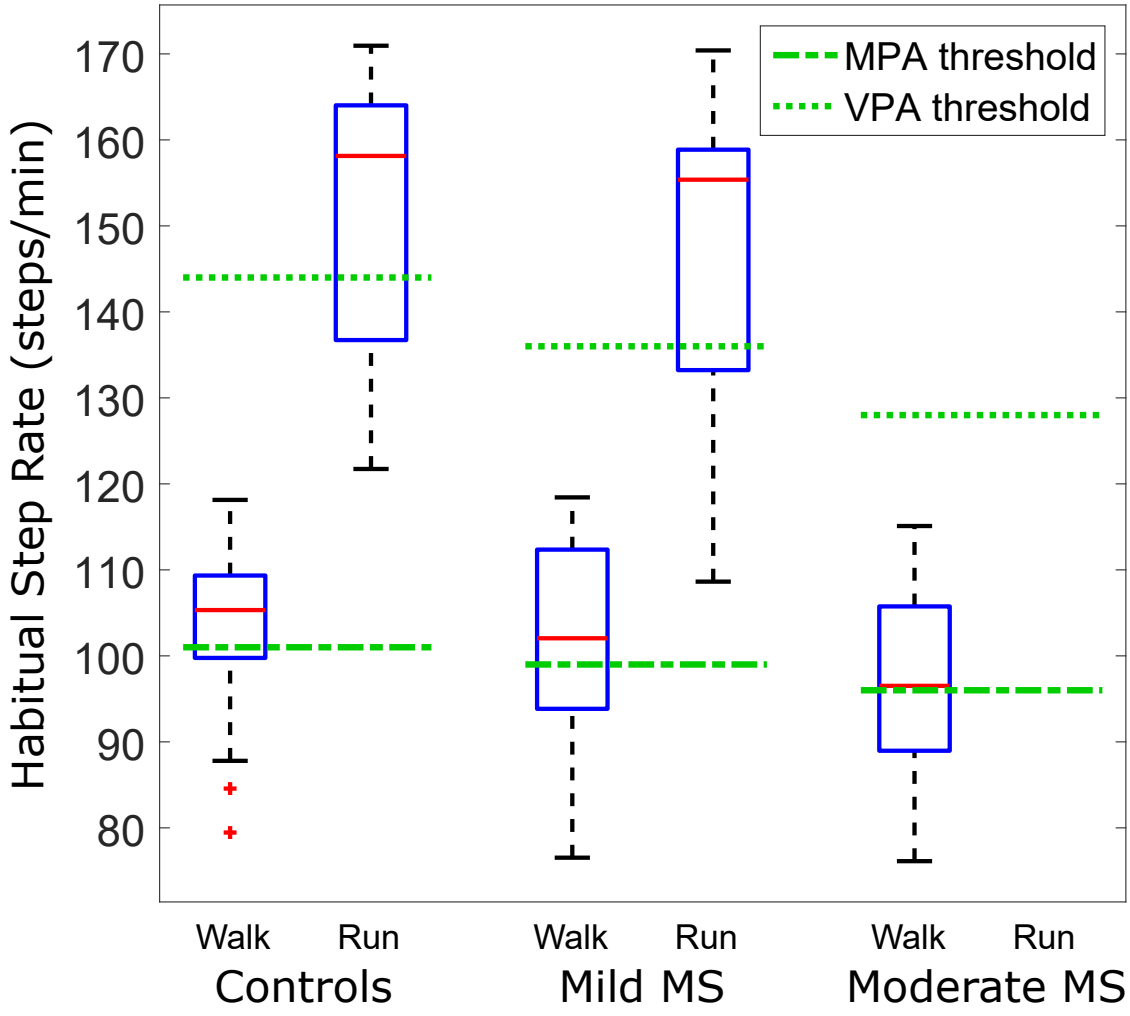


Figure 6.5: Variability in HWSR and HRSR within and between Disability Groups

moderate MS subjects, and all severe MS subjects. HRSRs fell below subject-appropriate VPA cut-points[87] in 25.0% of controls and 22.2% of mild MS subjects.

The HWSR results might seem to support the MPA and VPA cut-points, but closer inspection reveals several problems. HWSR varied substantially between subjects, as shown in Figure 6.5, so that the HWSR fell below the MPA cut-point in over a third of subjects. Consequently, their walking was not classified as MPA in the conventional analysis. Even when HWSR falls above the cut-point, walking might not be classified as MPA due to

within-subject step rate variability.

Deciding if this is an error – in other words, whether walking should be classified as MPA – depends on energy expenditure. If a subjects energy expenditure while walking does not exceed the U.S. DHHS requirement, then the walk should not be classified as MPA. Ultimately this is an empirical question which we intend to study directly in future work. However, energy expenditure at habitual walking speeds is remarkably consistent between persons[194], making it more likely that HWSR variability is caused by differences in age, weight[187], and leg length[86], as opposed to differences in energy expenditure.

Moving away from cut-points also provides improved information about the length and number of activity periods. The cut-point approach tends to overestimate the number of activity periods and underestimate their length[85] due to step rate variability during activities. As a result, the length and number of MPA, VPA, and MVPA periods obtainable from conventional analysis are not accurate. In contrast, the HMMs are contextually-aware; they "see" the step rates before and after the current point, reducing classification error. This is demonstrated in the bottom panel of Figure 2: the transient decrease in minute 30 is still identified as part of a single running activity.

6.3.4 Summary of Clinical Results

By applying a tried and true statistical method – the HMM – to habitual physical activity in MS, we have developed independent, real-world measures of walking capacity and walking behavior and demonstrated the limitations of contemporary methods of HPA processing. These methods fail to distinguish walking capacity from daily walking behaviors, resulting in imprecise outcome measures which are influenced by both factors. The MSR and HWSR measure capacity by detecting subjects' highest achieved step rate and habitual step rate, respectively, and their validity is supported by strong, statistically significant correlations to clinical walking outcomes. By this metric, the new statistics markedly outperform conventional HPA statistics. On the other hand, walking and running statistics from the HMMs are not

affected by capacity; they primarily measure behavior, making them appropriate outcomes for behavioral interventions. Moreover, our results cast fundamental doubts on the conventional approach to HPA, particularly the use of cut-points to classify MVPA, due to physiologic and behavioral variability between persons. These results illustrate the need for personalized, adaptive physical activity monitoring in the context of physical disability.

While these clinical results hold only in MS, we anticipate that similar results would be found in other forms of disability. Our primary goal is to improve MS outcomes, but we also aim to promote these techniques in HPA interpretation outside of MS. This will require continued work to validate the new statistics in a broader population. In particular, further study may be necessary in severe MS-related disability (EDSS$\geq$ 4.5), which comprised only 8.7% of our study population.

Having demonstrated the benefits of standard HMMs in a common clinical scenario, we now turn to the primary contribution of this chapter: Active Event Identification (AEI). HMMs improve activity classification through personalization, but AEI pushes several steps further. As we will see, AEI is model-adaptive; by judiciously eliciting input from patients, it learns to identify events in patient-defined terms, which can in turn inform its internal understanding of the world.

6.4 Active Event Identification (AEI)

In this section we present the AEI algorithm itself, along with supporting results. AEI is designed to learn about health events in subject-defined terms; it learns a correspondence between the *system's* understanding of the world and the *subject's* understanding of it. Often the subject has access to objective information that the system doesn't, so we can think of this correspondence as a reliable labeling of events by the subject. For example, subjects know whether they're walking or running, eating a meal, or having a panic attack, and can provide that information upon request. While the events themselves are objective –

they could be identified by an independent observer – the system measurements associated with those events may vary substantially from person to person, limiting the accuracy of more conventional classification methods. In other cases, monitoring may be focused on subjective experiences, such as pain or stress. The subject has privileged access to this information, so consulting him or her directly is the most straightforward way – perhaps the only way – to learn about these experiences' physiologic manifestations. At the same time, these consultations, or *queries*, come at a cost, namely the inconvenience to the subject. Minimizing patient inconvenience is critical to the success of monitoring systems, as we have repeatedly emphasized, so AEI must be judicious in its query selection.

It is important to note that in most cases, even the subjects themselves could not label events based on system observations alone. The subject knows when they're walking, for instance, but this doesn't imply that they could identify their walks based on step counts or other sensor data. Subjects can label events as they occur or by remembering them after the fact, but the match between system observations and subject observations is new information discovered by AEI.

Intuitively, the AEI algorithm learns to classify a stream of observations in subject-defined terms by querying at opportune times. Each observation o_t at time t is in O, and query responses r_t are in R. The responses may be viewed as additional observations, but they are only available when the system decides to query (and the subject responds). The relationship between observations, query responses, and hidden system states is shown by the probabilistic graphical model in Figure 6.6. This model is identical to an HMM with the addition of query responses, which are shaded in gray to indicate that they are not always obtained. These responses may also be viewed as the subject-defined state, whereas the usual states $s \in S$ are model-defined states.

When fitting personalized HMMs, we used the Baum-Welch algorithm[191], the most common approach to HMM training. Baum-Welch is an *offline* algorithm, meaning it is designed to run after all observations have been collected. In AEI, on the other hand,

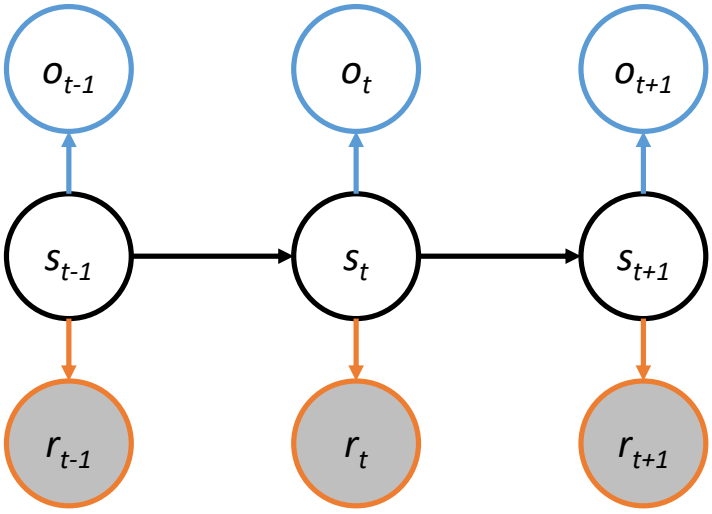


Figure 6.6: Probabilistic Graphical Model of AEI

subjects are queried about observations as they occur, so an *online*, real-time approach is required. Additionally, at each time t our algorithm must immediately decide whether to query. Queries should be as informative as possible – a notion we will formalize shortly – so the informativeness of each potential query must be assessed in real time. The first problem is solved by utilizing an online implementation of Baum-Welch, and the second is solved through active model learning. These topics are discussed in turn in the next two sections.

The goal of AEI is to correctly label observations in subject-defined terms. In other words, it must learn to predict subject responses. In the results sections, AEI is assessed in terms of the accuracy of those predictions, which improves over time. Further, its query selection should be better than random query selection in terms of improving its prediction accuracy. If not, active query selection is not useful.

This section begins with background regarding online Baum-Welch and active learning, then presents the AEI algorithm itself. The effectiveness of AEI is first explored with synthetic data. Finally, we evaluate AEI using our physical activity testbed, which is based on the fitted HMMs from persons with MS.

6.4.1 Online Expectation Maximization

There are a number of methods for online learning of HMMs, which is sometimes called symbol-wise learning[195]. AEI utilizes a straightforward, online adaptation of the Baum-Welch algorithm[196], an implementation of expectation maximization for HMMs. In this section we remind readers of the usual Baum-Welch update rules, then show how they are modified to suit the online case. A thorough presentation and derivation of Baum-Welch may be found in [191].

Baum-Welch begins with an initial point-estimate of HMM parameters $\theta^0 = \{\pi^0, \phi^0, \lambda^0\}$. The quantities $\alpha_i(t) = P(o_{1,t}, S_t = i \mid \theta)$ and $\beta_i(t) = P(o_{t+1,T} \mid S_t = i, \theta)$ are calculated recursively as follows for each state $i \in S$ in what is known as the forward-backward algorithm:

$$\alpha_i(1) = \pi_s \lambda_i(o_t) \tag{6.3}$$

$$\alpha_i(t+1) = \Big(\sum_{j=1}^{N} \alpha_j(t) \phi_{ji} \Big) \lambda_i(o_{t+1}) \tag{6.4}$$

$$\beta_i(T) = 1 \tag{6.5}$$

$$\beta_i(t) = \sum_{j=1}^{N} \phi_{ij} \lambda_j(o_{t+1}) \beta_j(t+1) \tag{6.6}$$

From α and β, the additional quantities $\gamma_i(t) = P(S_t = i \mid o_{1,T}, \theta)$ and $\xi_{ij}(t) = P(S_t = i, S_{t+1} = j \mid o_{1,T}, \theta)$ may be calculated:

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \tag{6.7}$$

$$\xi_{ij}(t) = \frac{\gamma_i(t)\phi_{ij}\lambda_j(o_{t+1})\beta_j(t+1)}{\beta_i(t)} \tag{6.8}$$

These quantities are then used to update θ to the maximum likelihood parameters with respect to the expected number of state transitions and observations from each state. The

update rules below assume that $\lambda_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, a univariate normal distribution.

$$\pi_i = \gamma_i(1) \tag{6.9}$$

$$\phi_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \tag{6.10}$$

$$\mu_i = \frac{\sum_{t=1}^{T} \gamma_i(t) o_t}{\sum_{t=1}^{T} \gamma_i(t)} \tag{6.11}$$

$$(\sigma^2)_i = \frac{\sum_{t=1}^{T} \gamma_i(t)(o_t - \mu_i)^2}{\sum_{t=1}^{T} \gamma_i(t)} \tag{6.12}$$

In this way, the parameters θ are repeatedly updated until a convergence criterion is reached, typically a threshold on the change in model likelihood.

This procedure is infeasible in an online setting, because it scales with the length of the observation sequence, which is always growing. Instead, we use a modified version of Baum-Welch which computes α, γ, and ξ for the current time only, then updates HMM parameters by incorporating this new information. More specifically, at the current time T, which is the last time observed so far, we compute the following:

$$\alpha_i(T) = \Big(\sum_{j=1}^{N} \alpha_j(T-1)\phi_{ji} \Big) \lambda_i(o_T) \tag{6.13}$$

$$\gamma_i(T) = \frac{\alpha_i(T)}{\sum_{j=1}^{N} \alpha_j(T)} \tag{6.14}$$

$$\xi_{ij}(T-1) = \frac{\alpha_i(T-1)\phi_{ij}\lambda_j(o_T)}{\sum_{k=1}^{N} \alpha_k(T)} \tag{6.15}$$

These updates follow from the usual Baum-Welch updates given that $\beta_i(T) = 1$ for all i; the important differences are (1) the new values are calculated only for time T, as we

have mentioned, and (2) $\alpha_j(T-1)$ was calculated with model parameters from the previous iteration, not the most recent model parameters. Model parameters θ are then updated to incorporate information at time T as follows:

$$\phi_{ij}^T = \frac{\sum_{t=1}^{T-1}\gamma_i(t)}{\sum_{t=1}^{T}\gamma_i(t)}\phi_{ij}^{T-1} + \frac{\xi_{ij}(T-1)}{\sum_{t=1}^{T}\gamma_i(t)} \tag{6.16}$$

$$\mu_i^T = \frac{\sum_{t=1}^{T-1}\gamma_i(t)}{\sum_{t=1}^{T}\gamma_i(t)}\mu_i^{T-1} + \frac{\gamma_i(T)o_T}{\sum_{t=1}^{T}\gamma_i(t)} \tag{6.17}$$

$$(\sigma^2)_i^T = \frac{\sum_{t=1}^{T-1}\gamma_i(t)}{\sum_{t=1}^{T}\gamma_i(t)}(\sigma^2)_i^{T-1} + \frac{\gamma_i(T)(o_T-\mu_i^T)^2}{\sum_{t=1}^{T}\gamma_i(t)} \tag{6.18}$$

Since parameters are updated as each new observation comes in, the parameter index is simply the time T, as our notation implies. In other words, at time T we update θ for the T^{th} time. The parameter updates appear more complicated than the usual forward-backward updates, but are actually easier to calculate, as the value $\sum_{t=1}^{T-1}\gamma_i(t)$ is simply stored in memory, and $\sum_{t=1}^{T}\gamma_i(t) = \sum_{t=1}^{T-1}\gamma_i(t) + \gamma_i(T)$. Importantly, we do not compute an entire forward pass, so $\alpha_i(T)$ only approximates $P(o_{1,t}, S_t = i \mid \theta^{T-1})$. This is due to the fact that that $\alpha_i(T-1)$ was computed using θ^{T-2}, $\alpha_i(T-2)$ was computed using θ^{T-3}, and so on. Nevertheless, this procedure has been shown to perform well in practice[196], and it also performs well in our datasets.

6.4.2 Active Learning

Deciding whether to query the subject is an example of *active learning*. This form of machine learning lies between supervised learning, in which all instances are labeled, and unsupervised learning, in which none of them are. In active learning, instances are initially unlabeled, but labels may be obtained from an oracle either in limited number, or at a cost[197]. In our application, the oracle is the patient, who has privileged information about current activities or experiences, and queries are budgeted over time.

The key to active learning lies in the method of assigning value to the potential labels that might be obtained. Once the value of labeling a particular instance is known, the learner typically requests the label of highest value. Our problem is more complicated, as we only allow AEI to obtain a label for the current time T, but we defer this discussion until the next section. The value of labeling can be based on several criteria, including uncertainty regarding the label, expected change to the model, expected reduction in classification error, and several other criteria[197]. In AEI, we wish to learn a correspondence between model states S and person-defined states R. This correspondence is specified by $\psi = P(R = r \mid S = s)$, a parameter of the augmented HMM depicted in Figure 6.6. The parameter ψ is similar to λ, but it is discrete rather than normally distributed. Intuitively, we'd like to choose queries which reduce our uncertainty about $\psi \in \Psi$. We formalize this intuition as the entropy of the posterior over Ψ given all other information. In other words, the expected value V of querying at time T is given by the following[198]:

$$V(T) = H(R_T \mid o_{1,T}, r_{1,T-1}, \theta) - H(R_T \mid \Psi, o_{1,T}, r_{1,T-1}, \theta) \tag{6.19}$$

To interpret this equation, we first note that $r_{1,T-1}$, the previously recorded subject responses, may be viewed as additional observations which are not always known. When responses become available, this new information is incorporated with a simple modification to the online EM update rules. If ψ were fixed, the expression $\lambda_i(o_T)$ in equations 6.13 and 6.15 could simply be replaced with $\lambda_i(o_T)\psi_i(r_T)$ to account for the probability of observing r_T. Instead, we maintain a distribution over Ψ to make active learning possible. As a result, we must use the expected value of Ψ, $\mathbb{E}(\Psi)$, based on the current posterior over Ψ. Setting $\psi^T = \mathbb{E}(\Psi)$, we update equations 6.13 and 6.15 as follows:

$$\alpha_i(T) = \Big(\sum_{j=1}^{N} \alpha_j(T-1)\phi_{ji} \Big) \lambda_i(o_T)\psi_i^T(r_T) \tag{6.20}$$

$$\xi_{ij}(T-1) = \frac{\alpha_i(T-1)\phi_{ij}\lambda_j(o_T)\psi_i^T(r_T)}{\sum_{k=1}^{N}\alpha_k(T)} \tag{6.21}$$

Returning to equation 6.19, the expression $H(R_T \mid o_{1,T}, r_{1,T-1}, \theta)$ is the entropy of the distribution $P(R_T = r \mid o_{1,T}, r_{1,T-1}, \theta)$, where $H(X)$ is given by $-\sum_{x\in X} P(X = x)\log P(X = x)$. Similarly, $H(R_T \mid \Psi, o_{1,T}, r_{1,T-1}, \theta)$ is a conditional entropy, where $H(X \mid Y) = \sum_{y\in Y} P(Y = y)\sum_{x\in X} P(X = x \mid Y = y)\log P(X = x \mid Y = y)$. We begin to evaluate equation 6.19 by substituting these expressions for H. Then, using our forward-backward algorithm notation from the previous section, we utilize an online approximation for $P(R_T = r \mid \Psi = \psi, o_{1,T}, r_{1,T-1}, \theta)$, which in turn leads to an expression for $P(R_T = r \mid o_{1,T}, r_{1,T-1}, \theta)$:

$$P(R_T = r \mid \Psi = \psi, o_{1,T}, r_{1,T-1}, \theta) = \sum_{i=1}^{N}\psi_i(r)\gamma_i(T) \tag{6.22}$$

$$P(R_t = r \mid o_{1,T}, r_{1,T-1}, \theta) = \sum_{\psi\in\Psi} P(\Psi = \psi \mid o_{1,T}, r_{1,T-1}, \theta)\sum_{i=1}^{N}\psi_i(r)\gamma_i(T) \tag{6.23}$$

By substituting equations 6.20, 6.21, and the expressions for H into equation 6.19, we end up with an equation which may be evaluated to determine $V(T)$ at each time T once o_T is known. The only remaining concern is the value of $P(\Psi = \psi \mid o_{1,T}, r_{1,T-1}, \theta)$, the posterior over Ψ, which must be updated to $P(\Psi = \psi \mid o_{1,T}, r_{1,T}, \theta)$ when a new response r_T is obtained. Note that $P(\Psi = \psi \mid o_{1,T}, r_{1,T-1}, \theta) = P(\Psi = \psi \mid o_{1,T-1}, r_{1,T-1}, \theta)$, since the new observation o_T alone does not affect our beliefs about Ψ given our online problem formulation. To simply the next few expressions, we write $P(\Psi = \psi \mid o_{1,T}, r_{1,T-1}, \theta)$ and $P(\Psi = \psi \mid o_{1,T}, r_{1,T}, \theta)$ as $p_\Psi^T(\psi)$. We can now use Bayes' rule to determine the value of $p_\Psi^T(\psi)$ with respect to the current belief state $\gamma(T)$:

$$p_{\Psi}^{T}(\psi) \propto P(r_T \mid \Psi = \psi, o_{1,T}, r_{1,T-1}, \theta) p_{\Psi}^{T-1}(\psi) \tag{6.24}$$

$$= \Big(\sum_{i=1}^{N} \psi_i(r_T)\gamma_i(T) \Big) p_{\Psi}^{T-1}(\psi) \tag{6.25}$$

Rather than evaluating the denominator found in Bayes' rule, we may simply normalize this distributionas shown in the next equation. Since γ is an online approximation, however, we utilize a conservative update rule for $p_{\Psi}^{T}(\psi)$ rather than the direct update in 6.25. The rule is similar to the online EM updates for ϕ, μ, and σ^2 (Equations 6.16 - 6.18), but we weight old vs new information based on η, the number of responses obtained thus far.

$$p_{\Psi}^{T}(\psi) = \frac{\eta - 1}{\eta} p_{\Psi}^{T-1}(\psi) + \frac{p_{\Psi}^{T-1}(\psi) \sum_{i=1}^{N} \psi_i(r_T)\gamma_i(T)}{\eta \sum_{\psi \in \Psi} p_{\Psi}^{T-1}(\psi) \sum_{i=1}^{N} \psi_i(r_T)\gamma_i(T)} \tag{6.26}$$

With this background in place, we now turn to the AEI algorithm itself.

6.4.3 The AEI Algorithm

Active event identification (AEI) is designed to teach our monitoring system a mapping between system states S and subject-defined states R. The term "subject-defined" implies that these states are inherently subjective, and in some cases this may be true. More generally, however, the subject has privileged information about their status which is not directly available to the monitoring system. They know (with certainty) whether they're walking, running, eating a meal, or experiencing a panic attack, and can teach the system to recognize those events. In some cases, the model may already "understand" these states in terms of their objective manifestations, but still need to learn their real-world labels. Importantly, however, learning this mapping can actually *improve* underlying model parameters by reducing uncertainty about the hidden model states S. In light of the arguments presented in Chapter 2, this aspect of AEI is critical. By selectively acquiring information from subjects, the

algorithm can improve its internal model to compensate for heterogeneity, drift, and other factors.

The mapping between S and R is determined by $\psi = P(R = r \mid S = s)$, as discussed in the previous section. AEI learns about ψ over time by selecting queries expected to reduce its uncertainty about Ψ, a population of possible instances of ψ. More specifically, it seeks to reduce the entropy of $p_\Psi^T(\psi)$, the posterior distribution over Ψ at time T.

The first step in implementing AEI involves specifying the population Ψ. As our previous notation implies, we have chosen Ψ to be finite. Each member $\psi \in \Psi$ is an $|S|$ by $|R|$ matrix whose elements ψ_{ij} specify the probability that the subject will provide response j from model state i. Note that while our distribution over Ψ is updated over time, each $\psi \in \Psi$ is time-invariant. We suppose that each state i has a correct label j, so that there exists a correct mapping f_c between model states and subject responses for which $f_c(i) = j$. Each possible ψ corresponds to a unique mapping of this form, the number of which is $|R|^{|S|}$. However, the subject is also capable of making an error with probability p_e. Together, we have the following structure for each ψ:

$$\psi_{ij} = \begin{cases} 1 - p_e & f(i) = j \\ \frac{p_e}{|R|-1} & f(i) \neq j \end{cases} \tag{6.27}$$

The initial distribution over Ψ – as well as the model parameters θ – may be initialized to incorporate prior knowledge when available. In this work, we initialize AEI with a uniform distribution over Ψ.

Lastly, we return to the fundamental question of active learning: when do we query? In an offline learning scenario, we could find $\text{argmax}_t\, V(t)$, the time maximizing the expected query value, and query at that time. In our online setting, however, only the current value is available. Rather than determining which query is best, the system must determine whether the current value is good enough. Motivated by the secretary problem[199], we incorporate a "minwait" parameter which specifies the minimum wait required between successive queries.

The algorithm keeps track of the maximum value of $V(t)$ during its required waiting period. When the wait is over, it decides to query whenever $V(T)$ exceeds this value.

A number of variations on this scheme might be considered. In fact, we have introduced a decay on the maximum value in our own experimental results. For the purposes of general AEI algorithm development, however, this simplest variation is most appropriate.

Active Event Identification Algorithm

procedure AEI(α, γ, ξ, θ, Ψ, $p^0_\Psi(\psi)$, minwait)
- $\sum(\gamma) \leftarrow 0$
- $\eta \leftarrow 0$
- wait $\leftarrow 0$
- threshold $\leftarrow 0$
- **while** monitoring **do**
  - $o_T \leftarrow$ **observation**
  - **update** α, γ, ξ **with** o_T ▷ (Eq 6.13 - 6.15)
  - **update** θ **with** α, γ, ξ, $\sum(\gamma)$ ▷ (Eq 6.16 - 6.18)
  - $v \leftarrow V(T)$ ▷ (Eq 6.19)
  - **if** wait $<$ minwait **then**
    - wait $\leftarrow$ wait $+ 1$
    - threshold $\leftarrow \max(\text{threshold}, v)$
  - **else if** $v \geq$ threshold **then**
    - **query** subject
    - $r_T \leftarrow$ **response**
    - $\psi^T = \mathbb{E}(\Psi)$
    - **update** α, γ, ξ **with** o_T, r_T ▷ (Eq 6.20, 6.14, 6.21)
    - **update** θ **with** α, γ, ξ, $\sum(\gamma)$ ▷ (Eq 6.16 - 6.18)
    - **update** $p^T_\Psi(\psi)$**with** r_T, γ, $\sum(\gamma)$, $p^{T-1}_\Psi(\psi)$ ▷ (Eq 6.26)
    - wait $= 0$
    - threshold $= 0$
    - $\sum(\gamma) \leftarrow \sum(\gamma) + \gamma$
    - $\eta \leftarrow \eta + 1$
  - **end if**
- **end while**

end procedure

For readability, the details of AEI parameter updates have been omitted. They may be found in the previous two sections; the specific equations required are specified in the comments to the right of each update statement.

Care must be taken when initializing AEI. We recommend uninformative values for α, γ, and ξ in particular, as the first few parameter updates depend heavily on their current values. For example, initializing α with all weight on a single state will invariably result in pathological HMM parameters, leading to errors. If prior information is included, $\sum(\gamma)$ may instead be initialized to a nonzero value to prevent this information from being immediately lost in the first parameter update. As suggested by Stenger et al [196], $\sum(\gamma)$ may be replaced with a fixed or variable learning rate depending on application requirements.

The loop in AEI continues indefinitely. Indeed, our strongly online implementation, which utilizes a constant amount of memory and computation, allows the algorithm to continue as long as hardware and real-world considerations will allow. The exceptions to this rule are the numeric values themselves, which can continually grow or shrink with successive iterations. This may be solved by normalizing the Baum-Welch parameters α, γ, and ξ, and by utilizing a learning rate in place of $\sum(\gamma)$.

6.4.4 AEI Simulation Results

In the first phase of validation, AEI was applied to synthetic data generated by Markov chains with Gaussian state to observation distributions. 50 chains with four states each were used to generate 50 corresponding observation sequences 10,000 samples in length. The mean and variance for each state were randomly selected from $\mathcal{U}(0,1)$ and $\mathcal{U}(0.025, 0.125)$, respectively, and were distinct for each of the 50 chains. The initial state vector was set to $\pi = (0.25, 0.25, 0.25, 0.25)$, and the following transition matrix ϕ was used:

$$\phi = \begin{bmatrix} 0.98 & 0.015 & 0.003 & 0.002 \\ 0.015 & 0.98 & 0.003 & 0.002 \\ 0.04 & 0.04 & 0.9 & 0.02 \\ 0.04 & 0.04 & 0.02 & 0.9 \end{bmatrix} \tag{6.28}$$

This transition matrix was chosen to meet several criteria. First, self-transitions are common, mimicking real-world events, which tend to span many sensor readings. Second, two of the states occur much frequently than the other two; in other words, there are two typical, commonly observed states and two unusual, infrequently observed states. Again, this was motivated by real-world events; we are most interest in accurately identifying the unusual states. Real-world health events such as sleep apnea, panic attacks, or physical activity occur infrequently, but the goal of AEI is to accurately identify them.

The AEI algorithm was run with five states, each with Gaussian state to observation distributions. The means of these distributions were evenly spaced between 0 and 1, with variance equal to 0.04. Initial values of α, γ, and ξ were then determined by introducing a 1000-sample "lag" used for a single pass of the forward-backward algorithm followed by a single HMM parameter update. The resulting values at $t = 1000$ were used to initialize AEI.

The most important initial setting for AEI, the distribution over Ψ, was chosen to be uniform over all possible $\psi \in \Psi$. In other words, no information about the correct $\psi^* \in \Psi$ was provided to the algorithm. Subject-defined events were determined by randomly mapping each state s_i of the synthetic Markov chain to a response $r_j \in \{1, 2, 3, 4\}$. The primary goal of AEI was to learn this mapping through its queries.

Figure 6.7 shows the application of AEI to one of the synthetic Markov chains. The top panel shows subject-defined events. By carefully inspecting this plot, we can observe that two of the states in the Markov chain – one frequent and one infrequent – have been mapped to the yellow event. The second panel shows the model defined states. We see that our online implementation of Baum-Welch has identified the four distinct model states within the first 2000 samples. The third panel is the most important of the figure; it shows the system's event predictions. By sample 3000, we see that it has learned to match the events shown in the first panel. Through selective querying, it has determined the correct mapping between S and R.

The fourth panel shows the predictions of an AEI-like algorithm that queries at *random*

Figure 6.7: Example of AEI Applied to Gaussian HMMs

times. We will refer to this algorithm as RQ (for "random query"). RQ's internal logic is the same as AEI with the exception of query selection. With its random queries, RQ is not able to identify the infrequent, brown-colored state even after 10,000 samples, and it identifies the green state much less quickly than AEI. The bottom panel shows the timing of AEI's queries. In total, both algorithms queried 73 times.

Figure 6.8 shows the accuracy of AEI in all 50 synthetic datasets. The left panel shows overall accuracy, defined as (correct event predictions ÷ total samples). The line shows the median accuracy among all 50 datasets, and the error bars show the interquartile range.

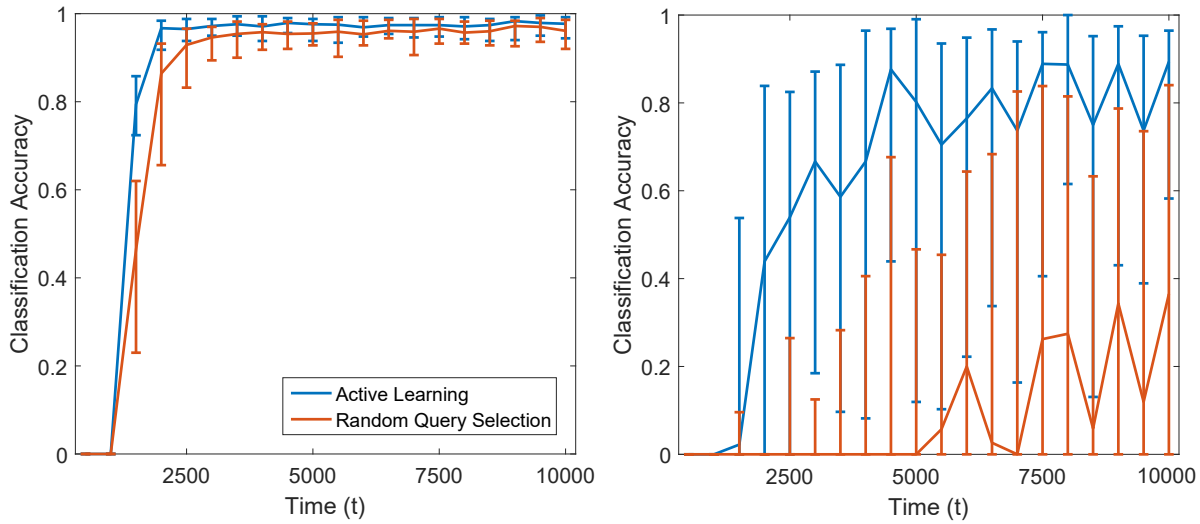


Figure 6.8: Accuracy of AEI Applied to Gaussian HMMs

It is clear that AEI outperforms RQ in terms of the speed of its initial improvement in accuracy. Beyond 5000 samples AEI's performance is marginally better than RQ. In the right panel, however, the difference between the two algorithms is dramatic. This panel shows classification accuracy for the *infrequently occuring responses only*, defined as (correct infrequent event predictions ÷ total infrequent event samples). By this metric, the advantage of AEI is clear. While RQ succeeds in learning the frequently occurring states, it often fails to learn the infrequently occurring states. This is a critical disadvantage in any real-world health monitoring scenario.

Intuitively, the event identification algorithms have limited opportunities to learn about the infrequent states. AEI is able to identify and seize these opportunities by quantifying the value of queries, whereas RQ is unable to identify them.

6.4.5 Testbed: Physical Activity in MS

The second phase of AEI validation was conducted using a physical activity *testbed*, or simulator, based on data from persons with MS. The fitted HMMs presented earlier in this chapter (Section 6.3) were used to generate new observations for a 14 day period. Use of this

testbed was necessary because the raw data alone is not sufficient for AEI, which requires subject-defined labels. While labels can be *estimated* from the raw data, the testbed approach is superior because the true states of the underlying Markov chain are known with certainty.

The possible subject-defined states were "walking", "running", or neither. Formally, our set of responses R was given by $R = \{\text{walking, running, not active}\}$. Walking and running states were both identified in 18 subjects out of the 126 in our cohort, many of whom were unable to run due to moderate or severe disability. Results from these 18 subjects were used to create the testbed. Each fitted HMM had between 5 and 7 states in total. As we have previously described, one of these states is a "not worn" state which always produces zeros as observations, and the remaining states are defined by negative binomial distributions. The fitted HMMs in the testbed also include transition matrices and initial state distributions, so generating new observations is straightforward. Responses were determined based on the underlying HMM state. Indentification of walking and running states is described in Section 6.3.2. The running state was mapped to the response "running", the walking state was mapped to "walking", and all other states were mapped to "not active". The 14 new days' worth of observations corresponded to $14 \times 60 \times 24 = 20160$ new samples.

Demographics for the testbed subjects are summarized in Table 6.5. All of the subjects with MS had mild disability (EDSS ≤ 2.5), and only one of the subjects, a control, was male. Age and walking outcomes were similar between groups.

AEI and RQ initialization were very similar to the previous section. This time a lag of 1440 samples (1 day) was used. As before, both event identification algorithms were given 5 distinct states, each with Gaussian state to observation densities. Their means were evenly spaced between 0 and 200, with variance equal to 1600. Again, the initial distribution over Ψ was uniform.

Figure 6.9 shows the application of AEI to data from one of the testbed subjects. This figure follows the same format as Figure 6.7: the top panel shows the subject-defined events (R), the second panel shows the model states identified by AEI, the third panel shows

Table 6.5: Demographics and Clinical Outcomes for Testbed Subjects

Variable	Control	Mild MS
Number of Subjects	10 (55.6%)	8 (44.4%)
Sex, F:M (% Female)	9:1 (90%)	8:0 (100%)
Age, mean ± SD	31.9 ± 10.1	40.3 ± 12.4
6MW Distance, mean ± SD	2119.6 ± 322.6	2003.2 ± 206.6
6MW Step Rate, mean ± SD	124.5 ± 11.8	131.4 ± 8.3
T25FW, mean ± SD	3.48 ± 0.49	3.65 ± 0.55
MSWS Score, mean ± SD	NA	15.1 ± 8.1
MFIS Total, mean ± SD	18.4 ± 7.3	19.8 ± 20.5

SD: Standard Deviation; 6MW: 6-Minute Walk; T25FW: Timed 25-Foot Walk; MSWS: MS Walking Scale; MFIS: Modified Fatigue Impact Scale

AEI's predictions of subject-defined events, and the fourth panel shows RQ's predictions of subject-defined events. By day 2, AEI has begun to understand events in subject-defined terms, and by day 4 it is highly accurate. RQ, on the other hand, has very poor accuracy throughout the 14 day period. In particular, it is never able to identify running, which occurs least often. Again, AEI has learned the correct mapping between S, shown in the second panel, and R, shown in the first panel, through selective querying. RQ queries just as often, but its queries are much less effective in terms of improving classification accuracy. In total, both algorithms queried 64 times in total over the 14 days. In a real-world scenario, the frequency of queries can be tuned using the "minwait" parameter.

Figure 6.10 shows the accuracy of AEI in all 18 of the testbed subjects. As before, accuracy increases quickly by the end of the first day of querying (Day 2), with better results observed for AEI compared to RQ. Differences may still be identified by Day 14, but they are very small. Once again, the right panel shows the dramatic difference between algorithms in terms of identifying *infrequently occurring* events. In this case, the infrequently occurring events are "walking" and "running", and the only frequently occurring event is "not active". Note that an algorithm whose prediction was always "not active" would perform well in terms of overall accuracy; this is exactly what is occurring in the left panel. For this reason, the right panel is much more indicative of the algorithm's ability to learn. RQ sometimes

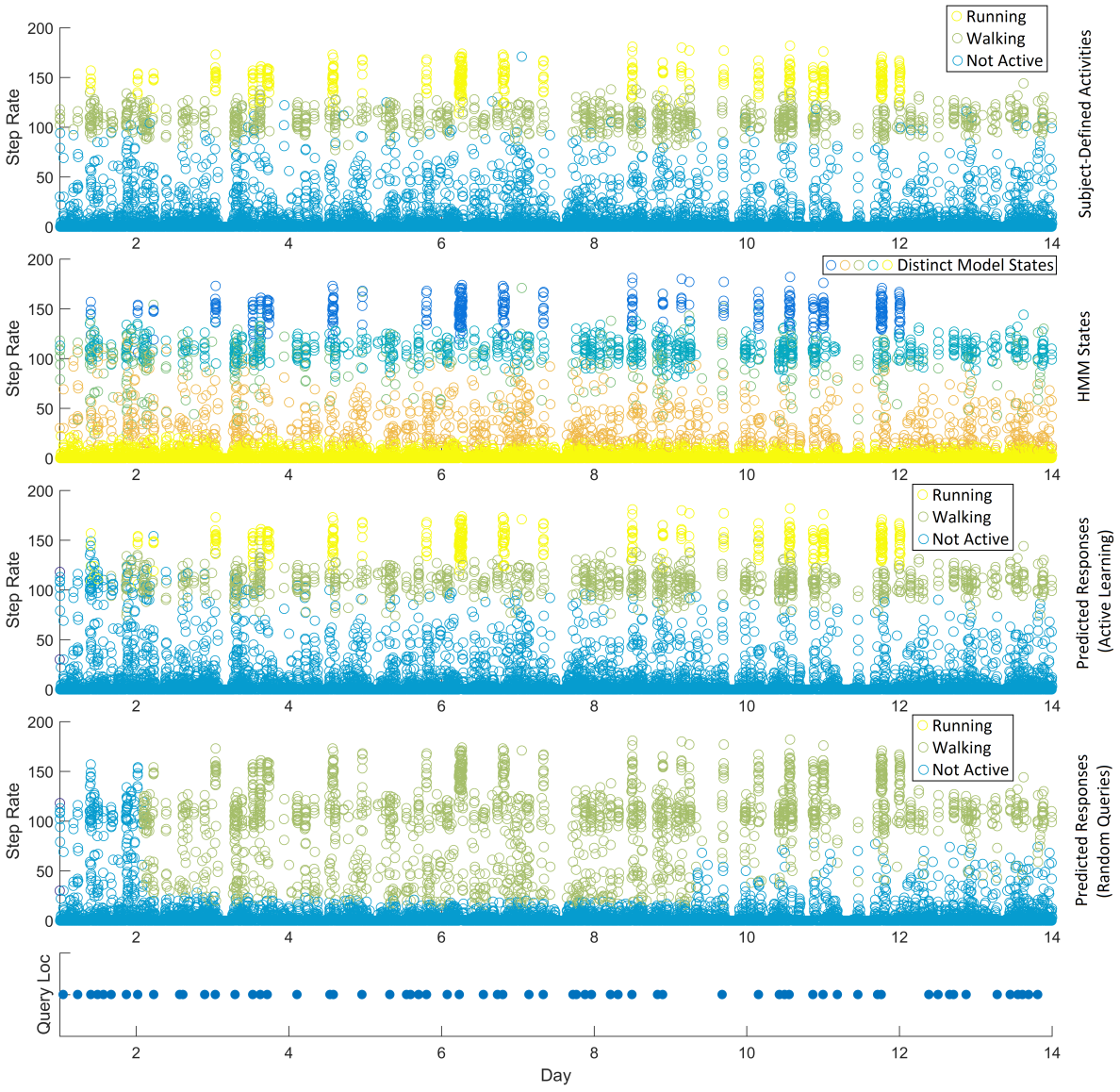


Figure 6.9: Example of AEI Applied to Physical Activity Testbed

performs well, as shown by the error bars, which indicate the interquartile range among all subjects. However, AEI consistently performs well, while RQ does not. On average, AEI correctly identifies over 90% of the minutes in which subjects are physically active by Day 6, whereas RQ identifies less than 75% of them even after Day 14.

This second round of validation further supports the effectiveness of AEI and the advantages it gains via active learning. Again, AEI is able to use information about the expected value of query responses in order to time its queries to optimize learning.

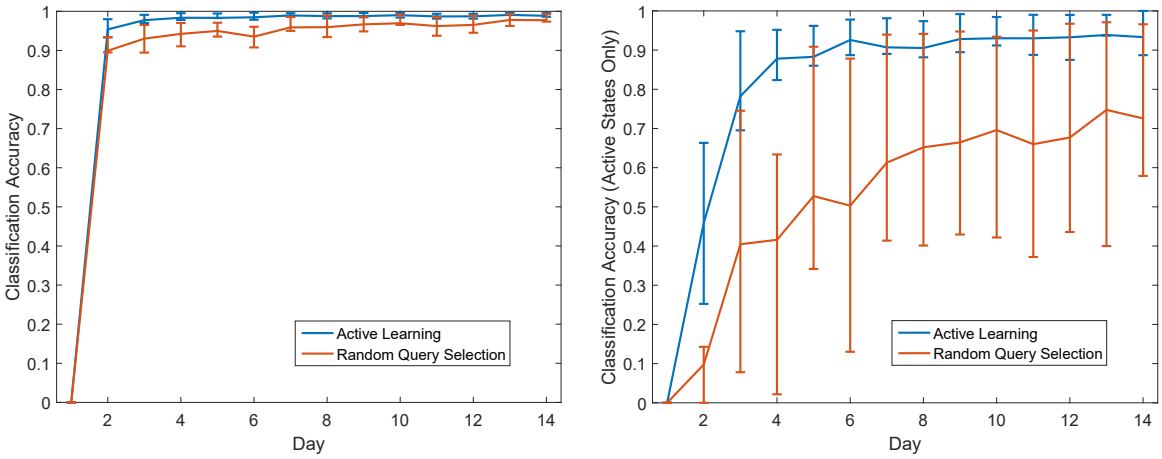


Figure 6.10: Accuracy of AEI Applied to Physical Activity Testbed

6.5 Summary and Future Work

This chapter began with clinically significant results that motivated and supported active event identification. A ubiquitous technique – the fitting of a Hidden Markov Model – helped to clarify the relationship between walking capacity and walking behaviors, potentially supporting a new approach to the monitoring of habitual physical activity in chronic disease. HMMs and other hidden variable models can be applied to a wide variety of health monitoring scenarios, such as apnea detection[200] or the identification of heart sounds[201]. Indeed, whenever physiologic measurements may be modeled as conditionally independent given an underlying state, an HMM-based approach to classification may be useful. The HMM is therefore an ideal foundation for an adaptive event classification system.

The culmination of this chapter was the presentation of active event identification (AEI), an adaptive algorithm which interacts with the subject in order improve its understanding of health-related events. AEI can use subject interactions to learn about meals, physical activities, heart attacks, seizures, and many other events. The subject knows when these events occur, and can therefore teach the system to recognize them given an appropriate paradigm for system-subject interaction. AEI optimizes these interactions by prompting at particularly valuable moments, allowing it to improve classification accuracy in many typical

monitoring scenarios without unduly burdening the patient. Even more exciting, AEI opens up an entirely new dimension in monitoring. By learning to recognize events in subject-defined terms, AEI can attempt to map subjective experiences to objective, physiologic manifestations. This aspect of AEI has applications in affective computing, psychology, and other fields that emphasize the subject perspective. There are countless unexplored questions in this domain, many of them related to the physiologic underpinnings of subjective experience.

Moreover, the subject is not the only "oracle" which can provide additional information about events. AEI can be applied to situations in which a high-powered must be used selectively, as described in Chapter 2, or care providers must decide whether to utilize a costly clinical diagnostic. By drawing upon the flexible, well-known HMM framework, AEI has innumerable applications.

Independent of AEI itself, these results illustrate the potential of active learning and patient interaction in health monitoring. We believe that patient interaction is an underutilized, exceedingly valuable source of clinically-relevant information. With the advent of mobile technologies, our ability to tap into this information is now limited by methodology, not the availability of the data itself. Active learning is a promising, versatile method to improve our understanding of clinical data, and it has a growing theoretical foundation. Promoting active learning and mobile-based patient interactions is an important goal of this work.

AEI is adaptive in every sense explored in this work. It adapts its decisions (query or not query), its outputs (event classifications), and its internal model to the growing history of objective observations and subject responses. These adaptations allow it to gracefully overcome many common difficulties encountered in health monitoring, as outlined in Chapter 2. With some judicious tuning of the learning rate, AEI can deal with heterogeneity and drift through continual model adaptation. It reduces patient burden by optimizing its interactions with the subject, as we have repeatedly emphasized, and it offers an elegant solution to the "problem" of subjectivity by learning about events in the patient's own terms. Importantly, however, it is only appropriate for monitoring problems which satisfy the

assumptions described in Section 6.1. These assumptions are not met, for example, by the scenarios presented in Chapter 4, in which consecutive measurements are highly correlated, or Chapter 5, in which underlying states fall on a spectrum of disability.

6.5.1 Limitations and Refinements

AEI has many limitations. In many ways, it is a work in progress. Most obviously, the validation process has just begun. While use of the physical activity testbed is an improvement over purely synthetic data, AEI remains to be validated in real, clinical data. This is quite difficult, as labeled data is hard to come by. Ironically, the scarcity of labeled data has partly motivated development of AEI, yet it is also a limiting factor. While it is possible to test AEI retrospectively with a curated, labeled dataset, the ideal method of validation is a prospective trial, which would serve not only to validate the algorithm, but also to assess is real-world utility.

Additionally, AEI has several limitations from a technical perspective. While Baum-Welch is never guaranteed to maximize the likelihood function, our online implementation may be particularly susceptible to undesirable local optima. As we have warned, the algorithm must be carefully initialized to avoid overfitting to the first few observations. A method to correct overfitting and/or pathologic HMM parameters must be developed prior to a prospective deployment.

From an optimal decision-making perspective, AEI is greedy. It attempts to maximize the value of its next query, but makes no attempt to assess the impact of querying on future query value. This may be interesting from a theoretical standpoint, but practically speaking, choosing queries which are optimal over the long-term is neither realistic nor necessary.

In its current form, AEI is extremely susceptible to the "curse of dimensionality", because the size of Ψ is $|R|^{|S|}$. This was not an issue in our validation studies due to the limited number of model states and responses, but must be corrected to make AEI generally applicable. The easiest solution to this problem might be to learn about model states in sequence. The system

could first determine which response is associated with s_1, for example, then move on to s_2 once its confidence about s_1 exceeds some threshold.

A more flexible and general approach is the use of ensemble learning, or Variational Bayes[202], which has been applied to active learning of HMMs in a different context[203]. This approach would allow us to place fewer assumptions on the structure of Ψ and assess the impact of queries on the model as a whole rather than on Ψ alone. Variational Bayes is a more principled way to maintain a population of models than the *ad hoc* approach we have used, though it comes with its own set of difficulties.

Lastly, our criterion for query selection could be explored in more detail. As we have pointed out, we have been motivated by the secretary problem, which itself has many variants. Further, there are numerous caveats and costs that must be accounted for in a real-world deployment. One particularly promising direction might be the incorporation of additional contextual information, such as location or phone usage, to inform the algorithm about the *cost* of querying. Generally speaking, there are a large number of potentially fruitful research directions related to AEI.

6.5.2 Active Behavioral Feedback

Lastly, there is an important connection between AEI and behavioral feedback, a topic of considerable interest in current clinical research. Section 2.4 overviews recent work in mobile interventions targeting behavior modification in type II diabetes in and several other chronic diseases. Indeed, encouraging healthy behaviors and implementing other disease prevention strategies might be the most impactful potential applications of mobile health technologies.

AEI relates to behavioral feedback in two related ways. First, it may be able to facilitate effective feedback by informing the monitoring system about subject health and behaviors. For instance, accurate information about the subject's meals and snacks can be leveraged by a feedback system to encourage healthy eating habits. Additionally, knowledge of the physiologic manifestations of events – say, the corresponding change in blood glucose – can

also help care providers better understand subject health, which contributes to behavior modification strategies.

Second, the queries selected by AEI are *themselves* a form of feedback. For better or worse, the appearance of a query reminds the subject that their behaviors are being monitored and interpreted. If the current event is physical exercise, the query might feel rewarding: it tells the patient that their positive behavior has been noticed. On the other hand, the situation might be reversed if the event were a nighttime snack, for instance. Depending on the circumstance or person, AEI's queries could have either positive or negative effects.

Chapter 7

Closing Remarks

The initial hype surrounding mobile health has passed, and we are now in the "trough of disillusionment". Despite the large number of remote monitoring trials that have been reported, the *impact* of mobile technologies remains unclear, particularly over the long-term. In our view, the core vision of the movement is sound: ubiquitous sensing and computing will make clinical medicine more exact, accessible, and holistic. However, there are a number of unforeseen obstacles now taking shape. Converting real-world data into clinical insights is a monumental task requiring collaboration between clinicians, engineers, and data scientists, one made more difficult by the complexities and heterogeneity of human physiology and behavior. Further, data collection and conceptualization ultimately depends on the cooperation of clinical trial participants, who tend to become less engaged over time.

Adaptive, personalized algorithms offer tremendous advantages in the face of these dilemmas. The overarching goal of this work has been to demonstrate these advantages through an extended case study in multiple sclerosis, a chronic and notoriously heterogeneous disease. By adjusting to the physiology and preferences of individual patients, adaptive algorithms can (1) maintain accurate classification, estimation, and/or decision-making despite these differences, (2) reduce the burden of monitoring placed on the patient, and (3) reduce or eliminate the need for large-scale data collection prior to deployment. Adaptation

is closely connected to the goals of precision medicine, which brings renewed focus and clarity to health monitoring by emphasizing variability between persons.

Over the course of this material, we have gradually worked toward algorithms that adapt in the strongest sense of the term. Our DTW-based physiologic signal monitoring algorithm is inherently personalized, as it evaluates the current signal with respect to a known baseline. It is not adaptive according to our definition, however, because currently there is no method to reevaluate this baseline over time. In Chapter 4 we suggest possible extensions to correct this shortcoming. Briefly, a baseline set of signal templates can be curated via distance-based clustering. Adaptive symptom reporting is adaptive, as the name implies, because it dynamically selects symptom-related queries in real-time based on an evolving history of observations. This approach allows it to accurately estimate disability while minimizing the burden placed on the patient. Active health event classification takes adaptation a step further by continually improving its internal model of the world. In this sense it is model-adaptive, or strongly adaptive, and can be deployed with minimal prior knowledge if the application requires it.

In may ways, Chapter 6 is the culmination of this work. Like ASR, AEI optimizes system-patient interactions through intelligent, adaptive prompting, and like the Warp Score, it gracefully adjusts to heterogeneity between and within persons. In the terms of Chapter 2, AEI addresses problems of patient burden, heterogeneity, and subjectivity through active learning and model adaptation. It can be applied to a wide variety of health events, with additional applications outside of health care: instead of the patient, the "oracle" queried by our system might be a high-powered sensor, an expensive diagnostic, or an expert annotator. This chapter is the most theoretical portion of the thesis, contributing to current research in active learning and self-adaptive systems.

Closing material specific to our research, including limitations and detailed plans for future work, is presented at the end of Chapters 3-5. Instead, we close with insights from two clinical research participants, one a patient and the other a provider, that remind us of the personal

nature of patients' perspectives and preferences on health monitoring. The first quote comes from a subject in our remote monitoring cohort, who explained that "answering the questions helped me face up to my mild disabilities and keep them in perspective". For this person, routine symptom reporting was a positive experience, perhaps even an empowering one, and indeed, they completed the surveys quite frequently. On the other hand, one of our providers reminds us that many patients prefer *not* to confront their disability on a daily basis; these patients rely on her to worry about their disease while they "put it on the shelf". In order for mobile health to escape the "trough of disillusionment", such differences between patients, whether physiologic or behavioral, must be understood and respected.

Bibliography

[1] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors (Basel, Switzerland)*, 13(12):17472–17500, 2013.

[2] The White House. Fact Sheet: President Obama's Precision Medicine Initiative.

[3] Francis S. Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795, feb 2015.

[4] National Research Council of E National Academies Disease, Committee on A Framework for Developing a New Taxonomy of Disease, Board of Life Sciences, Division of Earth and Life Sciences. *Toward Precision Medicine*. National Academies Press, Washington, D.C., dec 2011.

[5] M Stewart, J B Brown, A Donner, I R McWhinney, J Oates, W W Weston, and J Jordan. The impact of patient-centered care on outcomes. *The Journal of family practice*, 49(9):796–804, sep 2000.

[6] Klea D Bertakis and Rahman Azari. Patient-Centered Care is Associated with Decreased Health Care Utilization. *The Journal of the American Board of Family Medicine*, 24(3):229–239, 2011.

[7] Klea D Bertakis and Rahman Azari. Determinants and outcomes of patient-centered care. *Patient Education and Counseling*, 85(1):46–52, oct 2011.

[8] Ronald M. Epstein, Peter Franks, Cleveland G. Shields, Sean C. Meldrum, Katherine N. Miller, Thomas L. Campbell, and Kevin Fiscella. Patient-centered communication and diagnostic testing. *Annals of Family Medicine*, 3(5):415–421, sep 2005.

[9] Eskinder Eshetu Ali, Lita Chew, and Kevin Yi-Lwern Yap. Evolution and current status of mhealth research: a systematic review. *BMJ Innovations*, 2(1):33–40, jan 2016.

[10] Thomas L Webb, Judith Joseph, Lucy Yardley, and Susan Michie. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy., 2010.

[11] Alastair van Heerden, Mark Tomlinson, and Leslie Swartz. Point of care in your pocket: a research agenda for the field of m-health. *Bulletin of the World Health Organization*, 90(5):393–394, may 2012.

[12] Irem Tumer and Anupa Bajwa. A survey of aircraft engine health monitoring systems. In *35th Joint Propulsion Conference and Exhibit*, volume 99, pages 2528–2536, Reston, Virigina, jun 1999. American Institute of Aeronautics and Astronautics.

[13] S. W. Doebling, C. R. Farrar, and M. B. Prime. A Summary Review of Vibration-Based Damage Identification Methods. *The Shock and Vibration Digest*, 30(2):91–105, mar 1998.

[14] Michael G Pecht. Prognostics and Health Management of Electronics. 29(1):222–229, 2009.

[15] Peter C. Chang, Alison Flatau, and S. C. Liu. Review Paper: Health Monitoring of Civil Infrastructure. *Structural Health Monitoring*, 2(3):257–267, 2003.

[16] Mirza Mansoor Baig and Hamid Gholamhosseini. Smart health monitoring systems: An overview of design and modeling. *Journal of Medical Systems*, 37(2), 2013.

[17] Joseph Randall Moorman, Waldemar A. Carlo, John Kattwinkel, Robert L. Schelonka, Peter J. Porcelli, Christina T. Navarrete, Eduardo Bancalari, Judy L. Aschner, Marshall Whit Walker, Jose A. Perez, Charles Palmer, George J. Stukenborg, Douglas E. Lake, and Thomas Michael O'Shea. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: A randomized trial. *Journal of Pediatrics*, 159(6):900–906.e1, 2011.

[18] William A. Knaus, Douglas P. Wagner, Elizabeth A. Draper, Jack E. Zimmerman, Marilyn Bergner, Paulo G. Bastos, Carl A. Sirio, Donald J. Murphy, Ted Lotring, and Anne Damiano. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, dec 1991.

[19] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia Fu Chang, and Lei Hua. Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4):2431–2448, aug 2012.

[20] Riccardo Bellazzi, Fulvia Ferrazzi, and Lucia Sacchi. Predictive data mining in clinical medicine: A focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, sep 2011.

[21] Yoav Freund and Robert E Schapire. No Title. *Machine Learning*, 37(3):277–296, 1999.

[22] D Sculley. Online Active Learning Methods for Fast Label-Efficient Spam Filtering. In *Fourth Conference on Email and Anti-Spam*, pages 1–8, 2007.

[23] Erik P Vargo and Randy Cogill. Expectation-maximization for Bayes-adaptive POMDPs. *Journal of the Operational Research Society*, 66(10):1605–1623, oct 2015.

[24] Luigi De Gennaro, Cristina Marzano, Fabiana Fratello, Fabio Moroni, Maria Concetta Pellicciari, Fabio Ferlazzo, Stefania Costa, Alessandro Couyoumdjian, Giuseppe Curcio, Emilia Sforza, Alain Malafosse, Luca A. Finelli, Patrizio Pasqualetti, Michele Ferrara, Mario Bertini, and Paolo Maria Rossini. The electroencephalographic fingerprint of sleep is genetically determined: A twin study. *Annals of Neurology*, 64(4):455–460, oct 2008.

[25] *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5*. American Psychiatric Publishing, Washington, D.C, 5 edition edition, May 2013.

[26] Ambulatory heart rate changes in patients with panic attacks. *American Journal of Psychiatry*, 143(4):478–482, April 1986.

[27] R. Hoehn-Saric, D. R. McLeod, and W. D. Zimmerli. Psychophysiological response patterns in panic disorder. *Acta Psychiatrica Scandinavica*, 83(1):4–11, January 1991.

[28] W. T. Roth, A. Ehlers, C. B. Taylor, J. Margraf, and W. S. Agras. Skin conductance habituation in panic disorder patients. *Biological Psychiatry*, 27(11):1231–1243, June 1990.

[29] Alicia E. Meuret, Kamila S. White, Thomas Ritz, Walton T. Roth, Stefan G. Hofmann, and Timothy A. Brown. Panic attack symptom dimensions and their relationship to illness characteristics in panic disorder. *Journal of Psychiatric Research*, 40(6):520–527, September 2006.

[30] Eyal Dassau, B. Wayne Bequette, Bruce A. Buckingham, and Francis J. Doyle. Detection of a Meal Using Continuous Glucose Monitoring Implications for an artificial -cell. *Diabetes Care*, 31(2):295–300, February 2008.

[31] William J. Whelan, Danielle Hollar, Arthur Agatston, Hannah J. Dodson, and Dimitri S. Tahal. The glycemic response is a personal attribute. *IUBMB Life*, 62(8):637–641, jul 2010.

[32] John P. Kirwan, Thomas P. J. Solomon, Daniel M. Wojta, Myrlene A. Staten, and John O. Holloszy. Effects of 7 days of exercise training on insulin sensitivity and responsiveness in type 2 diabetes mellitus. *American Journal of Physiology. Endocrinology and Metabolism*, 297(1):E151–156, July 2009.

[33] David H. Spodick, Padma Raju, Richard L. Bishop, and Robert D. Rifkin. Operational definition of normal sinus heart rate. *The American Journal of Cardiology*, 69(14):1245–1246, 1992.

[34] P. Palatini. Need for a Revision of the Normal Limits of Resting Heart Rate. *Hypertension*, 33(2):622–625, feb 1999.

[35] M Lodén. Biophysical properties of dry atopic and normal skin with special reference to effects of skin care products. *Acta dermato-venereologica. Supplementum*, 192:1–48, 1995.

[36] David Cella, Susan Yount, Nan Rothrock, Richard Gershon, Karon Cook, Bryce Reeve, Deborah Ader, James F Fries, Bonnie Bruce, Mattias Rose, and PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical care*, 45(5 Suppl 1):S3–S11, may 2007.

[37] E. Basch, A. M. Deal, M. G. Kris, H. I. Scher, C. A. Hudis, P. Sabbatini, L. Rogak, A. V. Bennett, A. C. Dueck, T. M. Atkinson, J. F. Chou, D. Dulko, L. Sit, A. Barz, P. Novotny, M. Fruscione, J. A. Sloan, and D. Schrag. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *Journal of Clinical Oncology*, 34(6):JCO.2015.63.0830–, 2015.

[38] M S T Jock Murray. *Multiple sclerosis: the history of a disease*. Demos medical publishing, 2004.

[39] Richard Katz, Tsega Mesfin, and Karen Barr. Lessons From a Community-Based mHealth Diabetes Self-Management Program: It's Not Just About the Cell Phone. *Journal of Health Communication*, 17(sup1):67–72, 2012.

[40] James E. Aikens, Kara Zivin, Ranak Trivedi, and John D. Piette. Diabetes self-management support using mHealth and enhanced informal caregiving. *Journal of Diabetes and its Complications*, 28(2):171–176, 2014.

[41] Lora E. Burke, Mindi A. Styn, Susan M. Sereika, Molly B. Conroy, Lei Ye, Karen Glanz, Mary Ann Sevick, and Linda J. Ewing. Using mHealth Technology to Enhance Self-Monitoring for Weight Loss: A Randomized Trial. *American Journal of Preventive Medicine*, 43(1):20–26, 2012.

[42] Jasmine K. Zia, Jessica Schroeder, Sean A. Munson, James Fogarty, Linda Nguyen, Pamela Barney, Margaret M. Heitkemper, and Uri Ladabaum. Feasibility and Usability Pilot Study of a Novel Irritable Bowel Syndrome Food and Gastrointestinal Symptom Journal Smartphone App. *Clinical and Translational Gastroenterology*, 7(3):In Press, 2015.

[43] Lora E. Burke, Jun Ma, Kristen M. J. Azar, Gary G. Bennett, Eric D. Peterson, Yaguang Zheng, William Riley, Janna Stephens, Svati H. Shah, Brian Suffoletto, Tanya N. Turan, Bonnie Spring, Julia Steinberger, and Charlene C. Quinn. Current Science on Consumer Use of Mobile Health for Cardiovascular Disease Prevention A Scientific Statement From the American Heart Association. *Circulation*, 132(12):1157–1213, September 2015.

[44] Abby C. King, David K. Ahn, Brian M. Oliveira, Audie A. Atienza, Cynthia M. Castro, and Christopher D. Gardner. Promoting Physical Activity Through Hand-Held Computer Technology. *American Journal of Preventive Medicine*, 34(2):138–142, February 2008.

[45] Heleen Spittaels, I. De Bourdeaudhuij, J. Brug, and C. Vandelanotte. Effectiveness of an online computer-tailored physical activity intervention in a real-life setting. *Health Education Research*, 22(3):385–396, June 2007.

[46] Brianna S. Fjeldsoe, Yvette D. Miller, and Alison L. Marshall. MobileMums: A Randomized Controlled Trial of an SMS-Based Physical Activity Intervention. *Annals of Behavioral Medicine*, 39(2):101–111, February 2010.

[47] Minna Aittasalo, Marjo Rinne, Matti Pasanen, Katriina Kukkonen-Harjula, and Tommi Vasankari. Promoting walking among office employees evaluation of a randomized controlled intervention with pedometers and e-mail messages. *BMC Public Health*, 12(1):403, June 2012.

[48] Bang Hyun Kim and Karen Glanz. Text Messaging to Motivate Walking in Older African Americans: A Randomized Controlled Trial. *American Journal of Preventive Medicine*, 44(1):71–75, January 2013.

[49] Robert D. Reid, Louise I. Morrin, Louise J. Beaton, Sophia Papadakis, Jana Kocourek, Lisa McDonnell, Monika E. Slovinec D'Angelo, Heather Tulloch, Neville Suskin, Karen Unsworth, Chris Blanchard, and Andrew L. Pipe. Randomized trial of an internet-based computer-tailored expert system for physical activity in patients with heart disease. *European Journal of Preventive Cardiology*, 19(6):1357–1364, December 2012.

[50] Timothy W. Bickmore, Rebecca A. Silliman, Kerrie Nelson, Debbie M. Cheng, Michael Winter, Lori Henault, and Michael K. Paasche-Orlow. A Randomized Controlled Trial of an Automated Exercise Coach for Older Adults. *Journal of the American Geriatrics Society*, 61(10):1676–1683, October 2013.

[51] Julie B. Wang, Lisa A. Cadmus-Bertram, Loki Natarajan, Martha M. White, Hala Madanat, Jeanne F. Nichols, Guadalupe X. Ayala, and John P. Pierce. Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial. *Telemedicine and E-Health*, 21(10):782–792, October 2015. WOS:000362191800004.

[52] Hee-Seung Kim. A randomized controlled trial of a nurse short-message service by cellular phone for people with diabetes. *International Journal of Nursing Studies*, 44(5):687–692, July 2007.

[53] Kun-Ho Yoon and Hee-Seung Kim. A short message service by cellular phone in type 2 diabetic patients for 12 months. *Diabetes Research and Clinical Practice*, 79(2):256–261, February 2008.

[54] Zubaida Faridi, Lauren Liberti, Kerem Shuval, Veronika Northrup, Ather Ali, and David L. Katz. Evaluating the impact of mobile telephone technology on type 2 diabetic patients self-management: the NICHE pilot study. *Journal of Evaluation in Clinical Practice*, 14(3):465–469, June 2008.

[55] H. J. Yoo, M. S. Park, T. N. Kim, S. J. Yang, G. J. Cho, T. G. Hwang, S. H. Baik, D. S. Choi, G. H. Park, and K. M. Choi. A Ubiquitous Chronic Disease Care system using cellular phones and the internet. *Diabetic Medicine*, 26(6):628–635, June 2009.

[56] Robert SH Istepanian, Karima Zitouni, Diane Harry, Niva Moutosammy, Ala Sungoor, Bee Tang, and Kenneth A. Earle. Evaluation of a mobile phone telemonitoring system for glycaemic control in patients with diabetes. *Journal of Telemedicine and Telecare*, 15(3):125–128, April 2009.

[57] Charlene C. Quinn, Michelle D. Shardell, Michael L. Terrin, Erik A. Barr, Shoshana H. Ballew, and Ann L. Gruber-Baldini. Cluster-Randomized Trial of a Mobile Phone Personalized Behavioral Intervention for Blood Glucose Control. *Diabetes Care*, 34(9):1934–1942, September 2011.

[58] Karin Gruber, Patrick J. Moran, Walton T. Roth, and C. Barr Taylor. Computer-assisted cognitive behavioral group therapy for social phobia. *Behavior Therapy*, 32(1):155–165, 2001.

[59] Michelle G. Newman, Justin Kenardy, Steve Herman, and C. Barr. Comparison of palmtop-computer-assisted brief cognitive-behavioral treatment to cognitive-behavioral treatment for panic disorder. *Journal of Consulting and Clinical Psychology*, 65(1):178–183, 1997.

[60] Justin A. Kenardy, G. T, Derek W. Johnston, Michelle G. Newman, Aileen Thomson, and C. Barr. A Comparison of Delivery Methods of Cognitive-Behavioral Therapy for Panic Disorder: An International Multicenter Trial. *Journal of Consulting and Clinical Psychology*, 71(6):1068–1075, 2003.

[61] Susan J. Wenze, Michael F. Armey, and Ivan W. Miller. Feasibility and Acceptability of a Mobile Intervention to Improve Treatment Adherence in Bipolar Disorder A Pilot Study. *Behavior Modification*, 38(4):497–515, July 2014.

[62] Benjamin Ehrenreich, Bryan Righter, Di Andra Rocke, Lisa Dixon, and Seth Himelhoch. Are Mobile Phones and Handheld Computers Being Used to Enhance Delivery of Psychiatric Treatment?: A Systematic Review. *The Journal of Nervous and Mental Disease*, 199(11):886–891, November 2011.

[63] Matthew Price, Erica K. Yuen, Elizabeth M. Goetter, James D. Herbert, Evan M. Forman, Ron Acierno, and Kenneth J. Ruggiero. mHealth: A Mechanism to Deliver More Accessible, More Effective Mental Health Care. *Clinical Psychology & Psychotherapy*, 21(5):427–436, September 2014.

[64] Gede Pramana, Bambang Parmanto, Philip C. Kendall, and Jennifer S. Silk. The SmartCAT: An m-Health Platform for Ecological Momentary Intervention in Child Anxiety Treatment. *Telemedicine and e-Health*, 20(5):419–427, February 2014.

[65] Judit Bort-Roig, Nicholas D. Gilson, Anna Puig-Ribera, Ruth S. Contreras, and Stewart G. Trost. Measuring and Influencing Physical Activity with Smartphone Technology: A Systematic Review. *Sports Medicine*, 44(5):671–686, February 2014.

[66] X. Liang, Q. Wang, X. Yang, J. Cao, J. Chen, X. Mo, J. Huang, L. Wang, and D. Gu. Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis. *Diabetic Medicine*, 28(4):455–463, April 2011.

[67] Kingshuk Pal, Sophie V Eastwood, Susan Michie, Andrew J Farmer, Maria L Barnard, Richard Peacock, Bindie Wood, Joni D Inniss, and Elizabeth Murray. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. In *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd, March 2013.

[68] Kristin E. Heron and Joshua M. Smyth. Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, 15(1):1–39, February 2010.

[69] Fred D Lublin, Stephen C Reingold, Jeffrey A Cohen, Gary R Cutter, Per Soelberg Sørensen, Alan J Thompson, Jerry S Wolinsky, Laura J Balcer, Brenda Banwell, Frederik Barkhof, Bruce Bebo, Peter A Calabresi, Michel Clanet, Giancarlo Comi, Robert J Fox, Mark S Freedman, Andrew D Goodman, Matilde Inglese, Ludwig Kappos, Bernd C Kieseier, John A Lincoln, Catherine Lubetzki, Aaron E Miller, Xavier Montalban, Paul W O'Connor, John Petkau, Carlo Pozzilli, Richard A Rudick, Maria Pia Sormani, Olaf Stüve, Emmanuelle Waubant, and Chris H Polman. Defining the clinical course of multiple sclerosis: The 2013 revisions, jul 2014.

[70] Helen Tremlett, Donald Paty, and Virginia Devonshire. Disability progression in multiple sclerosis is slower than previously reported. *Neurology*, 66(2):172–177, January 2006.

[71] J. F. Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1452, nov 1983.

[72] Matthew H. Sutliff. Contribution of impaired mobility to patient burden in multiple sclerosis. *Current Medical Research and Opinion*, 26(1):109–119, January 2010.

[73] Dr Nicholas G. LaRocca. Impact of Walking Impairment in Multiple Sclerosis. *The Patient: Patient-Centered Outcomes Research*, 4(3):189–201, August 2012.

[74] Andrew D. Goodman, Theodore R. Brown, Keith R. Edwards, Lauren B. Krupp, Randall T Schapiro, Ron Cohen, Lawrence N. Marinucci, Andrew R. Blight, and on behalf of the MSF204 Investigators. A phase 3 trial of extended release oral dalfampridine in multiple sclerosis. *Annals of Neurology*, 68(4):494–502, October 2010.

[75] Andrew D. Goodman, Francois Bethoux, Theodore R. Brown, Randall T. Schapiro, Ron Cohen, Lawrence N. Marinucci, Herbert R. Henney, Andrew R. Blight, M. Agius, B. G. W. Arnason, F. A. Bethoux, C. T. Bever, J. D. Bowen, T. R. Brown, D. W. Dietrich, K. Edwards, M. S. Freedman, M. Freedman, N. J. Kachuck, M. D. Kaufman, M. Keilson, O. Khan, L. B. Krupp, T. P. Leist, J. W. Lindsey, F. D. Lublin, M. K. Mass, D. Mattson, D. McGowan, R. Naismith, C. OConnell, J. J. Oger, H. Panitch, M. A. Picone, K. W. Rammohan, R. T. Schapiro, S. R. Schwid,

T. Scott, C. Short, B. W. Thrower, T. L. Vollmer, A. Camac, J. A. Cooper, W. F. Chumley, A. Cross, R. T. Dunnigan, J. S. Gitt, M. Hillen, D. R. Jeffrey, B. O. Khatri, K. Kresa-Reahl, S. Moon, J. Preiningerova, M. Tullman, B. Weinstock-Guttman, and D. R. Wynn. Long-term safety and efficacy of dalfampridine for walking impairment in patients with multiple sclerosis: Results of open-label extensions of two Phase 3 clinical trials. *Multiple Sclerosis Journal*, 21(10):1322–1331, September 2015.

[76] J S Fischer, R a Rudick, G R Cutter, and S C Reingold. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 5(4):244–250, 1999.

[77] Myla D Goldman, Ruth Ann Marrie, and Jeffrey a Cohen. Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 14(3):383–390, 2008.

[78] Robert W Motl, Yoojin Suh, Swathi Balantrapu, Brian M Sandroff, Jacob J Sosnoff, John Pula, Myla D Goldman, and Bo Fernhall. Evidence for the different physiological significance of the 6- and 2-minute walk tests in multiple sclerosis. *BMC Neurology*, 12:6, 2012.

[79] Matthew M. Engelhard, Sriram Raju Dandu, Stephen D. Patek, John C. Lach, and Myla D. Goldman. Quantifying six-minute walk induced gait deterioration with inertial sensors in multiple sclerosis subjects. *Gait & Posture*, 49:340–345, sep 2016.

[80] D Gijbels, B O Eijnde, and P Feys. Comparison of the 2- and 6-minute walk test in multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 17(10):1269–72, oct 2011.

[81] Robert W. Motl, Edward McAuley, Erin M. Snook, and Rachael C. Gliottoni. Physical activity and quality of life in multiple sclerosis: intermediary roles of disability, fatigue, mood, pain, self-efficacy and social support. *Psychology, health & medicine*, 14(1):111–24, 2009.

[82] Robert W Motl, Myla D Goldman, and Ralph H B Benedict. Walking impairment in patients with multiple sclerosis: exercise training as a treatment option. *Neuropsychiatric Disease and Treatment*, 6:767–774, 2010.

[83] Robert W. Motl, Deirdre Dlugonski, Yoojin Suh, Madeline Weikert, Bo Fernhall, and Myla Goldman. Accelerometry and its association with objective markers of walking limitations in ambulatory adults with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation*, 91(12):1942–1947, 2010.

[84] Carolyn E. Schwartz, Armon Ayandeh, and Robert W. Motl. Investigating the minimal important difference in ambulation in multiple sclerosis: A disconnect between performance-based and patient-reported outcomes? *Journal of the Neurological Sciences*, 347(1-2):268–274, 2014.

[85] Vitali Witowski, Ronja Foraita, Yannis Pitsiladis, Iris Pigeot, and Norman Wirsik. Using hidden Markov models to improve quantifying physical activity in accelerometer data - A simulation study. *PLoS ONE*, 9(12):e114089, dec 2014.

[86] Robert L. Waters and Sara Mulroy. The energy expenditure of normal and pathologic gait, 1999.

[87] Stamatis Agiovlasitis and Robert W. Motl. Step-rate thresholds for physical activity intensity in persons with multiple sclerosis. *Adapted Physical Activity Quarterly*, 31(1):4–18, 2014.

[88] G Godin and R J Shephard. A simple method to assess exercise behavior in the community, 1985.

[89] C. E. Schwartz, T. Vollmer, H. Lee, and the North American Research Consortium on Multiple Sclerosis Outcomes Study Group. Reliability and validity of two self-report measures of impairment and disability for MS. *Neurology*, 52(1):63–63, jan 1999.

[90] J. C. Hobart, A. Riazi, D. L. Lamping, R. Fitzpatrick, and A. J. Thompson. Measuring the impact of MS on walking ability: The 12-Item MS Walking Scale (MSWS-12). *Neurology*, 60(1):31–36, jan 2003.

[91] Bernd C Kieseier and Carlo Pozzilli. Assessing walking disability in multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 18(7):914–924, jul 2012.

[92] A. D. Goodman, T. R. Brown, J. A. Cohen, L. B. Krupp, R. Schapiro, S. R. Schwid, R. Cohen, L. N. Marinucci, A. R. Blight, and For the Fampridine MS-F202 Study Group. Dose comparison trial of sustained-release fampridine in multiple sclerosis. *Neurology*, 71(15):1134–1141, October 2008.

[93] Andrew D Goodman, Theodore R Brown, Lauren B Krupp, Randall T Schapiro, Steven R Schwid, Ron Cohen, Lawrence N Marinucci, and Andrew R Blight. Sustained-release oral fampridine in multiple sclerosis: a randomised, double-blind, controlled trial. *The Lancet*, 373(9665):732–738, March 2009.

[94] Sarah Tyson and Louise Connell. The psychometric properties and clinical utility of measures of walking and mobility in neurological conditions: a systematic review. *Clinical Rehabilitation*, 23(11):1018–1033, November 2009.

[95] Lidwine Brigitta Mokkink, Francisca Galindo-Garre, and Bernard Mj Uitdehaag. Evaluation of the Multiple Sclerosis Walking Scale-12 (MSWS-12) in a Dutch sample: Application of item response theory. *Multiple Sclerosis Journal*, page 1352458516630821, feb 2016.

[96] Matthew M. Engelhard, Karen M. Schmidt, Casey E. Engel, J. Nicholas Brenton, Stephen D. Patek, and Myla D. Goldman. The e-MSWS-12: Improving the Multiple Sclerosis Walking Scale using Item Response Theory. *Quality of Life Research*, pages 1–10, jun 2016.

[97] P Ritvo, J S Fischer, D M Miller, H Andrews, D W Paty, and N G LaRocca. MSQLI-Multiple Sclerosis Quality of Life Inventory. *A user's manual. New York: National MS Society*, 1997.

[98] R. J. Mills, C. A. Young, J. F. Pallant, and A. Tennant. Rasch analysis of the Modified Fatigue Impact Scale (MFIS) in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(9):1049–1051, September 2010.

[99] Robert J. Gatchel, Jon D. Lurie, and Tom G. Mayer. Minimal Clinically Important Difference. *Spine*, 35(19):1739–1743, sep 2010.

[100] Rocco Haase, Thorsten Schultheiss, Raimar Kempcke, Katja Thomas, and Tjalf Ziemssen. Modern communication technology skills of patients with multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 19(9):1240–1, 2013.

[101] Cinzia Colombo, Graziella Filippini, Anneliese Synnot, Sophie Hill, Roberta Guglielmino, Silvia Traversa, Paolo Confalonieri, Paola Mosconi, and Irene Tramacere. Development and assessment of a website presenting evidence-based information for people with multiple sclerosis: the IN-DEEP project. *BMC Neurology*, 16(1):30, 2016.

[102] Chris H Polman, Stephen C Reingold, Brenda Banwell, Michel Clanet, Jeffrey A Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, Fred D Lublin, Xavier Montalban, Paul O'Connor, Magnhild Sandberg-Wollheim, Alan J Thompson, Emmanuelle Waubant, Brian Weinshenker, and Jerry S Wolinsky. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of neurology*, 69(2):292–302, feb 2011.

[103] MA Rizzo, OC Hadjimichael, J Preiningerova, and TL Vollmer. Prevalence and treatment of spasticity reported by multiple sclerosis patients. *Multiple Sclerosis*, 10(5):589–595, oct 2004.

[104] Rocco Haase, Thorsten Schultheiss, Raimar Kempcke, Katja Thomas, and Tjalf Ziemssen. Use and acceptance of electronic communication by patients with multiple sclerosis: a multicenter questionnaire study. *Journal of medical Internet research*, 14(5):e135, 2012.

[105] Lora E. Burke, Jun Ma, Kristen M J Azar, Gary G. Bennett, Eric D. Peterson, Yaguang Zheng, William Riley, Janna Stephens, Svati H. Shah, Brian Suffoletto, Tanya N. Turan, Bonnie Spring, Julia Steinberger, and Charlene C. Quinn. *Current Science on Consumer Use of Mobile Health for Cardiovascular Disease Prevention: A Scientific Statement from the American Heart Association*, volume 132. 2015.

[106] R. W. Motl, L. Pilutti, B. M. Sandroff, D. Dlugonski, J. J. Sosnoff, and J. H. Pula. Accelerometry as a measure of walking behavior in multiple sclerosis. *Acta Neurologica Scandinavica*, 2013.

[107] Robert W Motl, Deirdre Dlugonski, Yoojin Suh, Madeline Weikert, Bo Fernhall, and Myla Goldman. Accelerometry and its association with objective markers of walking limitations in ambulatory adults with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation*, 91(12):1942–1947, December 2010.

[108] Billie Giles-Corti and Robert J Donovan. The relative influence of individual, social and physical environment determinants of physical activity. *Social science & medicine*, 54(12):1793–1812, 2002.

[109] Nancy Humpel. Environmental factors associated with adults' participation in physical activity A review. *American Journal of Preventive Medicine*, 22(3):188–199, 2002.

[110] Madeline Weikert, Yoojin Suh, Abbi Lane, Brian Sandroff, Deirdre Dlugonski, Bo Fernhall, and Robert W. Motl. Accelerometry is associated with walking mobility, not physical activity, in persons with multiple sclerosis. *Medical Engineering and Physics*, 34(5):590–597, 2012.

[111] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, May 2004.

[112] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, February 1978.

[113] Lawrence R. Rabiner, A Rosenberg, and S Levinson. Considerations in Dynamic Time Warping Algorithms forDiscrete Word Recognition. {*IEEE*} *Transactions on Acoustics, Speech and SignalProcessing*, 26(S1):575–582, 1978.

[114] Eamonn Keogh, Li Wei, Xiaopeng Xi, Michail Vlachos, Sang-Hee Lee, and Pavlos Protopapas. Supporting Exact Indexing of Arbitrarily Rotated Shapes and Periodic Time Series Under Euclidean and Warping Distance Measures. *The VLDB Journal*, 18(3):611–630, June 2009.

[115] A. Akl and S. Valaee. Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, #x00026; compressive sensing. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2270–2273, March 2010.

[116] R. Muscillo, S. Conforto, M. Schmid, P. Caselli, and T. D'Alessio. Classification of Motor Activities through Derivative Dynamic Time Warping applied on Accelerometer Data. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007*, pages 4930–4933, August 2007.

[117] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H.A. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, volume 2, pages ii/973–ii/976 Vol. 2, March 2005.

[118] Liu Rong, Zhou Jianzhong, Liu Ming, and Hou Xiangfeng. A Wearable Acceleration Sensor System for Gait Recognition. In *2nd IEEE Conference on Industrial Electronics and Applications, 2007. ICIEA 2007*, pages 2654–2659, May 2007.

[119] E.G. Caiani, A. Porta, G. Baselli, M. Turiel, S. Muzzupappa, F. Pieruzzi, C. Crema, A. Malliani, and S. Cerutti. Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. *Computers in Cardiology 1998*, 25:73 – 76, 1998.

[120] Matthew M Engelhard, Sriram Raju Dandu, John C Lach, Myla D Goldman, James Q Miller, Ms Clinic, and Stephen D Patek. Toward Detection and Monitoring of Gait Pathology using Inertial Sensors under Rotation, Scale, and Offset Invariant Dynamic Time Warping.

[121] Eamonn J. Keogh and Michael J. Pazzani. Derivative Dynamic Time Warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–11. Society for Industrial and Applied Mathematics, Philadelphia, PA, apr 2001.

[122] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 262–270, 2012.

[123] Tsu-Wei Chen, Meena Abdelmaseeh, and Daniel Stashuk. Affine and Regional Dynamic Time Warpng. *arXiv:1505.06531 [cs]*, May 2015. arXiv: 1505.06531.

[124] Yu Qiao and Makoto Yasuhara. Affine Invariant Dynamic Time Warping and its Application to Online Rotated Handwriting Recognition. In *18th International Conference on Pattern Recognition, 2006. ICPR 2006*, volume 2, pages 905–908, August 2006.

[125] P. Bours and R. Shrestha. Eigensteps: A giant leap for gait recognition. In *2010 2nd International Workshop on Security and Communication Networks (IWSCN)*, pages 1–6, May 2010.

[126] Martina Mancini, Arash Salarian, Patricia Carlson-Kuhta, Cris Zampieri, Laurie King, Lorenzo Chiari, and Fay B. Horak. ISway: a sensitive, valid and reliable measure of postural control. *Journal of NeuroEngineering and Rehabilitation*, 9(1):59, August 2012.

[127] A Salarian, F.B. Horak, C. Zampieri, P. Carlson-Kuhta, J.G. Nutt, and K. Aminian. iTUG, a Sensitive and Reliable Measure of Mobility. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(3):303–310, June 2010.

[128] Rebecca I. Spain, Martina Mancini, Fay B. Horak, and Dennis Bourdette. Body-worn sensors capture variability, but not decline, of gait and balance measures in multiple sclerosis over 18 months. *Gait & Posture*, 39(3):958–964, March 2014.

[129] Jiaqi Gong, J. Lach, Yanjun Qi, and M.D. Goldman. Causal analysis of inertial body sensors for enhancing gait assessment separability towards multiple sclerosis diagnosis. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6, June 2015.

[130] Chieh Chien, James Yan Xu, Hua-i Chang, Xiaoxu Wu, and Greg J. Pottie. *Model Construction for Human Motion Classification Using Inertial Sensors*.

[131] Jiaqi Gong, Philip Asare, John Lach, and Yanjun Qi. Piecewise Linear Dynamical Model for Actions Clustering from Inertial Body Sensors with Considerations of Human Factors. In *Proceedings of the 9th International Conference on Body Area Networks*, BodyNets '14, pages 90–96, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[132] N.V. Boulgouris, K.N. Plataniotis, and D. Hatzinakos. Gait recognition using dynamic time warping. *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, pages 263–266, 2004.

[133] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3-D rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5-6):272–290, March 1997.

[134] Chris H. Polman, Stephen C. Reingold, Brenda Banwell, Michel Clanet, Jeffrey A. Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, Fred D. Lublin, Xavier Montalban, Paul O'Connor, Magnhild Sandberg-Wollheim, Alan J. Thompson, Emmanuelle Waubant, Brian Weinshenker, and Jerry S. Wolinsky. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, 69(2):292–302, February 2011.

[135] Sriram Raju Dandu, Matthew M. Engelhard, Myla D. Goldman, and John Lach. Determining physiological significance of inertial gait features in multiple sclerosis. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 266–271. IEEE, jun 2016.

[136] David Cella, Susan Yount, Nan Rothrock, Richard Gershon, Karon Cook, Bryce Reeve, Deborah Ader, James F. Fries, Bonnie Bruce, and Mattias Rose. The Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical care*, 45(5 Suppl 1):S3–S11, May 2007.

[137] Chih-Hung Chang and Bryce B. Reeve. Item Response Theory and its Applications to Patient-Reported Outcomes Measurement. *Evaluation and the Health Professions*, 28(3):264–282, sep 2005.

[138] Geoff N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, jun 1982.

[139] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2):100, 1969.

[140] Rafael Jaime De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.

[141] Susan E. Embretson and Steven Paul Reise. *Item Response Theory for Psychologists*. L. Erlbaum Associates, 2000.

[142] Chih-Hung Chang and Bryce B. Reeve. Item Response Theory and its Applications to Patient-Reported Outcomes Measurement. *Evaluation & the Health Professions*, 28(3):264–282, September 2005.

[143] Andrew D. Martin and Kevin M. Quinn. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, 10(2):134–153, may 2002.

[144] Xiaojing Wang, James O. Berger, and Donald S. Burdick. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1):126–153, 2013.

[145] Dalton F. Andrade and Heliton R. Tavares. Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, 95(1):1–22, 2005.

[146] S. J. Cano, L. E. Barrett, J. P. Zajicek, and J. C. Hobart. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Multiple Sclerosis Journal*, 17(2):214–222, February 2011.

[147] Craig A. Velozo, Ying Wang, Leigh Lehman, and Jia-Hwa Wang. Utilizing Rasch measurement models to develop a computer adaptive self-report of walking, climbing, and running. *Disability and Rehabilitation*, 30(6):458–467, January 2008.

[148] Pierre Michel, Pascal Auquier, Karine Baumstarck, Jean Pelletier, Anderson Loundou, Badih Ghattas, and Laurent Boyer. Development of a cross-cultural item bank for measuring quality of life related to mental health in multiple sclerosis patients. *Quality of Life Research*, 24(9):2261–2271, sep 2015.

[149] Pierre Michel, Karine Baumstarck, Badih Ghattas, Jean Pelletier, Anderson Loundou, Mohamed Boucekine, Pascal Auquier, and Laurent Boyer. A Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire Developed and Validated for Multiple Sclerosis: The MusiQoL-MCAT. *Medicine*, 95(14):e3068, apr 2016.

[150] Karon F. Cook, Alyssa M. Bamer, Dagmar Amtmann, Ivan R. Molton, and Mark P. Jensen. Six patient-reported outcome measurement information system short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. *Archives of Physical Medicine and Rehabilitation*, 93(7):1289–1291, 2012.

[151] André De Champlain and Marc E. Gessaroli. Assessing the Dimensionality of Item Response Matrices with Small Sample Sizes and Short Test Lengths. *Applied Measurement in Education*, 11(3):231–253, 1998.

[152] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2015. R package version 1.5.8.

[153] Sik-Yum Lee, Wai-Yin Poon, and P. M. Bentler. A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, 55(1):45–51, March 1990.

[154] Eiji Muraki. A Generalized Partial Credit Model: Application of an Em Algorithm. *ETS Research Report Series*, 1992(1):i–30, June 1992.

[155] R. Philip Chalmers. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6):1–29, 2012.

[156] Fritz Drasgow, Michael V. Levine, and Esther A. Williams. Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1):67–86, May 1985.

[157] Taehoon Kang and Troy T. Chen. *An Investigation of the Performance of the Generalized S-X[superscript 2] Item-Fit Index for Polytomous IRT Models. ACT Research Report Series, 2007-1.* ACT, Inc, June 2007.

[158] Seock-Ho Kim and Allan S. Cohen. Detection of Differential Item Functioning Under the Graded Response Model With the Likelihood Ratio Test. *Applied Psychological Measurement*, 22(4):345–355, December 1998.

[159] Matthew M Engelhard, Karen M Schmidt, Casey E Engel, J Nicholas Brenton, Stephen D Patek, and Myla D Goldman. The e-MSWS-12: Improving the Multiple Sclerosis Walking Scale using Item Response Theory. *Quality of Life Research*, 2016.

[160] C. McGuigan and M. Hutchinson. Confirming the validity and responsiveness of the Multiple Sclerosis Walking Scale-12 (MSWS-12). *Neurology*, 62(11):2103–2105, June 2004.

[161] Maria Orlando and David Thissen. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 24(1):50–64, March 2000.

[162] Antonio Scalfari, Anneke Neuhaus, Alexandra Degenhardt, George P. Rice, Paolo A Muraro, Martin Daumer, and George C. Ebers. The natural history of multiple sclerosis, a geographically based study 10: Relapses and long-term disability. *Brain*, 133(7):1914–1929, jul 2010.

[163] Ilse Baert, Jennifer Freeman, Tori Smedal, Ulrik Dalgas, Anders Romberg, Alon Kalron, Helen Conyers, Iratxe Elorriaga, Benoit Gebara, Johanna Gumse, Adnan Heric, Ellen Jensen, Kari Jones, Kathy Knuts, Benoît Maertens de Noordhout, Andrej Martic, Britt Normann, Bert O. Eijnde, Kamila Rasova, Carmen Santoyo Medina, Veronik Truyens, Inez Wens, and Peter Feys. Responsiveness and clinically meaningful improvement, according to disability level, of five walking measures after rehabilitation in multiple sclerosis: a European multicenter study. *Neurorehabilitation and neural repair*, 28(7):621–31, sep 2014.

[164] Emmanuelle Leray, Jacqueline Yaouanq, Emmanuelle Le Page, Marc Coustans, David Laplaud, Joël Oger, and Gilles Edan. Evidence for a two-stage disability progression in multiple sclerosis. *Brain*, 133(7):1900–1913, jul 2010.

[165] Helen Tremlett, Donald Paty, and Virginia Devonshire. Disability progression in multiple sclerosis is slower than previously reported, jan 2006.

[166] Helen Tremlett, Yinshan Zhao, Peter Rieckmann, and Michael Hutchinson. New perspectives in the natural history of multiple sclerosis, jun 2010.

[167] Jaana Paltamaa, Taneli Sarasoja, J Wikström, and E Mälkiä. Physical functioning in multiples sclerosis: A population-based study in central Finland. *Journal of Rehabilitation Medicine*, 38(6):339–345, 2006.

[168] Marcus Koch, Jop Mostert, Dorothea Heersema, and Jacques De Keyser. Progression in multiple sclerosis: Further evidence of an age dependent process. *Journal of the Neurological Sciences*, 255(1-2):35–41, 2007.

[169] Mr Harwell and Je Janosky. AN EMPIRICAL-STUDY OF THE EFFECTS OF SMALL DATASETS AND VARYING PRIOR VARIANCES ON ITEM PARAMETER-ESTIMATION IN BILOG. *Appl. Psychol. Meas.*, 15(3):279–291, sep 1991.

[170] Maria Orlando and Grant N Marshall. Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychological assessment*, 14(1):50–9, mar 2002.

[171] C. L. Hulin, R. I. Lissak, and F. Drasgow. Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Applied Psychological Measurement*, 6(3):249–260, jun 1982.

[172] David Thissen and Howard Wainer. Some standard errors in item response theory. *Psychometrika*, 47(4):397–412, dec 1982.

[173] Robert W. Motl and Erin M. Snook. Confirmation and extension of the validity of the Multiple Sclerosis Walking Scale-12 (MSWS-12). *Journal of the Neurological Sciences*, 268(12):69–73, May 2008.

[174] Molly C Cincotta, Matthew M Engelhard, Makela Stankey, and Myla D Goldman. Fatigue and fluid hydration status in multiple sclerosis: A hypothesis. *Multiple Sclerosis Journal*, 22(11):1352458516663854, oct 2016.

[175] L. Kleinman, M. W. Zodet, Z. Hakim, J. Aledort, C. Barker, K. Chan, L. Krupp, and D. Revicki. Psychometric evaluation of the fatigue severity scale for use in chronic hepatitis C. *Quality of Life Research*, 9(5):499–508, 2000.

[176] Myla D Goldman, Robert W Motl, and Richard A Rudick. Possible clinical outcome measures for clinical trials in patients with multiple sclerosis. *Ther Adv Neurol Disord*, 3(4):229–239, jul 2010.

[177] Håkan Nero, Martin Benka Wallén, Erika Franzén, Agneta Ståhle, and Maria Hagströmer. Accelerometer cut points for physical activity assessment of older adults with Parkinson's disease. *PLoS ONE*, 10(9):e0135899, sep 2015.

[178] Fei Sun, Ian J Norman, and Alison E While. Physical activity in older people: a systematic review. *BMC Public Health*, 13(1):449, dec 2013.

[179] Richard P. Troiano, David Berrigan, Kevin W. Dodd, Louise C. M??sse, Timothy Tilert, and Margaret Mcdowell. Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40(1):181–188, 2008.

[180] Amy E. Latimer-Cheung, Lara A. Pilutti, Audrey L. Hicks, Kathleen A. Martin Ginis, Alyssa M. Fenuta, K. Ann MacKibbon, and Robert W. Motl. Effects of exercise training on fitness, mobility, fatigue, and health-related quality of life among adults with multiple sclerosis: A systematic review to inform guideline development, 2013.

[181] Robert W Motl, Myla D Goldman, and Ralph H B Benedict. Walking impairment in patients with multiple sclerosis: exercise training as a treatment option. *Neuropsychiatric disease and treatment*, 6:767–774, 2010.

[182] Erin M Snook and Robert W Motl. Effect of exercise training on walking mobility in multiple sclerosis: a meta-analysis. *Neurorehabilitation and neural repair*, 23(2):108–116, feb 2009.

[183] R W Motl, P A Arnett, M M Smith, F H Barwick, B Ahlstrom, and E J Stover. Worsening of symptoms is associated with lower physical activity levels in individuals with multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 14(1):140–2, jan 2008.

[184] Rachel E. Klaren, Robert W. Motl, Deirdre Dlugonski, Brian M. Sandroff, and Lara A. Pilutti. Objectively quantified physical activity in persons with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation*, 94(12):2342–2348, 2013.

[185] Robert W Motl, Edward McAuley, and Erin M Snook. Physical activity and multiple sclerosis: a meta-analysis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 11(4):459–63, aug 2005.

[186] U S Department of Health, Human Services, U S Department of Health, Human Services, and Others. Physical activity guidelines for Americans, 2008.

[187] Vanderwa.Wh and C H Wyndham. Equation for Prediction of Energy Expenditure of Walking and Running. *Journal of Applied Physiology*, 34(5):559–563, 1973.

[188] Robert W. Motl, Erin M. Snook, Stamatis Agiovlasitis, and Yoojin Suh. Calibration of Accelerometer Output for Ambulatory Adults With Multiple Sclerosis. *Archives of Physical Medicine and Rehabilitation*, 90(10):1778–1784, October 2009.

[189] Brian M. Sandroff, Barry J. Riskin, Stamatis Agiovlasitis, and Robert W. Motl. Accelerometer cut-points derived during over-ground walking in persons with mild, moderate, and severe multiple sclerosis. *Journal of the Neurological Sciences*, 340(1-2):50–57, 2014.

[190] J.A. Levine. Measurement of energy expenditure. *Public Nealth nutrition*, 8(7A):1123–1132, oct 2005.

[191] Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *ReCALL*, 4(510):126, 1998.

[192] G.D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[193] Vitali Witowski, Ronja Foraita, Yannis Pitsiladis, Iris Pigeot, and Norman Wirsik. Using hidden markov models to improve quantifying physical activity in accelerometer data - a simulation study. *PloS One*, 9(12):e114089, 2014.

[194] R L Waters, B R Lunsford, J Perry, and R Byrd. Energy-speed relationship of walking: standard tables. *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, 6:215–222, 1988.

[195] Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin. A survey of techniques for incremental learning of HMM parameters, 2012.

[196] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J.M. Buhmann. Topology free hidden Markov models: application to background modeling. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 1:294–301, 2001.

[197] Burr Settles. Active Learning Literature Survey. *Machine Learning*, 15(2):201–221, 2010.

[198] Brigham Anderson and Andrew Moore. Active Learning for Hidden Markov Models: Objective Functions and Algorithms. *Proceedings of the 22nd international conference ...*, pages 9–16, 2005.

[199] Yogesh Girdhar and Gregory Dudek. Optimal online data sampling or how to hire the best secretaries. In *Proceedings of the 2009 Canadian Conference on Computer and Robot Vision, CRV 2009*, pages 292–298. IEEE, may 2009.

[200] Matthew T Clark, John B Delos, Douglas E Lake, Hoshik Lee, Karen D Fairchild, John Kattwinkel, and J Randall Moorman. Stochastic modeling of central apnea events in preterm infants. *Physiological measurement*, 37(4):463–484, apr 2016.

[201] S E Schmidt, C Holst-Hansen, C Graff, E Toft, and J J Struijk. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. *Physiological measurement*, 31(4):513–29, apr 2010.

[202] CA McGrory and DM Titterington. Variational Bayesian analysis for hidden Markov models. *Australian & New Zealand Journal of Statistics*, (61):1–23, 2009.

[203] Shihao Ji, Balaji Krishnapuram, and L Carin. Variational Bayes for continuous hidden Markov models and its application to active learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):522–532, 2006.