DISS. ETH NO. 29231

Enabling Objective Fatigue Quantification in Neurological Patients Through Mobile Technologies

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Liliana Leonor Barrios Aparicio
MSc UZH in Informatics
University of Zurich
born on 15.11.1989
citizen of the Bolivarian Republic of Venezuela

accepted on the recommendation of
Prof. Dr. Christian Holz, examiner
Prof. Dr. Cecilia Mascolo, co-examiner
Prof. Dr. Gunnar Rätsch, co-examiner
Dr. med. Andreas Lutterotti, co-examiner

2023

Abstract

The increasing computational capabilities of smartphones and wearable devices create the opportunity to change patient monitoring strategies in medical care, particularly for symptoms that, until now, have been challenging to assess, such as fatigue. Fatigue is a highly prevalent and debilitating symptom of several chronic conditions. Patients often describe it as extreme exhaustion that interferes with their daily life and manifests physically or cognitively. Yet, despite its high prevalence and devastating consequences, little is known about the symptom, and no effective treatment is available. Several factors affect our understanding of fatigue. Of those, we highlight three. First, patients' follow-up after a medical consultation is frequently limited to one-time point assessments separated by weeks or even months, creating a knowledge gap for physicians. Second, fatigue is commonly assessed through questionnaires, which are known to be unreliable and introduce bias, given their subjective nature. Third, fatigue's objective component, known as *fatigability*, is often neglected. Fatigability has been defined as the objective decline in performance during a cognitive or physical task. While there have been attempts to use digital technologies for better fatigue characterization, their focus is mainly on presenting a digital version of existing fatigue questionnaires.

This dissertation contributes technical innovation and exploration toward ubiquitous, more frequent, and remote fatigue monitoring in multiple sclerosis patients using smartphones and wearable devices. This technology has the potential to enable long-term and accurate disease progression monitoring. We devise new fatigability measurement methods and study their association with subjective fatigue. We hypothesize that the objective decline in performance (fatigability) can be measured using smartphones for both fatigue's physical and cognitive components. Moreover, we hypothesize that these measurements are valid outside the clinical setting and are associated with perceived fatigue measured with standard questionnaires. We conduct five different studies to test our hypotheses.

First, we propose a technique to objectively quantify motor fatigability using a commodity smartphone. Our method comprises a simple exertion task requiring rapid alternating finger tapping. Typically, motor fatigability is assessed using a handgrip dynamometer. This approach has been proven valid and accurate but requires special equipment and trained personnel. Our feasibility study with multiple sclerosis (MS) patients and healthy controls shows an association between our approach and the baseline handgrip method, providing a first step toward more frequent and remote monitoring. Second, we explore the association between tapping-derived metrics and perceived fatigue assessed with two standard clinical

scales. We conduct a two-week in-the-wild study with MS patients to evaluate the approach. Our novel smartphone-based fatigue metric, mean tapping frequency, objectively ranks perceived fatigue according to both fatigue scales. These results demonstrate that our approach is feasible and valid in uncontrolled environments.

Third, we introduce the cognitive fatigability assessment test (cFAST), a novel smartphone-based test to quantify cognitive fatigability. In our validation study, we classify MS patients as fatigued and non-fatigued using a clinical fatigue scale that allows the differentiation between physical and cognitive fatigue. The results reveal that our fatigability metric shows a statistically significant difference between the fatigued and non-fatigued groups. In particular, cognitively-fatigued patients decline in performance, while non-fatigued patients improve. Our results indicate that cFAST may potentially serve as a surrogate for subjective cognitive fatigue.

Continuous measurement of heart rate (HR) and heart rate variability (HRV) using commercially available wearable sensors provides the opportunity to improve remote patient monitoring. However, standard wearable devices use photoplethysmography (PPG) to derive HR and HRV data. It is yet unclear to which extent PPG signals can be used as a proxy for data collected using medical-grade devices. In a new study, we consider five consumer devices measuring HR and two devices measuring HRV and compare their signals with the output of standard electrocardiography (ECG) Holter monitor. Results from our study with 14 participants who followed a 55 minutes protocol show that PPG is a valid proxy for both HR and standard time- and frequency-domain measurements of HRV. Further, we demonstrate that wearable devices are suitable for monitoring both HR and HRV in daily life but might be limited during strenuous exercise. Finally, we introduce our fatigue data set, which comprises data from 55 MS patients and 25 controls. The data set was gathered during a two-week in-the-wild study, where participants used our monitoring infrastructure, including a smartphone application and companion wearable device. With this data, we provide a first glimpse into the capabilities of wearable devices for monitoring fatigue.

We conclude this dissertation with a discussion and outlook to the future. We believe remote monitoring will be part of routine medical care. In particular, objective fatigue quantification will play a significant role in understanding fatigue and developing effective therapies. We hope that our work provides a relevant contribution in that direction.

Zusammenfassung

Die immer größeren Rechenkapazitäten von Smartphones und tragbaren Geräten bieten die Möglichkeit, aktuelle Strategien zur medizinischen Patientenüberwachung anzupassen und zu verbessern. Insbesondere bei schwer beurteilbaren Krankheitssymptomen, wie zum Beispiel der chronischen Müdigkeit (Fatigue), könnten neue technische Hilfsmittel helfen, um die Patientenüberwachung zu erleichtern. Fatigue ist ein weit verbreitetes Symptom bei verschiedenen chronischen Erkrankungen. Patienten beschreiben Fatigue oft als extreme Erschöpfung, die ihren Tagesverlauf beeinträchtigt und sich körperlich oder kognitiv äussert. Doch trotz der hohen Prävalenz und den schwerwiegenden Auswirkungen ist nur wenig über Fatigue bekannt, und es gibt bis jetzt keine wirksame Behandlung. Mehrere Faktoren erschweren unser Verständnis von Fatigue. Erstens beschränkt sich die Nachbeobachtung von Patienten nach einer ärztlichen Konsultation häufig auf punktuelle Kontrollen im Abstand von Wochen oder sogar Monaten, was eine Wissenslücke für Ärzte darstellt. Zweitens wird Fatigue in der Regel anhand von Fragebögen bewertet, die bekanntermassen unzuverlässig sind und aufgrund ihres subjektiven Charakters zu Unregelmässigkeiten führen können. Drittens wird die objektive Komponente der Fatigue, genannt Fatigability, oft vernachlässigt. Fatigability ist als objektiver Leistungsabfall während einer bestimmten Zeit bei einer Aufgabenausführung definiert. Existierende Projekte, welche digitale Technologien einsetzen, um Fatigue zu erkennen, beschränken sich hauptsächlich auf die Darstellung bestehender Fragebögen. Daher stellt sich die Frage, ob es bessere Möglichkeiten gäbe, um digitale Technologien bei der Überwachung von Fatigue einzusetzen.

Diese Dissertation stellt neue Technologien und Strategien zur Überwachung von Fatigue mithilfe von Smartphones und tragbaren Geräten bei Multiple-Sklerose-Patienten vor. Unsere Methoden haben das Potenzial, eine langfristige und genaue Überwachung des Krankheitsverlaufs zu ermöglichen. Wir entwickeln neue Methoden zur Messung der Fatigue und untersuchen ihren Zusammenhang mit der subjektiven Fatigue. Wir stellen die Hypothese auf, dass der objektive Leistungsabfall (Fatigability) mit Hilfe von Smartphones sowohl für die körperlichen als auch für die kognitiven Komponenten der Fatigue gemessen werden kann. Ferner stellen wir die Hypothese auf, dass diese Messungen auch ausserhalb des klinischen Umfelds gültig sind und mit der wahrgenommenen Fatigue in Verbindung stehen, welche mit Standardfragebögen gemessen wird. Wir führten fünf verschiedene Studien durch, um diese Hypothesen zu testen.

Im ersten Teil dieser Dissertation stellen wir eine Technik zur objektiven Quantifizierung der motorischen Fatigue mit einem handelsüblichen Smartphone vor. Die Methode umfasst eine

einfache Anstrengungsaufgabe, die ein schnelles und abwechselndes Fingertippen erfordert. Normalerweise wird die motorische Fatigue mit einem Handgriffdynamometer gemessen. Dieser Ansatz hat sich als korrekt und genau erwiesen, erfordert jedoch eine spezielle Ausrüstung und geschultes Personal. Unsere Machbarkeitsstudie mit Multiple-Sklerose-Patienten (MS-Patienten) und gesunden Kontrollpersonen zeigt einen Zusammenhang zwischen unserem Ansatz und der grundlegenden Handgriffmethode, was einen ersten Schritt in Richtung einer häufigeren Fernüberwachung darstellt.

Im zweiten Teil untersuchen wir den Zusammenhang zwischen den durch das Fingertippen gewonnenen Messwerten und der wahrgenommenen Fatigue, gemessen mit den Standard-Fatigue-Skalen. Wir führten eine zweiwöchige Studie mit MS-Patienten durch, um den Ansatz zu evaluieren. Unsere neuartige Fatigue-Skala basierend auf der mittleren Fingertipprate stuft die wahrgenommene Fatigue objektiv nach den zwei Standard-Fatigue-Skalen ein. Die Ergebnisse zeigen, dass unser Ansatz in unkontrollierten Umgebungen praktikabel und valide ist.

Im dritten Teil stellen wir einen neuen Test zur Bewertung der kognitiven Fatigue (cFAST) vor, welcher nur auf einem Smartphone basiert. In unserer Validierungsstudie klassifizieren wir MS-Patienten als betroffen von Fatigue oder als nicht betroffen von Fatigue. Die Ergebnisse zeigen, dass unsere Fatigue-Skala einen statistisch signifikanten Unterschied zwischen den zwei Gruppen aufweist. Insbesondere bei Patienten mit kognitiver Fatigue sinkt die Leistung, während sie sich bei den nicht betroffenen Patienten verbessert. Unsere Ergebnisse deuten darauf hin, dass cFAST möglicherweise als Surrogat für die subjektive kognitive Fatigue dienen kann.

Im vierten Teil analysieren wir die kontinuierliche Messung der Herzfrequenz (HR) und der Herzfrequenzvariabilität (HRV) mit handelsüblichen tragbaren Sensoren, da diese Geräte eine bessere Fernüberwachung von Patienten ermöglichen. Handelsübliche tragbare Geräte verwenden jedoch die Photoplethysmographie (PPG), um Daten zu HR und HRV abzuleiten. Es ist noch unklar, inwieweit solche PPG-Signale als Ersatz für Daten von medizinisch zertifizierten Geräten genommen werden können. In unserer Studie haben wir fünf Consumer-Geräte zur Bewertung der Signalqualität der Herzfrequenz und zwei Geräte zur Messung der Herzfrequenzvariabilität (HRV) untersucht und mit einem Standard-Elektrokardiographie-Holter-Monitor verglichen. Die Ergebnisse unserer Studie zeigen, dass PPG ein gültiger Ersatz sowohl für die Herzfrequenz als auch für Standardmessungen im Zeit- und Frequenzbereich der HRV ist. Darüber hinaus zeigen wir, dass tragbare Geräte für die Überwachung von HR und HRV im Alltag geeignet sind, aber bei anstrengender körperlicher Betätigung eingeschränkt sein können.

Im letzten Teil stellen wir unseren Fatigue-Datensatz vor, der Daten von 55 MS-Patienten und 25 Kontrollpersonen umfasst. Der Datensatz wurde im Rahmen einer zweiwöchigen Studie erhoben, bei der die Teilnehmer unsere Smartphone-Anwendung und ein zusätzliches tragbares Gerät nutzten. Mit diesen Daten geben wir einen Einblick in die Möglichkeiten von tragbaren Geräten zur Überwachung der Fatigue. Wir schließen diese Dissertation mit

einer Diskussion und einem Ausblick für mögliche zukünftige Arbeit in diesem Forschungsbereich ab. Wir glauben, dass die Fernüberwachung Teil der medizinischen Routineversorgung sein wird. Wir glauben auch, dass die objektive Überwachung der Fatigue eine wichtige Rolle für das Verständnis der Fatigue und die Entwicklung wirksamer Therapien spielen wird. Wir hoffen, dass unsere Arbeit einen wichtigen Beitrag in diese Richtung leistet.

Acknowledgments

This work would not have been possible without the help and support of many people. Please note this is not an exhaustive list. First, I thank my advisors for guiding me during the process. Thank Prof. Friedemann Mattern for being incredibly supportive, allowing me to pursue a Ph.D. in Digital Health, and for still being present. Similarly, I thank Prof. Christian Holz for allowing me to join his team, for his support, for sharing his knowledge, and for guiding me in conducting research. I thank Dr. med. Andreas Lutterotti for believing in me and supporting the project from the beginning. Thanks for sharing your knowledge and for making this research possible. Without you, this project would not have been possible. Thanks to Prof. Gunnar Rätsch for supporting the project and agreeing to be my advisor. Furthermore, I thank Prof. Cecilia Mascolo for kindly agreeing to be part of my examination committee and reviewing this dissertation.

I want to thank all the members of the Distributed System Group for being incredibly supportive during our time working together and beyond. Despite all challenges, you made it possible for me to complete this journey. Thanks for your understanding, help, and support. I did not take it for granted. Jing Yang, Mihai Bâce, Vincent Becker, Lukas Burkhalter, Alexander Viand, Hossein Shafagh, Anwar Hithnawi, Vlad Coroamă, Leyna Sadamori, Wilhelm Kleiminger, and Barbara von Allmen, thank you all. Jing, it was an absolute pleasure meeting you and sharing the office with you. Despite the distance, I still consider you my office partner. Thanks for being present even from abroad. Thanks, Wilhelm Kleiminger, for motivating me to pursue a Ph.D. I would not have considered this as an option if it were not for your advice. Barbara, you are such a friendly and fun person to be around. I will never forget stopping at your office for a brief chat in the morning before starting to work. Thank you for always finding a way to help.

This dissertation is the result of working with many collaborators. Thank you, David Lindlbauer, for all the time you invested in me and for helping me shape this work into the HCI field. I enjoyed working closely with colleagues from the Neurology Department at University Hospital Zurich. Special thanks to Marc Hilty for bringing his positive energy and efforts to the project, Helen Hayward-Koennecke for helping shape our mobile app with her medical insides, and Rok Amon for his work and dedication. Thanks to all the other staff, doctors, nurses, and administrators that help make the research possible. And special thanks to all patients who kindly agreed to participate in our studies. Similarly, I thank the colleagues from USI Lugano for being so nice and open to collaborating. Thanks to Prof. Silvia Santini for guiding me at the beginning of this journey to shape this project

and for her support during the past years. Thanks to Shkurta Gashi and Elena Di Lascio for their advice and support.

At ETH, I had the honor of supervising several students' projects and theses. I learned from all of them. Thanks for wanting to be part of this work and for choosing me as your advisor: Sinan Demirci, Artur Gigon, Marina Draskovic, Lena Csomor, Wenjie Wang, Tianyi Xiao, Fabian Mächler, and last but certainly not least, Pietro Oldrati. Pietro, there are not enough lines to express my gratitude to you. You went from being an advisee to a colleague, and now it is my pleasure and honor to call you a friend. I have learned a lot from you. Thank you for everything. I look forward to seeing what is next in life for you. I am sure it will be fantastic.

Thanks to everyone who helped me shape the thesis, proof-reading, and helped me prepare for the defense. Thanks to Christoph Gebhardt and Rafael Wampfler for assisting in the last steps of this dissertation. Thanks, Lukas Burkhalter, for helping me with the German abstract.

Also, thanks to all my friends who have supported me these past years. You also played a role in this dissertation by making me disconnect from work, listening, or even proof-reading my work: Ana Guerra, Maria del Mar Linarez, Barbara Schaub, Jan Meier, Daniel Scherly, Lindsay Axford, Stefano Halabi, and Oriana Miranda.

Finally, I want to thank my family for always being there for me. My dear husband, Carlos, thanks for your time, help, and advice and for always having a joke to make difficult times more manageable. To my lovely son, Marcos Leonardo, I love you. Mom, thanks for being incredible, inspiring, hard-working, and strong. Thank you for all that you have done for us. Also, thanks for taking such good care of Marcos and always being happy to come to help us. Thanks to my brother, sister, and my dad for their support from abroad. Marilin, thanks for your strength and for being there for Mom and Dad. Erich, thanks for your company in the good and not-so-good times. Thanks to my mother-in-law for always being open to supporting us and ready to jump on a plane to come and give us a hand. Also, thanks to my cousin, Leonor, who is always one call away for whatever we need. Lastly, my dear brother, Mario, you were such a joyful person. You taught us to live and enjoy life now. Wherever you are, I know that you are celebrating with me. I miss you, I love you, and I dedicate this to you.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Motivation

User acceptance of consumer-wearable devices has increased significantly in recent years, especially for monitoring physical activity and sleep. In parallel, there is growing interest in mobile health (mHealth), the adoption of mobile computing technologies such as smartphones and wearable devices in health care [Kelli et al., 2017]. Traditional health care has often been limited to one-time assessments that evaluate the health status of patients during single medical consultations. Mobile and wearable technologies have the potential to overcome this limitation with their ability to perform more regular or continuous monitoring. This is particularly interesting in chronic conditions where monitoring patients' vital signs over extended periods could lead to a better understanding and management of symptoms that, until now, have been challenging to assess, such as fatigue.

Fatigue is a common symptom of many diseases caused by viral infections [del Rio and Malani, 2020; Townsend et al., 2020; Perrin et al., 2020], autoimmunity [Belza, 1995; Hewlett et al., 2005], cancer [Mitchell, 2010; Spelten et al., 2003], neurodegenerative [Friedman and Friedman, 1993] and cardiovascular disease [Falk et al., 2007; Casillas et al., 2006]. Patients often describe it as extreme exhaustion that presents either physically or cognitively. Up to 30% of individuals with COVID-19 suffer from this debilitating symptom even weeks following the acute disease [del Rio and Malani, 2020; Townsend et al., 2020; Perrin et al., 2020]. Despite its high prevalence, fatigue's pathogenesis remains uncertain, and no approved therapy is available yet [Kluger et al., 2013; Krupp, 2003; Rudroff et al., 2016]. The lack of objective measurements to quantify fatigue is a significant obstacle to better understanding the symptom.

Fatigue detection technologies can improve fatigue management and are not limited to chronic conditions. For example, they could also benefit other industries where workers' fitness is vital for safety (e.g., transportation). Several studies have shown a significant association between fatigue and an increased risk of accident and injury [Williamson et al., 2011]. Most organizations from the transportation domain implement controls and schedules which protect employees from extreme tiredness. However, there are no guarantees that employees comply with these regulations or that self-reported "fitness-for-duty" data policies are reliable [Dawson et al., 2014]. Therefore, fatigue detection has the potential to identify unacceptable tiredness levels and notify the employees or the responsible organization [Dawson et al., 2014].

We focus our attention on fatigue as a primary symptom of diseases, precisely, as a symptom of Multiple Sclerosis (MS). MS is an autoimmune disease characterized by recurrent inflammation in areas of the central nervous system [Mireia and Roland, 2005], which comprises the brain, spinal cord, and optic nerves. With an estimate of over 2.8 million patients affected worldwide [Society, 2018], MS is one of the leading causes of neurological disability in young adults. Fatigue is one of the most prevalent symptoms of MS; 75–95% of MS patients have reported fatigue at some point [John et al., 1994; Lauren and Dean, 1996; Anners et al., 2007; Kobelt et al., 2017]. There is no cure for MS. However, the latest MS disease-modifying therapies (DMTs) reduce the damage to the nerves and keep the disease stable by reducing inflammation and the number of relapses. These therapies have been shown to be effective. But, despite receiving DMTs and having the disease under control, fatigue continues to be a common complaint in patients, making MS a suitable candidate for studying fatigue. Furthermore, MS patients are typically young adults, which makes them an attractive population to explore the development of digital solutions.

Mobile applications and self-tracking are some of the tools MS patients adopt when confronted with an MS diagnosis. According to research, MS application functionality usually includes disease monitoring, fitness tracking, and life journaling [Giunti et al., 2017; Ayobi et al., 2017]. Pharmaceutical companies have as well developed different mobile applications targeting MS. Some examples include Floodlight – Genentech [2023], Aby – Biogen [2023], and SymTrac – Novartis [2023]. Although many MS-related applications exist, they focus mainly on self-management and displaying informational videos. Similarly, there are specific applications for fatigue quantification and management, but they rely on existing fatigue questionnaires and do not aim to quantify the symptom objectively [GAIA, 2023; Giunti, 2023]. Thus, we use MS as the target group in this dissertation to investigate and develop new fatigue quantification tools based on digital technologies. We hypothesize that such solutions can later be used in other diseases with fatigue as a primary debilitating symptom.

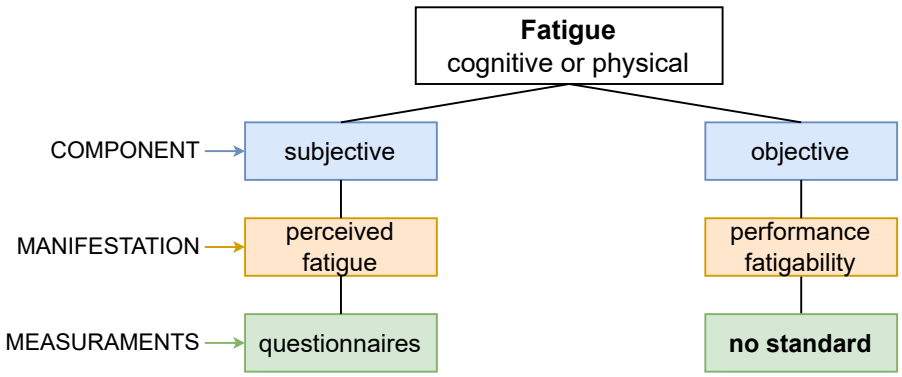

Figure 1.1.: Fatigue conceptualization overview. Fatigue's subjective component, which manifests as perceived fatigue, is measured with questionnaires, while fatigue's objective component, which manifests as performance fatigability, has no standard measurement approach.

1.2. Problem Setting

In research, fatigue has been defined as the subjective feeling of overwhelming exhaustion and tiredness and can manifest as a physical and cognitive symptom [Krupp et al., 2007]. As displayed in Figure 1.1, fatigue consists of a subjective and an objective component. The symptom is still poorly understood, and its severity is mainly assessed through its subjective component. This subjective component manifests in patients as the perception of fatigue (perceived fatigue). Currently, this is measured using questionnaires such as the Fatigue Severity Scale (FSS) [Krupp et al., 1989], Modified Fatigue Impact Scale (MFIS) [Téllez et al., 2005], and Fatigue Scale for Motor and Cognitive Functions (FSMC) [Penner et al., 2009]. Over a dozen fatigue questionnaires are available and used as patient-reported outcome measures in clinical trials [Penner and Paul, 2017]. Their heterogeneity and subjective nature are challenging for using them as outcome measures in clinical trials and comparing results' efficacy across different studies. The shortcomings of current fatigue assessments have been discussed in several studies [Schwid et al., 2002; Kluger et al., 2013; Braley and Chervin, 2010]. Schwid et al. [2002] state that current fatigue assessment methods rely on subjective self-reporting questionnaires, which can be confounded by other symptoms and require difficult retrospective analysis. Additionally, Kluger et al. [2013] pointed out the need for an agreed-upon definition of the term *"fatigue"* in many of those studies. The authors state that the lack of a unified taxonomy and assessment methods hampers progress in understanding fatigue. We follow Kluger et al. [2013]'s definition of fatigue as *"the subjective sensations of weariness and increased sense of effort"*.

The role of fatigue's objective component has gained more attention in recent

years (refer to Figure 1.1 right side). Researchers have suggested that fatigue is associated with *fatigability* [Dobkin, 2008; Steens et al., 2012; Loy et al., 2017; Wolkorte et al., 2015; Zijdewind et al., 2016]. Wolkorte et al. [2015] have highlighted the importance of including fatigability in the models to explain perceived fatigue in patients with MS. Kluger et al. [2013] define fatigability as *"the magnitude or rate of change in a performance criterion relative to a reference value over a given time of task performance"*. Fatigability is further divided into the motor and cognitive domains. Establishing an association between objective fatigability and subjective fatigue is an important goal for clinical research but has been proven difficult [Kluger et al., 2013]. Dobkin [2008] argued that fatigability could redefine our understanding of fatigue, because many symptoms of fatigue may be a consequence of demonstrable fatigability, but this has rarely been assessed. Existing methods to evaluate fatigability pose technical problems and require well-controlled experiments or expensive machinery [Dobkin, 2008]. As highlighted in Figure 1.1, no standard fatigability measurements exist. The most common approach to assess physical fatigability is with a handgrip dynamometer, while cognitive fatigability has been assessed through prolonged cognitive examination. Therefore, finding ubiquitous and inexpensive ways to measure fatigability could be beneficial to understand fatigue and support the unmet medical need for the development of objective measurements to quantify fatigue.

Other attempts to better characterize fatigue have been to study alteration of the autonomic nervous system (ANS), which regulates involuntary physiological processes including heart rate, blood pressure, respiration, and digestion [Waxenbaum et al., 2019; Florea and Cohn, 2014]. The ANS divides into the sympathetic nervous system, responsible for the body's "fight or flight" response, and the parasympathetic nervous system, associated with "rest and digest." MS-related inflammation also affects the ANS [Racosta and Kimpinski, 2016; Adamec and Habek, 2013], resulting in autonomic dysfunction (AD). Studies suggest a link between fatigue and autonomic dysfunction in MS [Keselbrener et al., 2000; Flachenecker et al., 2003; Merkelbach et al., 2001]. A common approach to assessing AD is measuring cardiac dysfunction through heart rate variability (HRV) metrics [Malik, 1996]. HRV consists of changes in the time intervals between consecutive heartbeats called interbeat intervals (IBIs) [Shaffer and Ginsberg, 2017]. However, AD measurements related to the cardiovascular system are often limited by standard assessment methods (e.g., electrocardiogram (ECG) or ECG Holter). Thus, state-of-the-art wearables capable of measuring IBIs would allow noninvasive and quantitative evaluations of MS autonomic dysfunction, which until now has been challenging to achieve.

In this dissertation, we seek to develop objective fatigue quantification measurements using mobile and wearable sensor data. Such measurements could aid in understanding how fatigue manifests in the patient's everyday life. Our goal is to enable the prediction of fatigue events, which in the future can lead to improvements in

patients' quality of life and the implementation of personalized health. Ultimately, we envision a system based on smartphone and wearable sensors that allows longitudinal surveillance.

1.3. Principal Contributions

In the following, we list the main technical contributions of the work presented in this thesis:

Rapid tapping on smartphones to assess motor fatigability

Our first contribution is the development of a new physical fatigability assessment. Our approach utilizes the ubiquitousness of smartphones in conjunction with a simple exertion task to assess the user's motor fatigability via alternate finger tapping. In an experiment with 20 MS patients and 35 controls, we compare our approach with a standard fatigability assessment done with a handgrip dynamometer (Figure 1.2). We show that the participants' performance decreases during the tapping task. Moreover, this performance decay correlates with the decrease in grip strength measured with the handgrip dynamometer for patients and control. We further show that this correlation is already present in the first 30 seconds of the tapping task, suggesting that performing the tapping task for 30 seconds is sufficient to measure motor fatigability.

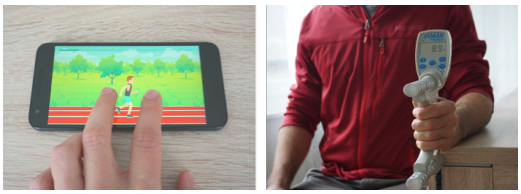

Figure 1.2.: Methods for validation of new objective fatigability metric.

Rapid tapping associated to fatigue in unsupervised settings

Our second contribution is developing a new objective and reliable measure of motor fatigability computed from raw tapping data and demonstrating its usability and validity when performed outside controlled settings and without medical supervision. We conducted a two-week in-the-wild study with 35 MS patients. The participants performed

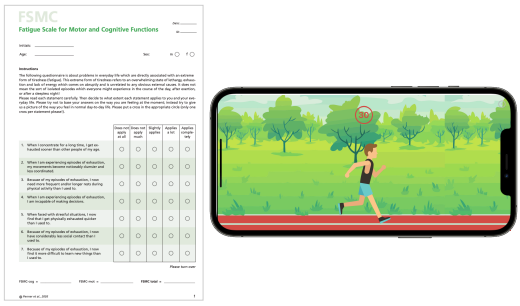

Figure 1.3.: Study methods for the association between perceived fatigue and objective fatigue.

a 30 s tapping task once per day. Using this data, we introduced a new metric to assess motor fatigability: *tapping frequency*. We showed that our new metric is a valid method to assess motor fatigability in the wild by comparing it to strength decline using a handgrip dynamometer. Following, we show an association between *tapping frequency* and two widely accepted and validated fatigue questionnaires in MS patients: FSS (AUC_{ROC} $\overline{X} = .81 \pm .05$) and FSMC ($AUC_{ROC}$ $\overline{X} = .76 \pm .05$.). Figure 1.3 depicts part of the study methods for this validation: (1) clinically validated fatigue scale (left) and tapping task (right). Our smartphone-based tapping task and the derived metric have the opportunity to be used regularly by patients outside the clinic and more frequently than currently done in the medical routine. Being an objective method, it also opens the potential for quantifying the direct effects of therapeutic interventions, which is a clear advantage over currently used questionnaires [Nourbakhsh et al., 2021].

Cognitive fatigability assessment test (cFAST)

Our third contribution is designing and implementing a new instrument for objectively quantifying cognitive fatigability, which we named cFAST. In a prospective multidisciplinary trial, we evaluated the new tool by studying its association with perceived cognitive fatigue measured through the validated FSMC cognitive subscale (Figure 1.4). Results from this study confirm an association between our objective assessment (cFAST) and subjective fatigue. Such an objective measure that serves as a surrogate of perceived fatigue could greatly help our understanding of fatigue and allow the objective monitoring of cognitive fatigue. Furthermore, to our knowledge, ours is the first study to introduce an instrument specifically designed to quantify cognitive fatigability. Thus, cFAST circumvents previous studies' limitations: (1) using standard cognitive tests to assess cognitive fatigability and (2) requiring long testing sessions. In addition, the proposed test was designed and implemented to investigate patient-reported outcomes in an uncontrolled, real-world application outside of the clinic, with the goal of providing important insights for patient management. Hence, the methodology can potentially be used in clinical trials, interventional studies, and routine patient care.

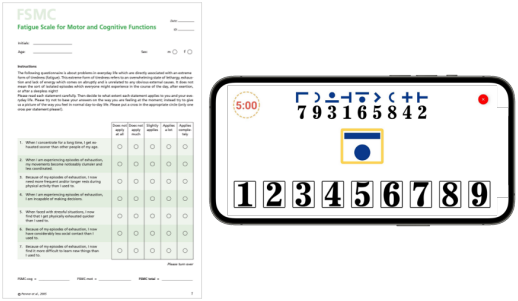

Figure 1.4.: Study methods: fatigue questionnaire (left), cFASt (right).

Wearable selection and Cronico fatigue data set

Our last contribution is the construction of our fatigue data set comprised of 55 MS patients and 25 controls. This data set was collected using smartphone and wearable data with our Cronico monitoring system during a two-week in-the-wild study (Figure 1.5). Our system includes two types of monitoring approaches: active and passive. In active monitoring, patients interact with our system to complete a task (e.g., fatigability task) or provide input, such as completing a fatigue questionnaire at a given time. In passive monitoring, our system collects contextual information about the patient's behavior and environment, such as the patient's vital signs, activity level, and weather conditions. As the first step to achieving this contribution, we had to find suitable commercially-available wearable devices that allowed continuous monitoring of patients' vitals over extended periods. Additionally, we needed to verify the validity of important parameters, such as heart rate and heart rate variability, that would enable us to quantify autonomic dysfunction and study its association with perceived fatigue. Following this, we designed and implemented our monitoring system. Chapter 7 presents preliminary results of the association between passive data and perceived fatigue. Cronico's fatigue data set is now being used for data exploration as part of an ETH – Personalized Health and Related Technologies (PHRT) grant.

Figure 1.5.: Devices used for gathering our fatigue dataset (Cronico app, two Everion devices and data syncronization app).

1.4. Dissertation Outline and Contributions Statement

This dissertation consists of three parts (I) physical fatigue, (II) cognitive fatigue, and (III) the use of wearables for unsupervised monitoring. Below, we show the dissertation outline in chapters with the corresponding contributions statements. The work presented is in the intersection of computer science and medicine. Andreas Lutterotti was the advisor from the medical domain while Christian Holz was the advisor for the computer science-related aspects of this dissertation. The contributions presented in this thesis were partly made by others. Unless otherwise mentioned, the work presented in this thesis is the author's original work.

Chapter 2 provides an overview of related work on physical fatigability, finger tapping for impairment quantification, cognitive fatigability, wearables for unsupervised monitoring, and smartphone-based application for health monitoring.

Chapter 3 introduces rapid alternating finger tapping on smartphones as a novel physical fatigability test. We compare our approach against maximal voluntary contraction (MVC) with a handgrip dynamometer in a controlled study including multiple sclerosis patients and a control group.

Contribution statement: As part of his Master thesis, Pietro Oldrati contributed to aspects of the analysis of the task. The data analysis methodology originated from discussions with David Lindlbauer. In addition, Marc Hilty and Helen Hayward-Koennecke contributed to the medical domain knowledge and patient recruitment.

Chapter 4 revisits the rapid tapping as fatigability test. It describes our two-week in-the-wild study on MS patients for studying the feasibility of establishing rapid tapping on smartphones as a proxy for subjective fatigue measured with two clinically-accepted fatigue scales.

Contribution statement: Pietro Oldrati contributed to parts of the analysis. The data analysis presentation originated from discussions with David Lindlbauer. Marc Hilty helped in the data recruitment and discussions.

Chapter 5 introduces a novel test to quantify cognitive fatigability, the cognitive fatigability assessment test (cFAST). Furthermore, the chapter shows the feasibility of establishing cFAST and a proxy for subjective cognitive fatigue through a controlled study with MS patients.

Contribution statement: Pietro Oldrati contributed to the specification and implementation of the test. Rok Amon and Marc Hilty helped with data gathering and medical discussions.

Chapter 6 presents a protocol and study for the validation PPG-derived HR and HRV metrics extracted from different off-the-shelf wearables devices and compares them to a standard ECG medical-grade Holter.

Contribution statement: Pietro Oldrati contributed to parts of the analysis. Silvia Santini contributed to discussions of data presentation.

Chapter 7 introduces Cronico, a multidimensional fatigue monitoring platform including a wearable for physiological sensing and a smartphone application with the developed fatigability tests as well as different questionnaires to quantify subjective symptoms (i.e., sleep, fatigue and stress). Furthermore, this chapter introduces the Cronico dataset, which contains data from our

two-week study for remote fatigue monitoring with the platform. The resulting dataset includes 55 MS patients and 25 controls, different data sources gathered with Cronico, and patients' medical data.

Contribution statement: Cronico's development started back in 2017. Pietro Oldrati contributed to its implementation and the implementation of the physiological data pre-processing pipeline. Marc Hilty contributed with patient recruitment, medical domain input, and contributed to the HRV data pipeline development.

All the mentioned collaborators received co-authorship on the related publications.

1.5. Publications

In the context of this thesis, the following peer-reviewed publications have been accepted:

Liliana Barrios, Pietro Oldrati, Silvia Santini, Andreas Lutterotti. Recognizing Digital Biomarkers for Fatigue Assessment in Patients with Multiple Sclerosis. *Proceedings of EAI International Conference on Pervasive Computing Technologies for Healthcare (EAI PervasiveHealth)*. New York, United States. May 21-24, 2018 [Barrios et al., 2018].

Liliana Barrios, Pietro Oldrati, Silvia Santini, Andreas Lutterotti. Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. *Proceedings of EAI International Conference on Pervasive Computing Technologies for Healthcare (EAI PervasiveHealth)*. Pages 251–261. Trento, Italy. May 20-23, 2019 [Barrios et al., 2019].

Liliana Barrios, Pietro Oldrati, David Lindlbauer, Marc Hilty, Helen Hayward-Koennecke, Christian Holz, Andreas Lutterotti. A Rapid Tapping Task on Commodity Smartphones to Assess Motor Fatigability. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Pages 1–10. Hawaii, United States. April 25-30, 2020 [Barrios et al., 2020].

Liliana Barrios, Pietro Oldrati, Marc Hilty, David Lindlbauer, Christian Holz, Andreas Lutterotti. Smartphone-Based Tapping Frequency as a Surrogate for Perceived Fatigue. An in-the-Wild Feasibility Study in Multiple Sclerosis Patients. *The Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*. Volume 5, Issue 3, Article No.: 89, pp 1-30. September 2021 [Barrios et al., 2021].

Marc Hilty, Pietro Oldrati, **Liliana Barrios**, Tamara Müller, Claudia Blumer, Magdalena Foege, PHRT consortium, Christian Holz, Andreas Lutterotti. Continuous monitoring with wearables in multiple sclerosis reveals an association of cardiac autonomic dysfunction with disease severity. *Multiple Sclerosis Journal – Experimental, Translational and Clinical*. June 2022 [Hilty et al., 2022].

Liliana Barrios, Rok Amon, Pietro Oldrati, Marc Hilty, Christian Holz, Andreas Lutterotti. Cognitive fatigability assessment test (cFAST): development of a new instrument to assess cognitive fatigability and pilot study on its association to perceived fatigue in multiple sclerosis. *SAGE journals, Digital Health.* August 2022 [Barrios et al., 2022].

Further planned publications that resulted from the data set gathered in the clinical studies conducted during my Ph.D. include:

[in preparation] Max Möbus, Pietro Oldrati, Marc Hilty, Christian Holz, **Liliana Barrios**. The Autonomic Nervous System as a Driver for Sleep Quality: The Impact of Multiple Sclerosis. *npj Digital Medicine*. 2023.

[in preparation] Max Möbus, Shkurta Gashi, Marc Hilty, PHRT consortium, Christian Holz. Sensor-Based Assessment of Fatigue in Patients with Multiple Sclerosis: a two-week intensive longitudinal study. *Lancet Digital Health*. 2023.

[in preparation] Shkurta Gashi, Max Möbus, Pietro Oldrati, Marc Hilty, PHRT Consortium, Christian Holz. Multiple Sclerosis Diagnosis from Smartphone and Wearable Sensor Data in Free-Living Environments. npj Digital Medicine 2023.

The Cronico dataset that resulted from our two-week in-the-wild study with MS patients is now being utilized for data exploration as part of a Personalized Health and Related Technologies (PHRT) grant.

Further publications that were conducted during the course of the Ph.D. research but are out of the scope of this dissertation are listed below:

Vincent Becker, Pietro Oldrati, **Liliana Barrios**, Gábor Sörös. TouchSense: Classifying and Measuring the Force of Finger Touches with an Electromyography Armband. *Proceedings of ACM International Conference Augmented Human.* Article No.: 34 Pages 1–3. February 2018. (Poster)

Vincent Becker, Pietro Oldrati, **Liliana Barrios**, Gábor Sörös. TouchSense: Classifying Finger Touches and Measuring their Force with an Electromyography Armband. *Proceedings of the 2021 ACM International Symposium on Wearable Computers*. Pages 1–8. Singapore. October 8 - 12, 2018.

Sarah Faltaous, Gabriel Haas, **Liliana Barrios**, Andreas Seiderer, Sebastian Felix Rauh, Han Joo Chae, Stefan Schneegass, Florian Alt. BrainShare: A Glimpse of Social Interaction for Locked-in Syndrome Patients. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Paper No.: LBW0155. Pages 1–6. May 2019. (Extended Abstract)

CHAPTER 2

Related Work

In this section, we review relevant related work from the fields of fatigue and fatigability measurements, applications of finger-tapping tests, and smartphone-based health monitoring and position our work in comparison to existing research in the field.

2.1. Motor Fatigability

Motor fatigability has been quantified as the decline in peak performance, power, or speed during physical activity [Schwid et al., 2003]. While there is no established methodology to measure fatigability, it has been assessed through walking (e.g., a 6-minute walking test [Goldman et al., 2008]), handgrip strength [Severijns et al., 2015], and a knee dynamometer [Surakka et al., 2004]. Most studies have applied maximal voluntary contractions (MVCs) within a given time limit to assess motor fatigability [Severijns et al., 2017; Severijns et al., 2015; Schwid et al., 1999; Djaldetti et al., 1996; Steens et al., 2012; Clarke, 1986], requiring patients to exert pressure on a handgrip over a given time.

Several ways to quantify fatigue during maximal voluntary contraction with a handgrip dynamometer have been proposed. According to Schwid et al. [1999], the simplest method is to compare the maximal strength at the beginning and at the end of the contraction, as suggested by Miller et al. [1993]. Bigland-Ritchie et al. [1983] found that force declines in a linear manner during sustained muscle contraction at a rate characteristic for each subject. Hence, the slope of the decline indicates the rate of fatigue. Nacul et al. [2018] studied handgrip strength

as an objective measure of disease status and severity in people with chronic fatigue syndrome (CFS). Their results show that CFS patients had significantly lower mean handgrip strength than healthy controls, suggesting that the mean handgrip could be used as an objective tool for diagnosis and measuring disease severity. Similarly, they found that MS patients have a lower mean handgrip strength than healthy controls. All of these assessment tasks require a special-purpose measurement apparatus and personnel for conducting observations [Dobkin, 2008; Severijns et al., 2017]. To overcome these limitations, the latest research is focused on developing alternative fatigability methods. Tanigawa et al. [2017] observed motor fatigue during fast tapping with the index finger on a custom button. Boukhvalova et al. [2018] hypothesized that tapping with the index finger on a smartphone may measure motor fatigability. However, they did not test their hypothesis. Finding ubiquitous and inexpensive ways to measure fatigability could be beneficial to understand fatigue. Particularly, we could reach a broader population that may not have access to existing assessments.

2.1.1. Finger Tapping and Impairment

Finger-to-thumb tapping is a standard test used in Parkinson's Disease (PD) patients to assess dysfunction of the *extrapyramidal motor system*, which leads to impairment in maintaining alternating movements. Several variants of finger tapping to quantify impairment in PD using digital technology exist [Prince et al., 2018; Taylor Tavares et al., 2005; Printy et al., 2014]. Prince et al. [2018] quantify PD-related disability with an alternating finger tap on a smartphone screen for 20 seconds (counting the total number of taps). Taylor Tavares et al. [2005] use a repetitive alternating finger-tapping (RAFT) task over 30 s on a physical keyboard to quantify motor impairment. Printy et al. [2014] conducted a battery of kinematic tasks using an iPhone 5C [1] and their custom application for data collection. They used the data to quantify impairment severity, in particular, bradykinesia. Similarly, Lou et al. [2003] use alternately pressing two spaced-out piano keys to measure fatigue in PD patients. However, this Fitts' law-style task on a real piano does not provide the benefits of more ubiquitous approaches. Finger tapping tasks used in PD patients assess PD-related impairment but are not suitable to assess motor fatigability given the PD-specific confounding.

2.1.2. Finger Tapping in MS

Finger-tapping assessments are not unique to PD. Several studies show that finger tapping can also measure MS-related impairment [Alusi et al., 2000; Scherer et al., 1997; Chipchase et al., 2003; Shirani et al., 2017]. Tapping the index finger

[1] Apple Inc.

with the thumb has also been suggested to quantify impairment in MS [Shirani et al., 2017]. Chipchase et al. [2003] conducted tapping with the index finger at maximal speed with the participant's hand resting on a surface. Using a counting device, Mazur-Mosiewicz and Dean [2011] conducted 10 tapping sessions of 10 s each. Their results indicate that the number of taps can differentiate MS patients and controls but found no correlation between finger tapping and fatigue severity. Alusi et al. [2000] found a good correlation between the nine-hole peg and tapping a key on a large calculator with the index finger, thus suggesting tapping as a useful objective assessment of upper limb function in tremulous patients with multiple sclerosis. Scherer et al. [1997] used alternating left and right index finger tapping to measure tapping speed on a standard PC keyboard (key F1 and key F12). Their task can detect minimal psycho-motor dysfunction in migraine and MS-related impairment. Furthermore, differences in a computerized single finger tapping concerning gender, hand dominance, and age are examined by Hubel et al. [2013]. Their results suggest that the task can be used as a diagnostic tool and that changes in tapping rate over time can be due to fatigue or other factors. Notermans et al. [1994] use a finger tapping test to measure ataxia – poor muscle control. Psychomotor vigilance testing (PVT) has been suggested as a potential standardized assessment tool for important aspects of MS-related fatigue [Rotstein et al., 2012]. Kay et al. [2013] introduced and validated PVT-Touch, a smartphone-based version of PVT. However, PVT is an alertness test and, thus, not within the scope of motor fatigability. In chapter 3, we introduce rapid alternating finger tapping to quantify fatigability on smartphones. Our short motor task is independent of visual stimuli, hence not influenced by reaction time. In a controlled study, we test and compare our method with the most commonly used physical fatigability approach, the handgrip dynamometer. In chapter 4, we evaluate the validity of the proposed approach in unsupervised settings by letting patients conduct the tapping trials at home.

2.2. Cognitive Fatigability

Cognitive fatigability measures the decline of cognitive performance during a task that requires sustained attention [Schwid et al., 2003]. It has been measured as an increase in reaction time, a decrease in accuracy, or comparing the performance during the first and last third of a task [Krupp and Elkins, 2000; Walker et al., 2019]. While a correlation between motor fatigability and perceived fatigue has been suggested in several studies [Dobkin, 2008; Steens et al., 2012; Loy et al., 2017; Wolkorte et al., 2015], less data is available on cognitive fatigability [Walker et al., 2012a; Morrow et al., 2015; Berard et al., 2018; Berard et al., 2020; van der Linden et al., 2003; Möller et al., 2014; Wang et al., 2014; Burke et al., 2018]. A possible cause is the complexity of inducing cognitive fatigability and the lack of consensus and dedicated tests to quantify it [Walker et al., 2019]. Prior studies used one of two strategies to

generate cognitive fatigability. Either they conducted a test battery, including the same test before and after fatiguing tasks, and compared their performance, or they employed a single prolonged cognitive task and measured the decline in performance within the task. Some of the used cognitive tests within fatigability research include: (1) the Paced Auditory Serial Addition Test (PASAT) [Tombaugh, 2006], (2) the Psychomotor vigilance task (PVT) [Basner and Dinges, 2011], and (3) the Stroop test [Stroop, 1935]. However, utilizing these non-specific cognitive performance tests to assess cognitive fatigability comes with certain drawbacks, such as long testing sessions.

2.2.1. Limitations of Cognitive Fatigability Studies

Fatigability in healthy subjects. It is typically studied through long examination sessions. van der Linden et al. [2003] induced fatigue through two hours of cognitively demanding tasks. Their study showed a significant difference in planning ability and increased errors between the non-fatigued and fatigued participants. Other cognitive fatigability studies in healthy subjects using the Stroop test employed a study length of 3 and 2 hours for young adults [Wang et al., 2014] and for older adults [Burke et al., 2018], respectively. However, long testing sessions are not unique to healthy subjects. Möller et al. [2014] administered two hourly test batteries for analyzing cognitive fatigability using three neuropsychological tests in subjects with mild traumatic brain injury.

Cognitive fatigability in MS. There is large heterogeneity when it comes to studying cognitive fatigability. DeLuca et al. [2008] studied fatigue in 15 MS and 15 controls by conducting four modified SDMT (mSDMT) trials over an hour of fMRI scanning where users were shown different symbol–digit pair probes at varying inter-stimulus. Participants had to respond “match” or “no match” to each probe by following a provided symbol-digit arrangement. The inter-stimulus interval randomly varied between 0 s, 4 s, 8 s or 12 s. Results from their study found no cognitive fatigability. Chen et al. [2020] also studied fatigability using a mSDMT within a Functional Magnetic Resonance Imaging (fMRI) setting. During examination, MS patients and controls completed a total of eight mSDMT (4 with high cognitive load and 4 with low cognitive load), each lasting 7.7 min. The authors did not study within trial performance, but across trial performance showed an increase in reaction time associated with subjective fatigue in MS patients. Berard et al. [2020] compared the performance during quintiles of a 20 min PVT session to quantify cognitive fatigability and found a greater increase in reaction time of patients compared to healthy controls. PVT is a simple reaction time task where participants have to press a button in response to the presence of a stimulus. However, its repetitive and

monotonous nature often results in participants reporting feelings of boredom [Pattyn et al., 2008], and thus the performance decline may be influenced by a lack of motivation rather than fatigability [Agyemang et al., 2021]. Finally, several authors employed the PASAT by comparing the decrease in accuracy between the beginning and end of the test [Walker et al., 2012a; Morrow et al., 2015; Berard et al., 2018; Agyemang et al., 2021; Berard and Walker, 2021; Bryant et al., 2004]. Even though the PASAT is applied in many studies, there is still significant methodological heterogeneity. First, some studies compared the performance between the first and the second half [Walker et al., 2012a] of the test, while others compared the performance between thirds [Morrow et al., 2015]. Second, despite there seems to be a general consensus of 3 s length inter-stimulus interval (ISI), this has not been uniformly applied in fatigability studies [Schwid et al., 2003; Berard et al., 2018; Bryant et al., 2004]. Third, it is known that MS patients may adopt a *chunking strategy*, particularly as task demands increase [Fisk and Archibald, 2001], meaning that they add two numbers, skip one, and add the following two, thus, reducing the overall difficulty of the task by decreasing the simultaneous cognitive load. Only recently, the first normative data on cognitive fatigability has been generated to account for the chunking strategy [Berard and Walker, 2021]. Fourth, the PASAT requires a medical examiner to conduct the test, making it more expensive to administer. Finally, patients have described the PASAT as unpleasant and causing anxiety [Walker et al., 2012b], limiting the applicability and repeatability of the tests.

As described above, there is no specific test for cognitive fatigability. Hence, in chapter 5, we introduce the cognitive fatigability assessment test (cFAST), which is a smartphone-based test designed to measure cognitive fatigability in a short period. We explain the details of the test development. Furthermore, we present results of the association between subjective fatigue and cognitive fatigability.

2.3. Wearables and Heart Rate Variability (HRV) Metrics for Unsupervised Monitoring

The role of autonomic dysfunction in symptoms like fatigue has been challenging to assess due to the limitations of current assessment methods. Heart rate variability (HRV) is believed to play a role in autonomic dysfunction. Studies have already shown an association between HRV and fatigue [Patel et al., 2011; Tran et al., 2009]. However, there still needs to be more consensus in terms of what are the relevant metrics. Furthermore, HRV is mainly quantified using an electrocardiogram (ECG), which requires specialized equipment and personnel to conduct the observations. The Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology released a report to provide standardization in the research and application of HRV [Richardson et al., 1996]. HRV metrics are

extracted from the ECG signals. The largest-amplitude portion of the ECG signal is called QRS complex [McSharry et al., 2003] and corresponds to the depolarization of the right and left ventricles of the heart. Normal-to-normal (NN) intervals are the intervals between adjacent QRS complexes. HRV refers to the oscillation between consecutive heartbeats (RR intervals) and the oscillation between instantaneous heart rates. Studies including HRV metrics derived from off-the-shelf wearables devices as opposed to those derived from standard ECG holters are still scarce.

2.3.1. Validation Studies on HR and HRV Derived From Wearable Devices

Several studies show that wearable devices can accurately measure mean HR based on PPG [Jo et al., 2016; Parak and Korhonen, 2014; Wallen et al., 2016]. Others focus on assessing the accuracy of HRV metrics extracted from chest strap monitors. Many authors [Bailón et al., 2013; Hernando et al., 2018; Giles et al., 2016; Nunan et al., 2008] have used the Polar chest strap, which is an electrode-based sensor, in their studies. For instance, Giles et al. [2016] show that the Polar V800 is able to produce RR interval recordings consistent with an ECG during rest, and that the HRV parameters derived from these recordings are also highly comparable. Nunan et al. [2008] compare the number of RR intervals recorded by the Polar S810 and a standard 12-lead ECG monitor and found that both devices have a good agreement when the wearer is lying down. Additionally, they found good agreement between the derived HRV metrics. Hernando et al. [2018] explore the reliability of Polar RS800 to measure HRV metrics during exercise. Their work shows that at high exercise intensity, low-frequency domain measurements have excellent reliability indices. However, high-frequency measurements have a low agreement.

Photoplethysmography (PPG) sensors. Less common are studies that examine the validity of HRV metrics derived from off-the-shelf wearable devices with PPG technology. Giardino et al. [2002] found good agreement between the HRV metrics obtained from a finger plethysmograph and an ECG with three leads. Vescio et al. [2018] developed a customized device that converts the PPG signal generated by a LED-photodiode couple placed on the earlobe into electric pulses. Their device was tested under stationary conditions with 10 participants. Their results show good agreement with the ECG recordings. In a recent review about heart rate variability based on wearable devices, Georgiou et al. [2018] reviewed 308 articles, and from those, only two articles considered measuring HRV with wearable devices using PPG technology. Their research concludes that there is a need for more robust studies in non-stationary conditions, with appropriate methodology, acquisition and analysis techniques to evaluate the ability of wearables to measure HRV based on PPG [Georgiou et al., 2018].

In summary, previous research has shown that the interbeat intervals (IBI) derived from PPG signals are comparable to the RR intervals obtained from ECG Holter monitors under non-ambulatory conditions. However, little is known about the quality of the signal provided by off-the-shelf wearable devices and their capability to measure HRV [Georgiou et al., 2018]. In chapter 6, we address this open question by evaluating two devices (Everion and Empatica) under different conditions.

2.4. Smartphone-Based Health Monitoring

Smartphones have been popular for monitoring chronic conditions due to their ubiquity. Much of the previous work has focused on self-reporting apps to track the development and manage these conditions (e.g., El-Gayar et al. [2013] and Preuveneers and Berbers [2008] for Diabetes, Lakshminarayana et al. [2017] for Parkinson's disease). MS patients have described their interest in the use of mobile apps for tracking their condition. MS remote monitoring is gaining traction thanks to the use of digital technologies [Marziniak et al., 2018]. Ayobi et al. [2017] described that when individuals faced the unpredictable and degenerative nature of MS, they regained a sense of control by intertwining self-care practices with different self-tracking technologies. Giunti et al. [2017] presented a systematic review of MS health applications and could only find a small number of MS-specific applications compared to other equally prevalent diseases.

Pharmaceutical companies and mHealth. Large pharmaceutical companies have as well shown their interest in the use of mobile devices for tracking MS. Floodlight is a smartphone-based digital assessment tool developed by Roche and Genentech [Genentech, 2023]. The application offers a series of questionnaires and tasks aimed at monitoring disease progression. Some of Floodlights' tasks include (a) Hand-function monitoring through "Pinching Test" and "Draw a Shape Test," (b) cognitive monitoring through "Smartphone-Based Symbol Digit Modalities Test," and (c) Gait monitoring with "Five-U-Turn Test" and "Two-Minute Walk Test" [Montalban et al., 2019]. Similarly, Aby from Biogen [2023], is a mobile application that offers a variety of resources to support patients living with MS, such as informational videos and self-reporting diaries. Many of the existing applications to support MS patients focus on the tracking of symptoms and providing information about the disease [Genentech, 2023; Biogen, 2019].

2.4.1. Fatigue Monitoring With mHealth in MS

There are a number of mobile applications for managing MS-related fatigue [Jongen et al., 2015; D'hooghe et al., 2018; Babbage et al., 2019; Giunti et al., 2020]. More

Stamina [Giunti et al., 2020] is a mobile application for the self-managing of MS-related fatigue. The app acts as a to-do list where users can input their daily tasks. The user's energy is represented through a visual metaphor (progress bar) and a symbolic unit (Stamina Credits) for quantifying the estimated effort per activity. The app's goal is to facilitate patients' energy management. Jongen et al. [2015] introduced MSmonitor, a web-based program for self-management and care of MS patients. Their pilot study data suggests that using MSmonitor led to increased health-related quality of life and helped patients self-manage their fatigue. MS Energize [Babbage et al., 2019] is an iPhone app focused on the self-management of fatigue for MS patients. The app works as a coach supporting patients in their fatigue management. Similarly, D'hooghe et al. [2018] introduced MS TeleCoach, a mobile application offering telemonitoring of fatigue and telecoaching of physical activity and energy management in persons with MS. Results from their 12-week study indicate an improvement in the fatigue level of the participants measured through the FSMC. Existing mobile applications for fatigue monitoring measure fatigue using questionnaires.

2.4.2. Sensing and Fatigue Monitoring

The work by Sehle et al. [2011] aims to quantify MS-related fatigue by using kinematic gait analysis objectively. They found a correlation between physical measurements and subjective fatigue scales. Kim et al. [2010] proposed a real-time digital fatigue score (RDFS) to overcome the retrospective assessment introduced by questionnaires by actively querying patients four times a day through notifications. Tong et al. [2019], aimed at predicting MS patients' FSS scores using data from connected devices, background information, and daily questions at weekly intervals. [Yu et al., 2013] created a portable wireless system that can differentiate between fatigued MS patients and matched healthy controls. Results from this work show that MS patients have alternations of the heart rate (HR) and HR frequency depending if the test required cognitive or physical effort.

In summary, previous research on fatigue has focused on using dedicated sensors or devices to investigate the symptom. In some cases, only one sensor or device was considered [Kim et al., 2010; Sehle et al., 2011]. In other instances [Yu et al., 2013], multiple sensors were employed simultaneously to develop a measuring system for short-time assessments. Closer to our work is Tong et al. [2019], who uses a series of wellness interconnected devices to predict fatigue based on the FSS. We build upon this basis and the advances of technology to propose a system based on a single off-the-shelf multi-sensor wearable device and a dedicated smartphone application. Wearable devices are lightweight and small size making them easy to handle and allowing for unobtrusive, continuous monitoring over a long-term period of time. In chapter 7, we introduce our infrastructure and clinical study setting for remote and

unobtrusive monitoring of MS patients using wearables and provide initial results of the capabilities of these devices for fatigue quantification.

Part I.

Motor Fatigue

CHAPTER 3

Rapid Tapping on Smartphones to Assess Motor Fatigability

This chapter is based on the following publication:

Liliana Barrios, Pietro Oldrati, David Lindlbauer, Marc Hilty, Helen Hayward-Koennecke, Christian Holz, Andreas Lutterotti. A Rapid Tapping Task on Commodity Smartphones to Assess Motor Fatigability. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Pages 1–10. Hawaii, United States. April 25-30, 2020.

Motor fatigability is defined as *"an objective decline in strength as routine use of muscle groups proceeds"* [Dobkin, 2008]. Dobkin [2008] argued that fatigability could redefine our understanding of fatigue because many symptoms of fatigue may be a consequence of demonstrable fatigability, but this has rarely been assessed. Current clinically-used methods only evaluate fatigue retrospectively using questionnaires like the Fatigue Severity Scale (FSS). Although objective measurements to quantify fatigability have been proposed, none is sufficiently researched to be established in clinical routine [Severijns et al., 2017]. For example, tests using isokinetic dynamometers that measure peak isometric torque [Kalron et al., 2011] on the knee or hand, or measures of electrically-induced torque have been proposed [Skurvydas et al., 2011]. Those devices, however, are expensive and bulky and typically require professional supervision to perform the tests properly [Dobkin, 2008]. Finding ubiquitous and inexpensive ways to assess fatigability would enable optimized treatment options that currently lack objective outcome parameters to prove their efficacy. Furthermore, regular assessment of fatigability in clinical routine would not only

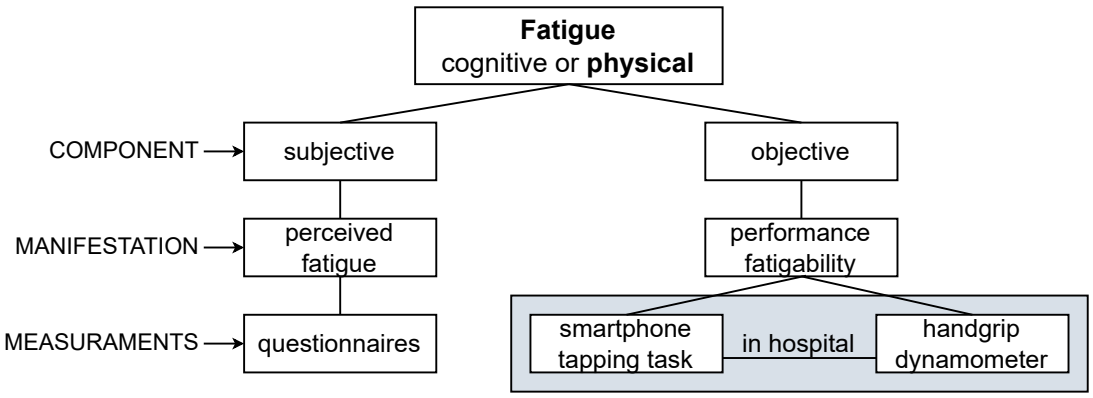

Figure 3.1.: State-of-the-art fatigue vs. fatigability. Fatigue is currently **only** measured by questionnaires which have several shortcomings like subjective and prone to recall bias. On the other hand, motor fatigability is often measured with a handgrip dynamometer. In this chapter, we conduct a controlled study (blue highlight) to evaluate the feasibility of establishing smartphone-based tapping as a valid physical fatigability method that overcomes the limitations of existing approaches relying on dedicated devices.

allow evaluation of disease progression but also add to the so-far limited options for quality of life measures.

In this chapter, we propose a commodity approach to measure fatigability. Our approach utilizes the prevalence of smartphones in conjunction with a simple tapping task, designed as an exertion technique to assess the user's motor fatigability. Finger tapping is commonly used to assess motor impairment [Prince et al., 2018; Taylor Tavares et al., 2005; Printy et al., 2014]. In our work, we re-purpose this task to quantify motor fatigability. The rapid tapping task requires barely any instructions (other than "please tap as fast as possible"), and can be performed on any commercially available smartphone. In an experiment with 20 MS patients and 35 healthy participants as the control group, we compare our approach with a standard fatigability assessment done with a handgrip dynamometer (Figure 3.1 blue highlight). Participants performed 500 alternating taps, which on average took patients roughly 2 min and healthy participants 75 s to complete. We show that participants' performance decreases during the tapping task and correlates ($\rho = 0.8$) with the decrease in grip strength measured with the handgrip dynamometer for patients and control. We further show that this correlation is also present in the first 30 s of performing the tapping task with $\rho = 0.78$ for patients and $\rho = 0.84$ for controls. Our results suggest that performing the simple tapping task for 30 s is sufficient to measure motor fatigability.

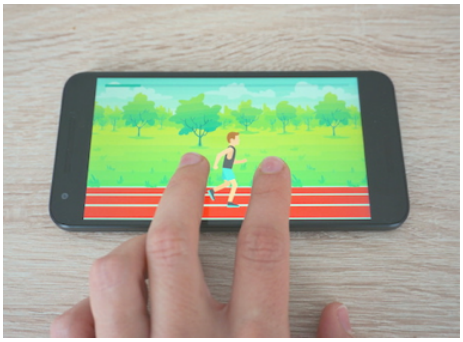

Figure 3.2.: User interface of tapping task used in study.

3.1. Assessing Fatigability Through Rapid Tapping

We aim to specify a task that can accurately quantify fatigability by meeting a set of requirements established by healthcare professionals (i.e., neurologists). The task should 1) be exhausting in terms of motor fatigue; however, it should not strain users' muscles for a prolonged time (i.e., enable quick recovery). The task should be 2) easy to learn and simple enough to perform without an experimenter present and 3) not require any specialized equipment to enable future in-the-wild studies. Lastly, it should 4) avoid the speed-accuracy trade-off, as described by Zhai et al. [2004]. For classical Fitts' law tasks such as pointing, users typically perform a task with high accuracy but slow (i.e., low exertion) or fast but with low accuracy. For many tasks, this means participants need to be well instructed to disregard errors and focus solely on speed. Even with clear instructions, participants might not completely disregard errors and thus might not perform the task as fast as possible. Since motor fatigability is measured when participants perform an exhausting task, other tasks, such as pointing, that are subject to this trade-off would not be well suited. We thus resorted to *rapid alternating tapping* as the task.

The simple user interface of the task is shown in Figure 3.2). Users perform alternating taps to complete the task. The avatar moves forward as users perform the task with speed depending on the tapping speed. A progress bar on the top indicates completion. We do not display any indication of accuracy to avoid the speed-accuracy trade-off. During our preliminary tests, we found that a goal of 500 alternating taps suffices to measure motor fatigability, as described in the Method section. We implemented the task on a commodity smartphone (Nexus 5X). However, porting it to other devices and operating systems would be trivial. To measure motor fatigability, we only require that the API has a measure for touch duration or time between taps.

Figure 3.3.: Apparatus of our experiment for the handgrip task (*left*) and tapping task (*right*).

3.2. Method

To analyze the validity of our exertion task as an indicator of motor fatigability, we compared our proposed tapping task with a standard handgrip dynamometer task performed by a control group and patients with MS. Each participant performed both tasks with their dominant and non-dominant hand. The ethics review board of the local University approved the study. MS patients performed the experiment at a local hospital under the supervision of healthcare professionals.

3.2.1. Participants

We recruited 35 participants as the control group (14 female, 21 male), ages 20–55 (μ = 31, σ = 7.7), all staff or students from a local university, and 20 MS patients (11 female, 9 male), aged 20–62 (μ = 43.1, σ = 12.9). The inclusion criteria for the control group included: no known or suspicions of illness, autoimmune disease, fatigue, or depression based on self-reports. MS patients were included if they had a confirmed diagnosis. The Expanded Disability Status Scale (EDSS) [Kurtzke, 1983] scores ranged from 0 to 8 (μ = 3 , σ = 2.5).

3.2.2. Apparatus

The control group performed the study in a quiet experimental room, sitting in a chair with armrests next to a desk, as shown in Figure 3.3. Patients performed the study in an examination room at the local hospital, also sitting in a chair with armrests next to a desk. Maximal voluntary handgrip contraction (MCV) was recorded using a digital Jamar handgrip dynamometer configured using the Jamar iOS tablet application. We used a Nexus 5X to run our fatigability application.

3.2.3. Design

We used a within-subject design with *Task* and *Hand* as independent variable with two levels each: *Handgrip* and *Tapping*; and *dominant* and *non-dominant* hand respectively. Order of *Task* and *Hand* was alternated, starting order was counterbalanced. Between each task, there was a resting period of three minutes to allow participant's muscles to recover.

3.2.4. Tasks

Handgrip

We used the handgrip dynamometer in the standard procedure to assess motor fatigability ([Severijns et al., 2017]). Participants sat upright, with both feet touching the ground and their forearms resting at a 90° angle on the armrests of the chair or the desk (see Figure 3.3). The dynamometer was held with the thumbs facing upwards in line with the forearm, and the grip size was adjusted to comfort. After a short period of familiarization, participants were asked to perform the MVC task for 30 s.

Tapping

Participants performed a rapid alternating tap on the smartphone screen while the hand was resting on a desk. The smartphone was placed in landscape mode. The exertion movement was performed with the index and middle fingers. Participants were asked to perform the tapping task as fast as possible without stopping until the app indicated completion. An initial assessment with six healthy participants performing 1500 alternating taps (naive about the end of the study) showed a clear decrease in performance after 500 taps. Hence, for the final experiment, participants were asked to perform 500 valid alternating taps, i.e., 250 taps per finger. A tap was considered valid and counted towards the goal of 500 taps if exactly one finger was on display.

3.2.5. Hypotheses

We performed the experiment concerning the following hypotheses. First, we expected a decrease in grip strength when using the handgrip dynamometer, as reported by previous studies[Severijns et al., 2017]. Secondly, we expected a decrease in performance as time progresses during the tapping task if the task is performed at maximal effort (speed). That is, users will take longer to alternate the fingers correctly on the screen, and the touch duration of each tap will increase with time. We

analyze our data concerning these hypotheses and explore the connection between the handgrip and the tapping task, to quantify and attribute rapid tapping to motor fatigability.

3.2.6. Procedure

Participants were briefly introduced to the setup and the experiment and completed a demographic questionnaire. Then, they completed a short training session for the tapping task, performing 40 alternating taps. Subsequently, they received instructions on how to use the handgrip dynamometer, including a demonstration by the experimenter. During the handgrip task, the experimenter instructed participants when to start and stop the MVC. After the introduction, participants completed all tasks with their dominant and non-dominant hands in counterbalanced order. They were asked to rest their arm and hand for three minutes between tasks.

3.2.7. Data Collection

During each 30-second trial of the handgrip dynamometer, we collected ten samples, which is the maximum sampling rate of the device we used. For the tapping task, we collected touch data from the smartphone using the Android API. We stored all timestamped touch-down coordinates and up events, from which we compute touch duration (i.e., how long did the finger touch the screen). Each sample in our dataset contains the finger position on the screen, touch duration, area size, and pressure. We define task performance for the tapping task as the average time participant's finger stayed on the screen (i.e., average touch duration). This means that the touch duration will be low for fast tapping (high performance), whereas the touch duration will increase for slow taps (low performance).

3.3. Results

In summary, our results show that the performance in the rapid tapping task correlates strongly with the fatigability measurements of the handgrip dynamometer. Tapping performance decreased significantly throughout a full trial (500 taps). Besides performance during a full trial, we analyze different subsets of the tapping task, precisely the first 10, 30, 60, and 90 seconds. For both groups, performing the tapping task for 30 s is sufficient to measure motor fatigability reliably.

3.3.1. Data Processing

We use touch duration as the primary performance metric to assess the tapping task. To account for outliers and noise in the tapping data, we performed the following data processing steps for each trial separately. We removed samples with a touch duration of more than three standard deviations away from the mean (1.2% for patients, 0.9% for healthy control). These outliers occurred when participants did not alternately lift their fingers but instead left one in contact with the screen and tapped only with the other. Outliers are evident, and thus we classified and removed them.

Tapping duration was low-pass filtered with a moving average of 20% of the trial data and normalized per participant and trial. We normalized tapping trials and handgrip trials separately, resulting in two motor fatigability slopes. The fatigability slope of the tapping task is positive (as duration increases), whereas that of the gripper is negative (as force decreases). To make trends comparable, our results are computed as *1 - normalized touch duration*.

To compare participants' performance in the handgrip task (10 samples per trial) and the tapping task (500 taps), we split the touch duration measurements into ten segments. Each segment contains samples from 10% of the total duration of the tapping trial (task). The final value for each segment is defined as the mean value of the data in that segment. To account for inertia when participants start both the handgrip and the tapping task, we discard the first segment and perform our analysis on the remaining nine segments.

3.3.2. Dominant vs. Non-dominant Hand

We analyzed the data from both tasks for participants' dominant and non-dominant hands. For the tapping task using the non-dominant hand, the data showed large variability. as shown in Figure 3.4 for patients.

In contrast to the dominant hand, the decrease in performance, while present, was less pronounced for the non-dominant hand. While data cleaning and statistical analysis as described in section *Performance Results* yielded a main effect for segments ($F_{8,117} = 10.592, p < .001$), Bonferroni adjusted Tukey's Post Hoc tests showed less statistically significant differences between segments as for the dominant hand. From observation, we believe this is due to challenges in coordinating the two fingers when performing the task. Since participants struggled to perform the task reliably, their speed decreased less. We, therefore, believe that the tapping task should be performed with the dominant hand. We thus performed all the following analyses on the data collected from dominant hand trials.

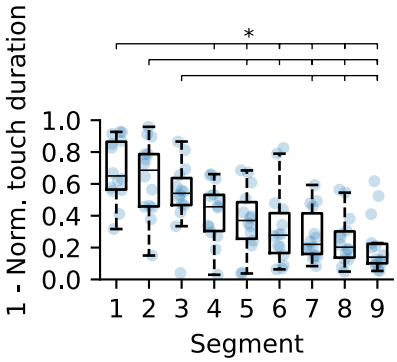

Figure 3.4.: Patients' full-duration tapping performance using their non-dominant hand showed larger variability than with the dominant hand.

3.3.3. Valid Trials

For a few trials, both groups (control and patients) did not follow instructions as expected when performing the handgrip and tapping task. The MVC task requires participants to evoke maximal potential from the beginning of the task. Hence, if participants maximally activate their muscles, no further increase in force is evoked during the task [Steens et al., 2012]. Similarly to Steens et al. [2012], we conservatively removed trials where participants failed at achieving maximal performance. For the 30-second handgrip task, we discarded the trials where consecutive measurements increased more than 50% of the maximum strength. This occurred in 2 of 35 trials for the control group and in 1 of 20 trials for patients.

To validate the tapping task, we fit a linear regression to the segment values and define trials as valid if 1) the slope of the regression is positive (i.e., touch duration increases), verifying an overall decrease in performance; and 2) consecutive segments do not have a duration decrease of more than 50% (i.e., participants' performance increases). Trials that fail these requirements suggest that participants did not perform the task as fast as possible, meaning they did not evoke maximal performance. The number of discarded trials for this condition depends on the analyzed time frame (Figure 3.9, *bottom*).

Analyzing the fully completed task, 63% and 90% of trials were valid for patients and the control group, respectively. Restricting the window to the first 30 seconds of the task, however, resulted in 84% valid trials for patients and 93% valid trials for the control group. For shorter durations (e.g., 10 seconds), not enough data is available to quantify fatigability accurately. For 60 seconds or longer, participants seem to pace themselves, recover, and then speed up again. We thus believe that the first 30 seconds of the tapping trial represent a suitable excerpt to assess motor fatigability.

3.3.4. Performance Results

We performed individual ANOVAs on the handgrip and tapping data with *segment* as independent variable (9 levels) for both the control group and patients. Statistically significant differences between segments demonstrate an actual, non-random decline in performance during a task. For the tapping task, we found a main effect of *segment* on average duration for the control group $F_{8,279} = 50.918,\ p < .001$ and for patients, $F_{8,99} = 14.211,\ p < .001$. To analyze the temporal progression of participants' performance, we performed a series of Bonferroni-corrected Tukey's Post Hoc tests. Results are illustrated in Figure 3.5 for patients and Figure 3.6 for the control group. For the control group, segments are mostly significantly different from segments after the subsequent one. Only after Segment 7, average performance flattens, and subsequent segments are no longer significantly different. Results for patients show a similar pattern. However, most segments are not significantly different from their direct successor, but 2 or 3 segments thereafter. Performance flattens after Segment 6. This decline in performance indicates that both tasks can successfully invoke motor fatigue for both the control group and patients.

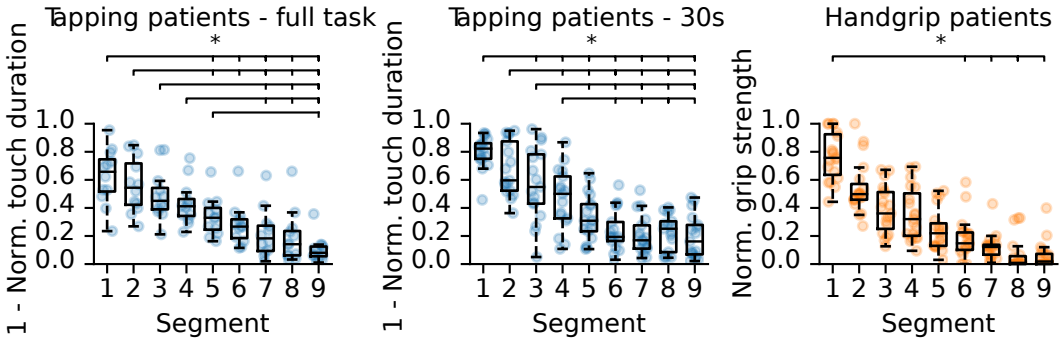

Figure 3.5.: Patient group data: full tapping task (*left*), the first 30 seconds (*center*), and the handgrip task (*right*).

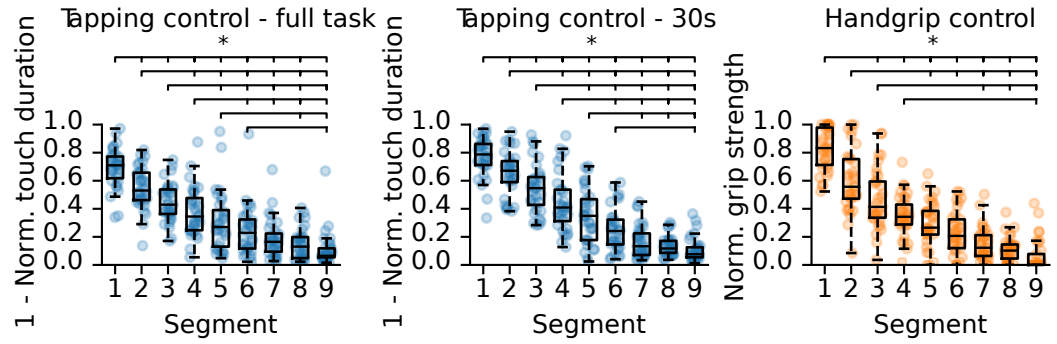

Figure 3.6.: Control group data: full tapping task (*left*), the first 30 seconds (*center*), and the handgrip task (*right*).

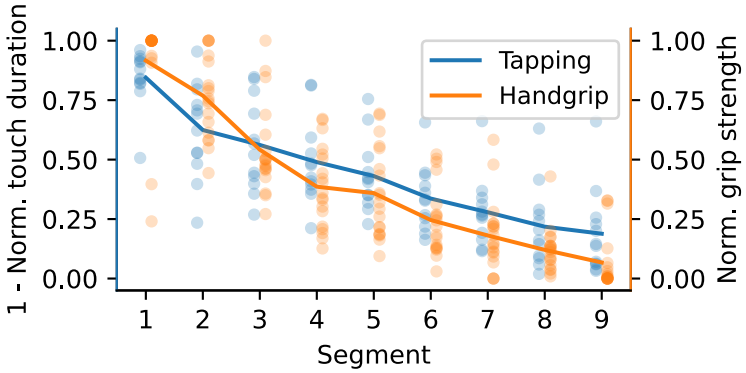

Figure 3.7.: Complete task recordings of the **patients group** for handgrip and tapping. The solid line indicates each segment's mean value.

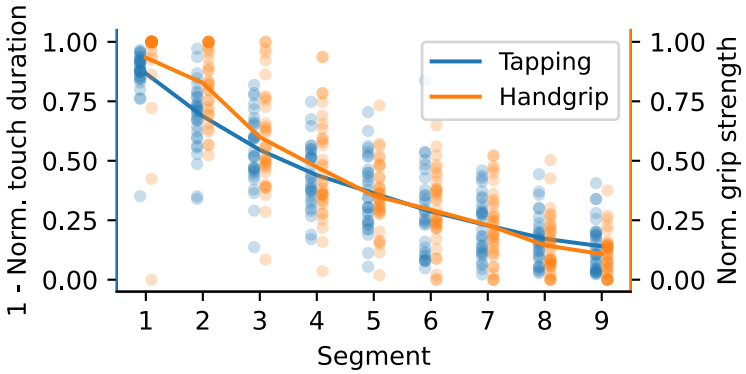

Figure 3.8.: Complete task recordings of the **control group** for handgrip and tapping. The solid line indicates each segment's mean value.

3.3.5. Handgrip vs. Tapping

To verify the relation between the handgrip and tapping task, we computed their Spearman's rank correlation coefficient ρ. Figure 3.7 illustrates the patient data, and Figure 3.8 shows the measurements of the control group. For both, the average correlation coefficient is $\rho = 0.8$.

Trial Duration

Each trial took on average 75 s for the control group (σ = 23.6 s) and 126.1 s for patients (σ = 81.4 s). To determine the optimal number of taps per trial leading to comparable results, we performed the same analysis as before on the first 10 s, 30 s, 60 s and 90 s of the recordings data as shown in Figure 3.9.

During the first 30 s, participants of the control group performed on average 249.1

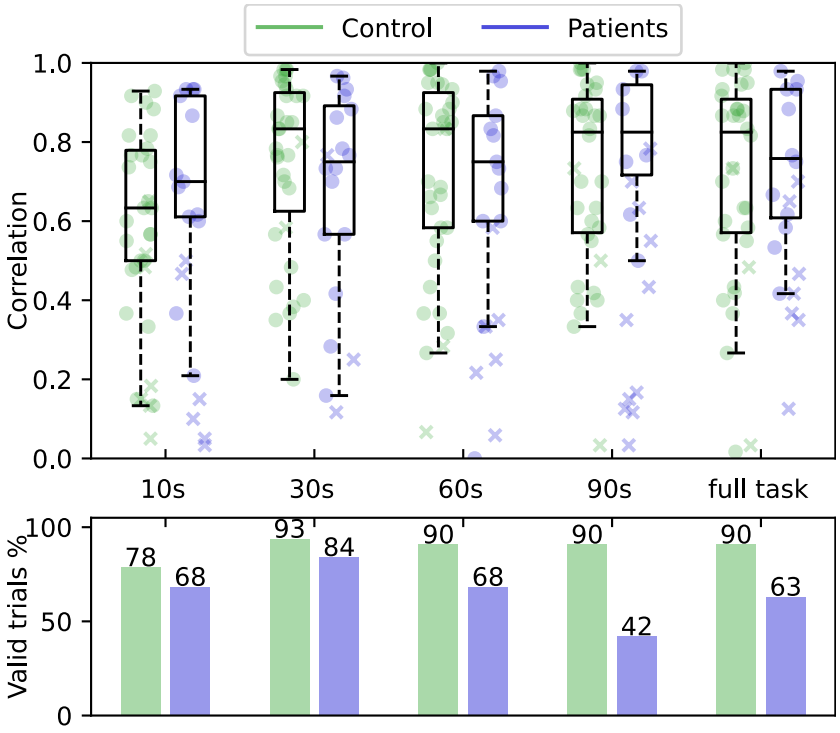

Figure 3.9.: (Top) Spearman's correlation between *1 - normalized touch duration* and handgrip (*top*) by task duration. Crosses represent invalid trials. (Bottom) The bar chart shows the percentage of valid trials.

taps (σ = 56.6 taps), while patients performed on average 170.6 taps (σ = 60.3 taps, leaving a large number of data points for analysis. We performed similar processing on the data (discarding outliers and invalid trials), but only used the first 30 seconds, and split them into 10 segments.

We performed the same analysis as with the full task duration. That is, two ANOVAs with *segment* as independent variable and touch duration as dependent variable, one for the control group and one for the patients. We again found a main effect of *segment* on the data of the control group $F_{8,279} = 79.606$, $p < .001$ and the patients $F_{8,144} = 28.39$, $p < .001$. A series of Tukey's Post Hoc tests revealed a similar pattern between segments as with the full data (Figure 3.5 and Figure 3.6). The statistically significant differences between the segments of the trials confirm that the non-random decline in performance is also present during a short task of 30s. Analyzing the correlation between the handgrip and this shortened tapping task revealed a correlation between the two tasks of $\rho = 0.78$ ($p < .001$) for patients and $\rho = 0.84$ ($p < .001$) for the control group. The similar correlation score indicates that a rapid tapping task of 30 seconds suffices to measure motor fatigue.

Finally, we analyze the agreement of our tapping task and handgrip dynamometer by comparing the rate of fatigue development captured by each method. Similarly

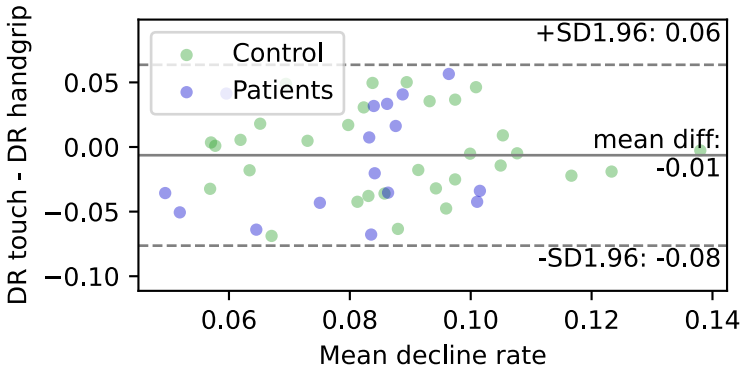

Figure 3.10.: Bland-Altman plot for mean decline rate (DR) of normalized touch duration (30 sec) and normalized handgrip strength shows a mean bias of -0.01 with LoA [0.06,-0.08].

to Lou et al. [2003], we use the slopes of the regression line of touch duration and handgrip strengths to assess fatigue rate. Figure 3.10 shows the Bland-Altman plot comparing the decline rate of the tapping task and the handgrip dynamometer. The plot shows no particular pattern in the data. The mean difference of almost zero (0.01) and all data within two within two standard deviations from the mean with limits of agreement between (LoA) [-0.08, 0.06] confirms the good agreement between both approaches. The normality of the differences was verified using the Shapiro-Wilk test ($p = .08$)

3.4. Discussion

The analysis of our evaluation showed that our simple rapid tapping task can be used to quantify motor fatigability. The task is easy to implement, runs on unmodified commodity smartphones, and, more importantly, the task is easy to perform for users. We thus believe that our method will allow moving beyond specialized hardware (e.g., handgrip dynamometers) and subjective feedback to make assessing motor fatigability ubiquitous and more accessible for all patients.

Through our experiment, we examined the appropriate task length to quantify motor fatigability using a tapping task. Our initial target was 500 taps, which resulted in varying completion tasks for participants. On average, patients took 168% longer than healthy controls to complete the task, which was not unexpected. The slowest patient completed the taps in 7.68 minutes, while the slowest healthy participant took 2.29 minutes. The variance in completion times shows the importance of limiting trial durations because performing the tapping task for up to 7 minutes causes physical strain and makes motivated compliance challenging. The results of our evaluation show that analyzing the first 30 seconds of our rapid tapping task is a suitable

assessment of motor fatigability, which may be important to enable more frequent and, ideally, continual monitoring in a straightforward manner.

The time to complete the tapping task for patients was partly governed by the severity of their condition as measured by the EDSS scores. Patients with higher EDSS scores tended to take longer. We did not, however, observe differences between patients in terms of measured motor fatigability with either task. This, however, needs to be investigated further since we need more patients for each score to perform reliable statistical analysis on this data. Preliminary results with three groups of MS disability based on the EDSS show these Spearman correlations: EDSS = 0, N = 4: .75, EDSS in [1,3], N = 4: .85, EDSS in [4,8], N = 8: .77; all $p < .001$

Other metrics for tapping performance

We initially investigated the suitability of alternative metrics to evaluate tapping performance, such as the number of taps per segment, tap pressure, and area size. The average duration between taps was noisier due to the occurrence of simultaneous or quasi-simultaneous taps (which we counted as invalid). We, therefore, decided to use touch duration as the primary performance metric. The number of taps shows slightly lower correlations than touch duration and a comparable number of invalid trials. Hence it is also a suitable metric. Analyzing pressure and area size, we found a low correlation with the handgrip measurements, possibly because participants' finger placement on the touchscreen is too person-specific.

3.5. Limitations and Future Work

Even though participants performed a training session before measurements were taken, some still did not start their trials with maximum speed. Instead of exhibiting fatigue, some participants showed an increase in performance, which resulted in a limited number of invalid trials (16% for patients, 7% for the control group). This indicates that while the task is generally well suited to measure fatigability, further interventions are needed to ensure that participants follow instructions closely. Furthermore, this highlights the importance of incorporating outlier removal in the computation of the fatigability metric.

We see potential in offering incentives to complete the tapping task with maximum effort, such as by further gamification or using scoring systems. We believe that such measures would decrease the number of outliers and potentially eliminate the need for dedicated outlier removal. The percentage of invalid trials and how this might vary under different environments and without supervision needs further investigation. We plan to explore other methods and analysis strategies to ensure

higher rates of valid trials. Additionally, intrinsic motivation is needed to perform the tapping task regularly, and we cannot estimate potential learning effects so far. Extending our research to longitudinal in-the-wild evaluations with within-subject comparisons will allow us to assess the use of fatigability to judge disease progression in MS populations. Moreover, comparative tests are needed to discriminate between fatigability and disability. We plan on using the 9-Hole Peg Test [Mathiowetz et al., 1985] to assess patients' fine motor skills and use the fatigue scale for motor and cognitive functions (FSMC) to categorize mild and severe fatigue in patients. Cognitive fatigability measure (e.g., the N-Back test) can help discriminate cognitive and motor fatigability.

3.6. Conclusion

We introduced a novel approach to assess motor fatigability on a commodity smartphone using a simple rapid tapping task. Our experiment with 20 multiple sclerosis patients and 35 healthy participants showed a significant correlation between the tapping tasks and grip strength measurements from a special-purpose handgrip dynamometer. We believe our work is a first step towards measuring motor fatigability without relying on specialized equipment, which can be expensive and require professional supervision. Our method may help quantify fatigue and complement the current use of subjective feedback through questionnaires, enabling patients to frequently and ubiquitously monitor their condition and react to changes accordingly.

CHAPTER 4

Rapid Tapping on Smartphones and its Association to Fatigue – In the Wild

This chapter is based on the following publication:

Liliana Barrios, Pietro Oldrati, Marc Hilty, David Lindlbauer, Christian Holz, Andreas Lutterotti. Smartphone-Based Tapping Frequency as a Surrogate for Perceived Fatigue. An in-the-Wild Feasibility Study in Multiple Sclerosis Patients. *The Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*. Volume 5, Issue 3, Article No.: 89, pp 1-30. September 2021.

In this chapter, we continue the focus on fatigue's physical aspect. As mentioned in the previous chapters, motor fatigability is typically measured through walking (e.g., 6-minute walking test [Goldman et al., 2008]), handgrip strength [Severijns et al., 2014], or using a knee dynamometer [Surakka et al., 2004]. These approaches have important limitations: 1) the requirement of expensive clinical equipment and trained professionals to conduct the tests, and 2) the restriction to a medical facility, and, consequently, these assessments are conducted only at a few or single time points. In chapter 3, we proposed a rapid tapping task on a smartphone as an inexpensive approach to assessing motor fatigability. Our controlled evaluation demonstrated a high correlation between smartphone tapping and fatigability measurements obtained with a handgrip dynamometer. However, studies on the association between fatigability measured by the tapping task and perceived fatigue are not available. Hence, it is still unknown if 1) a smartphone-based tapping task is a feasible proxy for fatigue and if 2) the task is valid in uncontrolled environments. Hence, in this chapter, we seek to investigate both of these questions (Figure 4.1).

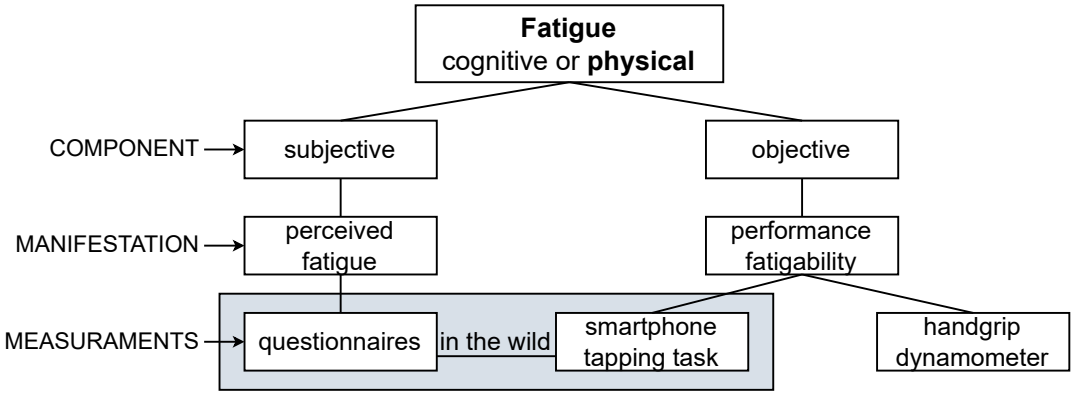

Figure 4.1.: Motor fatigability can be measured objectively with rapid tapping or a handgrip dynamometer. The association between fatigue and fatigability is not established yet. Hence, we conduct an empirical in-the-wild study (blue highlight) to evaluate the feasibility of using rapid tapping on a smartphone as a surrogate for fatigue.

This chapter's focus is two-fold: (I) To develop a new objective and reliable measure of motor fatigability computed from raw tapping data and demonstrate its usability and validity when performed outside controlled settings and without medical supervision (in-the-wild). We approach this goal by using the tapping task introduced in the previous chapter and conducting a two-week in-the-wild study with 35 MS patients. Participants performed a 30 s tapping task (i.e., a trial) once per day during the two weeks. Using this data, we introduce a new metric to assess motor fatigability: *tapping frequency*. We show that our new metric is a valid method to assess motor fatigability in-the-wild by comparing it to the previous approaches, specifically *touch duration* and strength decline using a handgrip dynamometer. (II) To evaluate the feasibility of establishing an objective and ubiquitous method as a surrogate to quantify perceived fatigue. To this end, we evaluate the performance of our proposed metric, *tapping frequency*, to classify fatigued and non-fatigued patients using *ROC* (receiver operating characteristic) curves and area under the *ROC* curve (AUC_{ROC}). We quantified perceived fatigue during the study with two widely accepted and validated fatigue questionnaires in MS patients: Fatigue Severity Scale (FSS) [Krupp et al., 1989] and Fatigue Scale for Motor and Cognitive Functions (FSMC) [Penner et al., 2009].

Our goal is to develop an objective metric to be used when monitoring patients with fatigue in medical routine or clinical trials, which until now has been hampered by the heterogeneity and subjective nature of questionnaires [Friedman et al., 2010]. We believe our results are an important step in understanding fatigue. Our smartphone-based tapping technique and evaluation metric have the opportunity to be used regularly by patients outside the clinic and more frequently than currently done

in the medical routine. Being an objective method, it also opens the potential for quantifying the direct effects of therapeutic interventions, which is a clear advantage over currently used questionnaires [Nourbakhsh et al., 2021].

4.1. Methods

To analyze the feasibility of establishing smartphone-based objective metrics as a surrogate for fatigue, we conducted a two-week in-the-wild study. We use the FSMC [Penner et al., 2009] to discriminate between motor-fatigued and non-motor-fatigued participants and the FSS [Krupp et al., 1989] to differentiate fatigued and non-fatigued participants. As a motor fatiguing task, we use the tapping task introduced in the previous chapter (Chapter 3). Participants performed the tapping task with their dominant hand each day of the two-week study. Through AUC_{ROC}, we evaluate the performance of our smartphone-based metrics to rank fatigued vs. non-fatigued participants in relation to the FSMC and FSS. Participants could exit the study at any point or continue for longer if desired. The local state ethics review board approved this study.

4.1.1. Participants

We recruited 35 MS patients at a specialized MS clinic (20 female, 15 male), aged 21–53 (M = 36.77, SD = 8.93). All MS patients had a confirmed diagnosis, signed written informed consent, and had Android smartphones. Seven of the 35 MS patients had hand impairments according to the Nine-Hole Peg Test (9-HPT) threshold (cf. section 4.1.2). The Expanded Disability Status Scale (EDSS) scores ranged from 0 to 6 (M = 2.31, SD = 1.7) and were obtained from the MS clinic at the beginning of the study.

4.1.2. Tasks and Baselines

Our study started with an on-boarding, during which we explained the study protocol to the participants. We also collected normative outcome measurements using the FSMC, Nine-Hole Peg Test (9-HPT), and handgrip dynamometer. Additionally, we asked participants to install our Android application on their smartphones. Our application included the tapping task. Furthermore, it sent daily notifications to the participants to remind them to complete the tapping trials during the in-the-wild study and to complete the FSS questionnaire directly in the app once per week.

Figure 4.2.: Study methods: smartphone-based fatigability task on the left, nine-hole peg test centered, and handgrip dynamometer on the right.

Clinical Baseline Methods

We use the 9-HPT to objectify hand function, the FSS and FSMC to categorize fatigue and a handgrip dynamometer as standard motor fatigability measurement. Neurological impairment was measured using the standard disability rating scale for MS patients (EDSS) [Kurtzke, 1983]. The EDSS scale ranges from 0 to 10 in steps of 0.5, with higher values representing higher disability levels.

- *Nine-hole Peg Test (9-HPT).* The 9-HPT is a standardized, quantitative assessment used to measure finger dexterity [Mathiowetz et al., 1985]. Figure 4.2 (middle) shows an image of the 9-HPT used in this study. Participants were asked to remove the pegs, one by one, from the container to the holes and then to place them back into the container using their dominant hand. As the final score, we used the average of the two trials. Patients with a total time greater than 23.17 s (normative value used at the local hospital derived with a standard procedure [Bertoni et al., 2015]) were classified as hand-impaired.
- *Handgrip Dynamometer.* We used sustained handgrip strength as a metric to assess motor fatigability [Severijns et al., 2017]. The test was conducted upright, with both feet on the ground and forearms resting on an armrest. The dynamometer was held with the thumbs facing upwards in line with the forearm, and the grip size was adjusted for comfort. Participants performed maximum voluntary contraction (MVC) for 30 s. The experimenter instructed participants when to start and stop the MVC. Maximum contraction in kilograms was recorded every 3 s for a total of 30 s, resulting in 10 consecutive measurements. Figure 4.2 (right) depicts the Jamar device we used.
- *Fatigue Scale for Motor and Cognitive Functions (FSMC) [Penner et al., 2009].* FSMC is used to assess MS-related cognitive and motor fatigue. The questionnaire consists of ten items corresponding to the cognitive sub-scale and ten to fatigue's physical aspects. Participants rated each of the items on

a 5-point Likert-type scale consisting of the following: (1) "Does not apply at all," (2) "Does not apply much," (3) "Slightly applies," (4) "Applies a lot," and (5) "Applies completely." FSMC offers cut-off values that determine fatigue levels in different aspects (general, cognitive, and physical). With the cut-off values, it is possible to rate the level of fatigue as mild, moderate, or severe. Appendix A.2, Table A.2 shows the different cut-off values for the distinct aspects of fatigue according to FSMC. Participants completed the FSMC before and after the two-week study. We used as the final score the mean of both completed questionnaires. In this study, we only focus on the physical aspect of fatigue, as the tapping task is a measurement of motor fatigability. We label participants as non-fatigued if their FSMC physical score is less than 22. Otherwise, they are considered fatigued. Appendix A.2 shows the items of the FSMC questionnaire.

- *Fatigue Severity Scale (FSS) [Krupp et al., 1989].* FSS is a widely-used questionnaire to assess fatigue in various diseases [Krupp et al., 1989; Valko et al., 2008]. The questionnaire consists of 9 questions about how fatigue interferes with the patient's activities. Patients rated the items on a 7-point Likert scale with values ranging from 1 = "strongly disagree" to 7 = "strongly agree." Higher scores indicate greater fatigue severity. The FSS final score is the mean of all items. We classified scores larger than 3.8 as fatigued participants. The FSS has no defined threshold to identify fatigued participants. Thresholds are usually defined depending on the study needs [Armutlu et al., 2007; Kaynak et al., 2006; Valko et al., 2008]. A score of 4 or higher is commonly used to identify severe fatigue [Armutlu et al., 2007; Kaynak et al., 2006]. For the FSMC, we chose to use the lower threshold (mild fatigue). Hence for the FSS, we chose 3.8 as a threshold, representing a more conservative score than the commonly used for severe fatigue. Using this threshold, we identified a correlation of $\rho = 0.85$ ($p < 0.0001$) between the FSS and FSMC scores which goes in line with the findings of Penner [Penner et al., 2009]. Through our mobile application, we reminded participants to complete the FSS questionnaire once per week. We used the mean of the completed questionnaires as the final score. Refer to Appendix A.1, Table A.1 for the complete FSS questionnaire.

Rapid Alternating Finger Tapping

Participants performed rapid finger tapping on their smartphone's screen with their dominant hand. We asked them to keep their hand resting on a flat surface while doing the task with the smartphone set to landscape mode. Figure 4.2 on the left shows an image of the tapping tasks. The exertion movement required participants to engage the index and middle fingers. Participants were asked to complete tapping

trials at their maximal performance (maximal speed) for 30 s. They had to tap as fast as possible without stopping until the app indicated completion. The application used was introduced in the previous chapter, with the only change of having a stop condition of 30 s. Furthermore, it did not offer immediate feedback to users if trials were conducted as expected.

4.1.3. Study Design

Our study included two phases: one in the hospital, and the other in-the-wild, highlighted in dark blue and light blue , respectively, in Figure 4.3.

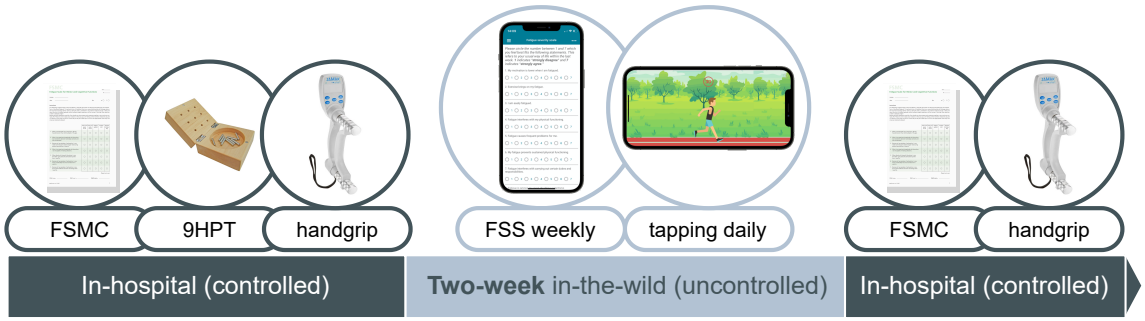

Figure 4.3.: Study design timeline with two phases: the hospital phase (dark blue) to gather baseline measurements, and in-the-wild phase (light blue), the core of this study. During the in-the-wild phase, participants complete tapping trials daily and the FSS questionnaires once per week. Pre and post in-hospital baselines and questionnaires were average to get the final scores.

In-hospital

We included a pre-and post-study phase, both guided by a healthcare professional. During the pre-phase, participants were briefly introduced to the study and completed the FSMC and a demographic questionnaire. We guided participants through installing and using our Android application on their smartphones. They also received instructions on how to complete the tasks, including a demonstration by the experimenter and a short familiarization session for each of the tasks: handgrip, tapping task, and 9-HPT. Participants completed all tasks with their dominant hand in counterbalanced order. Between tasks, participants rested their arm and hand for three minutes. During the post-phase, as shown in Figure 4.3, we collected the FSMC scale and handgrip measurements again. we averaged pre- and post-measurements to obtain the final scores. By combining two measures, we seek to reduce outliers and get more reliable baselines.

In-the-wild

The in-the-wild experiment started immediately after the initial in-hospital phase. We asked patients to complete tapping trials once daily after receiving the reminder notification. We did not set a specific time to complete the trials. Instead, we allowed notifications to be set randomly during the day to achieve higher fluctuations in the person's energy level. All trials conducted in-the-wild were completed with no supervision. Once per week, participants received a notification for completing the FSS questionnaire. The study duration was two weeks.

4.1.4. Data Collection

During each 30 s trial of the handgrip dynamometer, we recorded ten samples, which is the maximum sampling rate of the Jamar device we used. For the 9-HPT, we recorded participants' time (seconds) to complete the task. We recorded all touch events on the participant's smartphone throughout the tapping task using the Android API. We stored all timestamped touch-down coordinates and up events, from which we compute touch duration (i.e., how long did the finger touch the screen). Additionally, we computed the *tapping frequency* per second (i.e., number of taps recorded within a 1 s window), as a new feature to quantify fatigability. We define task performance for the tapping task in terms of the tapping frequency. This means tapping frequency will be high for fast tapping (high performance), whereas tapping frequency decreases for slow taps (low performance). During our analysis, we also incorporate the metrics introduced in Chapter 3 touch duration and its slope. Touch duration has the opposite behavior of tapping frequency. During a high performance, touch duration decreases and increases as performance decays (i.e., taps become slower). FSS scores were stored on the participants' phones.

4.1.5. Hypotheses

We analyzed the data concerning the following hypotheses:

1. We expect to find a comparable decrease in performance between the handgrip task and the in-the-wild tapping trials (tapping frequency), similar to the findings with touch duration (Chapter 3).
2. We expect to see a difference in tapping performance when comparing fatigued and non-fatigued participants suggesting an association between tapping performance and perceived fatigue.
3. We expect the smartphone-based tapping task to be feasible and provide valid results when conducted in unsupervised settings in-the-wild.

To verify H1, we compared the tapping performance against the handgrip performance through correlation. However, as this is an in-the-wild study, we compare each tapping trial with the mean handgrip of pre-and-post in-hospital phases. We verify H2 by evaluating features derived from the tapping task as a metric to classify fatigued and non-fatigued participants, according to the FSS and FSMC, using *ROC* (receiver operating characteristic) curves and area under the *ROC* curve (AUC_{ROC}). Finally, for H3, we use our validity algorithm to verify the trials in-the-wild.

4.1.6. Data Processing Pipeline

We use *tapping frequency* as the primary performance metric to assess motor fatigue with the smartphone tapping task. We define tapping frequency as the total number of taps registered during one second. Hence, we compute our feature with a one-second sliding window. We aim to monitor patients reliably in-the-wild. Thus, our data processing pipeline needs to handle noise and invalid tasks. An important difference to Chapter 3's data handling is that it did not include gap verification or handling, as that study was fully controlled and under supervision. Consequently, it was less prone to be affected by data gaps. However, the study described in this chapter is fully unsupervised. Hence, the relevance is verifying the data quality before conducting any analyses. Gaps can occur when a participant gets distracted by an incoming phone notification, call, or external factors. Our data processing pipeline includes three steps: (1) gap removal, (2) task validity, and (3) feature extraction.

1. Gap removal. We identify gaps within a tapping trial when no input is recorded on the smartphone's screen for over 843.5 ms. This threshold represents the 0.999th quantile of the time differences between consecutive taps in our in-the-wild dataset. We did not incorporate automatic gap detection in the app, as it would imply that we had prior knowledge of the tapping frequency of MS patients, which was not the case. Moreover, by setting a threshold without knowing how hand impairment could affect tapping, we could have erroneously stopped trials of participants with motor impairment. From our dataset, we have seen that gaps can occur at any time within a trial. If the gap occurs during the first half of a trial, we move the trial's start to after the gap. If the gap occurs after the second half of a trial, we move the trial's end time to before the gap occurs. We repeat this process recursively until all gaps within a trial have been removed. When removing gaps, we verify that the final trial's length is at least 27 s to ensure enough data for analysis. Shorter tapping trials are classified as invalid.

2. Tapping trial validity. We validate individual tapping trials by verifying that they are completed at maximal performance. First, we derive a continuous time

series at a constant sample rate and apply a 2^{nd} order Butterworth low-pass filter with a cutoff frequency of 0.5 Hz. Then, we proceed to find the time of maximal performance (i.e., maximal tapping frequency). We use a low-pass filter to avoid detecting outliers within the tapping frequency. Since the initial 3 seconds of tapping contains inertial behavior, we do not consider them when extracting the peak's performance time.

Two conditions must hold to verify sustained maximal performance during a tapping task: (1) The peak of maximal performance should occur during the first half of the task (i.e., the maximal tapping frequency should occur before 15 s). Later peaks in performance indicate that the person failed to start the task at maximal speed. (2) After the peak of maximal performance, we expect a negative slope in the tapping frequency data. To verify this condition, we fit a line to the tapping frequency data, taking the time of maximal peak performance as the task's start time. Following, we extract the line's slope. After the maximal peak in performance, a positive slope indicates that the person failed to perform maximal performance from the beginning of the trial. Figure 4.4 depicts examples of different cases of trial validity: invalid slope (left), invalid maximal performance location (center), and valid trial (right).

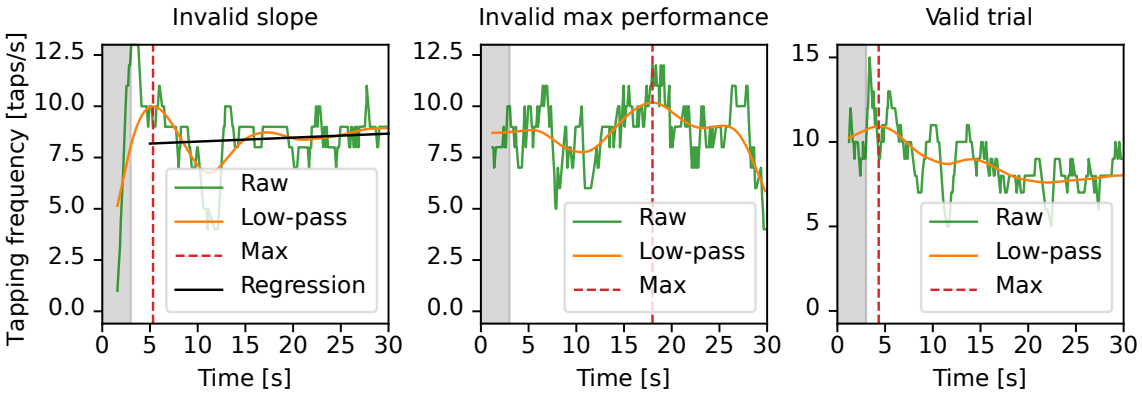

Figure 4.4.: A trial is invalid when the regression slope of the tapping frequency after the maximum is positive (*left*), or when the maximum of the filtered tapping frequency occurs after 15 s (*center*). Otherwise, the trial is considered valid (*right*), meaning the trial was completed at maximal performance. The first three seconds of the trial, depicted in grey, are discarded to avoid the influence of the initial inertia.

3. Feature extraction. We compute a set of features to evaluate the trial's performance. In particular, we compute the slope of the touch duration, the slope of the tapping frequency, and the mean and maximum tapping frequency.

4.2. Results

Our results show an association between the smartphone-based tapping performance metrics and perceived fatigue measured with the FSMC and FSS. We found a statistically significant difference between fatigued versus non-fatigued participants' performance. The difference between both groups (fatigued and non-fatigued) is significant during the whole study, indicating that the approach is valid in-the-wild and without supervision. Additionally, our new data processing pipeline and core metric, tapping frequency, increase the validity of trials by 19% in comparison to the approach presented in Chapter 3.

4.2.1. Tapping Frequency as a Valid Motor Fatigability Metric

We propose tapping frequency to quantify motor fatigability using the tapping task on commodity smartphones. We validate our approach by comparing it to two accepted fatigability methods: handgrip dynamometer and touch duration (see Chapter 3) on our in-the-wild dataset. We compute correlations between the mean handgrip and each single tapping trial at the participant level and report the combined correlations. We applied the data processing pipeline and feature generation as described in Section 4.1.6. We split the tapping data into ten segments to compare it to the handgrip dynamometer's ten measurements. Next, we discard the first data segment to account for inertia. We perform min-max normalization on the segmented data instead of standardization before computing the segments.

Using our trial validity definition, we classify 87% of the in-the-wild participants' trials as valid (Figure 4.6), which is a 19% increment compared to *touch duration* (Chapter 3) using our in-the-wild dataset. The correlation to the handgrip is comparable in both approaches. We used Spearman's correlation and obtained the following values, for touch duration $\bar{\rho} = 0.80$, CI: [0.39, 0.98] ($p < 0.05$ for all except 5 participants), and for our new approach (tapping frequency) $\bar{\rho} = 0.83$, CI: [0.54, 0.99] ($p < 0.05$ for all except 2 participants).

4.2.2. Fatigue Scores' Distribution

The FSMC and FSS score distributions of our study population are depicted in Figure 4.5. As tapping is a motor task, we focus our analyses on the physical aspect of the FSMC questionnaire. Following the FSMC cut-off values, we classified 18 patients as fatigued and 17 as non-fatigued. Two of the 35 patients did not complete a single FSS questionnaire. The FSS questionnaire was intended to be completed during the in-the-wild phase of the study. From the 32 patients who completed the FSS survey, we classified 17 patients as fatigued and 15 as non-fatigued. We observe

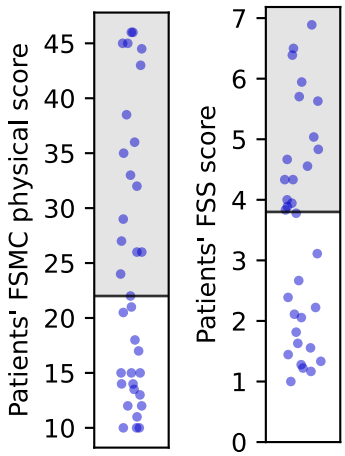

Figure 4.5.: Average physical FSMC and FSS scores of our study population. In grey, we depict the scores that we considered as *fatigued*.

a correlation of $\rho = 0.85$ ($p < 0.0001$) between the FSS and FSMC scores. With the conservative threshold of 4.0 for the FSS, 12 patients would classify as fatigued and 20 as non-fatigued.

4.2.3. Completed Trials and Validity

We collected a total of 487 tapping trials from 35 patients during our in-the-wild study. From those, 70 trials were classified by our validation algorithm as invalid. Figure 4.6 shows the valid and invalid trials per participant. One participant had more than half of the trials labeled as invalid (Figure 4.6 "Discarded patient"). We decided to discard data from this participant as the medical examiner noted during the in-hospital session that the participant had very long artificial nails that prevented them from tapping correctly. This resulted in a dataset of 34 patients with 473 tapping trials, of which 61 are labeled as invalid. The rest of the patients completed the study and achieved at least eight valid tapping trials during the whole study. The average validity during the study was 87% (min = 57.0%, max = 100.0%)

4.2.4. Tapping Frequency Outperforms Handgrip Strength When Analyzing Fatigue

We computed a series of non-parametric Kruskal-Wallis H-tests [Field and Hole, 2003] to identify statistically significant differences between fatigued and non-fatigued participants in terms of mean tapping frequency and handgrip strengths. The results are summarized in Figure 4.7. Following previous findings, we expect fatigued patients to show lower handgrip strength than non-fatigued patients [Nacul

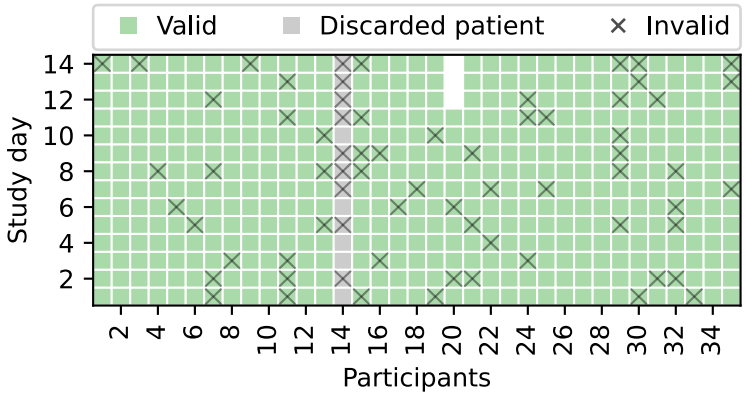

Figure 4.6.: Total tapping tasks completed per participant over the two-week study. One participant was discarded for having more than half of the trials invalid. 87% of trials were labeled as valid.

et al., 2018]. Figure 4.7 (top left) shows the mean tapping frequency distribution of the study population according to the FSMC classification, averaged over all the valid trials of the patients. We observe a statistically significant difference in tapping frequency comparing the fatigued and non-fatigued group with $H = 7.50$ ($p < 0.01$). However, there is no statistically significant difference in the mean handgrip strengths (Figure 4.7, bottom left).

While mean handgrip is confounded by gender, tapping frequency is not. Mean tapping frequency in female participants shows a significant difference between fatigued and non-fatigued with $H = 8.84$ ($p < 0.01$). However, this is not true among the male participants with $H = 2.72$ ($p = 0.09$). These results are shown in Figure 4.7 (top center). The smaller sample size may explain this compared to the female group. Only six male participants are classified as non-fatigued according to the FSMC scale. In terms of mean handgrip (Figure 4.7 bottom center), there is a statistically significant difference between genders when analyzing the non-fatigued $H = 11$ ($p < 0.001$) and fatigued $H = 7.25$ ($p < 0.01$) groups. However, the handgrip does not show a statistically significant difference between fatigued and non-fatigued participants.

The tapping performance and hand impairment analysis shows a statistically significant difference between non-impaired fatigued and non-fatigued participants with $H = 5.72$ ($p < 0.05$). However, there is no significant difference between the fatigued participants of the impaired and non-impaired groups. Only seven participants were classified as hand-impaired according to the 9-HPT. Of those, only two were non-fatigued. The small amount of non-fatigued and hand-impaired participants does not allow us to calculate whether there is a significant difference among the non-impaired population. In terms of handgrip strength, there is neither a difference between impaired and non-impaired participants nor a difference within these groups

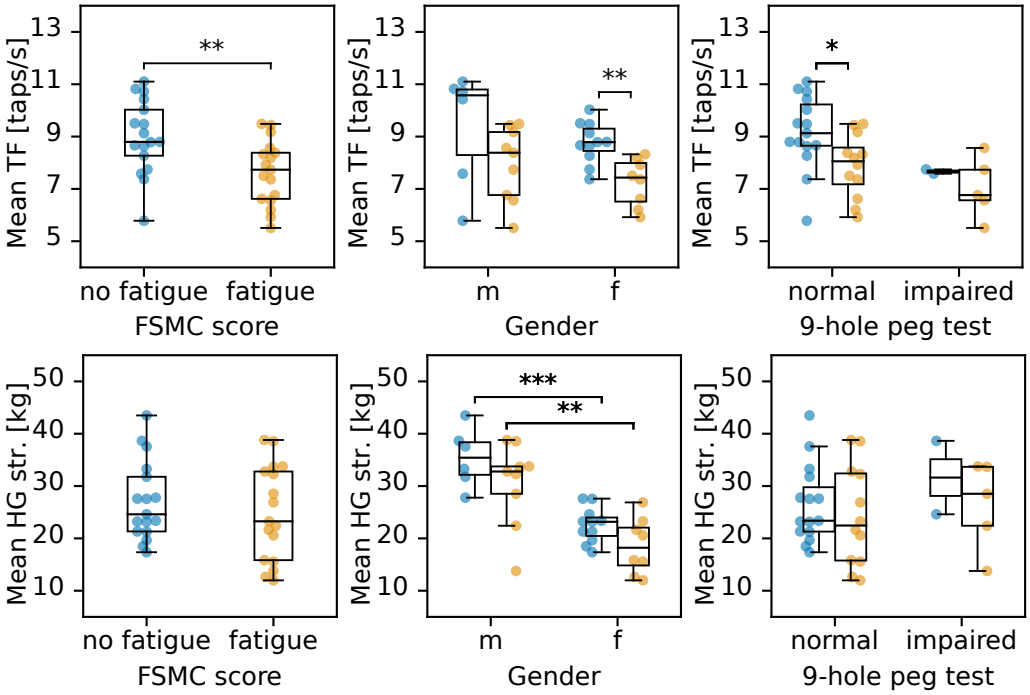

Figure 4.7.: Mean tapping frequency (*top*) and mean handgrip strength (*bottom*) in function of FSMC motor fatigue, gender, and impairment as defined by the 9-hole peg test. Fatigue is shown in orange and no fatigue in blue.

in terms of fatigue. Figure 4.7 (right) shows the box plots corresponding to this analysis.

We performed a similar analysis to explore mean tapping frequency and handgrip of fatigue and non-fatigued participants according to the FSS questionnaire, coming to similar conclusions presented in this section. However, when exploring the influence of gender on the mean tapping frequency and the FSS, we found a statistically significant difference between female and male participants. This is opposite to the FSMC, where no statistically significant difference was found for male participants. Refer to Appendix B.1 for the box plots related to the FSS scores. Furthermore, full descriptive statistics on the performed test and additional non-parametric tests are shown in Appendix B.3.

4.2.5. Tapping Frequency as a Surrogate for Perceived Fatigue

To explore the association between our metric and perceived fatigue, we computed the predictive power of the *tapping frequency* to rank fatigued participants according to the FSMC and FSS scores. We use *ROC* curves and AUC_{ROC} [Jin Huang and Ling, 2005] to evaluate the performance of our metric using it as threshold to classify between fatigued and not fatigued participants. AUC has the advantage that it

provides the features' overall classification performance without defining a threshold. Thresholds can be adapted depending on a specific purpose. In some cases, the focus is on high recall, while in others, on accuracy.

Evaluation Setting

To evaluate the robustness of our approach and compute confidence intervals for AUC_{ROC}, we use stratified Monte-Carlo sampling [Preacher and Selig, 2012] with 1000 iterations and randomly select (without replacement) in each iteration 2/3 of our participants' data (tapping trials) for evaluation. We partitioned the tapping data into six strata, following two partitioning criteria: (a) fatigued as a binary state according to FSMC or FSS, and (b) an age group, which can be one of three: [18,30), [30, 40), and [40,∞). Each participant and their data is fully assigned to one of the resulting six strata. Thus, when performing the stratified split, either a participant's data is fully considered or not at all. With this approach, we split at the participant level, ensure class balance, and account for age.

We report the average ($\overline{X}$) AUC_{ROC} with its respective confidence intervals. Additionally, we explore how the predictive power changes when combining more than one tapping trial. Thus, we combine consecutive, valid trials by averaging their features. For visual inspection, we include plots of the ROC curves corresponding to the 1000 splits and averaging three consecutive, valid trials. This section reports results for several features, specifically mean tapping frequency, maximum tapping frequency, and the slope of the tapping frequency. There is no established baseline for this classification task. Nevertheless, we consider the participant's age and the slope of the touch duration as baseline comparisons. Previous research shows that fatigue occurs more frequently in older patients, independently from disease severity [Colosimo et al., 1995]. Furthermore, we proposed touch duration declined rate (slope) as a fatigability metric.

FSMC - Motor Fatigue Ranking According to AUC_{ROC}

Our results show that maximum and mean tapping frequencies outperform the other features. When considering a single tapping trial ($t = 1$), tapping frequency ranks fatigue and non-fatigued participants with AUC_{ROC} $\overline{X} = .74 \pm .05$. Furthermore, we observe that the AUC_{ROC} increases when averaging consecutive trials' features. Tapping frequency reaches a maximum when combining three consecutive, valid trials, representing an improvement of 2 percentage points ($p.p$). Figure 4.8 (right) shows the ROC curves corresponding to the mean and maximum tapping frequency, best-performing features when averaging three successive valid trials. The slope of the tapping frequency ranks participants with AUC_{ROC} $\overline{X} = .65 \pm .05$ when $t = 3$.

Followed by touch duration slope with a AUC_{ROC} $\overline{X} = .60 \pm .05$ when $t = 3$, and age with AUC_{ROC} $\overline{X} = .57 \pm .05$. We computed slopes as features of motor fatigability as suggested in previous research [Bigland-Ritchie et al., 1983]. Similarly, we consider the participant's age as a feature, as previous research has shown that fatigue occurs more frequently in older patients, independently from disease severity [Colosimo et al., 1995]. Our suggested metric, tapping frequency, outperforms the baseline touch duration slope and age by 16 $p.p$ and 19 $p.p$, respectively. Our results show that tapping trial performance metrics outperform the motor fatigability metrics for assessing perceived fatigue.

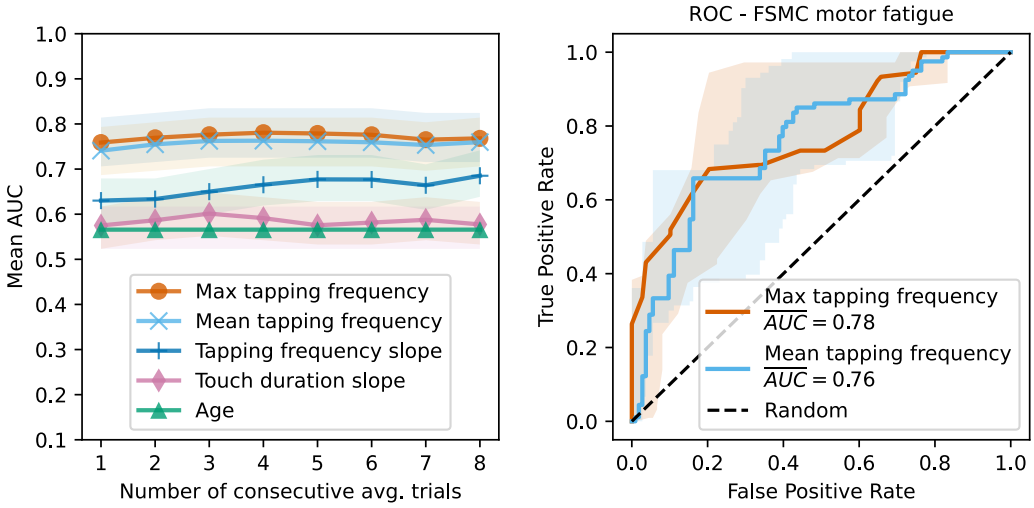

Figure 4.8.: Mean AUC_{ROC} when ranking motor fatigue according to FSMC of all participants (N=34) on the left. Mean tapping frequency shows the best performance in comparison to the other features. Also, reliability increases when averaging the features of consecutive valid trials (t). ROC curves for mean and maximum tapping frequency with $t = 3$ are displayed on the right. Data generated using Monte-Carlo simulation with 1000 iterations.

FSS - Fatigue Ranking According to AUC_{ROC}

Fatigue ranking in terms of the FSS questionnaire shows the same behavior as described for the FSMC ranking. Mean tapping frequency and maximum tapping frequency exhibit the best ranking performance. When considering a single tapping trial ($t = 1$), mean tapping frequency ranks fatigue and non-fatigued participants with AUC_{ROC} $\overline{X} = .80 \pm .05$. Furthermore, we observe that the AUC_{ROC} increases when averaging consecutive trials' features. Tapping frequency reaches a maximum when combining three consecutive, valid trials with AUC_{ROC} $\overline{X} = .81 \pm .05$. The next best feature is maximum tapping frequency with AUC_{ROC} $\overline{X} = .77 \pm .05$

when combing three trials ($t = 3$). Following is tapping frequency slope with with AUC_{ROC} $\overline{X}$ = .61 ± .05 when $t = 3$, age with with AUC_{ROC} $\overline{X}$ = .56 ± .05 when $t = 3$, and finally touch duration slope with AUC_{ROC} $\overline{X}$ = .50 ± .05. The touch duration slope shows a random behavior for ranking fatigue according to the FSS. Mean tapping frequency outperforms the fatigability baseline touch duration slope and age by 31 $p.p$ and 26 $p.p$, respectively (Figure 4.9) Similar to the FSMC ranking results (Section 4.2.5), tapping trial performance metrics outperform the motor fatigability metrics for assessing perceived motor fatigue.

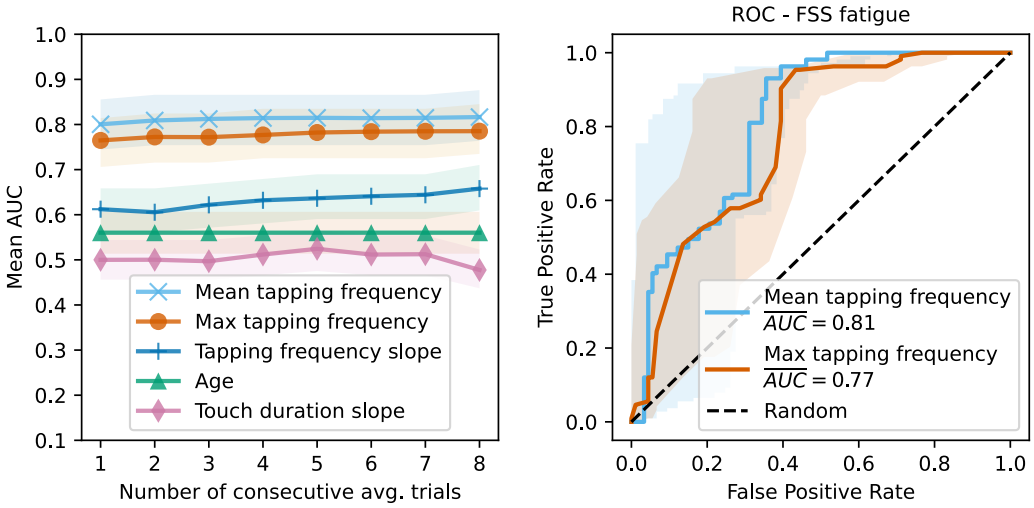

Figure 4.9.: Mean AUC_{ROC} for fatigue according to FSS of all participants (N=32) on the left. Mean tapping frequency shows the best performance in comparison to the other features. Also, reliability increases when averaging the features of consecutive valid trials (t). ROC curves for mean and maximum tapping frequency with $t = 3$ are displayed on the right. Data generated using Monte-Carlo simulation with 1000 iterations.

4.2.6. Participants' Adherence – Temporal Analysis

To verify that our approach is valid in-the-wild, we analyzed how the metric outcomes varied over the two-week study. We compute a sliding window and average the mean tapping frequency over three consecutive, valid trials. Afterward, we compute a series of Kruskal-Wallis H-tests to verify the statistically significant difference between fatigued and non-fatigued patients according to the FSMC and FSS held during the two-week study. Figure 4.10 and Figure 4.11 offer an overview of the results corresponding to the FSMC and FSS, respectively. The results show that the statistically significant difference between the fatigued and non-fatigued participants in terms of the mean tapping frequency holds in-the-wild for both questionnaires.

This confirms that our metric is valid in unsupervised settings and that the approach is suitable for monitoring fatigue remotely.

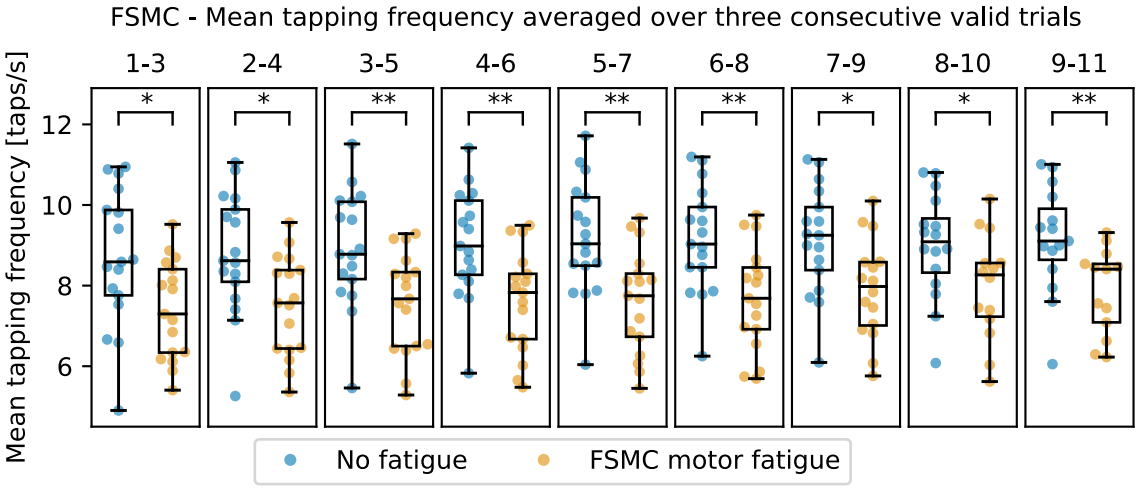

Figure 4.10.: Mean tapping frequency of three averaged valid asks during the course of the study grouped by motor fatigue as defined by the FSMC questionnaire.

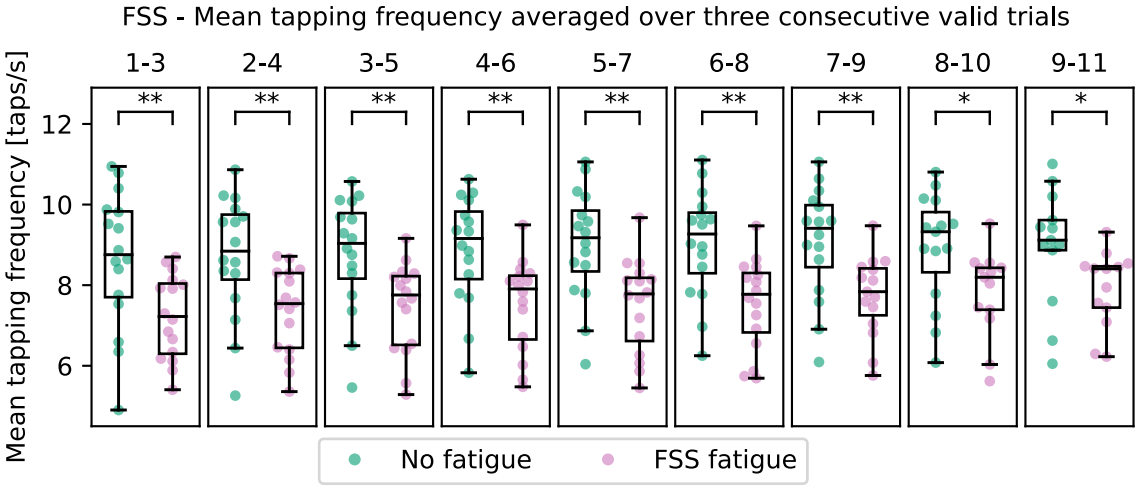

Figure 4.11.: Mean tapping frequency of three averaged valid asks during the course of the study grouped by fatigue as defined by the FSS questionnaire.

4.3. Discussion

In this section, we discuss the implications of our findings, the limitations of our work, and possible directions for future research.

4.3.1. Implication of Subjective/Objective Measurements of Fatigue

There is a clear unmet medical need to develop an objective measure to assess both motor and cognitive fatigue in MS patients. The availability of such a tool would be an essential component to creating new therapies and improving routine medical care by helping to assess the effect of an intervention and to differentiate from various confounding symptoms, e.g., sleepiness, mood alterations, and others. Despite being a debilitating symptom affecting 90% of all MS patients throughout the disease, there is still no approved therapy available. Therefore, different compounds have been tested in randomized placebo-controlled clinical trials (or are being used as off-label treatments). The results of these trials are inconsistent; some reported good efficacy of the therapies, whereas others did not show a benefit of the therapies [Hader et al., 1987; Cohen and Fisher, 1989; Ledinek et al., 2013; Krupp et al., 1995; Brioschi et al., 2009; Rammohan et al., 2002; Stankoff et al., 2005; Möller et al., 2011; Lange et al., 2009; Nourbakhsh et al., 2021]. The outcome measures in all these trials were subjective questionnaires. It is well accepted that the magnitude of the placebo effect is an important reason for the variability in efficacy [Nourbakhsh et al., 2021; Pucci et al., 2007; Sheng et al., 2013]. Hence, an objective measure would overcome this limitation for treatment development and provide a useful medical decision-making tool.

The smartphone-based tapping task is easy to administer, and because of its game-like character, it could potentially have a higher acceptance than standard questionnaires. While the tapping task takes less than a minute, the FSMC questionnaire involves 20 items that must be answered by deciding on five given choices on a Likert scale. Nevertheless, user acceptance needs to be assessed in future studies.

Further, the tapping task provides a direct (to the point) measurement. It could be administered several times a day to quantify fluctuations in performance fatigability, typical of fatigue. Questionnaires evaluate the symptom only retrospectively, usually for two weeks, and are therefore less precise and not sensitive to fluctuations or short-term changes in the severity of the symptom (i.e., following physical/medical interventions). Hence, a more continuous (higher frequency) assessment is advantageous for following patients over time and assessing new interventions' effects. However, one has to consider that fatigability relates to a specific task, while fatigue questionnaires cover a general feeling which affects the person as a whole. Thus, our approach complements existing fatigue quantification methods.

Our study provides a proof-of-concept for an association of motor fatigability, assessed by the tapping task, with subjective motor fatigue, assessed by the FSMC, which has been developed and validated in MS patients. Furthermore, the association between smartphone-based motor fatigability and perceived fatigue has also been confirmed with an independently validated fatigue questionnaire, FSS. Hence, the study

provides early evidence for an association between the objective smartphone-based motor fatigability measurement and perceived fatigue in MS patients. Nevertheless, further and more extensive studies are needed to establish the predictive value of the tapping task to subjective fatigue.

4.3.2. Tapping Frequency as Reliable Smartphone-based Motor Fatigability Metric

We believe that our proposed method is less prone to outliers compared to *touch duration*, introduced in Chapter 3. Touch duration could have an erroneous representation of the tapping task performance, given that the metric fails to account for the time when fingers perform their air motion. An example of this behavior is when the person is fast at lifting the fingers from the smartphone screen, but their finger's air motion is slow. Our metric, tapping frequency, does not suffer from this phenomenon, as it reflects the whole dynamics of the tapping task. Additionally, with our gap removal, we seek a more flexible approach.

Gap Removal for In-the-wild Studies

The gap removal intends to gain as much value from the data as possible while avoiding discarding complete trials, a key feature for in-the-wild studies. This is particularly useful for unsupervised settings where the person may get distracted while performing a trial. Phone notifications, calls, or external factors could cause distractions. Additionally, we noticed the utility of our validation algorithm when it detected problems with one patient. Later we learned that the patient had very long artificial nails that caused unreliable tapping. In summary, our new method makes fewer assumptions, increases validity by 19%, and shows a comparable correlation to the clinical baseline (handgrip).

4.3.3. Tapping Frequency – Difference Between Fatigued and Non-fatigued Patients

From the tapping frequency, we learned that non-fatigued participants delivered a higher mean tapping frequency than fatigued participants and that this difference is statistically significant. Patients defined as non-fatigued according to the FSS and FSMC questionnaires achieved higher maximum tapping frequencies. In contrast, when using the handgrip dynamometer, we notice no statistically significant difference between fatigued and non-fatigued patients. Moreover, tapping frequency is independent of gender, while the handgrip dynamometer is not. Hence, our approach

shows advantages and outperforms the handgrip dynamometer for monitoring motor fatigue.

4.3.4. Participants' Compliance to the Study Protocol

Through our experiment, we examined participants' adherence to the study protocol over two weeks. Compliance during the study was good. Analysis of the participants' two-week behavior shows no significant change in tapping frequency over time. All patients completed the two-week protocol, and the number of invalid trials did not show a particular pattern. Using our validity algorithm, we analyzed the completed tapping trials and found that only a small percentage was invalid. Our analysis shows that combining several tapping trials is advisable to achieve higher confidence in the results. We show that the average of three tapping trials is sufficient to classify fatigue.

4.3.5. External Validity of the Results

There is no standard objective method to measure overall fatigue, particularly perceived fatigue, other than standard questionnaires. Hence, to develop a new approach, one has to rely on these validated questionnaires as a reference. Therefore, as part of this study, we aimed to assess the association of motor fatigability, assessed with the tapping task, with perceived fatigue rated by standard questionnaires. The following steps have been taken to ensure the validity of the results. First, we validate tapping frequency as an objective measure of motor fatigability against a standard reference method (handgrip dynamometer). Second, the validity of an unsupervised assessment of the smartphone-based task has been confirmed in an in-the-wild study in MS patients. Third, we use the in-the-wild data to assess whether the results of the tapping task can be used as a surrogate for subjective fatigue, being classified using two different questionnaires, both validated in MS patients. Overall, the results provide early evidence for using the smartphone-based tapping task as a surrogate for perceived fatigue. However, more extensive and independent studies are needed to confirm the results and establish an objective task of motor fatigability as a surrogate for subjective fatigue.

4.3.6. User-interface, Interaction and Design Improvements

Informal feedback from the participants suggests that performing daily tasks can produce a lack of motivation and boredom. This can be addressed in further studies by introducing a gamification mechanism to keep the participant engaged and motivated. We recommend combining three tapping trials to achieve better results

and avoid demotivating the users. However, most importantly, we do not advise conducting the tapping task daily for prolonged periods. An alternative approach would require tapping trials for three consecutive days every 1-2 weeks. Further studies are necessary to estimate a suitable periodicity for the tapping task.

Immediate Validity Feedback

We only applied our validity algorithm during a post-processing phase. In future task design improvements, we recommend incorporating immediate feedback to the user to further reduce the total percentage of invalid trials. Trials can be automatically stopped when gaps exceed a defined threshold of 1 s. When this occurs, users can be notified of the specific problem (large gap) and can be asked to restart the tapping trial from the beginning.

Maximum Tapping Frequency and Shorter Trials

Our results indicate that maximal tapping frequency is also a suitable surrogate for fatigue. This has important implications as it would mean that our proposed validity algorithm would change, and potentially fewer trials will be discarded. Additionally, this would imply that the tapping trials could be shorter than 30 s. However, further studies are needed to evaluate the full implications of such changes. Further analysis suggests that the mean tapping frequency measured during only 15 s of a tapping trial produces comparable results, indicating that a shorter task may be viable. However, further studies are needed to confirm this hypothesis. In addition, we do not know how patients' behavior and intrinsic motivation will change when performing the task in a shorter time frame. Twenty seconds of tapping is a suitable compromise based on our observations. We do not recommend shorter trials as we know the initial 3 s of tapping accounts for task inertia and momentum. Moreover, applying the gap removal algorithm reduces the effective trial length, but trials must be sufficiently long to quantify fatigue.

4.3.7. Limitations and Future Challenges

Tapping and Impairment

A larger study population is needed to evaluate our metric's reliability in MS patients with hand impairment. Only two of seven hand-impaired patients were non-fatigued. Hence, at this point, we cannot conclude if there is a statistically significant difference between fatigued and non-fatigued patients within this specific population. However, we see this as a minor drawback of our approach. Our results show that our tapping

task is feasible and valid in our MS cohort and is, therefore, a promising tool for patients with other disease entities, such as post-COVID19 syndrome, which is not associated with hand impairment. Future studies should include larger numbers of MS patients combining the whole spectrum of disabilities and further expanding to other diseases, particularly those that do not entail hand impairment.

Recognizing Different Fatigue Levels

In this study, we used the FSMC as a 2-level assessment tool. However, the FSMC offers thresholds for the different fatigue levels: "mild," "moderate," and "severe." We used the FSMC for binary classification and considered patients fatigued once they exceeded the lowest threshold (mild fatigue). During future work, we plan to explore using our approach for classifying the multiple fatigue levels. A larger study population is needed to assess the feasibility of this approach.

Recommendation for Future Trials

First, single-tapping task measurements are usually unreliable as they could be classified as invalid. Averaging values of several trials leads to better results when analyzing fatigue. The frequency of the measurements is also an important point that should be taken into account. Even though we did not conduct specific interviews to get feedback about the usability of the task and study design, some patients gave informal feedback indicating that frequent testing may become tedious or tiresome.

4.4. Conclusion

We introduced a new metric as a proxy to quantify perceived fatigue objectively. Our metric, mean tapping frequency, is derived from a simple tapping task performed on commodity smartphones. The validity of the metric has been confirmed by a significant correlation with handgrip strength measurements, which is the current standard procedure for measuring motor fatigability. Additionally, we demonstrate that our approach is comparable to touch duration, which was introduced in the previous chapter. Our two-week in-the-wild study, in 35 MS patients, shows that mean tapping frequency can rank fatigued and non-fatigued with AUC_{ROC} $\overline{X} = .76 \pm 5$ according to the FSMC, and with AUC_{ROC} $\overline{X} = .81 \pm .05$, according to the FSS, indicating an association between fatigue and our smartphone-based assessment metric.

In summary, our results show that: (1) Tapping frequency is a valid motor fatigability metric. (2) Our data processing pipeline maintains task validity with an increase of

19% over *touch duration*. (3) Mean tapping frequency can discriminate fatigue rated by two clinical fatigue scales (FSS and FSMC). (4) Mean tapping frequency as an objective fatigue metric is valid in-the-wild. (5) Combining several trials improves the reliability of fatigue prediction. Future studies in MS patients with hand impairment are needed to establish the validity of our metric in this population. Furthermore, future longitudinal studies are needed to establish optimal time intervals between tapping trials and verify if our metric can be established as a surrogate for perceived fatigue.

Our goal was to study the feasibility of establishing an objective metric as a surrogate for perceived fatigue. We are confident that our work is a step toward ubiquitous and objective symptom quantification. Our simple model provides good interpretability and a higher chance of being adopted in clinical practice. Providing a novel tool to follow patients with fatigue continuously meets an important unmet medical need in MS and many areas of medicine where fatigue is a prevalent condition. An objective and reliable measure as a surrogate for fatigue facilitates further research on this devastating symptom, particularly the development of novel therapies. Additionally, the ability to monitor patients over time and independently from medical facilities (i.e., in-the-wild) provides an important advantage in assessing the effects of therapeutic interventions.

Part II.

Cognitive Fatigue

CHAPTER 5

Cognitive Fatigability Assessment Test – cFAST

This chapter is based on the following publication:

Liliana Barrios, Rok Amon, Pietro Oldrati, Marc Hilty, Christian Holz, Andreas Lutterotti. Cognitive fatigability assessment test (cFAST): development of a new instrument to assess cognitive fatigability and pilot study on its association to perceived fatigue in multiple sclerosis. *SAGE journals, Digital Health.* August 2022.

In the previous two chapters, we focused on fatigue's physical aspect. Now, we shift the focus to fatigue's cognitive aspect. As described in Chapter 2, there are no dedicated tests to quantify cognitive fatigability [Walker et al., 2019]. Existing studies use one of two strategies. Either they conduct a test battery or employ a single prolonged cognitive task to measure the decline in performance within the task. Amount the cognitive function tests utilized within fatigability research are: (1) the Paced Auditory Serial Addition Test (PASAT) [Tombaugh, 2006], (2) the Psychomotor vigilance task (PVT) [Basner and Dinges, 2011], and (3) the Stroop test [Stroop, 1935]. However, utilizing cognitive function tests to assess cognitive fatigability comes with drawbacks, such as long testing sessions. Hence, in this chapter, we propose a new test for measuring cognitive fatigability in a short period (i.e., 5 minutes) and refer to it as the Cognitive Fatigability Assessment Test (cFAST).

cFAST is inspired by the Symbol Digit Modality Test (SDMT) digit-symbol matching logic [Smith, 1982]. SDMT is a cognitive test that measures information processing speed to quantify disability. Studies showed that the SDMT is relatively resistant to practice effects [Benedict et al., 2012], in particular when rearranging the keys [Roar

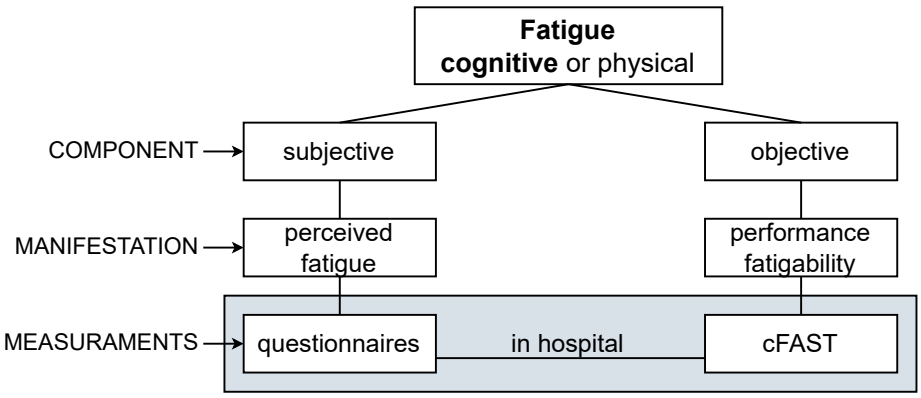

Figure 5.1.: Currently, there is no dedicated test to measure cognitive fatigability. Hence, we introduce cFAST, an specific test that aims at measuring cognitive fatigability. We validate our test by comparing it to standard fatigue questionnaires (FSMC cognitive subscale).

et al., 2016], making it an attractive tool for cognitive monitoring over time in clinical trials [Patel et al., 2017]. Moreover, it has also been validated for smartphones [van Oirschot et al., 2020]. Our solution uses a similar key-symbol matching strategy to measure fatigability instead of cognitive impairment.

We start the chapter by describing the development of our new objective and ubiquitous measurement of cognitive fatigability (cFAST). Our test is smartphone-based. We developed it through an iterative process involving patients, neuropsychologists, and neurologists. We opted for a smartphone-based implementation given the high acceptability and interest of MS patients in smartphone-based tools that allow them to monitor and manage their condition [Griffin and Kehoe, 2018; Apolinário-Hagen et al., 2018; Ayobi et al., 2017; Giunti et al., 2018; Van Kessel et al., 2017; Midaglia et al., 2019; Rice et al., 2021; Motl et al., 2017]. Following, we introduce our study to investigate the association between the newly developed objective measurement (cFAST) and perceived cognitive fatigue. We conducted a pilot study with MS patients who completed the cFAST and the FSMC [Penner et al., 2009]. Using the FSMC cognitive subscale, we assign the participants to the cognitive-fatigued (subscale≥22) and non-cognitive-fatigued (subscale<22) groups [Penner et al., 2009]. Figure 5.1 highlights the focus of our study within the context of fatigue and fatigability research. From the cFAST, we extracted a set of metrics and evaluated group differences with t-tests. Through Area Under the Receiver Operating Characteristics (AUROC), we assessed the performance of our proposed metrics to classify cognitively fatigued vs. non-cognitively fatigued patients. Furthermore, we investigated the relationship between our proposed test (cFAST) and metrics for disability.

5.1. Methods

5.1.1. Development of the Cognitive Fatigability Assessment Test (cFAST)

We aimed to develop a test to objectively quantify cognitive fatigability that meets the requirements: (1) engages cognitive processing speed and induces cognitive load, (2) is short, self-explanatory, and allows for remote monitoring, and (3) does not require medical supervision. We followed an iterative process during the design and development of the application. The medical professionals reviewed different prototypes to ensure an appropriate design based on clinical theory and practice is implemented. Additionally, we gathered informal feedback from MS patients regarding our prototypes before converging on our final design. Refer to Appendix C.1 for further details on the prototypes' designs and selection.

Figure 5.2 displays cFAST's user interface and highlights its elements. The test is designed to be carried out by holding the smartphone in landscape mode. The middle of the screen shows a large blue symbol (main symbol). The main symbol has to be mapped to its corresponding digit following the mapping rule displayed at the top of the screen. Selection occurs by tapping the numbers located at the bottom of the screen. Users have a limited time to find the corresponding number associated with the main symbol. A yellow progress bar around the symbol indicates how much time is left until the symbol is changed automatically. The main symbol changes under two circumstances: (1) after the user taps a number or (2) when the progress bar has entirely run out. Every time a new symbol appears, the associations and positions of the top mapping rule are randomized, and the progress bar is restarted. The randomization seeks to diminish the possibility of a learning effect associated with memorizing the digit-symbol mapping within the same test run. The progress bar is a pressure mechanism to motivate users to be fast and avoid resting periods. A timer located at the top left indicates how much time is left for the test to end. Users can exit the test at any moment by tapping the exit button located at the top right corner. If exited early, the test is considered invalid.

Our test is inspired by the SDMT (Symbol Digit Modalities Test) [Smith, 1982], as it is a widely used, accepted, and validated cognitive assessment test in MS. However, cFAST differs from the SDMT in several aspects:

1. cFAST is a cognitive fatigability test, while SDMT assesses cognitive impairment and working memory.
2. Contrary to the SDMT, cFAST does not allow participants to look ahead to match the following symbols. Hence, participants cannot anticipate the next answer to reduce their response time.

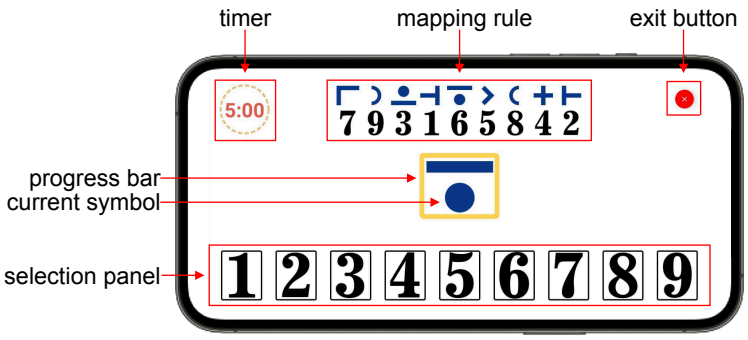

Figure 5.2.: cFAST user-interface with highlighted elements in red. *Note.* cFAST, cognitive fatigability assessment test.

3. There is a time limit to complete each selection in cFAST.
4. cFAST randomizes the matching rules after each answer, while SDMT has a fixed matching rule.
5. The duration of a cFAST session is 5 minutes, while the SDMT lasts 90 seconds. The increased duration is needed because cognitive fatigability is notoriously hard to elicit in a short time. However, cFAST is significantly shorter than existing attempts to measure cognitive fatigability.

All these design considerations seek to evaluate cognitive fatigability.

5.1.2. Application Logic

cFAST is designed with the aim of being conducted outside the clinic and without medical supervision. Therefore, the application logic is self-explanatory and contains a personalization phase to maximize the users' understanding and tailor it to their performance. This phase needs to be completed before being able to run cFAST. Figure 5.3 depicts the application logic diagram.

At the start of the personalization phase, users are prompted for a mandatory two-minute preparation step. Before starting the calibration, this step aims to familiarize users with the test's matching logic and rules. To this end, a confirmation step ensures that, during the preparation, users provided at least 70% correct digit-symbol matches out of a minimum of 20 answers. Contrary to the calibrated cFAST, there is no time limit to match individual symbols during preparation. Hence, symbols only change after the user presses a number from the selection panel. We refer to this method as *manual*. This functionality allows users to understand the test's matching logic without time pressure.

During preparation, users receive immediate feedback on whether their choice is

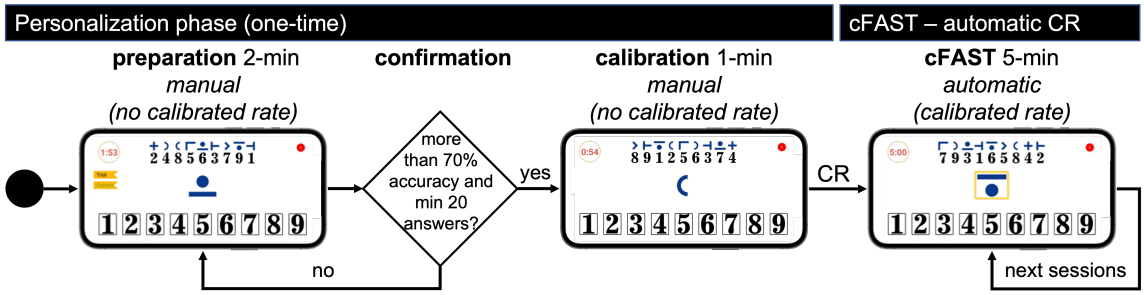

Figure 5.3.: cFAST application logic. In the personalization phase, users complete the preparation and confirmation to ensure they understand the test's matching logic and the calibration to derive the calibrated rate used in cFAST. After this phase, cFAST is personalized and ready to be used. *Note*. cFAST, cognitive fatigability assessment test; CR, calibrated rate.

correct or incorrect through a label located at the left side of the screen (Figure 5.4). Failed preparation trials indicate that the user has not sufficiently trained in operating the test yet or did not perform it as fast as possible and thus must repeat it. The motivation for providing immediate feedback is to help the user understand the matching mechanics of the test. This functionality is particularly beneficial for unsupervised settings where no medical examiner is present to clarify patient doubts. Users can start the calibration step only after preparation is passed successfully. The calibration step lasts one minute and uses the same logic as the preparation step but without providing feedback. At this point, we assume users understand the test's matching logic. Similar to preparation, calibration also employs a manual mechanism. However, its goal is to extract the users' reaction time, which we call *calibrated rate*. This rate is then used in cFAST. Thus, the manual function of the application has two goals: (1) during the preparation, it allows sufficient time for users to understand the test's matching logic, and (2) during calibration, it helps derive a personalized *calibrated rate*.

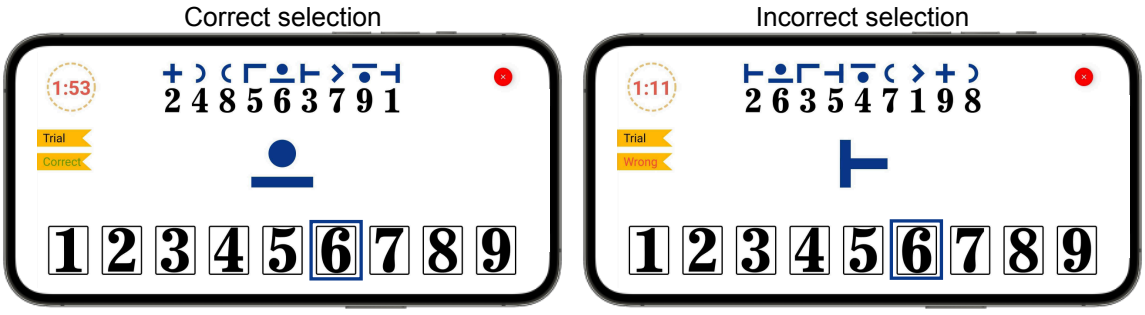

Figure 5.4.: Preparation step user interface. The blue rectangle indicates the answer provided by the user. After each number selection, the interface indicates with a label whether their attempt is **correct** (left) or **wrong** (right).

Deriving calibrated rate

The calibrated rate is a crucial feature of cFAST, and it is derived from the 1-minute calibration step of the personalization phase (Figure 5.3). The calibration step uses the same logic as the preparation step but without user feedback. During calibration, symbols are only changed once the user taps a number from the selection panel (manual mechanism). We use 85% percentile of the response time exhibited during the calibration step to extract the *calibrated rate*, meaning each user may perform the task at different rates but always with regards to their top performance. Thus, the calibrated rate is tailored to each user, accounting for patients' different levels of disability. Once the calibrated rate is derived, cFAST is personalized and ready to use.

Eliciting cognitive fatigability

During a cFAST session, users must repeatedly match a symbol with their corresponding number. However, tasks of this nature are typical examples of speed-accuracy trade-off [Zhai et al., 2004]. Participants tend to decide between performing the test with high accuracy but slow (i.e., low exertion) or fast but low accuracy. Either of these scenarios would significantly limit the fatigue-inducing effect of the test. With cFAST, we seek to reduce this trade-off by adding a limited timeframe (calibrated rate) for each selection. This timeframe is indicated through a yellow progress bar (Figure 5.5). With this approach, participants cannot spend unlimited time making a decision. Moreover, we hypothesize that the added pressure to make a fast selection contributes to the cognitive load required to induce cognitive fatigability.

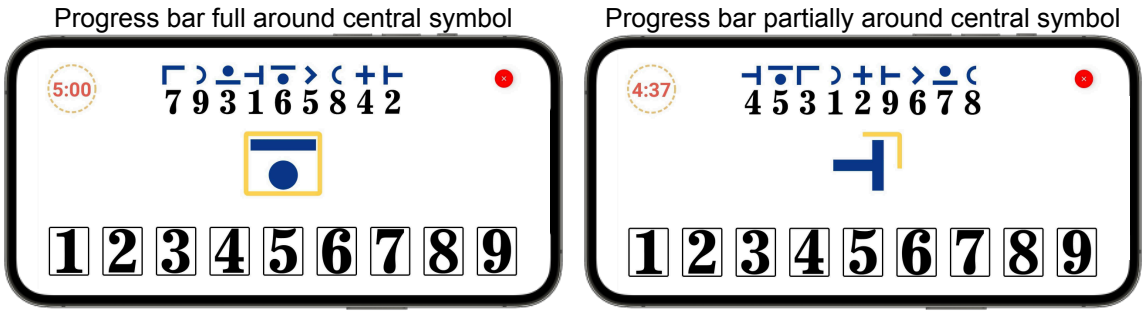

Figure 5.5.: cFAST user interface. The left side of the image displays the screen at the beginning of a 5-minute test. The yellow progress bar indicates the remaining time to complete a selection. The digit-symbol mapping is randomized after each selection to reduce learning effects. *Note*. cFAST, cognitive fatigability assessment test.

5.1.3. Participants

We recruited 48 patients from the MS outpatient clinic of the Department of Neurology, University Hospital Zurich, between September 2020 and April 2021. Participants provided written consent following the Declaration of Helsinki [World Medical Association, 2002]. The EDSS was obtained from the routine neurologic examinations performed at the hospital. This study was approved by the local ethics committee (Cantonal Ethics Committee Zurich, Switzerland). Inclusion criteria consisted of: (a) confirmed MS diagnosis and (b) age between 18 to 70. In addition, exclusion criteria included: diagnosis of depression, schizophrenia, bipolar disorders, ADHD, and regular intake of psychostimulants or anticonvulsant medications.

5.1.4. Study Design

Participants were briefly introduced to the study setup and completed a demographic questionnaire. Following, the study examiner showed them the application and the logic of the cFAST. Participants started with the 2-min preparation session. After successful completion, they performed the calibration step. Next, we asked participants to complete a first cFAST session of 5 minutes, considered a trial, to ensure they understood the test logic. There was a short break in which participants filled out the FSMC questionnaire. Next, participants performed a second cFAST session. Previous cognitive fatigability studies, including modified versions of the SDMT, do full trials and discard this data before conducting the test to ensure participants understand the test logic [Chen et al., 2020]. Hence, all data analyses presented in this chapter are based on the main cFAST and not the trial data.

5.1.5. Data Collection and Processing Pipeline

We collected touch data from the smartphone using a custom Android application we developed. Each sample in our dataset contains the symbol ID to be matched, the user's selection, if there was any, the current mapping rule, and the timestamp of the touch-down event. Our data processing pipeline includes three steps: (1) artifact detection, (2) cognitive adaptation removal, and (3) metrics extraction.

1. Artifact detection. We use *response time* as one of our primary performance metrics. Artifacts in *response time* typically appear when a user aims at tapping a digit to match the current symbol, but they run out of time. Hence, the newly displayed symbol is stored with a short *response time*, and the previous symbol is marked as a *missed answer* (Figure 5.6). These artifacts must be identified and removed to avoid double-counting errors and computing a misleading *response time*.

Therefore, in our preprocessing step, we remove any entry after a missed answer with a *response time* of less than the average minus two standard deviations of the entire cFAST session's *response time*. This results in subject-specific thresholds that account for the difference in average performance. With this method, we remove an average of 3.8 entries per session, with the average session containing 138 answers. Figure 5.6 right shows the same data after artifact removal.

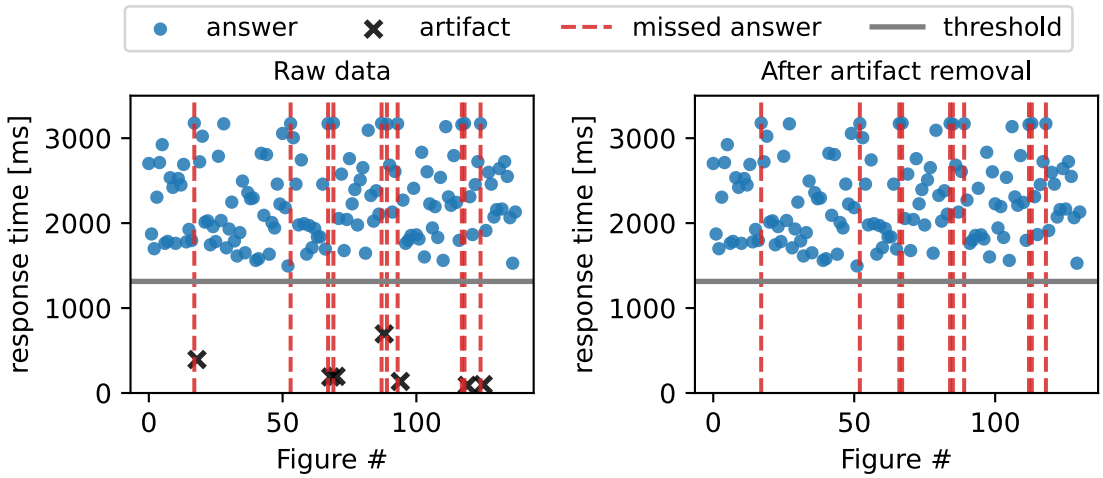

Figure 5.6.: Artifacts in *response time* typically appear when a user provides an answer shortly after running out of time. Therefore, the pressed digit is associated with the newly displayed figure. As a result, the previous entry is classified as a missed answer, and the current figure has a very short response time (left side). We detect and remove these artifacts to avoid misleading *errors* and *response time* values (right side).

2. Cognitive adaptation removal. Previous cognitive fatigue studies describe an adaptation phase occurring at the beginning of a cognitive task due to some unspecific modulations of training and adaptation and highlight the need to account for these effects when studying fatigue [Möckel et al., 2015; Wascher et al., 2014]. A common strategy to deal with the adaptation in cognitive fatigue studies is to omit the start of the task [Möckel et al., 2015; Wascher et al., 2014]. The adaptation phase is not unique to cognitive tasks as it has also been detected in motor fatigability tasks (refer to Chapter 3 and Chapter 4). A similar strategy is applied in motor tasks by removing the start of the task to account for the adaptation period [Schwid et al., 1999]. cFAST sessions exhibit an adaptation period in the initial part of the test, particularly for fatigue patients. Figure 5.7 depicts the average mean-normalized reaction times for all fatigued patients for 5 minutes cFAST in 30-second segments. During the first segments, we observe an increase in reaction time, followed by a decrease in reaction time in the third segment. We attribute these changes in performance to an adaptation

period before users are fully immersed in the test [Möckel et al., 2015; Wascher et al., 2014]. Hence, to make a fair comparison between the study participants, we discard the first 60 seconds of all cFAST tests (42 sessions) before extracting the metrics and performing the data analysis.

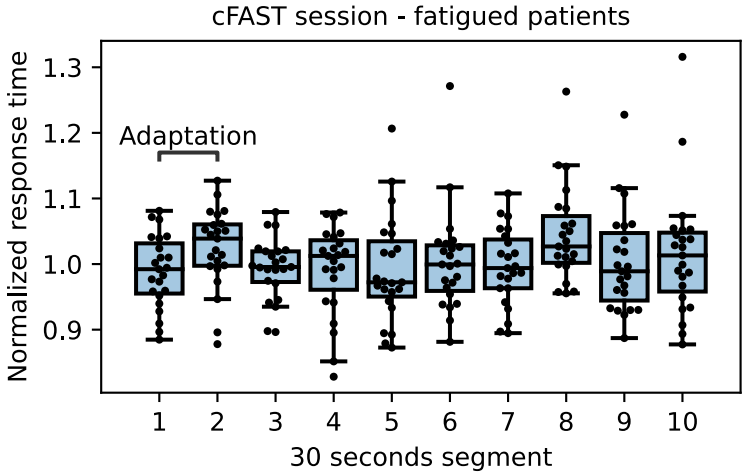

Figure 5.7.: cFAST session with average mean-normalized reaction time per 30 second segments for each fatigued participant. The first two segments (60 seconds) are discarded as we consider them part of the adaptation phase.

3. Metrics extraction. We define two sets of metrics to quantify performance during a cFAST test session: (1) general metrics, which represent the average performance during an entire test session, and (2) fatigability metrics, which measure the change in performance occurring between the first third and last third of a test session. Table 5.1 displays an overview of the proposed metrics with their definition.

5.1.6. Statistical Analyses

We use descriptive statistics to summarize and compare the study subpopulations. We evaluate the performance of our derived smartphone-based metrics to discriminate between cognitive-fatigued and non-cognitive fatigued subjects following the FSMC cognitive subscale (threshold=22) [Penner et al., 2009]. With t-tests, we explore group differences and consider $P < .05$ significant. Furthermore, through Area Under the Receiver Operating Characteristics (AUROC), we evaluate the performance of our derived smartphone-based metrics to classify cognitive fatigued vs. non-cognitive fatigued subjects, independently of age and EDSS. We assess the robustness of our approach and compute confidence intervals for AUROC using stratified Monte-Carlo

Table 5.1.: Metrics description.

Type	Name	Description
general	*response time*	average time in milliseconds to tap a digit from the selection panel after the appearance of a new symbol.
	calibrated rate	time duration in milliseconds for each new symbol – derived from the calibration phase (corresponds to progress bar duration).
	correct	total correct matches
	errors	total errors including wrong matches and missed answer.
fatigability	$\Delta correct$	percent change in correct between the first and the last third of the task.
	$\Delta response\ time$	percent change in response time between the first and the last third of the task.
	$\Delta errors$	percent change in errors between the first and the last third of the task.

sampling [Preacher and Selig, 2012] with 1000 iterations and randomly select (without replacement) in each iteration 1/2 of our participants' data (cFAST sessions) for evaluation. We partition the cFAST data into eight strata, following two partitioning criteria: (a) cognitive fatigued as a binary state according to FSMC cognitive subscale (threshold=22) and (b) an EDSS group, which can be one of four: [0,1), [1, 2), [2, 3), and [3,∞). The idea of this partition is to find a metric that works best in the whole spectrum of disability. Each participant and their data is fully assigned to one of the resulting eight strata. Thus, when performing the stratified split, either a participant's data is fully contained in the split or not at all. Hence, with our approach, we split at the participant level, ensure class balance, and account for disability. Additionally, as age also influences cognitive performance [Möller et al., 2014], we create eight additional strata following two partitioning criteria: (a) cognitive fatigued as a binary state and (b) age group, which can be one of four: (18, 30), [30,40), [40,50) and [50, 70]. This partition aims at reducing the influence of age in the metrics by assigning weights according to the group sizes. Furthermore, we use one-way analysis of covariance (ANCOVA) with EDSS as a covariant to rule out the effect of disability when analyzing fatigue.

Finally, we further explore how cFAST and our proposed metrics relate to disability by measuring the performance of the metrics to rate disability according to EDSS. To this end, we split the study participants in two groups according to EDSS and

analyzed the difference in performance between both groups. We classify patients with EDSS>1.5 (n=24) as disabled and patients with EDSS≤1.5 (n=24) as not disabled. For this evaluation, we partition our dataset into four strata, following two partitioning criteria: (a) disabled as a binary state according to the EDSS (0 for EDSS≤1.5 and 1 for EDSS>1.5), and (b) cognitive fatigued as a binary state according to FSMC cognitive subscale (threshold=22). Additionally, we use the same age groups as we did for the cognitive fatigue evaluation. We report the average AUROC with 95% confidence intervals. In addition, we include plots of the ROC curves for visual inspection.

5.2. Results

5.2.1. Participant Characteristics

We recruited 48 study participants, and from those, we excluded six due to comorbidities, including iron deficiency, personality disorder, hypothyroidism, and narcolepsy type 1. Table 5.2 summarizes the study participants divided into the two subgroups of interest (i.e., no cognitive fatigue and cognitive fatigue according to the FSMC subscore). Of the recruited MS patients, 21 did not have cognitive fatigue, and 27 were cognitively fatigued. Of those we included in our analysis, 19 participants did not have fatigue, and 23 were fatigued. Figure 5.8 shows the flow chart of the study and an overview of the excluded patients. The gender distribution of the participants in the two groups, the mean and standard deviation of their age, EDSS, and the FSMC subscales are listed in Table 5.2. As expected, we found a significant difference in all the FSMC scores. However, we found no statistically significant difference between the age and gender distributions of the two groups.

5.2.2. Correlation to Clinical Data

Our analysis indicates a significant Spearman rank correlation between several proposed general metrics and the clinical data. Table 5.3 shows an overview of all the computed correlations. The *response time* and *correct* metrics showed the highest correlation with EDSS (ρ=0.6, P<.001 and ρ=-0.6, P<.001, respectively). Then, *calibrated rate* follows with ρ=0.5, P=.001. On the other hand, *errors* did not significantly correlate to EDSS (ρ=-0.07, P=.67). We also found a significant correlation when analyzing the relationship between our metrics and the FSMC cognitive subscore. Again, *response time* and *correct* showed the highest correlation to the FSMC subscore (ρ=0.39, P=.01 and ρ=-0.38, P=.01, respectively). Neither *calibrated rate* (ρ=0.27, P=.09) nor *errors* (ρ=0.1, P=.51) significantly correlated to the FSMC cognitive subscore. Age also correlates to the proposed general performance metrics.

Table 5.2.: Demographic Characteristics of Participants

		No fatigue	Cognitive fatigue	*P*
Number		19	23	
Age, mean (SD)		36.89 (12.15)	38.22 (12.20)	.73
Gender, n (%)				
	m	8 (42)	6 (26)	.44
	w	11 (58)	17 (74)	
MS type, n(%)				
	PMS	1(5)	3 (13)	.61
	RRMS	18(95)	20 (87)	
Disease duration, mean (SD)		9.63 (5.88)	12.52 (8.51)	.20
DMT, n(%)				
	None	1 (5)	1 (4)	
	Interferon beta-1a	1 (5)	0 (0)	
	Dimethyl fumarate	2 (11)	1 (4)	
	Teriflunomide	1 (5)	1 (4)	
	Glatiramer acetate	1 (5)	1 (4)	
	Fingolimod	1 (5)	1 (4)	
	Natalizumab	6 (32)	8 (35)	
	Rituximab	1 (5)	3 (13)	
	Ocrelizumab	5 (26)	7 (31)	
Fatigue medication, n (%)				
	None	19 (100)	22 (96)	1.00
	Modafinil	0 (0)	1 (4)	
EDSS, mean (SD)		1.00 (1.18)	2.41 (1.95)	.006
FSMC, mean (SD)				
	Total	30.84 (8.00)	64.30 (16.29)	<.001
	cognitive	14.26 (3.25)	31.70 (8.81)	<.001
	Motor	16.58 (5.60)	32.61 (8.68)	<.001

Notes: Data are mean (SD) or n (%). PMS: progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis; Disease duration is measured in years since first manifestation; EDSS: expanded disability status scale; FSMC: Fatigue Score for motor functions and cognition; DMT: disease modifying therapy.

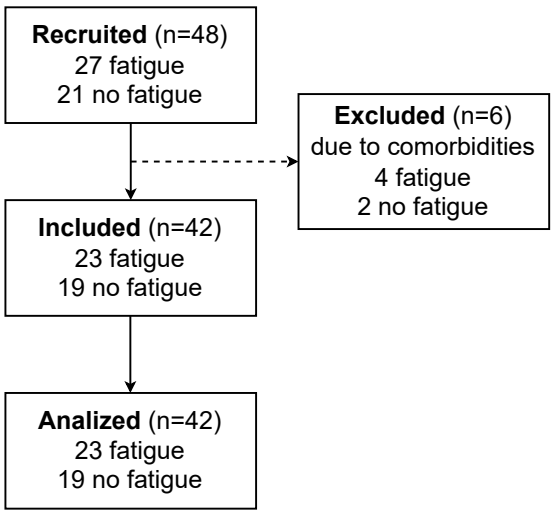

Figure 5.8.: Flow chart of the study and overview of excluded participants.

Table 5.3.: Spearman rank correlation coefficient ρ: metrics vs. clinical data

	EDSS	**FSMC cognitive score**	**Age**
response time	0.6 (<.001)	0.39 (.01)	0.61 (<.001)
calibrated rate	0.5 (.001)	0.27 (.09)	0.51 (.001)
correct	−0.6 (<.001)	−0.38 (.01)	−0.66 (<.001)
errors	−0.07 (.67)	0.1 (.51)	0.01 (.93)
$\Delta correct$	−0.03 (.86)	−0.21 (.17)	0.08 (.59)
$\Delta response\ time$	0.21 (.17)	0.24 (.13)	0.08 (.59)
$\Delta errors$	−0.13 (.40)	0.13 (.42)	−0.2 (.19)

Note. Data are ρ (P). EDSS, expanded disability status scale; FSMC, fatigue score for motor functions and cognition.

Among the correlating metrics, we found *correct* (ρ=-0.66, P<.001), *response time* (ρ=0.61, P<.001), and *calibrated rate* (ρ=0.51, P=.001). We found no significant correlation between the fatigability metrics and the clinical data.

5.2.3. cFAST Relationship to Perceived Fatigue

We investigated the relationship between our metrics and perceived fatigue by determining statistically significant differences between the cognitive-fatigued and non-fatigued groups. Table 5.4 depicts a complete overview of the metrics' mean value and standard deviation for both groups, as well as the t-test results. In addition, Table C.2 in Appendix C includes the non-parametric testing results using Mann-Whitney U.

Table 5.4.: Metrics comparison between fatigued and non fatigued patients with mean (SD), independent samples t-test (two-tailed) to assess whether there is a statistically significant difference between the groups, and Cohen's d effect size.

	No fatigue	**Cognitive fatigue**	**t**	***P***	***Cohen's d***
*response time**	2083.3 (358.31)	2586.88 (961.28)	2.16	.04	0.669
*calibrated rate***	3289.47 (1229.75)	3922.91 (1396.06)	1.54	.13	0.478
correct	109.11 (15.97)	90.96 (24.21)	−2.8	.008	−0.868
errors	7.58 (6.07)	8.04 (4.13)	0.29	.77	0.091
$\Delta correct$	3.51 (11.19)	−2.73 (9.95)	−1.91	.06	−0.593
$\Delta response\ time$	−0.96 (5.5)	2.69 (4.94)	2.27	.03	0.703
$\Delta errors$	−0.46 (2.05)	0.03 (1.86)	0.81	.42	0.252

*response time is not normally distributed for the subgroup cognitive fatigue.
**calibrated rate is not normally distributed for the subgroups.

We found a significant difference between both groups regarding *response time* (t=2.16, *P*=.04, d=0.669). The group with cognitive fatigue had an average *response time* of 2586.88 (SD=961.28) *ms*, compared to the 2083.3 (SD=358.31) *ms* of non-fatigued participants. We did not find a statistically significant difference in *calibrated rate* (t=1.54, *P*=.13). Furthermore, we found that *correct* differed significantly between the groups (t=-2.8, *P*=.008, d=-0.868). The non-fatigued participants gave an average of 109.11 (SD=15.97) *correct* answers, while the fatigued group had an average of 90.96 (SD=24.21) *correct* answers. However, *errors* was not significantly different between the groups (t=0.29, *P*=.77).

Regarding the fatigability metrics, we found that $\Delta response\ time$ significantly differed between the groups (t=2.27, *P*=.03, d=0.703). On average, fatigued participants had a $\Delta response\ time$ of 2.69 (SD=4.94) *ms*, while non-fatigued participants had an average $\Delta response\ time$ of -0.96 (SD=5.5) *ms*. $\Delta errors$ and $\Delta correct$ did not show a statistically significant difference between the groups (t=0.81, *P*=.42 and t=-1.91, *P*=.06, respectively).

To analyze the temporal progression of participants' performance during a cFAST session, we performed a series of paired t-tests. Figure 5.9 on the left depicts the average normalized *response time* in the three thirds of the session for non-fatigued MS patients. While Figure 5.9, on the right, shows the results for MS patients with cognitive fatigue. For the group with no fatigue, the results are primarily flat and with a slight trend to improve over time, while for the fatigued group, we see a significant increase in *response time* (*P*=.02) between the first and last third of the session.

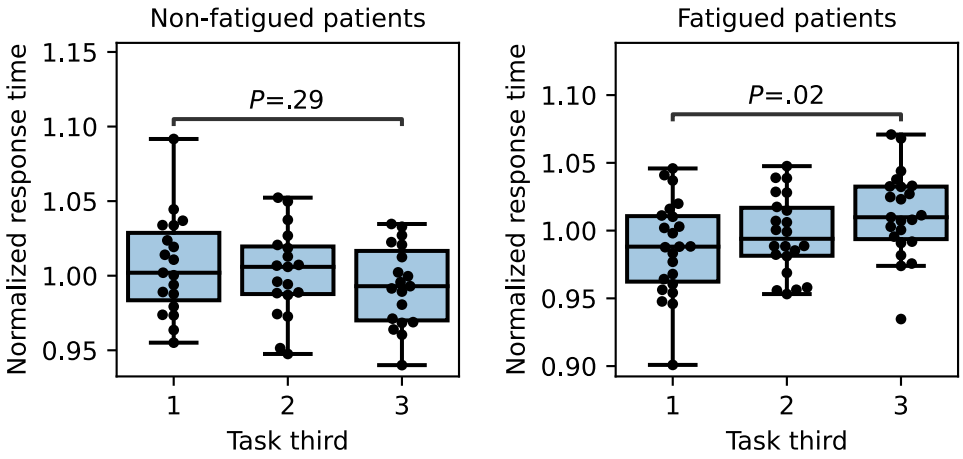

Figure 5.9.: Average normalized *response time* during the three-thirds of the cFAST session data after preprocessing for non-fatigued patients (left) and fatigued patients (right). A significant increase in the response time between the first and the last third of the task is present for fatigued patients only. The thirds were compared using a paired t-test.

5.2.4. cFAST Relationship to Disability

The disabled group has a mean EDSS of 3.26 (SD=1.54), and the non-disabled group has a mean EDSS of 0.54 (SD=0.67). The detailed demographics of these groups are described in Appendix C Table C.1. Table 5.5 shows a complete overview of the metrics' average value and standard deviation for both groups and the t-test results. Table C.3 in Appendix C includes the non-parametric testing results using Mann-Whitney U.

We found a significant difference in *response time* between the groups (t=2.47, *P*=.02, d=0.844). Participants without disability had an average *response time* of 2080.23 (SD=317.39) *ms*, compared to the 2696.61 (SD=1030.37) *ms* exhibited by the disabled patients. Similarly, *calibrated rate* was significantly lower for participants without disability (t=2.38, *P*=.02, d=0.737), with an average of 3211.22 (SD=840.63) *ms* against the 4151.0 (SD=1658.95) *ms* of disabled participants. Consequently, *correct* followed the same trend (t=-3.19, *P*=.003, d=-0.989). On average, disabled MS patients provided 88.11 (SD=25.44) *correct* answers, compared to the higher 108.3 (SD=15.1) of participants without disability. We found no significant difference in *errors* (t=-0.17, *P*=.86).

We performed the same analysis with the fatigability metrics. $\Delta errors$, $\Delta response\ time$, and $\Delta errors$ showed no statistically significant difference between the not disabled and disabled groups (respectively t=-0.3, *P*=.77, t=1.98, *P*=.33 and t=-0.6, *P*=.55).

Table 5.5.: Metrics comparison between disabled and not disabled patients with mean (SD), independent samples t-test (two-tailed) to assess whether there is a statistically significant difference between the groups, and Cohen's d effect size.

	Not disabled (n=23)	**Disabled (n=19)**	**t**	***P***	***Cohen's d***
*response time**	2080.23 (317.39)	2696.61 (1030.37)	2.47	.02	0.844
*calibrated rate***	3211.22 (840.63)	4151.0 (1658.95)	2.38	.02	0.737
correct	108.3 (15.1)	88.11 (25.44)	−3.19	.003	−0.989
errors	7.96 (5.69)	7.68 (4.26)	-0.17	.86	-0.053
$\Delta correct$	0.55 (10.68)	−0.47 (11.35)	−0.3	.77	−0.093
$\Delta response\ time$	0.29 (5.55)	1.95 (5.33)	1.98	.33	0.304
$\Delta errors$	−0.03 (2.05)	−0.4 (1.83)	−0.6	.55	−0.187

*response time is not normally distributed for the subgroup disabled. Levene's Test P<.05 equal variance not assumed.

**calibrated rate is not normally distributed for the subgroups.

5.2.5. Predictive Power of the cFAST Metrics to Classify Cognitive Fatigue

To further explore the association between cognitive fatigability and perceived fatigue, we assessed the predictive power of our metrics to classify cognitive fatigue participants according to the FSMC cognitive subscale. Table 5.6 shows the results corresponding to the mean AUROC with its respective confidence intervals. The results indicate that the best features for fatigue independently of the EDSS are the fatigability metrics. $\Delta response\ time$ had the highest AUROC with 0.74 (95% CI 0.64-0.84). Following, $\Delta correct$ and $\Delta errors$ with an average AUROC of 0.72 (95% CI 0.63-0.85) and 0.65 (95% CI 0.53-0.77), respectively. From the general metrics, *response time* performed the best with a mean AUROC of 0.63 (95% CI 0.50-0.76). The *correct* metric had an AUROC of 0.62 (95% CI 0.50-0.74). *Calibrated rate* produced an AUROC of 0.59 (95% CI 0.44-0.74). The *errors* metric showed an AUROC of 0.58 (95% CI 0.44-0.72). Age had an average AUROC of 0.58 (95% CI 0.47-0.69). Lastly, EDSS had an AUROC of 0.53 (95% CI 0.43-0.63).

5.2.6. Predictive Power of the cFAST Metrics to Classify Disability

To evaluate the best cFAST metrics to classify disability independently of fatigue, we performed the same analysis as we did for cognitive fatigue. Results suggest that the *general* metrics are better than the $fatigability$ metrics for disability in AUROC. A complete overview of these results is shown in Table 5.7. *Response time* produced an average AUROC of 0.64 (95% CI 0.50-0.78), followed by age with an average AUROC of 0.63 (95% CI 0.53-0.73). Following, *correct* showed an average AUROC

Table 5.6.: AUROC score corresponding for cognitive fatigue classification according to the FSMC cognitive subscale for the proposed metrics (sorted by AUROC in descending order).

Metric type	Metric name	↓AUROC (95% CI)
Fatigability	*Δresponse time*	0.74 (95% CI 0.64-0.84)
Fatigability	*Δcorrect*	0.72 (95% CI 0.63-0.85)
Fatigability	*Δerrors*	0.65 (95% CI 0.53-0.77)
General	*response time*	0.63 (95% CI 0.50-0.76)
General	*correct*	0.62 (95% CI 0.50-0.74)
General	*calibrated rate*	0.59 (95% CI 0.44-0.74)
General	*errors*	0.58 (95% CI 0.44-0.72)
Demographic	Age	0.58 (95% CI 0.47-0.69)
Demographic	EDSS	0.53 (95% CI 0.43-0.63)

of 0.63 (95% CI 0.49-0.77). *Calibrated rate* had an average AUROC of 0.59 (95% CI 0.43-0.75). *Δerrors* had a mean AUROC of 0.55 (95% CI 0.41-0.69). Following, *Δcorrect* produced an AUROC of 0.52 (95% CI 0.38-0.66). AUROC of *errors* for disabled patients was 0.51 (95% CI 0.38-0.64). Finally, *Δresponse time* was the worst metric for disability with an average AUROC of 0.50 (95% CI 0.36-0.64).

5.2.7. Differences in Predictive Power Between the Best Fatigue and Disability Metrics

Figure 5.10 on the left shows a visual representation of the ROC curves corresponding to the FSMC classification for *Δresponse time*, best-performing feature to classify cognitive fatigue and *response time*, best-performing feature to classify disability. *Δresponse time* outperforms *response time* by 11 percentage points in classifying fatigue according to the FSMC. The center of the figure shows boxplots of *Δresponse time* for the groups fatigued and non-fatigued, as well as the t-test results. The image displays the statistically significant difference between the fatigue and non-fatigued groups (t=2.27, P=.03). Similarly, the right shows the boxplots corresponding to the *response time*. There is a statistically significant difference between the groups (t=2.16, P=.04). The difference is significant also without the outlier in the fatigue group.

We conducted a one-way analysis of covariance (ANCOVA) to examine whether *response time* differed between fatigue and non-fatigue groups when controlling for EDSS. For this analysis, we did remove the outlier in *response time* in the fatigue group as the outlier violated the normality assumptions of ANCOVA. We verified the test assumptions: Shapiro-Wilk test indicates the data is normally distributed for the

Table 5.7.: AUROC score corresponding to disability classification according to the EDSS split with threshold 1.5 for the proposed metrics (sorted by AUROC in descending order).

Metric type	**Metric name**	**↓AUROC (95% CI)**
Fatigability	*Δresponse time*	0.74 (95% CI 0.64-0.84)
Fatigability	*Δcorrect*	0.72 (95% CI 0.63-0.85)
Fatigability	*Δerrors*	0.65 (95% CI 0.53-0.77)
General	*response time*	0.63 (95% CI 0.50-0.76)
General	*correct*	0.62 (95% CI 0.50-0.74)
General	*calibrated rate*	0.59 (95% CI 0.44-0.74)
General	*errors*	0.58 (95% CI 0.44-0.72)
Demographic	Age	0.58 (95% CI 0.47-0.69)
Demographic	EDSS	0.53 (95% CI 0.43-0.63)

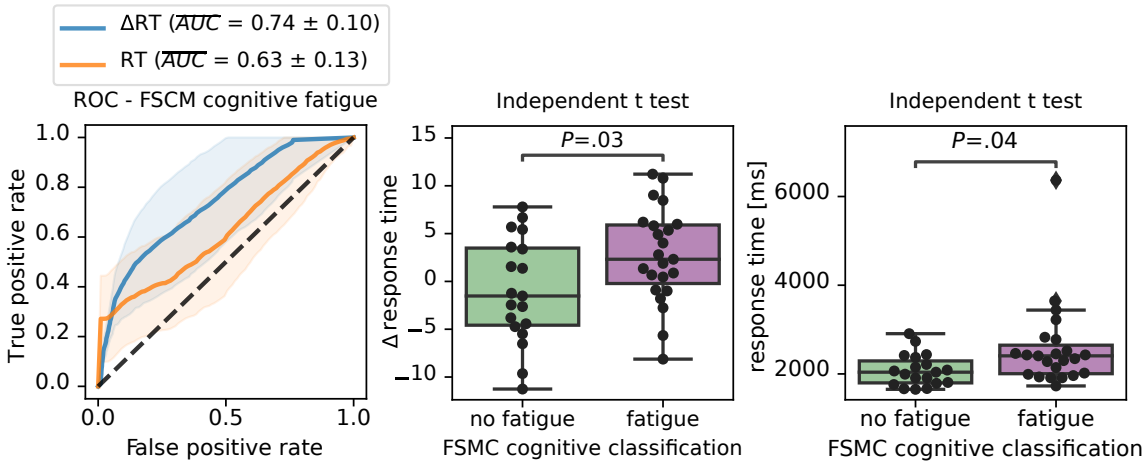

Figure 5.10.: Mean AUROC for cognitive fatigue according to FSMC cognitive subscale (N=42). ROC curves for *Δresponse time* (*ΔRT*) and *response time (RT)* are displayed on the left. Data generated using Monte-Carlo simulation with 1000 iterations. The center of the figure shows the t-test results for *Δresponse time*, the feature with the highest AUROC for fatigue. *Δresponse time* and *response time* show a statistically significant difference between the fatigue groups.

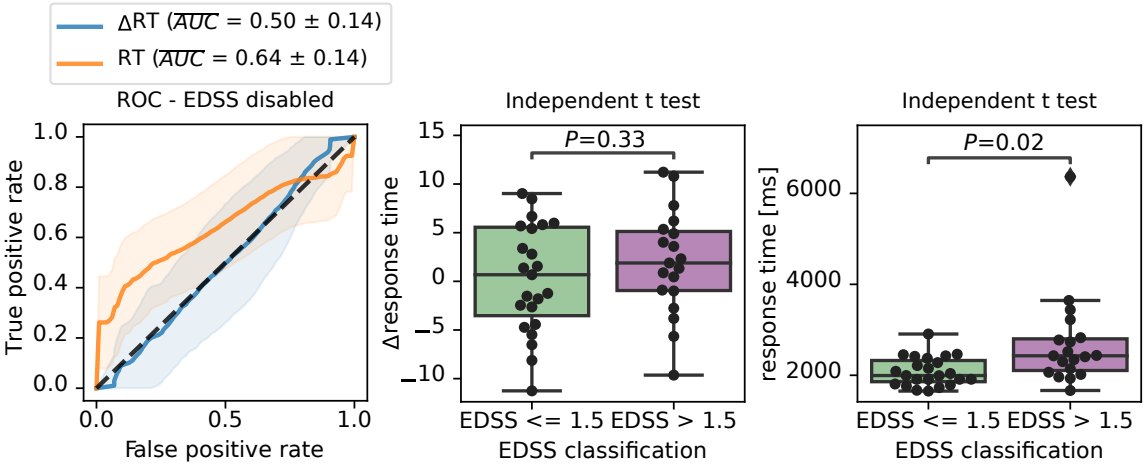

Figure 5.11.: Mean AUROC for disability according to EDSS (N=48). ROC curves for response time (RT) *Δ response time* (*ΔRT*) are displayed on the left. The center shows the t-test results for *Δ response time*, the feature with the highest AUROC for fatigue, and *response time* (right side). *Δresponse time* does not show a statistically significant difference between the disability groups, while response time does.

group with no fatigue W(19)=.926 (*P*=.15) but not for the fatigued group W(22)=.899 (P=.03). However, as the distribution is close to normal and ANCOVAs are robust to this assumption violation, no further steps were taken. Visual analysis with a scatter plot indicates similar regression slopes. An F test indicates no interaction between EDSS and fatigue group F=(1,37)=.24 (*P*=.64). Finally, Levene's Test confirms the homogeneity of variance with F(1,39)=1.27 (*P*=.27). ANCOVA analysis reveals that after controlling for EDSS (disability), there was no significant difference in response time between the fatigue groups F(1,38)=1.42, *P*=.24. For a similar analysis on *correct*, refer to Appendix C.1.

Figure 5.11 shows data corresponding to disability classification according to the EDSS threshold. The left side of Figure 5.11 shows a visual representation of the ROC curves corresponding to the disability classification for *Δresponse time* (*ΔRT*), the best-performing feature to classify cognitive fatigue and *response time*, the best-performing feature to classify disability. In this case, *response time* outperforms *Δresponse time* by 14 percentage points.

5.3. Discussion

We described the development process and pilot study of a new test (cFAST) for cognitive fatigability. Our result provides early evidence that the cFAST measurement could help identify patients with cognitive fatigue, as assessed by the FSMC cognitive

subscale. So far, only a few studies assess cognitive fatigability with specific tasks in MS patients [DeLuca et al., 2008; Chen et al., 2020]. Moreover, previous results are contradictory, with some showing fatigability while others not [DeLuca et al., 2008; Chen et al., 2020]. Cognitive fatigability studies are in their infancy, and research could benefit from new approaches and validation studies. Our approach differs from previous methods in that it is tailored to patients' disabilities with its calibration mechanism that also enforces rapid decision-making, which we believe contributes to eliciting cognitive fatigability within a single test session and in a short period. In addition, our smartphone-based test is easy to administer, portable, and designed to be applied outside clinical settings, potentially allowing for remote and frequent monitoring. Concerning cognitive testing, healthy controls, and MS patients perceive the PASAT as unpleasant and less likable, while the SDMT is preferred and found appropriate for cognitive testing [Walker et al., 2012b]. Thus, we believe cFAST will have good acceptance as it follows a similar logic to the SDMT and does not require patients to perform arithmetic operations under pressure like the PASAT. However, user acceptance of the cFAST must be assessed in future studies.

5.3.1. *Fatigability* Metrics Relate to Fatigue, While *General* Metrics Relate to Disability

We derived two sets of metrics from cFAST: *fatigability* and *general* metrics. Our initial group-level analysis with a t-test revealed statistically significant differences between fatigued and non-fatigued patients with several *general* and *fatigability* metrics. Overall, we found more significant differences between the groups with the *general* metrics than fatigability *metrics*. However, the ANCOVA analysis revealed that EDSS is associated with the metrics *response time* and *correct*. Furthermore, the statistical difference in the fatigue groups in terms of these metrics is due to disability and not fatigue. Hence, after controlling for EDSS, the statistical difference between the groups disappears. We further analyzed how the groups' differences related to patients' disabilities. To this end, we divided our study population into two groups according to EDSS, disabled (EDSS>1.5) and non-disabled (EDSS≤1.5). This grouping revealed statistically significant differences with the *general* metrics but not with the *fatigability* metrics. This result suggests that *general* metrics are related to and confounded by disability, while this is not true for the *fatigability* metrics. We conducted the AUROC analysis controlling for disability with Monte-Carlo simulations and stratified splits to further rule out the effect of disability from the fatigue analysis. These results confirmed our hypothesis that *fatigability* metrics are better predictors of fatigue than *general* metrics. $\Delta response\ time$, the best-performing metric to classify fatigue (with an average AUROC of 0.74), is 11 percentage points above *response time*, the best-performing *general* metric for fatigue. Conversely, *general* metrics dominate the disability classification, with *response time*

being the best metric (average AUROC of 0.64), 9 percentage points above the best fatigability metric $\Delta errors$. Analysis of the *fatigability* metrics revealed that, on average, performance during the tests tends to worsen for fatigued patients, while patients without fatigue tend to improve. Previous work on fatigability showed decline towards the end of sustained cognitive activity in MS patients while controls did not [Schwid et al., 2003; Bryant et al., 2004]. Our findings go in line with these results. However, our analysis focused only on MS patients to decrease disease-specific confoundings.

5.3.2. Consideration for Remote and Unsupervised Monitoring

We designed and implemented cFAST to achieve remote monitoring. Hence, cFAST seeks to be self-explanatory. For instance, trials aim at familiarizing the users with the core test logic of matching numbers to symbols following the shown mapping rule. Thanks to the feedback displayed after every answer, users can quickly realize when they are making mistakes. The immediate feedback, together with the requirement of at least 70% correct answers out of a minimum of 20, helps us determine if the user has correctly understood the test logic and the need to perform it quickly. As described in the methods section, we derived the pace of the cFAST, calibrated rate, from the calibration phase. The speed requirement seeks to induce cognitive fatigability in a short period. Calibrated rate is derived for each patient, personalizing the test and adjusting for the different disability spectrums and baseline performance of the patients.

5.3.3. Limitations and Future Work

A limitation of our study is the lack of a gold standard cognitive fatigability assessment to validate our approach. Currently, there is no established method to quantify cognitive fatigability. Until now, existing research has used cognitive tests protracted for extended periods as an attempt to induce and quantify fatigability. However, these approaches tend to be long, tedious, and costly. Moreover, the results from these experiments are inconclusive. Hence, we directly compared our metrics to a widely accepted and validated fatigue questionnaire within MS research, the FSMC. The FSMC has the advantage of offering a subscale to evaluate cognitive fatigue independently of physical fatigue. Another limitation of our study is our sample size, which is limited to 42 study participants. We are aware that more extensive evaluations are needed to determine if the test can be established as a surrogate for perceived cognitive fatigue for clinical decision-making. In particular, our pilot study uses a cross-sectional design. Thus, we are not able to define the clinical significance of the changes in the fatigability scores in individual patients. Future studies are

needed to address this question. Finally, we designed cFAST to be suitable for remote and unsupervised monitoring. However, in this study, the evaluation was conducted within the hospital in a controlled environment. Further studies are needed to confirm the results, including longitudinal outside-the-hospital evaluations in larger MS cohorts and within-subjects comparison. Nevertheless, we believe our study offers a detailed evaluation of our newly developed cognitive fatigability test.

As part of future work and prior to the clinical implementation, more data has to be generated to further evaluate the generalization of the adaptation phase. Additionally, our study highlights the need for implementing changes to improve data quality in an unsupervised setting. First, we recommend incorporating a statement in the cFAST instructions about the importance of conducting the test in a distraction-free environment (i.e., activate 'do not disturb' modality, use a quiet room). Second, we recommend automatically dismissing test sessions if no input is recorded within a certain period after the start. Distractions in uncontrolled environments (e.g., incoming phone calls or messages) can result in empty test sessions or significant periods without data, thus producing erroneous values for the proposed set of metrics. Moreover, future studies should examine whether cFAST could aid clinicians in distinguishing between confounding such as depression, sleepiness, or others. Finally, we need to investigate further the frequency patients need to conduct the calibration phase in unsupervised settings. However, we believe that calibration has to be performed only once and that the calibrated rate can be recomputed, if necessary, directly from the existing patients' cFAST sessions. Nonetheless, this requires further studies, including longitudinal data.

5.4. Conclusion

In this chapter, we introduced cFAST, a novel smartphone-based test to quantify cognitive fatigability tailored to the user's disability by its calibration mechanism. With cFAST, we aim at having an objective surrogate of fatigue that allows monitoring individual patients over time in uncontrolled environments (e.g., at home). We do not seek to have a diagnostic tool but rather a solution for clinicians to make informed and timely decisions as to whether a patient's condition is improving or deteriorating and act accordingly. Results from our pilot study provide evidence supporting the validity of our approach and show that the fatigability metrics could potentially be used as a surrogate for perceived cognitive fatigue and motivate further research in this area.

Part III.

Wearables for Unsupervised Monitoring

CHAPTER 6

Accuracy of Heart Rate Sensors Based on PPG for In-The-Wild Analysis

This chapter is based on the following publication:

Liliana Barrios, Pietro Oldrati, Silvia Santini, Andreas Lutterotti. Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. *Proceedings of EAI International Conference on Pervasive Computing Technologies for Healthcare (EAI PervasiveHealth)*. Pages 251–261. Trento, Italy. May 20-23, 2019.

Part *I* and *II* of this dissertation focused on active monitoring of the symptom, particularly for quantifying physical and motor fatigue. Part *III* is about passive monitoring. In passive monitoring, the goal is that the system collects contextual information about the patient's behavior and the environment without the patient noticing, such as the patient's vital signs, level of activity, and weather conditions. We believe that combining passive and active monitoring will enable us to build different models for fatigue prediction. However, as a first step into passive monitoring, we need to find suitable commercially-available wearable devices that allow continuous monitoring of patients' vitals over extended periods. Furthermore, we need to verify the validity of important parameters, such as heart rate, that have already been associated with fatigue. Hence in this chapter, we introduce a validation study of commercially-available wearables. Standard wearable devices use photoplethysmography (PPG) to derive data on heart rate (HR) and heart rate variability HRV. However, it still needs to be determined to which extent PPG signals can be used as a proxy for data collected using medical-grade devices, particularly for HRV. To address this challenge, we consider five consumer devices to assess the signal quality of HR and two devices

measuring HRV and compare them with standard electrocardiography (ECG) Holter monitor. We collected data from fourteen participants who followed a 55 minutes protocol for at least two sessions. Using this data set, we show that PPG is a valid proxy for both HR and standard time- and frequency-domain measurements of HRV.

6.1. Methods

Our goal is to evaluate the performance of PPG sensors found in commodity wearable devices under different settings for measuring heart rate HR and inter-beat intervals (IBI). To this end, we conducted a series of experiments. Fourteen volunteers took part in the study, seven males with a median age of 33 (range 23-54) and seven females with a median age of 36 (range 26-51). Their mean height is 170 cm, and their mean weight is 67 kg. Volunteers gave full written informed consent to participate in the study. All procedures were approved by the ETH Zurich local committee (EK 2018-N-89).

6.1.1. Study Design

To assess the validity of the PPG sensors, we use a similar protocol to the one suggested by Jo et al. [2016]. The protocol starts with 5 minutes of resting on a stationary bike followed by five activities: biking (60 W), biking (120 W), walking (5 km/h), jogging (8 km/h) and running (10 km/h). Each activity lasts 5 minutes, and between each activity, there is a resting period of 5 minutes. The left side of Figure 6.1 depicts our study protocol.

6.1.2. Experiment I - Accuracy of PPG Based HR Monitors

The goal of this experiment is to compare the level of agreement of the Empatica E4 [Empatica Inc., 2018](version 1) and Everion [Biovotion AG, 2018](VSM1-3.0, M4 version 03.11.00) with the mean HR derived from popular fitness trackers: Fitbit Charge HR [Fitbit, 2018], Polar OH1 [Polar, 2018], and Wahoo Ticker Fit [Wahoo Fitness, 2018]. Participants wore two Empatica E4 devices (one on each wrist), two Everion devices (one per arm), and a medical-grade Holter monitor, the General Electric Seer 1000 [General Electric Healthcare, 2018] with five leads, as depicted on the right side of Figure 6.1. The fitness trackers were placed on the arm of the participants without a predefined position. Six participants took part in this experiment, three male and three female. Each participant completed our validation protocol two times on different days.

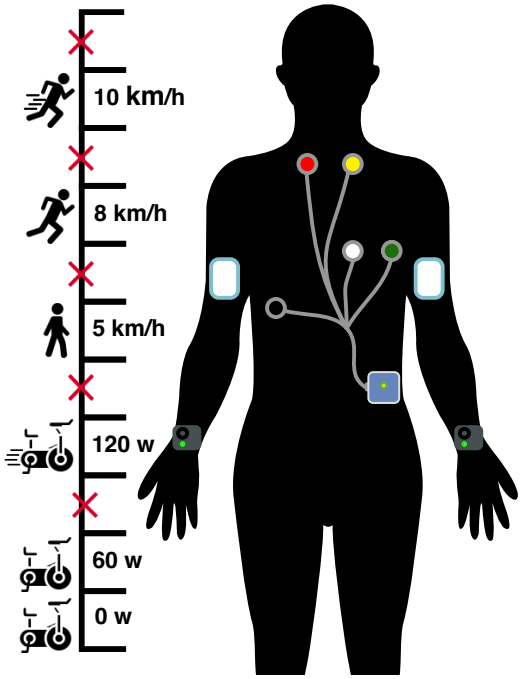

Figure 6.1.: Sensor validation protocol at the left side and sensor placement at the right side. Empatica E4 devices on the wrists, Everion devices on the arms, and the Holter monitor attached with five electrodes to the chest.

6.1.3. Experiment II - Comparing Everion, Empatica, and Holter

The goal of this experiment is to assess the heart rate and interbeat intervals derived from PPG sensors as a valid proxy for HRV. We considered two off-the-shelf sensors capable of measuring HR through photoplethysmography (PPG) and electrodermal activity (EDA): Empatica E4 and Everion devices. Fourteen subjects participated in the experiment and completed the protocol two times on different days. Participants wore two Empatica E4 devices (one on each wrist), two Everion devices (one per arm), and a medical-grade Holter monitor, the General Electric Seer 1000 [General Electric Healthcare, 2018] with five leads, to record ECG signals. The sensor placement is depicted on the right side of Figure 6.1. Moreover, we explore the variance between successive measurements. To this end, two participants performed three extra sessions. We compute the mean HR difference between Everion, Empatica, and Holter per activity and perform an ANOVA analysis per session.

6.1.4. Data Collection

Table 6.1 shows details regarding the data types, frequency, and export method. During the experiment, Polar data was stored locally on the devices and later exported

Table 6.1.: Collected sensor data.

Device	Variable	Frequency	Export
Empatica E4	HR (bpm)	1 Hz	Empatica
	IBI	-	
Everion	HR (bpm)	1 Hz	Bluetooth LE
	IBI	-	
Seer 1000	HR (bpm)	1 Hz	CardioDay V2.5
	RR-intervals	-	
Polar	HR (bpm)	1 Hz	PolarFlow
Fitbit	HR (bpm)	1/3 Hz (varies)	Fitbit.com
Wahoo	HR (bpm)	1 Hz	Wahoo Fitness

using the Polar Flow smartphone application. Similarly, Wahoo data was exported to a CSV file using the Wahoo Fitness application. Fitbit data was downloaded from fitbit.com [googlefitbit, 2016]. Empatica data was stored locally on the devices and later exported with the Empatica Connect software. Everion data was streamed via Bluetooth Low Energy to an Android phone during the experiment and later exported as a CSV file. Finally, QRS complexes were obtained from the ECG Holter with the software CardioDay [General Electric Healthcare, 2019] from GE Healthcare.

6.1.5. Data Analysis

Before evaluating the level of agreement of the different devices involved in our experiments, interbeat interval sequences derived from the ECG and PPG devices were aligned through cross-correlation. Additionally, we did an outlier analysis and excluded data points resulting from potential errors or artifacts caused during data acquisition, i.e., HR equal to zero during the experiment.

Metrics

We use different metrics to measure the performance and level of agreement of the different devices. We report the mean and standard deviation of the HR. We evaluate the existence of bias, with its limits of agreement [LoA], using the Bland-Altman [Bland and Altman, 1986] plot. The Bland-Altman plot [Bland and Altman, 1986] is a plot of the difference between two methods against their mean, allowing the investigation of any possible relationship between the measurement error and the true value. In this plot, none of the values are considered to be the true value. Thus, the mean value is used as the best estimate. In our analysis, we consider the HR derived from the Holter versus HR derived from the wearable devices. Additionally, we

compute the intraclass correlation coefficient (ICC) with its 95% confidence interval, Pearson correlation ($corr$) and squared error R^2.

Following Koo and Li [2016] guidelines for selecting and reporting ICC, we computed ICC and its 95% confident intervals using IBM SPSS statistics [Armonk, Released 2017] based on single measurement type, absolute agreement definition and 2-way mixed-effects model. ICC results are interpreted as in [Koo and Li, 2016]: values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.

Heart rate variability

We derive different time and frequency domain measures of HRV from the IBI and NN time series provided by Empatica, Everion, and ECG Holter. An overview of the metrics is depicted in Table 6.2. According to the recommendations by Richardson et al. [1996], 5 minutes is an appropriate length for short-term recordings of HRV. When analyzing the spectrum for short-term recordings time varies between 2 and 5 minutes. We use fast Fourier transformation (FFT) to derive frequency domain HRV measurements from the IBI interval time series. In accordance with Richardson et al. [1996], we divide the power spectrum for frequency domain HRV analysis into the following bands: VLF (0.00 - 0.04 Hz), LF (0.04 - 0.15 Hz) and HF (0.15 - 0.40 Hz). For the calculation of HRV parameters, we select identical segments larger than 180 s of NN intervals from the ECG and a wearable device. Then, we apply cubic interpolation. Finally, we analyze the spectrum with Welch's periodogram using the following parameters: hamming window, an overlap of 50%, and linear detrend.

6.2. Results

6.2.1. Accuracy of PPG Based HR monitors

We compare the performance of the Empatica and the Everion versus commonly used fitness trackers. Table 6.3 depicts different metrics comparing each wearable with the Holter. The table is organized by device with its associated sample size. Additionally, we split each case into activities. To make a fair comparison between the fitness trackers, Empatica and Everion, we did not use quality filters on the data as this functionality is not available on the fitness trackers. Figure 6.2 depicts the Bland Altman plot for all devices.

Table 6.2.: Overview of the heart rate variability metrics computed.

Metric	**Domain**	**Definition**
RMSS	time	Square root of the mean squared differences of successive NN intervals.
SDNN	time	Standard deviation of the NN intervals.
NN50	time	Number of interval differences of successive NN intervals greater than 50 ms.
pNN50	time	Proportion derived by dividing NN50 by the total number of NN intervals.
VLF	frequency	Very low frequency.
LF	frequency	Low frequency.
HF	frequency	High frequency.
LFnu	frequency	Normalized low frequency.
HFnu	frequency	Normalized high frequency.
LF:HF	frequency	Ratio.

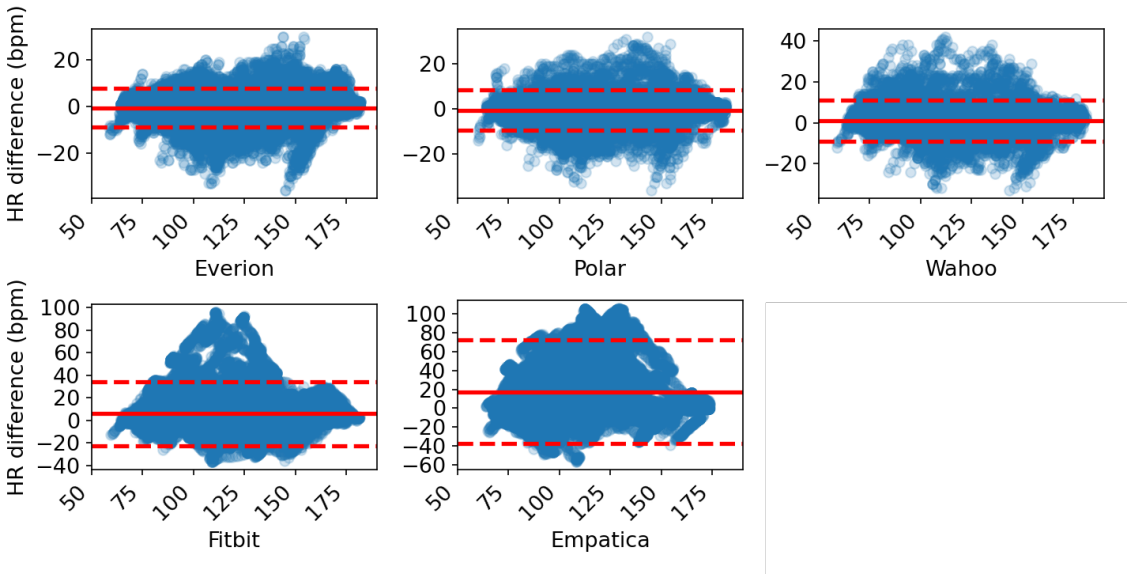

Figure 6.2.: Bland-Altman plot with LoA for each HR monitor. Wrist-based devices show largest bias (Empatica 17.35 [-37.16, +71.86] and Fitbit 5.89 [-22.19, +33.97]) than armband-based monitors (Everion -0.46 [-8.67, +7.75], Polar -0.51 [-9.38, 8.36], and Wahoo 1.01 [-8.95, +10.96]).

Table 6.3.: Experiment I - Heart rate analysis per activity

Device	Activity	Size	ICC [95% CI]	Corr	Bias [95% LoA]
Empatica	init	6268	.464 [+.422,+.502]	0.50	+2.83 [-21.04, +26.69]
(79241)	rest	33676	.466 [+.349,+.558]	0.51	+7.21 [-26.46, +40.88]
	bike (60 W)	7856	.167 [-.002,+.314]	0.29	+17.95 [-24.07, +59.96]
	bike (120 W)	7968	.223 [-.027,+.422]	0.42	+32.64 [-33.95, +99.24]
	walk	8096	.433 [+.331,+.517]	0.47	+6.60 [-26.32, +39.52]
	jog	7690	.026 [-.015,+.068]	0.06	+35.76 [-25.11, +96.63]
	run	7687	.016 [-.014,+.046]	0.05	+50.08 [-18.77 ,+118.94]
	avg		**0.256 [+.149, +.346]**	**0.33**	**+21.87 [-25.10, +68.84]**
Everion	init	6268	.957 [+.953,+.960]	0.96	-0.67 [-8.59, +7.25]
(78821)	rest	33590	.972 [+.967,+.976]	0.97	-1.06 [-9.59, +7.47]
	bike (60 W)	7856	.981 [+.980,+.982]	0.98	-0.27 [-5.15, +4.61]
	bike (120 W)	7666	.988 [+.988,+.989]	0.99	-0.18 [-6.31, +5.95]
	walk	8096	.993 [+.993,+.993]	0.99	-0.23 [-4.05, +3.58]
	jog	7683	.965 [+.962,+.969]	0.97	+0.89 [-8.29, +10.07]
	run	7662	.952 [+.950,+.954]	0.95	+0.29 [-11.73, +12.30]
	avg		**0.972 [+.970,+.974]**	**0.97**	**-0.17 [-7.67, +7.32]**
Fitbit	init	3134	.767 [+.645,+.838]	0.82	+3.80 [-11.34, +18.95]
(39624)	rest	16812	.827 [+.821,+.832]	0.84	+1.43 [-20.21, +23.07]
	bike (60 W)	3928	.499 [+.170,+.682]	0.63	+9.59 [-16.21,+35.38]
	bike (120 W)	3984	.353 [+.079,+.542]	0.48	+18.81[-31.24, +68.87]
	walk	4048	.729 [+.541,+.824]	0.81	+5.43 [-13.62, +24.47]
	jog	3854	.703 [+.441,+.822]	0.78	+8.09 [-17.09, +33.27]
	run	3852	.782 [+.468,+.887]	0.86	+8.22 [-13.93, +30.38]
	avg		**0.665 [+.452, +.776]**	**0.74**	**+7.91 [-17.66, +33.48]**
Polar	init	3134	.959 [+.936,+.953]	0.95	-1.08 [-9.93, +7.77]
(39624)	rest	16812	.969 [+.959,+.976]	0.97	-1.46 [-10.27, +7.35]
	bike (60 W)	3928	.972 [+.970,+.973]	0.97	-0.13 [-6.24,+5.99]
	bike (120 W)	3984	.984 [+.983,+.985]	0.98	+0.28 [-7.10, +7.66]
	walk	4048	.989 [+.988,+.989]	0.99	+0.01 [-4.86, +4.89]
	jog	3854	.964 [+.957,+.969]	0.97	+1.19 [-8.41, +10.80]
	run	3852	.950 [+.947,+.953]	0.95	+0.69 [-11.89,+13.26]
	avg		**0.969 [+.963,+.971]**	**0.97**	**-0.07 [-8.38, +8.24]**
Wahoo	init	3094	.913 [+.875,+.936]	0.92	+2.06 [-8.41,+12.53]
(38492)	rest	16406	.965 [+.964,+.967]	0.97	+0.58 [-9.22,+10.38]
	bike (60 W)	3567	.971 [+.966,+.974]	0.97	+0.60 [-4.88, +6.08]
	bike (120 W)	3732	.984 [+.982,+.986]	0.99	+0.80 [-6.21, +7.81]
	walk	3975	.972 [+.964,+.977]	0.97	+1.14 [-6.36 ,+8.65]
	jog	3854	.939 [+.918,+.952]	0.95	+2.13 [-10.29, +14.54]
	run	3852	.937 [+.930,+.944]	0.94	+1.30 [-12.56, +15.18]
	avg		**0.954 [+.943,+.962]**	**0.96**	**+1.23 [-8.28, +10.74]**

Wrist-based devices

Figure 6.2 shows that the wrist-based devices, Empatica and Fitbit, have the largest bias, 17.35 bpm and 5.898 bpm, respectively. The analysis per activity (Table 6.3) shows that the bias increases with the activity level in both cases. The Fitbit is more affected during bike activities, and the Empatica during jogging and running. Figure 6.3 shows an overview of the level of agreement of each device per activity. These results are consistent with the bias analysis. Wrist-based devices are more affected than armband-based devices. The Empatica shows the lowest agreement (poor agreement) within all different activities, especially during a jog and run. However, the results for Empatica are improved after filtering the data, resulting in good agreement during the initial and rest activities. Fitbit has the lowest agreement during bike activities, good agreement during rest, and moderate agreement during the initial activity. A possible explanation for the poor agreement during bike activities with these devices is the posture of the wrist on the bike. Bending of the wrist can generate loose contact between the skin and the heart rate monitor resulting in low-quality measurements.

Armband-based devices

In Figure 6.2, we can observe that the bias and data distribution is similar for all armband-based devices, i.e., Everion, Polar, and Wahoo. The three devices show smaller bias in comparison to the wrist-based devices. There is no particular trend in the bias depending on the activity. Similarly, Figure 6.3 shows that all devices have a similar level of agreement in terms of ICC. Everion, Polar, and Wahoo show excellent reliability with regards to the Holter in all the activities, showing that armband devices are less susceptible to artifacts due to movement.

Users' preferences

After the experiments, participants completed a short questionnaire indicating their preferred style of a wearable device (armband, wristband) for continuous monitoring during (i) day, (ii) sleep, and (iii) 24/7. A Cochran's Q test did not indicate any differences among the three proportions, $p = .717$, showing that the user's preference is not affected by the duration of the monitoring phase.

6.2.2. Comparing Everion, Empatica and Holter

To compare both devices, we started by computing metrics corresponding to the mean HR derived from the Empatica E4 and Everion relative to the medical-grade

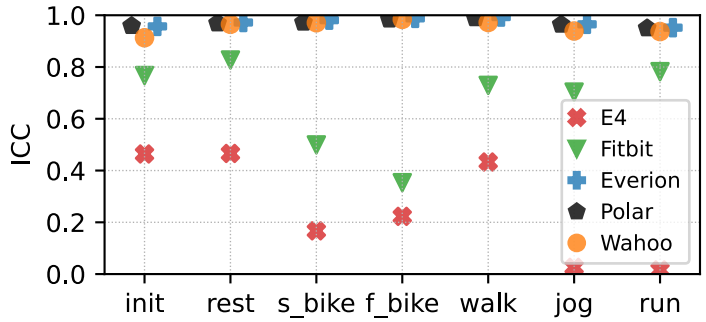

Figure 6.3.: Level of agreement according to ICC for each device in experiment II. Notably the level of agreement of the wrist-based devices is lower than for the armband-based devices with Empatica being more affected as the activity level increases.

Holter. Empatica and Everion have mechanisms to assess the quality of the retrieved heart rate, i.e., low or high quality. We refer to high-quality datasets as *Everion Best* and *Empatica Best*. Everion provides a heart rate quality parameter to filter values depending on their quality. In our analysis *Everion Best* corresponds to the HR quality parameter of 99%. Additionally, we include in our analysis a dataset with HR quality 90%; we refer to these as *Everion q90*. In the case of Empatica, we consider HR quality to be high when IBI is present in the data. Empatica applies a filter to its IBI data, thus wrong beats are not included in the output IBI file [Inc., 2023]. Table 6.4 shows the metrics corresponding to the datasets *Everion Best* and *Empatica Best*.

Figure 6.4 shows the Bland-Altman plot for the *Empatica Best* and *Everion Best* datasets. For Empatica, the 95% limits of agreement ranged from -34.66 to +51.13 with a mean difference of 8.23 bpm. While for the Everion, the 95% limits of agreement ranged from -6.55 to +6.06 with a mean difference of -0.24 bpm. The data distribution of both devices shows no specific pattern. Further analysis showed that the data of both devices is normally distributed with the majority of the data points located within two standard deviations.

Performance per activity

To understand why the Empatica has a larger bias, Figure 6.5 shows the bias per activity. The bias increases significantly during activities involving exercise (μ bias 30 bpm) and remains low (μ bias 1 bpm) during the initial activity, rest, and walking. Similarly, during non-strenuous activities, Empatica's mean HR and standard deviation are similar to the Holter's. However, as the activity level increases, the difference between Empatica's mean HR and the Holter also increases. This behavior is not observed with Everion. Figure 6.5 shows that overall activities Everion's mean

Table 6.4.: Experiment II - Heart rate analysis per activity

Case	Activity	Size	ICC [95% CI]	Corr	Bias [95% LoA]
Everion Best	init	3808	.979 [+.978,+.981]	0.98	-0.26 [-5.83, +5.31]
(63592)	rest	21364	.988 [+.984,+.991]	0.99	-1.07 [-7.67,+5.53]
	bike (60 W)	9445	.989 [+.989,+.990]	0.99	+0.05 [-3.91, +4.01]
	bike (120 W)	8299	.995 [+.995,+.995]	1.00	+0.17 [-4.00, +4.34]
	walk	6722	.995 [+.995,+.995]	0.99	-0.10 [-3.36, +3.16]
	jog	8093	.973 [+.971,+.974]	0.97	+0.34 [-7.55, +8.23]
	run	5861	.974 [+.971,+.976]	0.97	+0.77 [-8.59, +10.12]
	avg		**0.985 [+.983,+.986]**	**0.99**	**-0.01 [-5.84,+5.81]**
Empatica Best	init	3852	.755 [+.741,+.769]	0.76	+0.01 [-16.74, +16.76]
(35705)	rest	20883	.834 [+.825,+.842]	0.84	+1.36 [-18.09, +20.82]
	bike (60 W)	4948	.031 [-.006,+.067]	0.05	+20.16 [-28.61, +68.94]
	bike (120 W)	4401	.118 [-.023,+.247]	0.22	+34.47 [-37.06, +106.00]
	walk	1287	.597 [+.558,+.634]	0.61	+2.41 [-27.74, +32.56]
	jog	233	.073 [-.044,+.195]	0.15	+32.14 [-24.77, +89.04]
	run	101	.120 [-.056,+.300]	0.22	+34.12 [-33.35, +101.59]
	avg		**0.361 [+.285,+.436]**	**0.41**	**17.81 [-26.62,+62.24]**

HR behave similarly to the Holter, with no significant difference between the values. Additionally, we can observe that Everion's bias increases as the level of activity increases. The largest bias occurs during rest, where Everion underestimates the mean HR in average by 1 bpm. However, in overall activities, the bias of the Everion is small with a mean value of -0.01 bpm.

Figure 6.6 shows at the bottom the ICC per activity for *Everion Best*, *Everion q90*, and *Empatica Best*. The Empatica's ICC significantly decreases during the bike, jog, and run activities, indicating that the device is unsuitable for monitoring HR during strenuous exercise. The Everion, on the other hand, shows high ICC in both datasets. Moreover, the top of Figure 6.6 shows that the *Everion q90* dataset is around three times larger than *Everion Best*. Thus, relaxing the heart rate quality threshold allows us to have a larger dataset with similar accuracy. In the case of Empatica, we can see that the size of the dataset gets greatly affected as movement increases, with the sample size being only 0.5% of the original data for the run activity.

Finally, we analyzed the mean HR difference per activity on five successive sessions and found no statistically significant differences between group means as determined by one-way ANOVA F(4,30) = .338, p = .850.

6.2.3. Heart Rate Variability Analysis

We started our analysis by extracting IBI segments from the time-series. From the Empatica, we obtained a total of 137 IBI segments with an average length

Table 6.5.: Experiment II - Heart rate variability analysis per activity for the Everion device.

Activity	Metric	ICC [95% CI]	Corr	R^2	Bias [95% LoA]
Init	RMSS	+.899 [+.742, +.967]	+0.91	+0.82	+0.22 [-12.22, +12.65]
μ len: 241 s	SDNN	+.876 [+.697, +.953]	+0.88	+0.75	+2.13 [-14.57, +18.84]
μ peaks: 318/314	PNN50	+.967 [+.912, +.988]	+0.99	+0.94	+0.76 [-3.99, +5.50]
# seg: 17	LF	+.982 [+.952, +.993]	+0.98	+0.96	+61.63 [-416.36, +539.63]
	HF	+.918 [+.792, +.969]	+0.98	+0.87	+143.34 [-754.77, +1041.46]
	LF:HF	+.625 [+.215, +.846]	+0.92	+0.49	+2.27 [-5.85, +10.38]
	LFnu	+.745 [+.420, +.900]	+0.81	+0.63	+0.24 [-16.67, +17.15]
	HFnu	+.745 [+.420, +.900]	+0.81	+0.63	-0.24 [-17.15, +16.67]
Rest	RMSS	+.935 [+.904, +.956]	+0.95	+0.88	-0.91 [-7.35, +5.53]
μ len: 249 s	SDNN	+.982 [+.974, +.988]	+0.98	+0.97	+0.10 [-7.89, +8.09]
μ peaks: 363/357	PNN50	+.954 [+.933, +.968]	+0.96	+0.92	+0.04 [-2.97, +3.06]
# seg: 111	LF	+.946 [+.946, +.946]	+0.95	+0.89	+37.20 [-724.00, +798.40]
	HF	+.914 [+.877, +.940]	+0.94	+0.86	-6.90 [-489.31, +475.52]
	LF:HF	+.614 [+.413, +.744]	+0.67	+0.30	+1.31 [-4.05, +6.67]
	LFnu	+.701 [+.525, +.808]	+0.75	+0.26	+3.26 [-10.40, +16.93]
	HFnu	+.701 [+.525, +.808]	+0.75	+0.26	-3.26 [-16.93, +10.40]
Bike (60 W)	RMSS	-.018 [-.292, +.303]	-0.02	-3.42	-2.70 [-12.81, +7.41]
μ len: 294 s	SDNN	+.793 [+.596, +.899]	+0.81	+0.53	-2.10 [-13.50, +9.31]
μ peaks: 489/483	PNN50	+.394 [+.045, +.662]	+0.46	-1.44	-0.14 [-1.12, +0.85]
# seg: 28	LF	+.945 [+.886, +.974]	+0.94	+0.88	-6.15 [-126.78, +114.49]
	HF	+.053 [-.271, +.389]	+0.09	-8.16	-40.86 [-245.64, +163.93]
	LF:HF	-.012 [-.208, +.250]	-0.03	-0.98	+2.59 [-3.65, +8.82]
	LFnu	-.065 [-.345, +.266]	-0.08	-1.32	+7.61 [-23.17, +38.40]
	HFnu	-.065 [-.345, +.266]	-0.08	-1.32	-7.61 [-38.40, +23.17]
Bike (120 W)	RMSS	+.064 [-.161, +.427]	+0.15	-6.94	-6.15 [-17.04, +4.73]
μ len: 272 s	SDNN	+.835 [+.552, +.946]	+0.87	+0.55	-4.48 [-24.52, +15.57]
μ peaks: 541/533	PNN50	+.355 [-.112, +.728]	+0.59	-4.47	-0.34 [-1.36, +0.67]
# seg: 13	LF	+.266 [-.168, +.669]	+0.47	-5.93	-57.90 [-225.22, +109.43]
	HF	-.136 [-.464, +.347]	-0.34	-19.56	-81.61 [-332.89, +169.66]
	LF:HF	+.053 [-.316, +.507]	+0.23	-0.37	+2.43 [-5.03, +9.88]
	LFnu	+.031 [-.386, +.512]	+0.05	-0.58	+14.53 [-39.90, +68.96]
	HFnu	+.060 [-.386, +.512]	+0.05	-0.58	-14.53 [-68.96, +39.90]
Walk	RMSS	+.144 [-.091, +.478]	+0.42	-6.66	-5.74 [-12.31, +0.83]
μ len: 291 s	SDNN	+.755 [+.358, +.910]	+0.81	+0.38	-3.54 [-14.15, +7.07]
μ peaks: 444/446	PNN50	-.008 [-.361, +.412]	-0.03	-30.34	-0.79 [-3.74, +2.15]
# seg: 17	LF	+.417 [-.017, +732]	+0.51	-1.59	-77.18 [-407.37, +253.02]
	HF	+.046 [-.193, +.386]	+0.30	-97.74	-158.47 [-503.17, +186.23]
	LF:HF	+.238 [-.104, +.614]	+0.63	-1.32	+4.95 [-1.01, +10.92]
	LFnu	+.128 [-.111, +.455]	+0.49	-22.01	17.61 [-9.04, +44.26]
	HFnu	+.128 [-.111, +.455]	+0.49	-22.01	-17.61 [-44.26, +9.04]

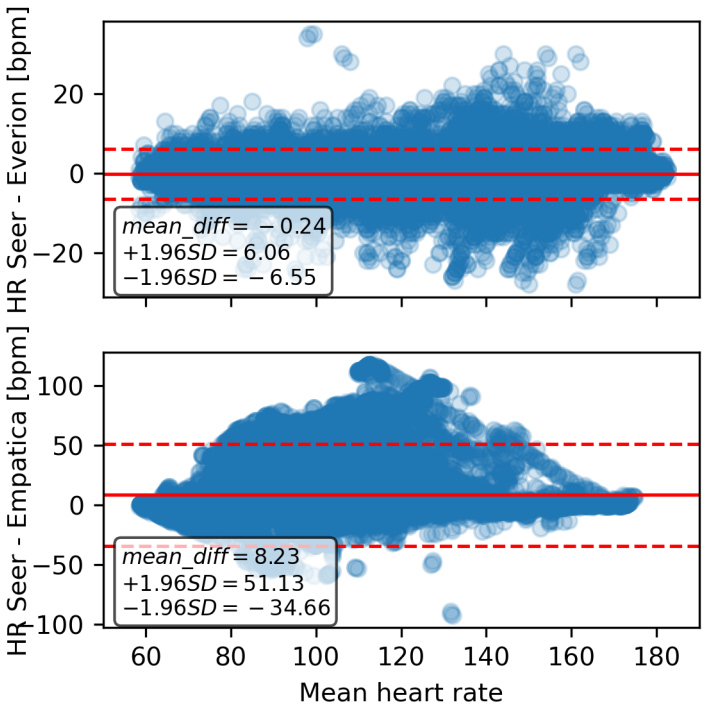

Figure 6.4.: Bland-Altman plot for Empatica/Holter and Everion/Holter. Empatica's mean bias (8.23 LoA [51.13,-34.66]) is larger than Everion's mean bias (-0.24 LoA [6.06, -6.55]). Overall the data seems to be well distributed showing no particular pattern.

of 49 s. Per activity, we gathered the following number of segments (with mean duration): init 12 (μ 44 s), rest 81 (μ 50 s), bike slow 23 (μ 44 s), bike fast 20 (μ 53 s), walk 1 (μ 34 s). Only two IBI segments in the whole dataset are longer than 2 minutes. The recommended length for short-term HRV analysis ranges from 3-5 minutes [Richardson et al., 1996; Shaffer and Ginsberg, 2017]. Thus, we are unable to compute short-term HRV analysis for this device. Future work can overcome this limitation by applying techniques to approximate the missing IBI signal.

Table 6.5 depicts the results from our HRV analysis comparing the Everion with the ECG Holter. The table is organized per activity. For each activity, we extracted IBI segments larger than 3 minutes. Figure 6.7 depicts the level of agreement between the HRV metrics derived from the Everion and Holter during each activity. There is good agreement during the initial and rest activities in all metrics. Agreement decreases with higher activity levels and varies depending on the metric. HF is more affected by increasing activity levels.

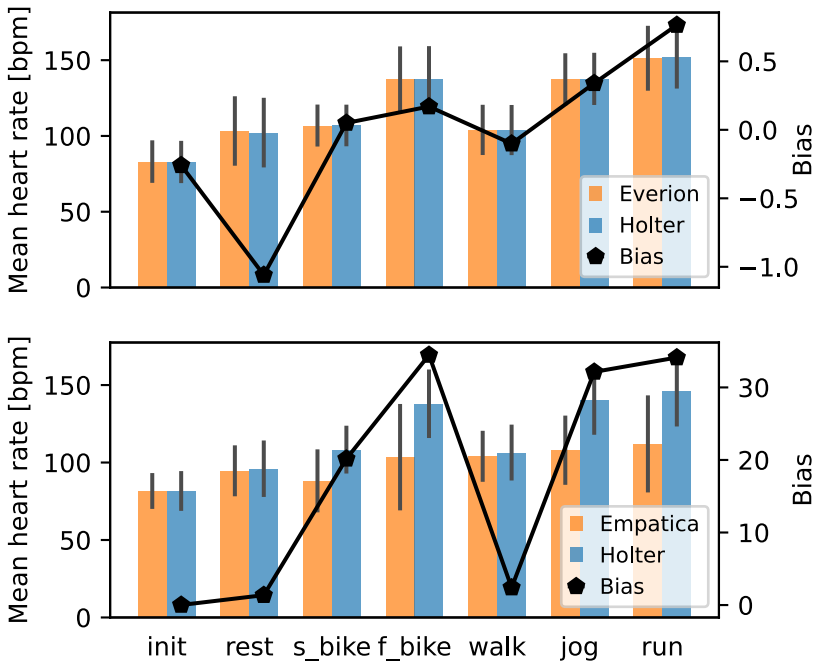

Figure 6.5.: Mean HR, standard deviation and bias per activity of Empatica, Everion and Holter. In particular, Empatica's bias increases significantly during strenuous activities and remains low while being less active. The difference between Empatica's mean HR and Holter shows a similar behavior to the bias, increasing with exercise. Everion's bias increases with the level of activity but overall remains low. Everion's mean HR and standard deviation show similar behaviors as the Holter.

Sedentary activities

During the initial and *rest* activity, there is good agreement in all HRV measurements. For the *initial* activity, we found 17 segments larger than 3 minutes. The mean length of the segments is 241 s. The highest agreement occurs on the frequency domain metric LF with *ICC* between +.952 and +.993, indicating excellent agreement. The lowest agreement occurs on the ratio LF:HF with mean *ICC* ranging from +.215 to +.846, indicating poor agreement. Time domain measurements indicate better agreement with the Holter, ranging from moderate to excellent. For the *rest* activity, we collected 111 segments with an average duration of 249 s. Overall the results are satisfactory in this activity. Excellent agreement occurs in all time-domain metrics and LF. Followed by good agreement in HF, moderate agreement in the normalized LF and HF, and poor agreement for the ratio LF:HF.

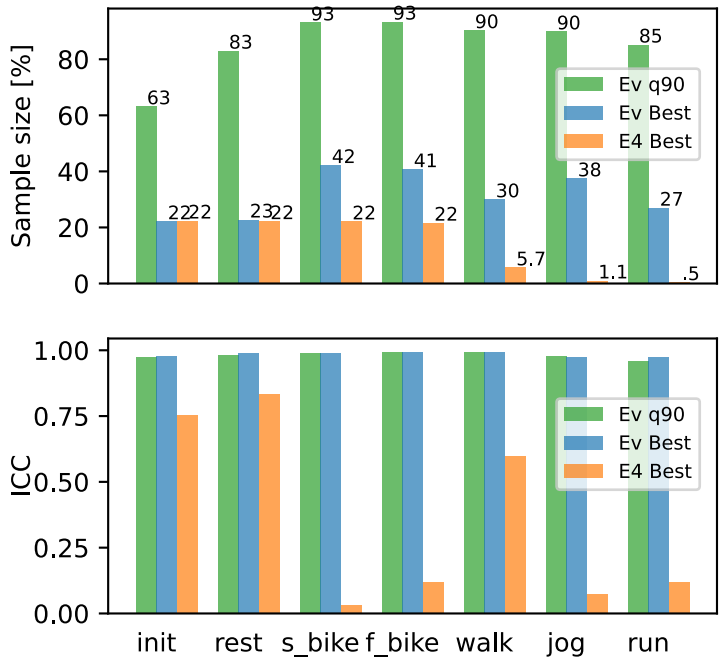

Figure 6.6.: The bottom image shows the ICC corresponding to each activity. In particular, Empatica's accuracy is significantly lower during strenuous activities. In the case of Everion, both datasets are comparable showing high ICC over all activities. The top figure shows the fraction of the sample size of each dataset in relation to its original dataset. The Everion q90 dataset is up to four times the size of Everion Best. Empatica best is considerably smaller than its original dataset, and it is more affected while jogging and running.

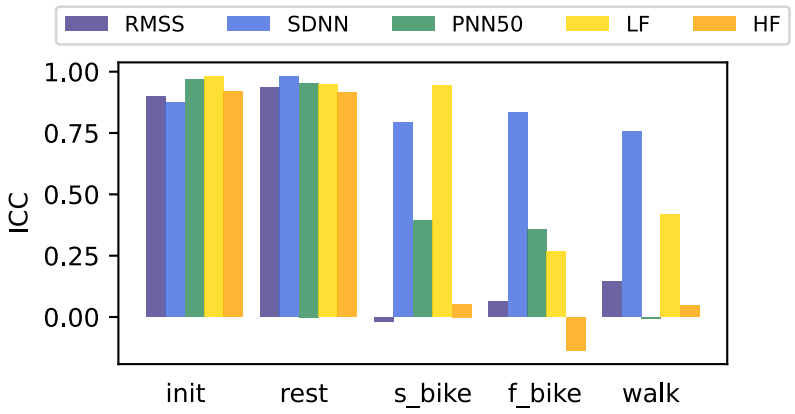

Figure 6.7.: Intraclass correlation for different HRV metrics. As the activity level increases the ICC decreases, more notably in the HF band.

Moderate/High intensity activities

In the activity *bike* (60 W), the average length of the 28 considered segments is 294 s. The highest agreement happens in LF with *ICC* ranging from +.886 to +.974, indicating good agreement. Followed by moderate agreement in SDNN with *ICC* ranging from +.596 to +.899. The rest of the metrics show poor agreement in relation to the Holter. In the bike (120 W) activity, only 13 segments are larger than 3 minutes, with an average length of 272 s. In this activity, only one metric shows moderate agreement, SDNN with *ICC* ranging from +.552 to +.946. The rest of the metrics show poor agreement. We consider 17 segments with an average length of 291 s for the *walk* activity. In this activity, all metrics show poor reliability. The activities jog and run are not included in the analysis as we were unable to extract IBI segments longer than 3 minutes from them.

6.2.4. HRV Monitoring During Ambulatory Activities

The evaluation of HRV during ambulatory activities is still a subject of study. LF shows higher agreement than HF as the activity level increases. This is consistent with the results found in [Hernando et al., 2018]. Moreover, our findings agree with those reported by Michael et al. [2017]. In general, the reliability of time domain and frequency domain measurements decreases as exercise intensity increases. Moreover, our analysis shows similar results in the normalized spectral analysis, with LFnu increasing during low-moderate intensity exercise and decreasing during higher-intensity exercise. In contrast, HFnu shows the opposite response as shown in Figure 6.8. Additionally, Figure 6.8 shows that the normalized metrics of the Everion and the Holter follow similar trends, even though their level of agreement is low.

6.3. Discussion

6.3.1. Users' preference

In general, devices perform better when there is less movement, as in the *rest* period and *initial* activity. However, both wrist-located devices perform poorly on the bike, jog, and run activities, indicating that (i) the wrist's posture may affect the accuracy of wrist-based monitoring, (ii) wrist-based monitors are more susceptible to movement in comparison with non-wrist devices such as the Polar OH1, Wahoo Ticker, and Everion. From our short questionnaire, we learned that 58.3% of the participants prefer wrist-based devices. Thus, when designing experiments, there is a trade-off to be made between comfort/users' preferences and reliability.

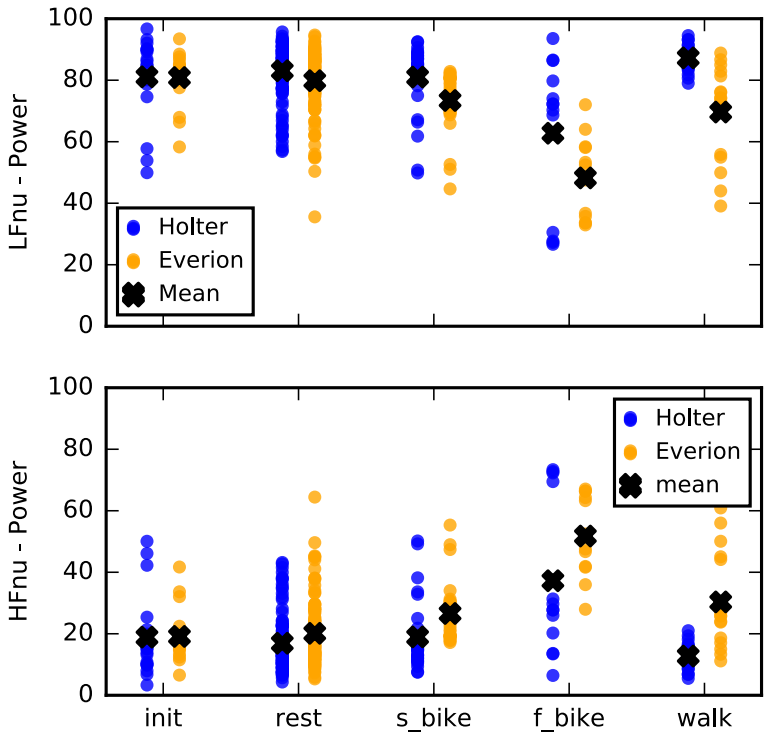

Figure 6.8.: LFnu increases during low-moderate intensity exercise and decreases during higher intensity exercise, while HFnu demonstrates the opposite behavior.

6.3.2. Everion, Empatica and Holter

Everion and Empatica provide useful mechanisms to ensure a high quality of the resulting dataset. We recommend making use of these parameters. Overall we consider both of Everion's datasets comparable and showing excellent agreement with respect to the Holter in all activities. Empatica provides good agreement for the *initial* activity and textitrest periods and moderate agreement while walking. Our results indicate that Empatica is less suitable for tracking mean HR during ambulatory conditions or high-intensity activities.

6.3.3. Everion and Holter During Ambulatory Conditions

Even though many studies, including ours, use the Holter monitor as a baseline, there may be better solutions to monitor HR during strenuous sports. The Holter's cables and electrodes are very susceptible to movement. Therefore, wearable devices may provide better reliability under these conditions. Figure 6.9 shows an example of this case. We can observe that the Holter signal becomes noisy as the subject engages with the jog and running activities. However, this may not be the case when

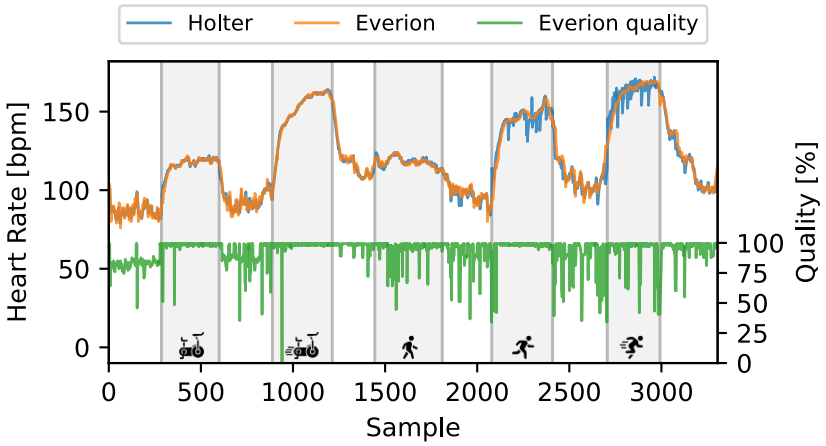

Figure 6.9.: Signals collected using the Everion and medical-grade Holter monitor. In particular, the Everion device shows very good agreement with the data of the Holter monitor. The Holter shows less reliability (noise) during the jog and run activities.

monitoring IBI. Figure 6.10 shows Everion's and Holter's IBI of one subject during the rest and bike activities. Everion and Holter show good agreement during rest, but Everion shows more noise during the bike activity. Further experiments are required to determine if wearable devices can provide more reliability than a Holter monitor during sports activities.

6.4. Conclusion

We gathered a dataset with multiple off-the-shelf wearable devices comprising several sensors to track physiological data and made our dataset publicly available to the research community. We focus on evaluating the agreement between mean HR and HRV metrics derived from the PPG sensor in wearable devices and a standard ECG Holter monitor under different physiological conditions. We show that armband-based devices dominate in precision when monitoring mean HR in all considered settings. Additionally, we show that the Everion device is a valid proxy for HRV metrics during periods not involving strenuous physical activity. Therefore, we hypothesize that the Everion is a potential candidate for continuous monitoring of physiological data in persons with sedentary lifestyles, such as office workers, patients, etc. Furthermore, we look into the future and challenge whether the Holter monitor is a better baseline than wearable devices for monitoring HR and HRV during strenuous activities and conclude that further exploration is needed. Finally, we show that participants have a preference for wrist-based devices and that their choices are not significantly affected by the predetermined duration of the monitoring.

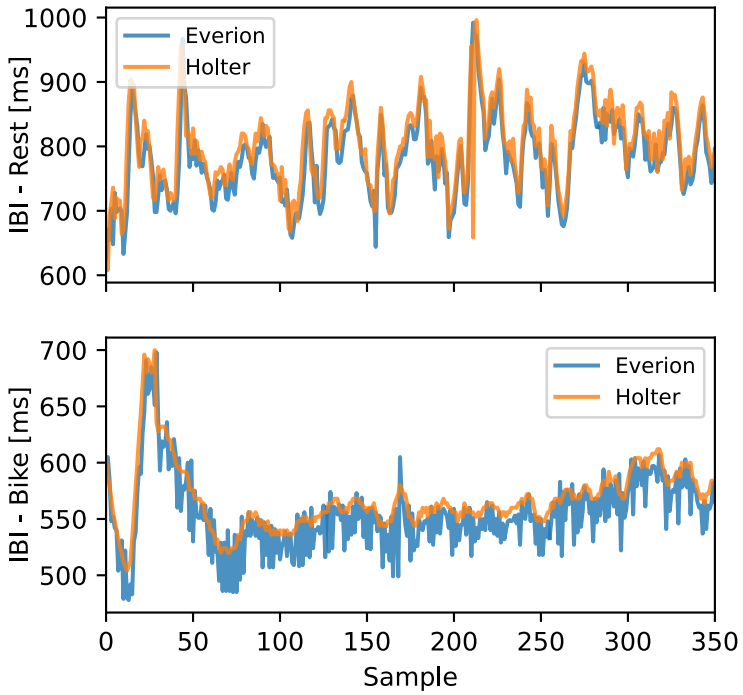

Figure 6.10.: Comparing of Everion and Holter interbeat intervals (IBI) during rest and bike activities. The devices show very good agreement during rest. Everion signals show more variation during the bike session.

Thus, there is a trade-off between comfort and reliability when deciding between armband-based or wrist-based devices.

CHAPTER 7

Cronico – Multidimensional Platform for Unsupervised Fatigue Monitoring

This chapter uses the interbeat intervals (IBIs) data processing pipeline developed and described in the following publication:

Marc Hilty, Pietro Oldrati, Liliana Barrios, Tamara Müller, Claudia Blumer, Magdalena Foege, PHRT consortium, Christian Holz, Andreas Lutterotti. Continuous monitoring with wearables in multiple sclerosis reveals an association of cardiac autonomic dysfunction with disease severity. *Multiple Sclerosis Journal – Experimental, Translational and Clinical*. June 2022.

In the previous chapter, we evaluated different wearables devices and selected Everion as a good candidate for continuous monitoring of physiological data, in particular, HR and HRV. In this chapter, we introduce a multidimensional plaform for the continuous patient monitoring with focus on fatigue. Our platform comprises a wearable device (Everion) and a mobile application, which we refer to as Cronico. Using this platform, we conducted an in-the-wild study to build a comprehensive dataset that allows further studies on fatigue using digital technologies. We provide initial results of using steps and HRV variability for fatigue quantification. The collected dataset is currently used by fellow researchers and results from their analyses is out of the scope of this dissertation.

Table 7.1.: Cronico's survey type, notification frequency and description.

Type	Frequency	Description
Fatigue VAS	*4/day*	Visual analog scale from 1 to 10, where 10 is the highest fatigue and 1 no fatigue. Notification triggered with experience sampling method (ESM) [Gabriel et al., 2019] during four time intervals.
Alerness scale	*1/day*	Samn-parelli scale with seven statements "Completely exhausted unable to function, ready to drop", "extremely tired, very difficult to concentrate," "moderately tired," "a little tired," "ok, somehow fresh," "very lively, responsive but not at peak," "fully alert, wide awake." [Samn and Perelli, 1982]
Sleep protocol	*1/day*	Morning and evening sleep protocol from the German association for sleep medicine. [Hoffmann et al., 1997]
FSS	*1/week*	Fatigue severity scale [Krupp et al., 1989]
Stress VAS	*1/day*	Stress visual analog scale from 0 "no stress at all" to 10 "extremelly stressed. Notification in the evening to evaluate overall stress of the day." [Lesage et al., 2012]

7.1. Cronico Development

We developed a smartphone application for collecting ground truth data about the persons' fatigue level and to enable disease characterization. To this end, Cronico incorporates the surveys listed in Table 7.1. The table also includes the surveys respective notification frequency during the study. Furthermore, Cronico's latest release include other data gathering features such as the physical and cognitive fatigability tasks described in the previous chapters. Refer to Table 7.2 for further details. Refer to the Appendix, Figure E.1 for further details on Cronicos' user interface.

7.2. Methods

To analyze the feasibility of using wearables as a proxy for subjective fatigue we conducted a two-week in-the-wild study. We use the FSMC [Penner et al., 2009] to discriminate between motor-fatigued and non-motor-fatigued participants and the FSS [Krupp et al., 1989] to differentiate fatigued and non-fatigued participants. For data collection we used the cronico application accompanied by two Everion sensors for continuous monitoring (24/7) of physiological signal. Table 7.3 depicts

Table 7.2.: Cronico's other data types

Type	Frequency	Description
Tapping test	*1/day*	30 s of rapid alternating finger tapping to measure physical fatigability (introduced in Chapter 4)
cFAST	*3/week*	5 min cognitive fatigability assessment test (introduced in Chapter 5).
Diary	-	Free text entry for patients to report important events that occured during their day.
Activity	-	Activity list to report specific activity types from a given list with their start and end time. Can be triggered on demand, but also activity tagging is triggered automatically with the Fatigue VAS notification for better understanding of patients data. Custom activities can be entered on demand.
Weather	-	Weather conditions of the user's location, triggered on location change.

the sensors included in the Everion device, and Table 7.4 depicts the different features provided by the Everion.

7.2.1. Participants Recruitment

We recruited MS patients with a confirmed diagnosis with age 18 or older without other diseases. Patients were recruited at the Neuroimmunology Outpatient Clinic of the University Hospital of Zurich, Switzerland, between 29 November 2019 and 29 July 2021, and provided informed written consent. Controls were recruited from hospital staff, family and friends of the investigators. We excluded controls with chronic illness, regular medication intake, and fatigue. Controls followed the same protocol as patients. The study protocol was reviewed and approved by the Cantonal Ethics Committee of Zurich and uploaded to kofam.ch (SNCTP000003494).

7.2.2. Everion Device and Data Synchronization

After comparing different devices' performance in (Chapter 6), we opted for the Everion device for unsupervised monitoring of physiological signals. To achieve 24/7 monitoring, we designed our study protocol to include two Everion sensors per participant. This way, we would avoid data gaps while charging the sensors (typically 2-3 hours). Participants were required to always keep one sensor in the charging station while wearing the other sensor on their upper arm. We use a color coding

Table 7.3.: Everion sensor data type and description measurement unit.

Type	Description	Sampling	Unit
Acceleration (ACC)	Three-axis accelerometer sensor	51.2 Hz	g
Barometer (BAR)	Barometer sensor	1 Hz	mbar
Electrodermal Activity (EDA)	Impedance sensor	1 Hz	Alternating current
Temperature (TEMP)	Infrared sensors	1 Hz	$°C$
Photoplethysmography (PPG)	Green, red, infrared and photodiode	1 Hz	

Table 7.4.: Sensor-derived features provided by the Everion device.

Data Type	Unit	Range	Sensor
Heart Rate	Beats per minute (bpm)	30 - 240	PPG
Blood Oxygenation (SpO2)	%	65 - 100 (at rest) 80 - 100 (under motion)	PPG
Skin Blood Perfusion	-	0 - 0.5.1	PPG
Respiration Rate	Breaths per minute (BPM)	6 - 30	PPG
Skin Temperature	$°C$	0 - 60	TEMP
Blood Pulse Wave	-	0 - 5.1	PPG
Heart Rate Variability	ms	0 - 255	PPG
Inter-Beat Interval	ms	1 - 4095	PPG
Energy Expenditure	kCal/day	0 - 65535	
Steps	steps/day	0 - 65535	ACC
Health Scores & Activity	steps/day	0 - 100	ACC
Electrodermal Activity	kOhm	0 - 21.8	EDA
Barometric pressure	Mbar	500 - 1100	BAR
Accelerometer	g	+/- 8	ACC
Activity	-	0 - 21.8	ACC
Quality metrics	-	0 - 100	-

Notes: Activity refers to *resting*, *biking*, *walking*, *running*. Quality metrics refer to some parameters containing a quality metric, which specifies whether the data should be kept for further analysis, i.e., greater than 50 is good, and lower than that should not be considered for further analysis.

Figure 7.1.: Study methods for in-the-wild study phase: two color-coded everion devices for continuous monitoring of physiological data with one charger(left), stationary phone for everion data download and sync (center) and cronico app for gathering ground truth feeling.

scheme to facilitate data synchronization and handling of the sensors. One sensor was set to be used during the day (Figure 7.1 Everion with red label), while the other was meant to be used during the evening (Figure 7.1 Everion with blue label). After wearing a sensor for the whole day, study participants would replace the sensor from the charging station, start the data download using the dedicated app ((Figure 7.1 center) and start wearing the sensors that was located in the charging station (now fully charged). Participants would repeat this process for 14 days.

7.2.3. Study Design

Our study included two phases: one in the hospital, and the other in-the-wild, highlighted in dark and light blue , respectively, in Figure 7.2.

In-hospital

We included a pre-and post-study phase, both guided by a healthcare professional and conducted at the local hospital. During the pre-phase, participants were briefly introduced to the study and completed the FSMC and a demographic questionnaire and baseline measurements. We guided participants through installing and using our Android application on their smartphones and explain how to use the Everion device with the companion data sync application. Figure 7.1 shows the devices and applications. They also received instructions on how to complete the tasks, including a demonstration by the experimenter and a short familiarization session for each of the tasks: handgrip, tapping task, and 9-HPT. We used the 9-HPT to objectify hand

Figure 7.2.: Study design timeline with two phases: the hospital phase (dark blue) to gather baseline measurements, and in-the-wild phase (light blue), the core of this study. During the in-the-wild phase, participants used the cronico app and the Everion sensor 24/7.

function, the FSS and FSMC to categorize fatigue, and a handgrip dynamometer as standard motor fatigability measurement. Neurological impairment was measured using the standard disability rating scale for MS patients (EDSS) [Kurtzke, 1983]. Additionally, we use the Composite Autonomic Symptom Score (COMPASS-31) to measures autonomic symptom severity to enable disease characterization [Sletten et al., 2012]. Participants completed all tasks with their dominant hand. Between tasks, participants rested their arm and hand for three minutes. During the post-phase, as shown in Figure 7.2, we collected the FSMC scale and handgrip measurements again. Both measures, pre and post, were averaged to obtain the final scores.

In-the-wild

We asked patients to use our cronico application including all surveys listed in Table 7.1, furthermore some of the patients also completed the tapping trials as described in Section 4 as they were part of both studies. Patients had the opportunity to decide whether they wanted to use the protocol only with the Everion and not including tapping, only tapping and not including the Everion or use both Everion and tapping. The study duration was two weeks.

7.2.4. Feature Extraction

Steps. To extract individual continuous walks, we looked at the step detection of the Everion. The sensor outputs the number of steps detected at each second. We define a *continuous walk* to consist of steps that occur continuously, with gaps of at most 60 seconds. We allow short gaps where users briefly stop during their walk, e.g., to cross a road. Moreover, we discard instances shorter than 180 seconds to remove short walks that could result from noise. Then, for each of the remaining

walks, we compute the following measures: walk duration, total steps per walk, and average cadence (steps/min). Following, for every user and each of these measures, we calculate the average, the maximum, and the sum (where appropriate) over the whole study duration. Hence, we are left with the following metrics:

- Total steps
- Average steps per walk
- Maximum steps per walk
- Total walk duration
- Average walk duration
- Maximum walk duration

IBIs and HRV metrics. To extract the HRV metrics we used the data processing pipeline develop for our previous work [Hilty et al., 2022]. This pipeline included the following steps. First, IBIs were checked for artifacts based on the method proposed by Berntson et al. [1990]. Artifacts were removed, and missing IBIs were linearly interpolated. Subsequently, the data was divided into non-overlapping 5-min segments. Segments that, due to artifact correction, contained more than four interpolated heartbeats in a row were discarded as they are to be considered unreliable [Berntson et al., 1990]. Furthermore, we excluded all 5-min segments with excessive activity as this reduces the reliability of the measurements. Refer to Chapter 7 for further explanations on monitoring HRV during active periods. To calculate HRV metrics based on the IBIs, we followed the recommendations of the HRV Task Force for time domain (RMSSD, SDNN, pNN20, pNN50), frequency domain (HF, LF) and nonlinear domain (SD1, SD2) [Richardson et al., 1996]. For each valid segment, metrics for time, frequency, and nonlinear domains were calculated using the open-source Python library pyHRV [Caridade Gomes, 2019]. According to research, most metrics were omitted due to redundancy or due to evidence of them being prone to artifacts, or artificial noise [Ciccone et al., 2017; Antali et al., 2021]. Table 7.5 shows an overview of the HRV metrics we considered [Kim et al., 2018; Shaffer and Ginsberg, 2017].

7.2.5. Statistical Analyses

Mann–Whitney U tests were applied to explore group differences. We calculated Spearman's rank correlation to explore the association between variables. We considered $p < 0.05$ to be significant. Finally, effect sizes and 95% CIs were reported using the standardized mean difference (SMD). We checked if the HRV metrics are

Table 7.5.: Heart rate variability metrics.

Type	Description
SDNN	Standard deviation of all IBIs.
SD1	The root mean square of the difference between adjacent IBIs, multiplied by a constant.
NN50	Number of pairs of adjacent IBI differing by more than 50 ms in the entire recording.
pNN50	NN50 divided by the number of total successive differences.
NN20	Number of pairs of adjacent IBI differing by more than 20 ms in the entire recording.
pNN20	NN20 divided by the number of total successive differences.
LF/HF	The ratio of LF to HF power (LF/HF ratio)

Notes: LF power, absolute power of the low-frequency band (0.04-0.15 Hz). *HF power*, absolute power of the high-frequency band (0.15-0.4 Hz) [Shaffer and Ginsberg, 2017].

confounded by age by controlling its effects on fatigue through analysis of covariance (ANCOVA).

7.3. Preliminary Results

This section presents preliminary results from Cronico's in-the-wild fatigue dataset. Particularly, for the steps analysis, we focus on patients with low disability (EDSS < 3.5) to reduce the effects of disability when analyzing fatigue. We present preliminary results on steps and HRV analysis.

7.3.1. Participants

Table 7.6 shows the demographics of our fatigue data set participants. We have recruited 55 and 25 age and gender-matched controls. The majority (47/55, 85.45%) of patients had a relapsing-remitting disease course, while only 8 (14.55%) had a progressive (primary or secondary) clinical form. There was an average of 12.63 (SD 2.12) days with Everion data per study participant.

7.3.2. Removing Short Walks

We collected a total of 44800 walks for patients and 16942 walks for controls. From those, we removed walks that lasted less than 180 s, 14105 walks for controls, and 39318 for patients. For our final analysis, we had 5482 walks for patients and 2837

Table 7.6.: Demographic Characteristics of Participants

		Control	Patient	*P*
Number		25	55	
Age, mean (SD)		33.66 (10.39)	36.27 (9.71)	0.295
Gender, n (%)				0.687
	m	11 (44)	20 (36.36)	
	w	14 (56)	35 (63.64)	
Ethinicity, n (%)	Asian	3 (12.00)		0.044
	African		1 (1.85)	
	Caucasian	19 (76.00)	50 (92.59)	
	Hispanic	1 (4.00)		
	Middle-Eastern	2 (8.00)	3 (5.56)	
MS type, n(%)				
	Progressive		8 (14.55)	
	Relapsing-remitting		47 (85.45)	
Disease duration, mean (SD)			12.52 (8.51)	
DMT, n(%)				
	Dimethylfumarat		7 (12.73)	
	Natalizumab		12 (21.82)	
	None		8 (14.55	
	Ocrelizumab		14 (25.45)	
	Ozanimod		2 (3.64)	
	Rituximab		4 (7.27)	
	Siponimod		1 (1.82)	
	Teriflunomide		1 (1.82)	
	aHSCT		6 (10.91)	
EDSS, mean (SD)			2.41 (1.95)	
FSMC, mean (SD)				
	Total		52.22 (21.84)	
	Cognitive		25.93 (11.68)	
	Motor		26.28 (11.05)	

Notes: Data are mean (SD) or n (%). Disease duration is measured in years since first manifestation; EDSS: expanded disability status scale; FSMC: Fatigue Score for motor functions and cognition; DMT: disease modifying therapy.
Chi-squared tests for the following variables may be invalid due to the low number of observations: MS type, Ethinicity.

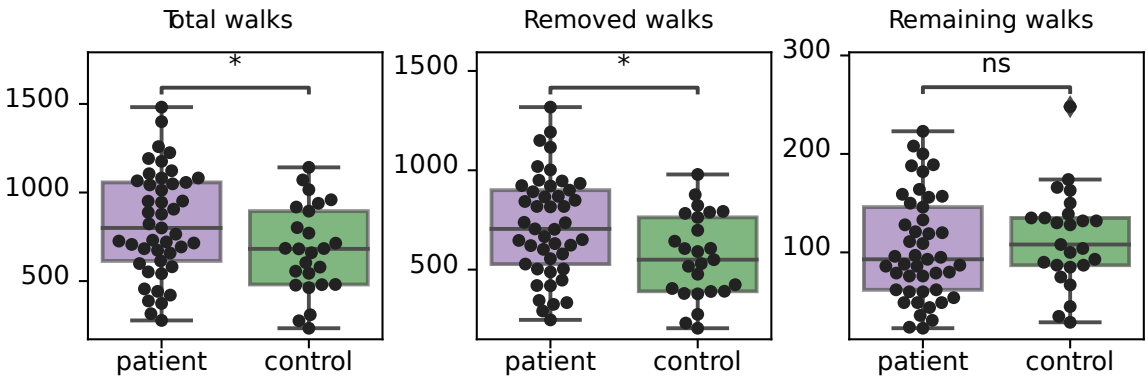

Figure 7.3.: Total collected walks (left), removed walks (center) and remaining walks (right). After removing short walks, there is no longer a statistically significant difference between patients and controls.

walks for controls. Figure 7.3 compares the number of walks of patients and controls at the different stages. There is a statistically significant difference between the total number of walks of patients and controls ($U = 400.0$, $p = .047$). There is also a statistically significant difference between the removed patients' walks and controls' walks ($U = 358.0$, $p = .012$). After removing short walks, there is no longer a statistically significant difference between the patients and controls regarding the number of walks ($U = 639.5$, $p = .348$).

7.3.3. Steps Correlate to EDSS

As Table 7.8 shows, EDSS strongly correlated with several steps features. This is expected as EDSS accounts for walking disability from scores 4 and above. Hence, to remove the confounding related to walking impairment, in the steps analysis, we only include patients with EDSS≤3.5. Ten patients with walking impairment were no longer considered for the steps analysis. Table 7.7 depicts the updated demographics for the steps analysis after removing patients with walking disability. Table 7.9 depicts the correlation between the walk-derived features, the FSMC subscales, and EDSS for patients without walking impairment. EDSS is no longer correlated with the proposed steps metrics in this subset of patients.

7.3.4. Step Metrics Correlation to FSMC Subscales

As displayed in Table 7.9, FSMC physical fatigue subscore is significantly correlated to several features. The only two features that do not strongly correlate to the FSMC physical scale are *mean walk duration* and *total walk duration* with $\rho = -0.27$ ($p = .07$) and $\rho = -0.26$ ($p = .09$), respectively. Only two proposed metrics

Table 7.7.: Demographic Characteristics only considering patients with EDSS≤3.5

		Control	Patient	*P*
Number		25	45	
Age, mean (SD)		33.66 (10.39)	34.4 (9.15)	0.770
Gender, n (%)				0.531
	m	11 (44)	15 (33.33)	
	w	14 (56)	30 (66.67)	
Ethinicity, n (%)	Asian	3 (12)		0.090
	African		1 (2.27)	
	Caucasian	19 (76.00)	40 (90.91)	
	Hispanic	1 (4.00)		
	Middle-Eastern	2 (8.00)	3 (6.82)	
MS type, n(%)				
	Progressive		1 (2.22)	
	Relapsing-remitting		43 (95.56)	
Disease duration, mean (SD)			5.25 (6.07)	
DMT, n(%)				
	Dimethylfumarat		7 (15.56)	
	Natalizumab		10 (22.22)	
	None		7 (15.56)	
	Ocrelizumab		13 (28.89)	
	Ozanimod		2 (4.44)	
	Rituximab		1 (2.22)	
	Teriflunomid		1 (2.22)	
	aHSCT		4 (8.89)	
EDSS, mean (SD)			1.63 (0.99)	
FSMC, mean (SD)				
	Total		48.17 (19.83)	
	Cognitive		24.41 (11.09)	
	Motor		23.74 (9.65)	

Notes: Data are mean (SD) or n (%). Disease duration is measured in years since first manifestation; EDSS: expanded disability status scale; FSMC: Fatigue Score for motor functions and cognition; DMT: disease modifying therapy.
Chi-squared tests for the following variables may be invalid due to the low number of observations: MS type, Ethinicity.

Table 7.8.: Spearmann rank correlation coefficient – Steps features, FSMC and EDSS. N=55 all patients considered.

Feature	Total	Physical (*P*)	Cognitive (*P*)	EDSS (*P*)
max steps/walk	**-0.33 (.02)**	**-0.36 (.01)**	-0.21 (.12)	**-0.34 (.01)**
mean steps/walk	-0.24 (.08)	**-0.28 (.04)**	-0.15 (.28)	**-0.29 (.03)**
max walk/duration	**-0.30 (.03)**	**-0.31 (.02)**	-0.23 (.10)	**-0.34 (.01)**
mean walk/duration	-0.22 (.1)	-0.26 (.05)	-0.13 (.34)	-0.23 (.09)
total steps	**-0.32 (.02)**	**-0.35 (.01)**	-0.22 (.11)	**-0.3 (.03)**
total walk duration	-0.25 (.07)	-0.26 (.05)	-0.17 (.22)	-0.26 (.05)

Table 7.9.: Spearmann rank correlation coefficient – Steps features, FSMC and EDSS. N=45, only patients with EDSS <=3.5 considered.

Feature	Total	Physical (*P*)	Cognitive (*P*)	EDSS (*P*)
max steps/walk	**–0.30 (.04)**	**-0.3 (.046)**	-0.26 (.09)	-0.08 (.61)
mean steps/walk	-0.21 (.17)	**-0.33 (.03)**	**-0.31 (.04)**	-0.01 (.96)
max walk/duration	**–0.35 (.02)**	**-0.31 (.04)**	**-0.36 (.02)**	-0.18 (.23)
mean walk/duration	-0.27 (.07)	-0.27 (.07)	-0.24 (.11)	-0.04 (.81)
total steps	**–0.38 (.01)**	**-0.31 (.04)**	-0.29 (.05)	-0.12 (.42)
total walk duration	-0.33 (.03)	-0.26 (.09)	-0.24 (.11)	-0.15 (.33)

significantly correlated to the FSMC cognitive fatigue subscores. Those are *mean steps/walk* with $\rho = -0.31$ ($p = .04$), and *max walk/duration* with ρ =-0.36 ($p = .02$). Three metrics are significantly correlated to FSMC general fatigue subscore. Those are *max steps/walk* with ρ=–0.30 ($p = .04$), *max walk/duration* ρ=–0.35 ($p = .02$) and *total steps* $\rho = -0.38$ ($p = .01$).

7.3.5. Steps Metrics and Physical Fatigue

We evaluated the group differences regarding our steps metrics and the FSMC physical subscore for two subgroups, mild and severe fatigue. Table 7.10 depicts the group differences for the fatigue (N=20) and non-fatigued patients (N=25) according to **mild** fatigue. Four metrics shows statistically significant difference between the groups: *max steps/walk* with $U = 145.0$ $(p = .02)$, *mean steps/walk* with $U = 149.0$ $(p = .02)$, mean walk/duration with $U = 141.0$ $(p = .01)$, and finally *total steps* with $U = 139.0$ $(p = .01)$. *Max walk/duration* and *Total walk duration* did not show a statistically significant difference between the fatigue groups with $U = 176.0$ $(p = .09)$ and $U = 171.0$ $(p = .07)$ respectively.

Table 7.10.: Group differences according to steps metrics and the FSMC **mild** physical fatigue definition.

Feature	No Fatigue N=20	Fatigue N=25	U (*P*)
max steps/walk	4617.67 (1879.35)	3389.25 (2412.4)	**145.0 (.02)**
mean steps/walk	651.6 (282.79)	484.44 (191.2)	**149.0 (.02)**
max walk/duration	3284.45 (1185.94)	2711.0 (1611.58)	176.0 (.09)
mean walk/duration	576.71 (152.65)	472.57 (141.59)	**141.0 (.01)**
total steps	66322.2 (28177.83)	48716.62 (37573.46)	**139.0 (.01)**
total walk duration	61579.3 (27460.22)	49329.28 (36602.73)	171.0 (.07)

Group differences with Mann–Whitney U test . FSMC thresholds: no fatigue < 22.

Table 7.11.: Group differences according to steps metrics and the FSMC **severe** physical fatigue definition.

Feature	No Fatigue N=20	Fatigue N=12	U (*P*)
max steps/walk	4617.67 (1879.35)	3030.17 (2449.1)	**56.0 (.01)**
mean steps/walk	651.6 (282.79)	474.56 (212.59)	70.0 (.05)
max walk/duration	3284.45 (1185.94)	2304.67 (1617.04)	**56.0 (.01)**
mean walk/duration	576.71 (152.65)	461.92 (166.43)	**62.0 (.03)**
total steps	66322.2 (28177.83)	46783.22 (48870.62)	**59.0 (.02)**
total walk duration	61579.3 (27460.22)	47434.0 (45844.18)	73.0 (.07)

Group differences with Mann–Whitney U test. FSMC thresholds: no fatigue < 22, fatigue ≥ 32.

Table 7.11 depicts the fatigue groups according to the FSMC severe physical fatigue threshold. In total, 20 patients were classified as non-fatigued, while 12 were labeled severely fatigued. In this case, an additional metric shows statistical significance. That is *max walk/duration* with $U = 56.0$ $(p = .01)$. Additionally, *mean steps/walk* shows a group difference close to statistical significance with $U = 70.0$ $(p = .05)$. *Total walk duration* shows no statistically significance between the severely fatigued groups with $U = 73.0$ $(p = .07)$.

Figure 7.4 displays box plots corresponding to the fatigue classification according to the FSMC physical subscale with the threshold for severe fatigue. We have 12 patients with severe physical fatigue and 20 with no fatigue, and 25 controls. The figure shows that when using *total steps* as metric, there is a statistically significant difference between non-fatigued and fatigued patients with $U = 59.0$ ($p = .019$). Furthermore, a statistically significant difference exists between the fatigue and control group with $U = 70.0$ $p = 0.01$. Finally, the figure also reveals no statistically

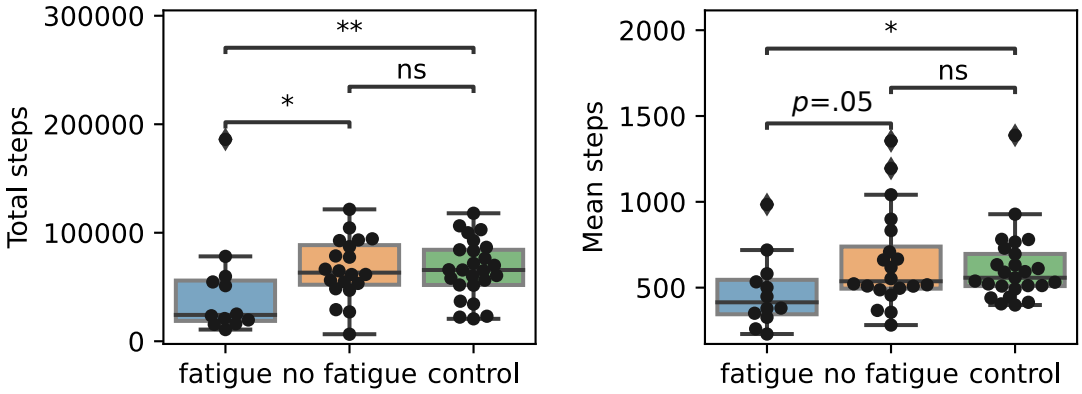

Figure 7.4.: Steps metrics relation to FSMC physical fatigue (severe)

significant difference between the patients with no fatigue and the control group $U = 248.0$ $p = 0.97$, suggesting non-fatigued MS patients and controls have similar walking patterns.

Similar results are shown in Figure 7.4 right side. When using *mean steps* as metric, there is a difference close to significant between the fatigue and no fatigue groups ($U = 70.0$, $p = .05$). Controls and patients with no fatigue show no statistically significant difference ($U = 249.0$ $p = .99$). Finally, there is a statistically significant difference between the fatigue and control group ($U = 78.0$ $p = .02$).

7.3.6. Steps Metrics and Cognitive Fatigue

When evaluating groups' differences between cognitive and non-cognitive fatigue patients according to the threshold "mild," none of the proposed steps metrics shows a statistically significant difference. When considering severe cognitive fatigue (fatigued = 12, non-fatigue =18), three steps metrics display the statistically significant difference between groups: max walk/duration ($U = 55.0$, $p = .03$), total steps ($U = 60.0$, $p = .04$), and total walk duration ($U = 59.0$, $p = .04$). These results suggest that step metrics are more related to physical than cognitive fatigue.

7.3.7. HRV and Cognitive Fatigue

We evaluated group differences between the fatigue and non-fatigued patients according to the FSMC cognitive subscale and two thresholds: mild and severe. We evaluated our metrics during different times of the day. To this end, we use a percentage of the day. We defined 24h as 100% of the day. In this subsection, we present results for the *early morning*, which we calculated as the first 15% of the day since

Table 7.12.: Group differences according to HRV-derived metrics (early morning and late evening) and FSMC **mild** cognitive fatigue definition.

	Feature	No Fatigue N=19	Fatigue N=36	U (*P*)	ANCOVA* (*P*)
Early morning	SDNN	73.86 (22.71)	60.82 (17.92)	**478.0 (.02)**	**0.048**
	NN20	203.06 (27.58)	182.0 (40.77)	**461.0 (.04)**	0.109
	PNN20	559.98 (10.81)	50.79 (11.43)	**487.0 (.01)**	**0.012**
	NN50	81.61 (39.42)	58.37 (35.2)	**473.0 (.02)**	0.066
	PNN50	24.94 (13.64)	16.65 (10.61)	**480.0 (.01)**	**0.035**
	LF/HF	2.52 (0.99)	2.96 (1.33)	285.0 (0.32)	0.329
	SD1	34.97 (12.28)	28.34 (10.39)	**480.0 (.01)**	.082
Late night	SDNN	65.16 (22.51)	53.28 (18.41)	**464.0 (.03)**	0.082
	NN20	193.77 (33.6)	164.66 (51.45)	**473.0 (.02)**	0.066
	PNN20	57.51 (12.99)	45.86 (14.9)	**489.0 (.01)**	**0.012**
	NN50	74.04 (42.0)	50.3 (41.03)	**483.0 (.01)**	0.105
	PNN50	22.96 (14.72)	14.45 (12.46)	**486.0 (.01)**	0.06
	LF/HF	2.54 (1.15)	2.9 (1.47)	304.0 0.51	0.442
	SD1	33.68 (13.35)	27.02 (12.69)	**486.0 (.01)**	0.149

* Analysis of covariance (ANCOVA) with *age* as covariate. Group differences with Mann–Whitney *U* test. FSMC cognitive fatigue thresholds: no fatigue < 22, fatigue ≥ 22.

the person woke up. We also show results for the late evening, which represents 15% of the day before the person falls asleep.

Table 7.12 displays the groups' difference for fatigue according to the FSMC *mild* cognitive fatigue definition. Mann–Whitney *U* tests reveal that all metrics but one show statistically significant differences. LF/HF shows no statistically significant difference between fatigue groups, neither in the morning ($U = 285.0$, $p = .32$) nor in the evening ($U = 304.0$, $p = .51$). Three metrics show statistically significant differences during the early morning, even after controlling for *age* with ANCOVA. Those are SDNN $U = 478.0$ ($p = .02$), PNN20 $U = 487.0$ ($p = .01$) and PNN50 $U = 480.0$ ($p = .01$). NN20, NN50, and SD1 are statistically significantly different during the early morning, but there are no longer statistically significant differences after ANCOVA. Late at night, only one metric is statistically significant after ANCOVA. That is PNN20 with $U = 489.0$ ($p = .01$)

Table 7.13 displays the group differences for fatigue according to the FSMC *severe* cognitive fatigue definition. Mann–Whitney *U* tests revealed that all metrics but one show statistically significant differences. LF/HF shows no statistically significant difference between the fatigue groups, neither in the morning ($U = 140.0$, $p = .16$),

Table 7.13.: Group differences according to HRV-derived metrics (early morning and late evening) and FSMC **severe** cognitive fatigue definition.

	Feature	No Fatigue N=19	Fatigue N=20	U (*P*)	ANCOVA* (*P*)
Early morning	SDNN	73.86 (22.71)	57.26 (18.39)	**281.0 (0.01)**	0.091
	NN20	203.06 (27.58)	172.32 (44.55)	**280.0 (0.01)**	0.105
	PNN20	59.98 (10.81)	47.61 (11.98)	**299.0 (0.0)**	**0.013**
	NN50	81.61 (39.42)	52.47 (36.42)	**284.0 (0.01)**	0.147
	PNN50	24.94 (13.64)	14.67 (10.1)	**289.0 (0.01)**	0.077
	LF/HF	2.52 (0.99)	3.23 (1.55)	140.0 (0.16)	0.253
	SD1	34.97 (12.28)	26.4 (9.17)	**288.0 (0.01)**	0.115
Late night	SDNN	65.16 (22.51)	48.02 (17.75)	**290.0 (0.01)**	0.068
	NN20	193.77 (33.6)	150.96 (56.89)	**290.0 (0.01)**	0.053
	PNN20	57.51 (12.99)	41.47 (15.7)	**297.0 (0.0)**	**0.01**
	NN50	74.04 (42.0)	42.8 (41.51)	**291.0 (0.0)**	0.159
	PNN50	22.96 (14.72)	11.99 (11.71)	**294.0 (0.0)**	0.093
	LF/HF	2.54 (1.15)	2.92 (1.66)	176.0 (0.7)	0.656
	SD1	33.68 (13.35)	24.07 (10.66)	**299.0 (0.0)**	0.108

* Analysis of covariance (ANCOVA) with *age* as covariate. Group differences with Mann–Whitney *U* test. FSMC cognitive fatigue thresholds: no fatigue < 22, fatigue ≥ 34.

nor in the late evening ($U = 176.0$, $p = .7$). In this case, only one metric kept the statistically significant difference between the group. This was the PNN20 in the morning and night with $U = 299.0$ ($p = .0$) and $U = 97.0$ ($p = .0$), respectively.

7.3.8. HRV and Physical Fatigue

Our group difference analysis revealed no statistically significant differences with our metrics during the evening when classifying fatigue according to the FSMC *mild* physical fatigue definition. Using the FSMC mild physical fatigue, 20 were labeled with no fatigue, and 35 were labeled as fatigued. During the morning, we found statistically significant differences between the fatigue groups with the following metrics:(1) PNN20 with $U = 479.0$ $(P = .02)$ and ANCOVA $p = .035$, (2) NN50 with $U = 463.0$ (P.049) and ANCOVA $p = .118$, (3) PNN50 with $U = 472.0$ ($p = .03$) and ANCOVA 0.035, and finally (4) SD1 with $U = 466.0$ ($p = .04$) and ANCOVA $p = 0.039$. When considering *severe* physical fatigue (Table 7.14), we do observe statistical significance in the groups during the evening and morning metrics. However, in this case, none of the metrics is statistically significant after conducting

Table 7.14.: Group differences according to HRV-derived metrics (early morning and late evening) and FSMC **severe** physical fatigue definition.

	Feature	No Fatigue N=20	Fatigue N=21	U (*P*)	ANCOVA* (*P*)
Early morning	SDNN	73.94 (25.17)	59.04 (18.78)	284.0 (0.06)	0.091
	NN20	199.33 (32.03)	179.17 (46.29)	269.0 (0.13)	0.732
	PNN20	59.52 (12.18)	49.26 (12.52)	**299.0 (0.02)**	0.103
	NN50	81.02 (43.37)	56.79 (38.56)	**291.0 (0.04)**	0.437
	PNN50	25.16 (15.18)	15.83 (10.69)	**298.0 (0.02)**	0.203
	LF/HF	2.47 (0.86)	3.14 (1.55)	158.0 (0.18)	0.238
	SD1	35.74 (14.57)	27.48 (9.8)	**295.0 (0.03)**	0.222
Late night	SDNN	64.74 (24.63)	51.02 (20.24)	**290.0 (0.04)**	0.295
	NN20	186.67 39.66)	160.78 (60.68)	270.0 (0.12)	0.689
	PNN20	55.94 (15.09)	44.08 (17.2)	**293.0 (0.03)**	0.191
	NN50	71.53 (46.46)	50.9 (47.3)	282.0 (0.06)	0.743
	PNN50	22.56 (16.46)	14.31 (13.56)	**288.0 (0.04)**	0.442
	LF/HF	2.55 (1.16)	2.88 (1.68)	198.0 (0.76)	0.645
	SD1	33.79 (15.56)	26.69 (13.75)	**296.0 (0.03)**	0.519

* Analysis of covariance (ANCOVA) with *age* as covariate. Group differences with Mann–Whitney *U* test. FSMC thresholds: no fatigue < 22, fatigue ≥ 32.

ANCOVA. Similarly, Table 7.15 shows the groups' differences analysis but when using the FSS to label fatigue.

7.3.9. PNN20 and Fatigue

Figure 7.5 left shows box plots corresponding to the *PNN20* metrics comparing controls, fatigue, and non-fatigued patients according to the FSMC cognitive fatigue classification using as threshold: ≥ 34 fatigued < 22 no fatigue. We observe a statistically significant difference between the cognitive and non-cognitive fatigue groups with $U = 297.0$ $p = 0.003$. There is a close to significant difference between the cognitively fatigued patients and controls with $U = 48.0$ $p = 0.05$. Finally, there is no statistically significant difference between the control and non-fatigued patients ($U = 109.0$, $p = 0.26$)

Figure 7.5 right shows box plots corresponding to the *PNN20* metrics comparing controls, fatigue, and non-fatigued patients according to the FSMC physical fatigue classification using as threshold: ≥ 32 fatigued < 22 no fatigue. We observe a statistically significant difference between the physical and non-physical fatigue

Table 7.15.: Group differences according to HRV-derived metrics (early morning and late evening) and FSS for general fatigue.

	Feature	No Fatigue N=22	Fatigue N=21	U (*P*)	ANCOVA* (*P*)
Early morning	SDNN	68.3 (21.45)	60.4 (20.18)	283.0 (0.21)	0.795
	NN20	197.02 (30.42)	168.11 (38.59)	**338.0 (0.01)**	0.167
	PNN20	56.22 (10.6)	48.64 (12.67)	307.0 (0.07)	0.372
	NN50	71.63 (36.8)	52.58 (31.91)	**315.0 (0.04)**	0.571
	PNN50	21.07 (12.32)	15.64 (10.76)	304.0 (0.08)	0.717
	LF/HF	2.64 (1.02)	3.17 (1.29)	179.0 (0.21)	0.382
	SD1	31.75 (11.01)	27.77 (11.08)	296.0 (0.12)	0.843
Late night	SDNN	60.54 (21.37)	51.39 (19.14)	290.0 (0.16)	0.66
	NN20	186.88 (37.15)	147.07 (48.9)	**341.0 (0.01)**	0.107
	PNN20	53.31 (12.18)	42.73 (16.15)	**328.0 (0.02)**	0.243
	NN50	63.62 (39.91)	42.84 (36.32)	**322.0 (0.03)**	0.553
	PNN50	18.75 (12.73)	12.97 (12.26)	**319.0 (0.03)**	0.688
	LF/HF	2.67 (1.13)	3.04 (1.43)	205.0 (0.54)	0.357
	SD1	30.52 (11.96)	25.93 (13.35)	**316.0 (0.04)**	0.838

* Analysis of covariance (ANCOVA) with *age* as covariate. FSS thresholds: no fatigue < 4, fatigue ≥ 4.

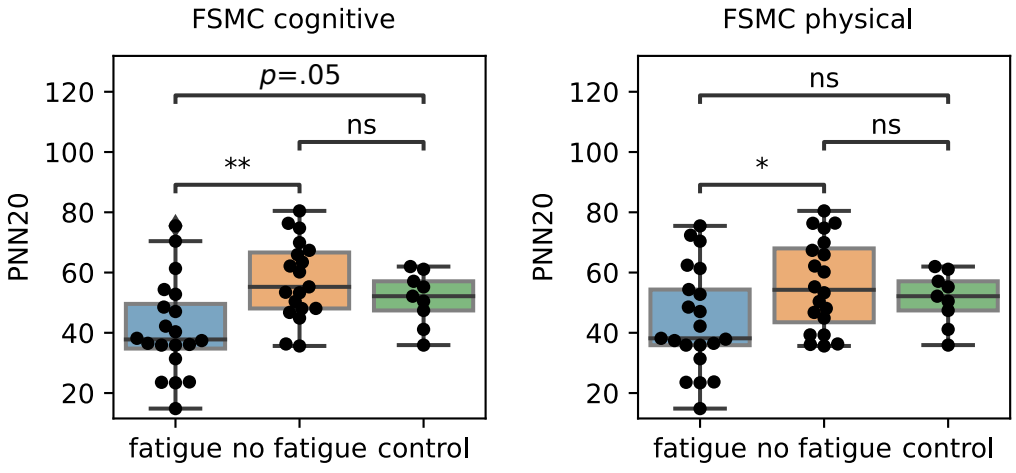

Figure 7.5.: *PNN*20 shows a statistically significant difference between the fatigue groups, but there is no difference between controls and the patients groups.

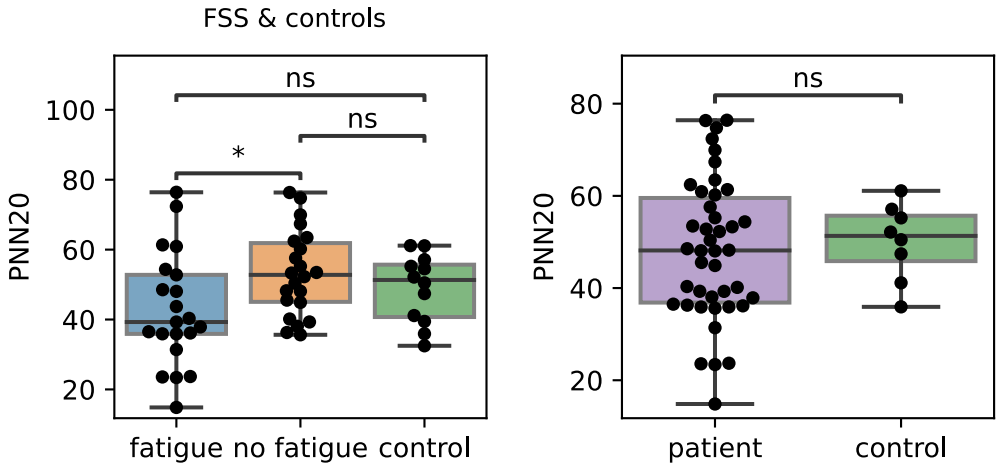

Figure 7.6.: *PNN*20 shows a statistically significant difference between the fatigue groups, but there is no difference between controls and the patients groups.

groups with $U = 293.0$ $p = 0.03$. There is no significant difference between the physically fatigued patients and controls with $U = 65.0$ $p = 0.19$. Finally, there is no statistically significant difference between the control and non-fatigued patients ($U = 104.0$ $p = 0.52$)

Figure 7.6 shows box plots corresponding to the *PNN20* metrics comparing controls, fatigue, and non-fatigued patients according to the FSS fatigue classification using 4 as the fatigue threshold. In general, we observe a statistically significant difference between the fatigue and non-fatigue groups with $U = 328.0$, $p = 0.02$. However, there is no difference between the fatigue and control group ($U = 87.0$, $p = 0.152$) nor between the no fatigue and control group (U=155.0, $p = 0.42$). When analyzing only the group difference in terms of *PNN*20 between patients and controls, we found no statistically significant difference ($U = 274.0$, $p = 0.75$)

7.4. Discussion

7.4.1. Steps and Fatigue

Our analysis reveals that steps derived features are more related to physical than cognitive fatigue. In this sense, it is crucial to consider the role of impairment when analyzing fatigue in MS patients with movement-related metrics. We observed that our metrics are significantly correlated to EDSS. Hence, when using steps, one needs to consider the patients' impairment level to avoid quantifying impairment instead of fatigue. We showed that total and mean steps are reliable metrics for evaluating fatigue in MS patients with low disability. In particular, we observed

group differences in the fatigue groups and that controls and non-fatigued MS patients have similar behaviors.

7.4.2. HRV and Cognitive Fatigue

Our results show that the HRV metrics are more related to cognitive than physical fatigue. Table 7.13 shows that all HRV metrics but one show a statistically significant difference between the fatigue groups when classifying severe cognitive fatigue. LF/HF was the only metric that showed no statistically significant difference between groups. This metric was not significant in the morning or night. Further analysis with ANCOVA revealed that *age* is significantly correlated to all HRV metrics. After controlling for *age* as the covariate, only $PNN20$ stayed significantly with $p = 0.013$ for the morning and $p = 0.01$ for the night. Table 7.14 shows that when classifying severe physical fatigue, not all metrics are statistically significant in the morning or evening. In this case, $PNN20$ was no longer significant after controlling for *age* as the covariate. Similarly, in Table 7.15, we observed that only a subset of metrics differentiates between the fatigue and non-fatigued group according to the FSS classification. This is to be expected, as the FSS measures general fatigue and does not have specific thresholds for cognitive fatigue. As well for this scale, all metrics were no longer significant after controlling for *age* as the covariate. The role of age is something that needs further analysis. We cannot draw a conclusion in this regard with the current data set. However, age and disability go hand-in-hand; as one ages, it is only natural that disability increases. Therefore, further studies need to explore if completely removing the effect of *age* is an appropriate solution or if this approach is too strict.

7.4.3. HRV and Implications of Time of Measurement

Our results indicate that the time of measuring HRV has an important role in detecting significance. We explore the group difference during the early morning, late evening, day, and sleep. The best time to measure HRV and study group difference for fatigue was during the early morning and late evening, as highlighted in Table 7.13, Table 7.14 and Table 7.15. Furthermore, day-time measurements only revealed a statistically significant difference in the $PNN20$ metric, while HRV metrics during sleep showed the worst performance. None of the HRV metrics revealed a statistically significant difference between the groups during sleep.

7.5. Conclusion

We introduced our Cronico monitoring infrastructure and our Cronico fatigue data set. Our data set included data from 55 patients and 25 controls during a two-week in-the-wild study. Our preliminary results on the role of HRV as a passive metric for fatigue quantification revealed that HRV metrics are more related to cognition than physical fatigue. $PNN20$ was the best-performing metric for differentiating cognitive and non-cognitively fatigued patients. We also showed that the time of measurement plays an important role in the HRV metrics and that *age* is strongly related to HRV. On the other hand, steps are related to physical fatigue but not cognitive fatigue.

CHAPTER 8

Conclusion

In this thesis, we explored how computer science can support patient monitoring and clinical research. We focus on the prevalence of smartphones and wearable devices to study fatigue and enable more regular assessments. The technical foundations are based on data analysis, representation, and feature extraction to model fatigability. Our proposed methods allow for studying the association between perceived fatigue and objective fatigability, a task that has so far been difficult due to the limitations of the existing medical approaches. In the following, we summarise the contributions and limitations of this dissertation and outline potential directions for future research.

8.1. Principal Contributions

In Chapter 3, we presented an exertion technique on commodity devices that involves simple alternating rapid finger tapping to assess motor fatigability. The proposed task is fast, easy to implement, and can easily be integrated in the patient's daily routine We derived a metric from the data of such tapping tasks to represent motor fatigability: the increase of time a user keeps a finger on the screen, which we refer to as *touch duration*. We detail the specifics of the metric and our processing pipeline to extract it from the data collected through the mobile app. An evaluation of our approach on 20 MS patients and 35 controls showed that our metric strongly correlates with data collected from a standard handgrip dynamometer. We further show that this correlation is also present in the first 30 seconds of performing the tapping task with $\rho = 0.78$ for patients and $\rho = 0.84$ for controls. Our work is a first step towards measuring motor fatigability without relying on specialized equipment, which can be expensive and require professional supervision. Our method may help quantify

fatigue and complement the current use of subjective feedback through questionnaires, enabling patients to frequently and ubiquitously monitor their condition and react to changes accordingly.

In Chapter 4, we build upon the developed tapping task to study the association between fatigability and perceived fatigue in the wild. We introduced a new smartphone-based metric, *tapping frequency*, to quantify motor fatigability with the tapping task. We showed that *tapping frequency* is better than the previously introduced metric *touch duration*, as it shows an invalid trend in fewer trials while maintaining a similar correlation. We provided a proof of concept of the applicability of the tapping task and our metric in uncontrolled environments in the wild. There is a statistically significant difference between fatigued and non-fatigued groups during the whole study. Furthermore, we showed that combining several trials improves the reliability of the fatigue prediction. Our results on the association between tapping task and perceived fatigue during uncontrolled conditions showed that mean tapping frequency ranks motor fatigue according to the FSMC with $AUC_{ROC}\ \overline{X} = .76\ \pm .05$ and that it ranks fatigue according to the FSS with $AUC_{ROC}\ \overline{X} = .81\ \pm .05$. Our fatigue data set comprising 35 MS patients is available to the research community[1]. We introduced a simple model that provides good interpretability and thus a higher chance of adoption in clinical practice.

In Chapter 5, we introduced the cognitive fatigability assessment test (cFAST), a novel smartphone-based test to quantify cognitive fatigability. Our pilot study with 42 MS patients, 23 fatigued and 19 non-fatigued as defined by the FSMC, provides evidence that cFAST produces a quantifiable drop in task performance in a short period. Furthermore, our results indicate that cFAST has the potential to serve as a surrogate for subjective cognitive fatigue. When classifying cognitive fatigue, our cognitive fatigability metric Δ *response time* has a mean AUROC of 0.74 ±.05. Furthermore, Δ *response time* shows a statistically significant difference between the fatigued and non-fatigued groups (t=2.27, P=.03). In particular, we observed that cognitively-fatigued patients declined in performance while non-fatigued patients improved. cFAST is significantly shorter than the existing cognitive fatigability assessments and does not require specialized equipment. Thus, it could enable frequent and remote monitoring and substantially aid clinicians in better understanding and treating fatigue.

In Chapter 6, we shifted focus from fatigability quantification to evaluating devices for unsupervised monitoring. This was our first step towards building a platform for assessing the autonomic nervous system and studying its role in fatigue. We evaluated the agreement between several HR and HRV metrics derived from the PPG sensor in wearable devices and the same metrics derived from a standard ECG Holter

[1] https://www.research-collection.ethz.ch/handle/20.500.11850/494324

monitor under different physiological conditions. We collected an activity dataset[2] comprised of fourteen participants (7 female and 7 male) using six different wearable devices. We thoroughly evaluated the sensors with three different experiments, for a total of 30 sessions (55 minutes each). Our results showed that armband-based devices dominate in precision when monitoring mean HR in all considered settings. Additionally, we showed that the Everion device is a valid proxy for HRV metrics during periods not involving strenuous physical activity. Therefore, the Everion is a potential candidate for continuous monitoring of physiological data in everyday life but not while doing strenuous activities.

Finally, in Chapter 7, we introduced the Cronico infrastructure and Cronico fatigue data set comprising data from 56 MS patients and 25 controls. We gathered our data set in a two-week in-the-wild study. Study participants used our mobile application to complete fatigability tasks and provide perceived fatigue sensation to be used as ground truth. Following the findings in Chapter 6, we incorporated the use of Everion to quantify physiological signals from our participants. Preliminary results on the association between passive sensing and fatigue showed metrics derived from step count as potential candidates. However, the role of HRV and autonomic dysfunction in fatigue quantification needs further exploration. Our resulting data set is now being analyzed as part of an ETH PHRT grant for further exploration of fatigue models and disease progression.

8.2. Limitations and Outlook

A discussion of the specific limitations of the different contributions can be found in the corresponding chapters. Below we list general limitations and future research directions.

Sample size and impairment. A recurrent limitation in our studies is the small sample size. Larger clinical studies to establish the proposed methods as part of the clinical routine are needed. In particular, it is necessary to include patients representing the full spectrum of disabilities in terms of EDSS. Furthermore, the person's dexterity may influence results from our proposed physical and cognitive fatigability tests. Hence, the reliability of our metrics in MS patients with hand impairment has to be assessed in future studies.

Validation in other diseases and differentiation to confounding symptoms Until now, all our evaluations were done with MS patients. This target group was chosen due to the high prevalence of fatigue in these patients and the degree to which the disease course can be controlled with existing disease-modifying therapies.

[2] https://www.research-collection.ethz.ch/handle/20.500.11850/374755

Nevertheless, fatigue is a common symptom of other chronic diseases, some of which also have a larger incidence. Hence, future work should explore the validity of our proposed methods in other conditions. We believe our methods have the potential to be applied outside of MS. The limitation of impairment is less relevant for patients with other disease entities, such as long COVID, which is not associated with hand impairment. Moreover, future studies are needed to investigate if our tools can distinguish between fatigue and confounding symptoms such as depression, sleepiness, or others.

Fatigue fluctuation and multilevel fatigue classification. In our studies, we used a cross-sectional study design. Thus, we are not able to define the clinical significance of the changes in the fatigability scores in individual patients. Future studies are needed to determine if the changes in the measured fatigability scores correlate with fluctuations in fatigue severity and how this information can be used in patient monitoring. A possible approach to achieve this goal is conducting a longitudinal intervention study to measure the effectiveness of a fatigue intervention. Furthermore, in our results, we limited our analyses to a binary classification problem, partly due to the small sample size. However, fatigue scales such as the FSMC identify different levels of fatigue: none, mild, moderate, and severe. Therefore, future studies should explore the possibility of achieving multiclass fatigue classification.

Considerations for remote monitoring. Our studies highlighted the need for implementing changes to improve data quality and compliance in unsupervised settings. Strategies to verify that the tests are conducted in a distraction-free environment are needed. A simple method can be to automatically dismiss test sessions if no input is recorded after a certain period. Distractions in uncontrolled environments could influence test results, for example, by affecting the tapping frequency during the physical test or generating many missed or wrong answers during cFAST. Finally, we need to investigate the test re-test reliability of our methods and establish a suitable periodicity to conduct the tests.

User interface and design improvements. Informal feedback from the participants suggests that performing daily tasks can produce a lack of motivation and boredom. This can be addressed in further studies by using gamification to keep the participant engaged and motivated. Furthermore, as measuring fatigability requires users to elicit maximal effort, exploring additional incentives during the test may be worthwhile.

Autonomic dysfunction and fatigue. We only presented preliminary evidence of the association between physiological data and perceived fatigue. Future work should focus on verifying if autonomic dysfunction is related to fatigue. A challenge to address when looking at this data is to consider the role of age, as this also influences autonomic response. Additionally, our data only contains data from

two weeks. More extended studies are needed to better understand the role of the autonomic nervous system in fatigue.

8.3. Closing Remarks

In this dissertation, we have shown the potential of using smartphones and wearable devices to quantify physical and cognitive fatigue objectively. Although we only show preliminary results on physiological data analysis, we believe that by combining the proposed objective measurements and the passive data, it is possible to develop models to quantify fatigue more accurately. An objective and reliable measure as a surrogate for fatigue facilitates further research on this devastating symptom, particularly the development of novel therapies. Additionally, the ability to monitor patients over time and independently from medical facilities (i.e., in the wild) provides an important advantage in assessing the effects of therapeutic interventions and devising strategies for managing the symptom. We hope this thesis has made a step forward in the objective fatigue quantification and remote monitoring of disease progression in MS patients. Additionally, we hope our methods apply to other medical conditions with fatigue as a primary debilitating symptom.

Bibliography

[Adamec and Habek, 2013] Ivan Adamec and Mario Habek. Autonomic dysfunction in multiple sclerosis. *Clin. Neurol. Neurosurg.*, 115 Suppl 1:S73–8, December 2013. DOI: https://doi.org/10.1016/j.clineuro.2013.09.026.

[Agyemang et al., 2021] Cathy Agyemang, Jason A Berard, and Lisa A S Walker. Cognitive fatigability in multiple sclerosis: How does performance decline over time on the paced auditory serial addition test? *Mult. Scler. Relat. Disord.*, 54:103130, September 2021. DOI: https://doi.org/10.1016/j.msard.2021.103130.

[Alusi et al., 2000] S H Alusi, J Worthington, S Glickman, L J Findley, and P G Bain. Evaluation of three different ways of assessing tremor in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 68(6):756–760, 2000. DOI: http://dx.doi.org/10.1136/jnnp.68.6.756.

[Anners et al., 2007] Lerdal Anners, Celius G. Elisabeth, Krupp B. Lauren, and Dahl A. Alv. A prospective study of patterns of fatigue in multiple sclerosis. *European Journal of Neurology*, 14:1338–1343, 2007. DOI: http://dx.doi.org/10.1111/j.1468-1331.2007.01974.x.

[Antali et al., 2021] Flóra Antali, Dániel Kulin, Konrád István Lucz, Balázs Szabó, László Szűcs, Sándor Kulin, and Zsuzsanna Miklós. Multimodal assessment of the pulse rate variability analysis module of a Photoplethysmography-Based telemedicine system. *Sensors*, 21, Aug 2021. DOI: https://doi.org/10.3390/s21165544.

[Apolinário-Hagen et al., 2018] Jennifer Apolinário-Hagen, Mireille Menzel, Severin Hennemann, and Christel Salewski. Acceptance of mobile health apps for disease management among people with multiple sclerosis: Web-Based survey study. *JMIR Form Res*, 2(2):e11977, December 2018. DOI: https://doi.org/10.2196/11977.

[Armonk, Released 2017] NY: IBM Corp Armonk. Ibm spss statistics for windows, version 25.0., Released 2017.

[Armutlu et al., 2007] Kadriye Armutlu, Nilufer Cetisli Korkmaz, Ilke Keser, Vildan Sumbuloglu, Derya Irem Akbiyik, Zafer Guney, and Rana Karabudak. The validity and reliability of the fatigue severity scale in turkish multiple sclerosis patients. *International Journal of Rehabilitation Research*, 30(1), 2007. DOI: https://doi.org/10.1097/MRR.0b013e3280146ec4.

[Ayobi et al., 2017] Amid Ayobi, Paul Marshall, Anna L. Cox, and Yunan Chen. Quantifying the body and caring for the mind: Self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6889–6901, New York, NY, USA, 2017. ACM. DOI: https://doi.org/10.1145/3025453.3025869.

[Babbage et al., 2019] Duncan R. Babbage, Kirsten van Kessel, Juliet Drown, Sarah Thomas, Ann Sezier, Peter Thomas, and Paula Kersten. Ms energize: Field trial of an app for self-management of fatigue for people with multiple sclerosis. *Internet Interventions*, 18:100291, 2019. DOI: https://doi.org/10.1016/j.invent.2019.100291.

[Bailón et al., 2013] Raquel Bailón, Nuria Garatachea, Ignacio de la Iglesia, Jose Casajús, and Pablo Laguna. Influence of running stride frequency in heart rate variability analysis during treadmill exercise testing. *IEEE Transactions on Biomedical Engineering*, 60(7):1796–1805, 2013. DOI: https://doi.org/10.1109/TBME.2013.2242328.

[Barrios et al., 2018] Liliana Barrios, Pietro Oldrati, Silvia Santini, and Andreas Lutterotti. Recognizing digital biomarkers for fatigue assessment in patients with multiple sclerosis. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare – Demos, Posters, Doctoral Colloquium*, PervasiveHealth'18, New York, NY, USA, 8 2018. EAI. DOI: http://dx.doi.org/10.4108/eai.20-4-2018.2276340.

[Barrios et al., 2019] Liliana Barrios, Pietro Oldrati, Silvia Santini, and Andreas Lutterotti. Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, pages 251–261, New York, NY, USA, May 2019. Association for Computing Machinery. DOI: https://doi.org/10.1145/3329189.3329215.

[Barrios et al., 2020] Liliana Barrios, Pietro Oldrati, David Lindlbauer, Marc Hilty, Helen Hayward-Koennecke, Christian Holz, and Andreas Lutterotti. A rapid tapping task on commodity smartphones to assess motor fatigability. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery. DOI: https://doi.org/10.1145/3313831.3376588.

[Barrios et al., 2021] Liliana Barrios, Pietro Oldrati, Marc Hilty, David Lindlbauer, Christian Holz, and Andreas Lutterotti. Smartphone-Based Tapping Frequency as a Surrogate for Perceived Fatigue: An in-the-Wild Feasibility Study in Multiple Sclerosis Patients. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–30, September 2021. DOI: https://doi.org/10.1145/3478098.

[Barrios et al., 2022] Liliana Barrios, Rok Amon, Pietro Oldrati, Marc Hilty, Christian Holz, and Andreas Lutterotti. Cognitive fatigability assessment test (cFAST): Development of a new instrument to assess cognitive fatigability and pilot study on its association to perceived fatigue in multiple sclerosis. *Digit Health*, 8:1–17, August 2022. DOI: https://doi.org/10.1177/20552076221117740.

[Basner and Dinges, 2011] Mathias Basner and David F Dinges. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, 34(5):581–591, May 2011. DOI: https://doi.org/10.1093/sleep/34.5.581.

[Belza, 1995] B L Belza. Comparison of self-reported fatigue in rheumatoid arthritis and controls. *The Journal of rheumatology*, 22(4), 1995.

[Benedict et al., 2012] Ralph H B Benedict, Audrey Smerbeck, Rajavi Parikh, Jonathan Rodgers, Diego Cadavid, and David Erlanger. Reliability and equivalence of alternate forms for the symbol digit modalities test: implications for multiple sclerosis clinical trials. *Mult. Scler.*, 18(9):1320–1325, September 2012. DOI: https://doi.org/10.1177/1352458511435717.

[Berard and Walker, 2021] Jason A Berard and Lisa A S Walker. Increasing the clinical utility of the paced auditory serial addition test: Normative data for standard, dyad, and cognitive fatigability scoring. *Cogn. Behav. Neurol.*, 34(2):107, June 2021. DOI: https://doi.org/10.1097/WNN.0000000000000268.

[Berard et al., 2018] Jason A Berard, Andra M Smith, and Lisa A S Walker. A longitudinal evaluation of cognitive fatigue on a task of sustained attention in early Relapsing-Remitting multiple sclerosis, 2018. DOI: https://doi.org/10.7224/1537-2073.2016-106.

[Berard et al., 2020] Jason A Berard, Zhuo Fang, Lisa A S Walker, Alyssa Lindsay-Brown, Leila Osman, Ian Cameron, Roxana Cruce, Greg O Cron, Mark S Freedman, and Andra M Smith. Imaging cognitive fatigability in multiple sclerosis: objective quantification of cerebral blood flow during a task of sustained attention using ASL perfusion fMRI. *Brain Imaging Behav.*, 14(6):2417–2428, December 2020. DOI: https://doi.org/10.1007/s11682-019-00192-7.

[Berntson et al., 1990] G G Berntson, K S Quigley, J F Jang, and S T Boysen. An approach to artifact identification: application to heart period data. *Psychophysiology*, 27(5):586–598, September 1990. DOI: https://doi.org/10.1111/j.1469-8986.1990.tb01982.x.

[Bertoni et al., 2015] Rita Bertoni, Ilse Lamers, Christine C Chen, Peter Feys, and Davide Cattaneo. Unilateral and bilateral upper limb dysfunction at body functions, activity and participation levels in people with multiple sclerosis. *Multiple Sclerosis Journal*, 21(12):1566–1574, 2015. DOI: https://doi.org/10.1177/1352458514567553.

[Bigland-Ritchie et al., 1983] B. Bigland-Ritchie, R. Johansson, O. C. Lippold, and J. J. Woods. Contractile speed and emg changes during fatigue of sustained maximal voluntary contractions. *Journal of Neurophysiology*, 50(1):313–324, 1983. DOI: https://doi.org/10.1152/jn.1983.50.1.313.

[Biogen, 2019] Biogen. Above ms - meet aby, 2019.

[Biogen, 2023] Biogen. Aby. https://www.aby-app.ca/, 2023. Accessed: 2023-3-2.

[Biovotion AG, 2018] Everion monitor. https://www.biovotion.com/, 2018.

[Bland and Altman, 1986] John Martin Bland and Douglas Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327:307 – 310, 1986.

[Boukhvalova et al., 2018] Alexandra K. Boukhvalova, Emily Kowalczyk, Thomas Harris, Peter Kosa, Alison Wichman, Mary A. Sandford, Atif Memon, and Bibiana Bielekova. Identifying and quantifying neurological disability via smartphone. *Frontiers in Neurology*, 9:740, 2018. DOI: https://doi.org/10.3389/fneur.2018.00740.

[Braley and Chervin, 2010] Tiffany J. Braley and Ronald D. Chervin. Fatigue in Multiple Sclerosis: Mechanisms, Evaluation, and Treatment. *Sleep*, 33(8):1061–1067, 2010. DOI: https://doi.org/10.1093/sleep/33.8.1061.

[Brioschi et al., 2009] A. Brioschi, S. Gramigna, E. Werth, F. Staub, C. Ruffieux, C. Bassetti, M. Schluep, and J. M. Annoni. Effect of modafinil on subjective fatigue in multiple sclerosis and stroke patients. *European Neurology*, 62(4):243–249, 2009. DOI: https://doi.org/10.1159/000232927.

[Bryant et al., 2004] Deborah Bryant, Nancy D Chiaravalloti, and John DeLuca. Objective measurement of cognitive fatigue in multiple sclerosis. *Rehabil. Psychol.*, 49(2):114–122, May 2004. DOI: https://doi.org/10.1037/0090-5550.49.2.114.

[Burke et al., 2018] Sarah E Burke, Immanuel Babu Henry Samuel, Qing Zhao, Jackson Cagle, Ronald A Cohen, Benzi Kluger, and Mingzhou Ding. Task-Based cognitive fatigability for older adults and validation of mental fatigability subscore of pittsburgh fatigability scale. *Front. Aging Neurosci.*, 10:327, October 2018. DOI: https://doi.org/10.3389/fnagi.2018.00327.

[Caridade Gomes, 2019] Pedro Miguel Caridade Gomes. *Development of an open-source Python toolbox for heart rate variability (HRV)*. PhD thesis, Hochschule für angewandte Wissenschaften Hamburg, 2019.

[Casillas et al., 2006] J.M. Casillas, S. Damak, J.C. Chauvet-Gelinier, G. Deley, and P. Ornetti. Fatigue in patients with cardiovascular disease. *Annales de Réadaptation et de Médecine Physique*, 49(6):392 – 402, 2006. DOI: https://doi.org/10.1016/j.annrmp.2006.04.003.

[Chen et al., 2020] Michelle H Chen, Glenn R Wylie, Brian M Sandroff, Rosalia Dacosta-Aguayo, John DeLuca, and Helen M Genova. Neural mechanisms underlying state mental fatigue in multiple sclerosis: a pilot study. *J. Neurol.*, 267(8):2372–2382, August 2020. DOI: https://doi.org/10.1007/s00415-020-09853-w.

[Chipchase et al., 2003] SY Chipchase, NB Lincoln, and KA Radford. Measuring fatigue in people with multiple sclerosis. *Disability and Rehabilitation*, 25(14):778–784, 2003. DOI: https://doi.org/10.1080/0963828031000093477.

[Ciccone et al., 2017] Anthony B Ciccone, Jacob A Siedlik, Jill M Wecht, Jake A Deckert, Nhuquynh D Nguyen, and Joseph P Weir. Reminder: RMSSD and SD1 are identical heart rate variability metrics. *Muscle Nerve*, 56(4):674–678, Oct 2017. DOI: https://doi.org/10.1002/mus.25573.

[Clarke, 1986] David H. Clarke. Sex differences in strength and fatigability. *Research Quarterly for Exercise and Sport*, 57(2):144–149, 1986. DOI: https://doi.org/10.1080/02701367.1986.10762190.

[Cohen and Fisher, 1989] Ronald A. Cohen and Marc Fisher. Amantadine Treatment of Fatigue Associated With Multiple Sclerosis. *Archives of Neurology*, 46(6):676–680, 06 1989. DOI: https://doi.org/10.1001/archneur.1989.00520420096030.

[Colosimo et al., 1995] C. Colosimo, E. Millefiorini, M. G. Grasso, F. Vinci, M. Fiorelli, T. Koudriavtseva, and C. Pozzilli. Fatigue in ms is associated with specific clinical features. *Acta Neurologica Scandinavica*, 92(5):353–355, 1995. DOI: https://doi.org/10.1111/j.1600-0404.1995.tb00145.x.

[Dawson et al., 2014] Drew Dawson, Amelia K. Searle, and Jessica L. Paterson. Look before you (s)leep: Evaluating the use of fatigue detection technologies within a fatigue risk management system for the road transport industry. *Sleep Medicine Reviews*, 18(2):141–152, 2014. DOI: https://doi.org/10.1016/j.smrv.2013.03.003.

[del Rio and Malani, 2020] Carlos del Rio and Preeti N. Malani. Covid-19—new insights on a rapidly changing epidemic. *JAMA*, 323, 04 2020. DOI: https://doi.org/10.1001/jama.2020.3072.

[DeLuca et al., 2008] John DeLuca, Helen M Genova, Frank G Hillary, and Glenn Wylie. Neural correlates of cognitive fatigue in multiple sclerosis using functional MRI. *J. Neurol. Sci.*, 270(1-2):28–39, July 2008. DOI: https://doi.org/10.1016/j.jns.2008.01.018.

[D'hooghe et al., 2018] Marie D'hooghe, Geert Van Gassen, Daphne Kos, Olivier Bouquiaux, Melissa Cambron, Danny Decoo, Andreas Lysandropoulos, Bart Van Wijmeersch, Barbara Willekens, Iris-Katharina Penner, and Guy Nagels. Improving fatigue in multiple sclerosis by smartphone-supported energy management: The ms telecoach feasibility study. *Multiple Sclerosis and Related Disorders*, 22:90 – 96, 2018. DOI: https://doi.org/10.1016/j.msard.2018.03.020.

[Djaldetti et al., 1996] Ruth Djaldetti, Ilan Ziv, Anat Achiron, and Eldad Melamed. Fatigue in multiple sclerosis compared with chronic fatigue syndrome. *Neurology*, 46(3):632–635, 1996. DOI: https://doi.org/10.1212/WNL.46.3.632.

[Dobkin, 2008] Bruce H. Dobkin. Fatigue versus activity-dependent fatigability in patients with central or peripheral motor impairments. *Neurorehabilitation and Neural Repair*, 22(2):105–110, 2008. DOI: https://doi.org/10.1177/1545968308315046.

[El-Gayar et al., 2013] Omar El-Gayar, Prem Timsina, Nevine Nawar, and Wael Eid. Mobile applications for diabetes self-management: Status and potential. *Journal of Diabetes Science and Technology*, 7(1):247–262, 2013. DOI: https://doi.org/10.1177/193229681300700130.

[Empatica Inc., 2018] E4 wristband, April 2018.

[Falk et al., 2007] Kristin Falk, Karl Swedberg, Fannie Gaston-Johansson, and Inger Ekman. Fatigue is a prevalent and severe symptom associated with uncertainty and sense of coherence in patients with chronic heart failure. *European Journal of Cardiovascular Nursing*, 6(2):99–104, 2007. DOI: https://doi.org/10.1016/j.ejcnurse.2006.05.004.

[Field and Hole, 2003] Andy Field and Graham J Hole. *How to Design and Report Experiments*. SAGE Publications Ltd, 2003.

[Fisk and Archibald, 2001] J D Fisk and C J Archibald. Limitations of the paced auditory serial addition test as a measure of working memory in patients with multiple sclerosis. *J. Int. Neuropsychol. Soc.*, 7(3):363–372, March 2001. DOI: https://doi.org/10.1017/s1355617701733103.

[Fitbit, 2018] Fitbit charge hr, 2018.

[Flachenecker et al., 2003] P Flachenecker, A Rufer, I Bihler, C Hippel, K Reiners, K V Toyka, and J Kesselring. Fatigue in MS is related to sympathetic vasomotor dysfunction. *Neurology*, 61(6):851–853, September 2003. DOI: https://doi.org/10.1212/01.wnl.0000080365.95436.b8.

[Florea and Cohn, 2014] Viorel G Florea and Jay N Cohn. The autonomic nervous system and heart failure. *Circ. Res.*, 114(11):1815–1826, May 2014. DOI: https://doi.org/10.1161/CIRCRESAHA.114.302589.

[Friedman and Friedman, 1993] J Friedman and H Friedman. Fatigue in parkinson's disease. *Neurology*, 43(10):2016–2016, 1993. DOI: https://doi.org/10.1212/wnl.43.10.2016.

[Friedman et al., 2010] Joseph H. Friedman, Guido Alves, Peter Hagell, Johan Marinus, Laura Marsh, Pablo Martinez-Martin, Christopher G. Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn Stebbins, and Anette Schrag. Fatigue rating scales critique and recommendations by the movement disorders society task force on rating scales for parkinson's disease. *Movement Disorders*, 25(7):805–822, 2010. DOI: https://doi.org/10.1002/mds.22989.

[Gabriel et al., 2019] Allison S Gabriel, Nathan P Podsakoff, Daniel J Beal, Brent A Scott, Sabine Sonnentag, John P Trougakos, and Marcus M Butts. Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods*, 22(4):969–1006, October 2019. DOI: https://doi.org/10.1177/1094428118802626.

[GAIA, 2023] GAIA. elevida - das Online-Programm für menschen mit MS fatigue. https://elevida.de/, 2023. Accessed: 2022-12-1.

[Genentech, 2023] Inc. Genentech. Floodlight MS. https://www.portal.roche.de/products/neurologie/digitale-gesundheitsloesungen/floodlight-ms.html, 2023. Accessed: 2023-3-2.

[General Electric Healthcare, 2018] Holter recorder seer* 1000, 2018.

[General Electric Healthcare, 2019] Cardioday 2.5 holter ecg, 2019.

[Georgiou et al., 2018] Konstantinos Georgiou, Andreas V. Larentzakis, Nehal N. Khamis, Ghadah I. Alsuhaibani, Yasser A. Alaska, and Elias J. Giallafos. Can wearable devices accurately measure heart rate variability? a systematic review. *Folia Medica*, 60(1):7 – 20, 2018.

[Giardino et al., 2002] Nicholas Giardino, Paul Lehrer, and Robert Edelberg. Comparison of finger plethysmograph to ecg in the measurement of heart rate variability. *Psychophysiology*, 39(2):246–53, 2002.

[Giles et al., 2016] David Giles, Nick Draper, and William Neil. Validity of the polar v800 heart rate monitor to measure rr intervals at rest. *European Journal of Applied Physiology*, 116(3):563–571, 2016.

[Giunti et al., 2017] Guido Giunti, Estefania Guisado-Fernandez, and Brian Caulfield. Connected health in multiple sclerosis: A mobile applications review. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 660–665, 2017. DOI: https://doi.org/10.1109/CBMS.2017.27.

[Giunti et al., 2018] Guido Giunti, Jan Kool, Octavio Rivera Romero, and Enrique Dorronzoro Zubiete. Exploring the specific needs of persons with multiple sclerosis for mhealth solutions for physical activity: Mixed-Methods study. *JMIR Mhealth Uhealth*, 6(2):e37, February 2018.

[Giunti et al., 2020] Guido Giunti, Octavio Rivera-Romero, Jan Kool, Jens Bansi, Jose Luis Sevillano, Anabel Granja-Dominguez, Guillermo Izquierdo-Ayuso, and Diego Giunta. Evaluation of more stamina, a mobile app for fatigue management in persons with multiple sclerosis: Protocol for a feasibility, acceptability, and usability study. *JMIR Res Protoc*, 9(8), 2020. DOI: https://doi.org/10.2196/18196.

[Giunti, 2023] Guido Giunti. More stamina - an evidence-based fatigue management solution for persons with multiple sclerosis. https://www.morestaminaapp.com, 2023. Accessed: 2023-3-2.

[Goldman et al., 2008] 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 Journal*, 14(3):383–390, 2008. DOI: https://doi.org/10.1177/1352458507082607.

[googlefitbit, 2016] googlefitbit, 2016.

[Griffin and Kehoe, 2018] Nicola Griffin and Maria Kehoe. A questionnaire study to explore the views of people with multiple sclerosis of using smartphone technology for health care purposes. *Disabil. Rehabil.*, 40(12):1434–1442, June 2018.

[Hader et al., 1987] W. Hader, P. Duquette, A. Auty Winnipeg, S. Hashimoto, J. Noseworthy, G. Sawa, D. Brunet, R. Nelson, T. Gray, G. Klein, G. Francis, Y. Lapierre, B. Weinshenker, W. Barkas, S. Philips, M. Girard, and T. J. Murray. A randomized controlled trial of amantadine in fatigue associated with multiple sclerosis. *Canadian Journal of Neurological Sciences*, 14(3):273–278, August 1987. DOI: https://doi.org/10.1017/S0317167100026603.

[Hernando et al., 2018] David Hernando, Nuria Garatachea, Rute Almeida, Jose Casajús, and Raquel Bailón. Validation of heart rate monitor polar rs800 for heart rate variability analysis during exercise. *Journal of Strength and Conditioning Research*, 32(3):716–725, March 2018.

[Hewlett et al., 2005] Sarah Hewlett, Zoë Cockshott, Margaret Byron, Karen Kitchen, Sue Tipler, Denise Pope, and Maggie Hehir. Patients' perceptions of fatigue in rheumatoid arthritis: Overwhelming, uncontrollable, ignored. *Arthritis Care & Research*, 53(5):697–702, 2005. DOI: https://doi.org/10.1002/art.21450.

[Hilty et al., 2022] Marc Hilty, Pietro Oldrati, Liliana Barrios, Tamara Müller, Claudia Blumer, Magdalena Foege, consortium, PHRT, Christian Holz, and Andreas Lutterotti. Continuous monitoring with wearables in multiple sclerosis reveals an association

of cardiac autonomic dysfunction with disease severity. *Multiple Sclerosis Journal - Experimental, Translational and Clinical*, 8(2), June 2022. DOI: https://doi.org/10.1177/20552173221103436.

[Hoffmann et al., 1997] R M Hoffmann, T Müller, G Hajak, W Cassel, and Arbeitsgruppe Diagnostik der Deutschen Gesellschaft für Schlafforschung und Schlafmedizin DGSM. Abend-Morgenprotokolle in schlafforschung und Schlafmedizin—Ein standardinstrument für den deutschsprachigen raum. *Somnologie*, 1(3):103–109, October 1997.

[Hubel et al., 2013] Kerry A. Hubel, Bruce Reed, E. William Yund, Timothy J. Herron, and David L. Woods. Computerized measures of finger tapping: Effects of hand dominance, age, and sex. *Perceptual and Motor Skills*, 116(3):929–952, 2013. DOI: https://doi.org/10.2466/25.29.PMS.116.3.929-952.

[Inc., 2023] Empatica Inc. How is ibi.csv obtained? https://support.empatica.com/hc/en-us/articles/201912319-How-is-IBI-csv-obtained-, December 2023.

[Jin Huang and Ling, 2005] Jin Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[Jo et al., 2016] Edward Jo, Kiana Lewis, Dean Directo, Michael J Kim, and Brett A Dolezal. Validation of biofeedback wearables for photoplethysmographic heart rate tracking. *Journal of sports science & medicine*, 15(3):540–547, 08 2016.

[John et al., 1994] Fisk D. John, Pontefract Alexandra, Ritvo G. Paul, Archibald J. Catherine, and Murray J. Thomas. The impact of fatigue on patients with multiple sclerosis. *The Canadian Journal of Neurological Science*, 21:9–14, 1994. DOI: https://doi.org/10.1017/S0317167100048691.

[Jongen et al., 2015] Peter Joseph Jongen, Ludovicus G Sinnige, Björn M van Geel, Freek Verheul, Wim I Verhagen, Ruud A van der Kruijk, Reinoud Haverkamp, Hans M Schrijver, J Coby Baart, Leo H Visser, Edo P Arnoldus, H Jacobus Gilhuis, Paul Pop, Monique Booy, Wim Lemmens, Rogier Donders, Anton Kool, and Esther van Noort. The interactive web-based program msmonitor for self-management and multidisciplinary care in multiple sclerosis: concept, content, and pilot results. *Patient preference and adherence*, 9:1741–1750, 12 2015. DOI: https://doi.org/10.2147/PPA.S93783.

[Kalron et al., 2011] Alon Kalron, Anat Achiron, and Zeevi Dvir. Muscular and gait abnormalities in persons with early onset multiple sclerosis. *J Neurol Phys Ther.*, 35(4):164–169, 2011. DOI: https://doi.org/10.1097/NPT.0b013e31823801f4.

[Kay et al., 2013] Matthew Kay, Kyle Rector, Sunny Consolvo, Ben Greenstein, Jacob Wobbrock, Nathaniel Watson, and Julie Kientz. Pvt-touch: Adapting a reaction time test for touchscreen devices. IEEE, 5 2013. DOI: https://doi.org/10.4108/icst.pervasivehealth.2013.252078.

[Kaynak et al., 2006] H. Kaynak, A. Altintaş, D. Kaynak, Ö. Uyanik, S. Saip, J. Ağaoğlu, G. Önder, and A. Siva. Fatigue and sleep disturbance in multiple sclerosis. *European Journal of Neurology*, 13(12):1333–1339, 2006. DOI: https://doi.org/10.1111/j.1468-1331.2006.01499.x.

[Kelli et al., 2017] Heval Mohamed Kelli, Bradley Witbrodt, and Amit Shah. The future of mobile health applications and devices in cardiovascular health. *Euro Med J Innov*, pages 92–97, 2017.

[Keselbrener et al., 2000] L Keselbrener, S Akselrod, A Ahiron, M Eldar, Y Barak, and Z Rotstein. Is fatigue in patients with multiple sclerosis related to autonomic dysfunction? *Clin. Auton. Res.*, 10(4):169–175, August 2000.

[Kim et al., 2010] Edward Kim, Jesus Lovera, Laura Schaben, Dennis Bourdette, and Ruth Whitham. Novel method for measurement of fatigue in multiple sclerosis: Real-time digital fatigue score. *J Rehabil Res Dev*, 47(5):477–84, 2010. DOI: https://doi.org/10.1682/JRRD.2009.09.0151.

[Kim et al., 2018] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: A Meta-Analysis and review of the literature. *Psychiatry Investig.*, 15(3):235–245, March 2018.

[Kluger et al., 2013] Benzi M. Kluger, Lauren B. Krupp, and Roger M. Enoka. Fatigue and fatigability in neurologic illnesses. *Neurology*, 80(4):409–416, 2013. DOI: https://doi.org/10.1212/WNL.0b013e31827f07be.

[Kobelt et al., 2017] Gisela Kobelt, Alan Thompson, Jenny Berg, Mia Gannedahl, Jennifer Eriksson, the MSCOI Study Group, and the European Multiple Sclerosis Platform. New insights into the burden and costs of multiple sclerosis in europe. *Multiple Sclerosis Journal*, 23(8):1123–1136, 2017. DOI: https://doi.org/10.1177/1352458517694432.

[Koo and Li, 2016] Terry Koo and Mae Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 06 2016.

[Krupp and Elkins, 2000] L B Krupp and L E Elkins. Fatigue and declines in cognitive functioning in multiple sclerosis. *Neurology*, 55(7):934–939, October 2000.

[Krupp et al., 1989] Lauren B. Krupp, Nicholas G. LaRocca, Joanne Muir-Nash, and Alfred D. Steinberg. The Fatigue Severity Scale: Application to Patients With Multiple Sclerosis and Systemic Lupus Erythematosus. *Archives of Neurology*, 46(10):1121–1123, 10 1989. DOI: https://doi.org/10.1001/archneur.1989.00520460115022.

[Krupp et al., 1995] L. B. Krupp, P. K. Coyle, C. Doscher, A. Miller, A. H. Cross, L. Jandorf, J. Halper, B. Johnson, L. Morgante, and R. Grimson. Fatigue therapy in multiple

sclerosis. *Neurology*, 45(11):1956–1961, 1995. DOI: https://doi.org/10.1212/WNL.45.11.1956.

[Krupp et al., 2007] Lauren B Krupp, Nancy McLinskey, and William S MacAllister. Fatigue in multiple sclerosis. *Multiple Sclerosis Therapeutics*, pages 805–818, 2007.

[Krupp, 2003] Lauren B. Krupp. Fatigue in multiple sclerosis. *CNS Drugs*, 17(4):225–234, 2003. DOI: https://doi.org/10.2165/00023210-200317040-00002.

[Kurtzke, 1983] John F. Kurtzke. Rating neurologic impairment in multiple sclerosis. *Neurology*, 33(11):1444–1444, 1983. DOI: https://doi.org/10.1212/WNL.33.11.1444.

[Lakshminarayana et al., 2017] Rashmi Lakshminarayana, Duolao Wang, David Burn, K. Ray Chaudhuri, Clare Galtrey, Natalie Valle Guzman, Bruce Hellman, Ben James, Suvankar Pal, Jon Stamford, Malcolm Steiger, R. W. Stott, James Teo, Roger A. Barker, Emma Wang, Bastiaan R. Bloem, Martijn van der Eijk, Lynn Rochester, and Adrian Williams. Using a smartphone-based self-management platform to support medication adherence and clinical consultation in parkinson's disease. *npj Parkinson's Disease*, 3(1):2, 2017. DOI: https://doi.org/10.1038/s41531-016-0003-z.

[Lange et al., 2009] Rüdiger Lange, Marek Volkmer, Christoph Heesen, and Joachim Liepert. Modafinil effects in multiple sclerosis patients with fatigue. *Journal of Neurology*, 256(4):645–650, 2009.

[Lauren and Dean, 1996] Krupp B. Lauren and Pollina A. Dean. Mechanisms and management of fatigue in progressive neurological disorders. *Current Opinion on Neurology*, 9:456–460, 1996. DOI: https://doi.org/10.1097/00019052-199612000-00011.

[Ledinek et al., 2013] Alenka Horvat Ledinek, Mojca Cizek Sajko, and Uros Rot. Evaluating the effects of amantadin, modafinil and acetyl-l-carnitine on fatigue in multiple sclerosis – result of a pilot randomized, blind study. *Clinical Neurology and Neurosurgery*, 115:S86–S89, 2013. DOI: https://doi.org/10.1016/j.clineuro.2013.09.029.

[Lesage et al., 2012] F-X Lesage, S Berjot, and F Deschamps. Clinical stress assessment using a visual analogue scale. *Occup. Med.*, 62(8):600–605, December 2012.

[Lou et al., 2003] Jau-Shin Lou, Greg Kearns, Theodore Benice, Barry Oken, Gary Sexton, and John Nutt. Levodopa improves physical fatigue in parkinson's disease: A double-blind, placebo-controlled, crossover study. *Movement Disorders*, 18(10):1108–1114, 2003. DOI: https://doi.org/10.1002/mds.10505.

[Loy et al., 2017] Bryan D. Loy, Ruby L. Taylor, Brett W. Fling, and Fay B. Horak. Relationship between perceived fatigue and performance fatigability in people with multiple

sclerosis: A systematic review and meta-analysis. *Journal of Psychosomatic Research*, 100:1 – 7, 2017. DOI: https://doi.org/10.1016/j.jpsychores.2017.06.017.

[Malik, 1996] Marek Malik. Heart rate variability. *Ann. Noninvasive Electrocardiol.*, 1(2):151–181, April 1996.

[Marziniak et al., 2018] Martin Marziniak, Giampaolo Brichetto, Peter Feys, Uta Meyding-Lamadé, Karen Vernon, and Sven G Meuth. The use of digital and remote communication technologies as a tool for multiple sclerosis management: Narrative review. *JMIR Rehabil Assist Technol*, 5(1):e5, Apr 2018. DOI: https://doi.org/10.2196/rehab.7805.

[Mathiowetz et al., 1985] Virgil Mathiowetz, Karen Weber, Nancy Kashman, and Gloria Volland. Adult norms for the nine hole peg test of finger dexterity. *The Occupational Therapy Journal of Research*, 5(1):24–38, 1985. DOI: https://doi.org/10.1177/153944928500500102.

[Mazur-Mosiewicz and Dean, 2011] Anna Mazur-Mosiewicz and Raymond S. Dean. *Halstead-Reitan Neuropsychological Test Battery*, pages 727–731. Springer US, Boston, MA, 2011. DOI: https://doi.org/10.1007/978-0-387-79061-9_1311.

[McSharry et al., 2003] Patrick McSharry, Gari Clifford, Lionel Tarassenko, and Leonard Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003.

[Merkelbach et al., 2001] S Merkelbach, U Dillmann, C Kölmel, I Holz, and M Muller. Cardiovascular autonomic dysregulation and fatigue in multiple sclerosis. *Mult. Scler.*, 7(5):320–326, October 2001.

[Michael et al., 2017] Scott Michael, Kenneth S Graham, and Oam Davis, Glen M. Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals-a review. *Frontiers in physiology*, 8:301; 301–301, 05 2017.

[Midaglia et al., 2019] Luciana Midaglia, Patricia Mulero, Xavier Montalban, Jennifer Graves, Stephen L Hauser, Laura Julian, Michael Baker, Jan Schadrack, Christian Gossens, Alf Scotland, Florian Lipsmeier, Johan van Beek, Corrado Bernasconi, Shibeshih Belachew, and Michael Lindemann. Adherence and satisfaction of smartphone- and Smartwatch-Based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study. *J. Med. Internet Res.*, 21(8):e14863, August 2019.

[Miller et al., 1993] R. G. Miller, R. S. Moussavi, A. T. Green, P. J. Carson, and M. W. Weiner. The fatigue of rapid repetitive movements. *Neurology*, 43(4):755–755, 1993. DOI: https://doi.org/10.1212/WNL.43.4.755.

[Mireia and Roland, 2005] Sospedra Mireia and Martin Roland. Immunology of multiple sclerosis. *Annual Review of Inmmunology*, 65(23):683–747, 2005. DOI: https://doi.org/10.1146/annurev.immunol.23.021704.115707.

[Mitchell, 2010] Sandra A. Mitchell. Cancer-related fatigue: state of the science. *Clinical Translation*, 2 (5):364–383, 2010.

[Möckel et al., 2015] Tina Möckel, Christian Beste, and Edmund Wascher. The effects of time on task in response Selection–An ERP study of mental fatigue. *Sci. Rep.*, 5:10113, June 2015.

[Möller et al., 2011] F Möller, J Poettgen, F Broemel, A Neuhaus, M Daumer, and C Heesen. Hagil (hamburg vigil study): a randomized placebo-controlled double-blind study with modafinil for treatment of fatigue in patients with multiple sclerosis. *Multiple Sclerosis Journal*, 17(8):1002–1009, 2011.

[Möller et al., 2014] Marika Christina Möller, Catharina Nygren de Boussard, Christian Oldenburg, and Aniko Bartfai. An investigation of attention, executive, and psychomotor aspects of cognitive fatigability. *J. Clin. Exp. Neuropsychol.*, 36(7):716–729, June 2014.

[Montalban et al., 2019] Xavier Montalban, Patricia Mulero, Luciana Midaglia, Jennifer Graves, Stephen L. Hauser, Laura Julian, Mike Baker, Jan Schadrack, Christian Gossens, Alf Scotland, Florian Lipsmeier, Corrado Bernasconi, Shibeshih Belachew, and Michael Lindemann. Floodlight: Smartphone-based self-monitoring is accepted by patients and provides meaningful, continuous digital outcomes augmenting conventional in-clinic multiple sclerosis measures (p3.2-024). *Neurology*, 92(15 Supplement), 2019.

[Morrow et al., 2015] Sarah A Morrow, Heather Rosehart, and Andrew M Johnson. Diagnosis and quantification of cognitive fatigue in multiple sclerosis. *Cogn. Behav. Neurol.*, 28(1):27–32, March 2015.

[Motl et al., 2017] Robert W Motl, Elizabeth A Hubbard, Rachel E Bollaert, Brynn C Adamson, Dominique Kinnett-Hopkins, Julia M Balto, Sarah K Sommer, Lara A Pilutti, and Edward McAuley. Randomized controlled trial of an e-learning designed behavioral intervention for increasing physical activity behavior in multiple sclerosis. *Mult Scler J Exp Transl Clin*, 3(4):2055217317734886, October 2017.

[Nacul et al., 2018] Luis Carlos Nacul, Kathleen Mudie, Caroline C. Kingdon, Taane G. Clark, and Eliana Mattos Lacerda. Hand grip strength as a clinical biomarker for me/cfs and disease severity. *Frontiers in Neurology*, 9:992, 2018. DOI: https://doi.org/10.3389/fneur.2018.00992.

[Notermans et al., 1994] N C Notermans, G W van Dijk, Y van der Graaf, J van Gijn, and J H Wokke. Measuring ataxia: quantification based on the standard neurological examination. *Journal of Neurology, Neurosurgery & Psychiatry*, 57(1):22–26, 1994. DOI: https://doi.org/10.1136/jnnp.57.1.22.

[Nourbakhsh et al., 2021] Bardia Nourbakhsh, Nisha Revirajan, Bridget Morris, Christian Cordano, Jennifer Creasman, Michael Manguinao, Kristen Krysko, Alice Rutatangwa, Caroline Auvray, Salman Aljarallah, Chengshi Jin, Ellen Mowry, Charles McCulloch, and Emmanuelle Waubant. Safety and efficacy of amantadine, modafinil, and methylphenidate for fatigue in multiple sclerosis: a randomised, placebo-controlled, crossover, double-blind trial. *The Lancet Neurology*, 20(1):38–48, 2021. DOI: https://doi.org/10.1016/S1474-4422(20)30354-9.

[Novartis, 2023] Novartis. First of its kind disease tracking and management resource launched for people with MS – SymTrac. https://www.novartis.com/, 2023. Accessed: 2023-3-2.

[Nunan et al., 2008] David Nunan, Djordje Jakovljevic, Gay Donovan, Lynette Hodges, Gavin Sandercock, and David Brodie. Levels of agreement for rr intervals and short-term heart rate variability obtained from the polar s810 and an alternative system. *European Journal of Applied Physiology*, 103(5):529–537, 2008. DOI: https://doi.org/10.1007/s00421-008-0742-6.

[Parak and Korhonen, 2014] Jakub Parak and Ilkka Korhonen. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3670–3673, 2014. DOI: https://doi.org/10.1109/EMBC.2014.6944419.

[Patel et al., 2011] M Patel, S K L Lal, D Kavanagh, and P Rossiter. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Syst. Appl.*, 38(6):7235–7242, June 2011. DOI: https://doi.org/10.1016/j.eswa.2010.12.028.

[Patel et al., 2017] Viral P Patel, Lisa A S Walker, and Anthony Feinstein. Deconstructing the symbol digit modalities test in multiple sclerosis: The role of memory. *Mult. Scler. Relat. Disord.*, 17:184–189, October 2017. DOI: https://doi.org/10.1016/j.msard.2017.08.006.

[Pattyn et al., 2008] Nathalie Pattyn, Xavier Neyt, David Henderickx, and Eric Soetens. Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiol. Behav.*, 93(1-2):369–378, January 2008. DOI: https://doi.org/10.1016/j.physbeh.2007.09.016.

[Penner and Paul, 2017] Iris-Katharina Penner and Friedemann Paul. Fatigue as a symptom or comorbidity of neurological diseases. *Nat. Rev. Neurol.*, 13(11):662–675, November 2017. DOI: https://doi.org/10.1038/nrneurol.2017.117.

[Penner et al., 2009] Iris-Katharina Penner, Chiara Raselli, Markus Stöcklin, Klaus Opwis, Ludwig Kappos, and Pasquale Calabrese. The fatigue scale for motor and cognitive functions (fsmc): validation of a new instrument to assess multiple sclerosis-related

fatigue. *Multiple Sclerosis Journal*, 15(12):1509–1517, 2020/07/30 2009. DOI: https://doi.org/10.1177/1352458509348519.

[Perrin et al., 2020] Ray Perrin, Lisa Riste, Mark Hann, Andreas Walther, Annice Mukherjee, and Adrian Heald. Into the looking glass: Post-viral syndrome post covid-19. *Medical hypotheses*, 144:110055–110055, 11 2020. DOI: https://doi.org/10.1016/j.mehy.2020.110055.

[Polar, 2018] Polar oh1. https://www.polar.com/en/products/accessories/oh1-optical-heart-rate-sensor, 2018.

[Preacher and Selig, 2012] Kristopher J. Preacher and James P. Selig. Advantages of monte carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2):77–98, 2012. DOI: https://doi.org/10.1080/19312458.2012.679848.

[Preuveneers and Berbers, 2008] Davy Preuveneers and Yolande Berbers. Mobile phones assisting with health self-care: A diabetes case study. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '08, pages 177–186, New York, NY, USA, 2008. ACM. DOI: https://doi.org/10.1145/1409240.1409260.

[Prince et al., 2018] John Prince, Siddharth Arora, and Maarten de Vos. Big data in parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiological Measurement*, 39(4):044005, 2018. DOI: https://doi.org/10.1088/1361-6579/aab512.

[Printy et al., 2014] Blake P. Printy, Lindsey M. Renken, John P. Herrmann, Isac Lee, Bryant Johnson, Emily Knight, Georgeta Varga, and D. Diane Whitmer. Smartphone application for classification of motor impairment severity in parkinson's disease. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014. DOI: https://doi.org/10.1109/EMBC.2014.6944176.

[Pucci et al., 2007] E Pucci, P Branãs, R D'Amico, G Giuliani, A Solari, and C Taus. Amantadine for fatigue in multiple sclerosis. *The Cochrane database of systematic reviews*, 2007(1):CD002818–CD002818, 01 2007. DOI: https://doi.org/10.1002/14651858.CD002818.pub2.

[Racosta and Kimpinski, 2016] Juan Manuel Racosta and Kurt Kimpinski. Autonomic dysfunction, immune regulation, and multiple sclerosis. *Clin. Auton. Res.*, 26(1):23–31, February 2016. DOI: https://doi.org/10.1007/s10286-015-0325-7.

[Rammohan et al., 2002] K W Rammohan, J H Rosenberg, D J Lynn, A M Blumenfeld, C P Pollak, and H N Nagaraja. Efficacy and safety of modafinil (provigil) for the treatment of fatigue in multiple sclerosis: a two centre phase 2 study. *Journal of neurology, neurosurgery, and psychiatry*, 72(2):179–183, 02 2002. DOI: https://doi.org/10.1136/jnnp.72.2.179.

[Rice et al., 2021] Dylan R Rice, Tamara B Kaplan, Gladia C Hotan, Andre C Vogel, Marcelo Matiello, Rebecca L Gillani, Spencer K Hutto, Andrew S Ham, Eric C Klawiter, Ilena C George, Kristin Galetta, and Farrah J Mateen. Electronic pill bottles to monitor and promote medication adherence for people with multiple sclerosis: A randomized, virtual clinical trial. *J. Neurol. Sci.*, 428:117612, September 2021. DOI: https://doi.org/10.1016/j.jns.2021.117612.

[Richardson et al., 1996] P Richardson, W McKenna, M Bristow, B Maisch, B Mautner, J O'Connell, E Olsen, G Thiene, J Goodwin, I Gyarfas, I Martin, and P Nordet. Report of the 1995 world health Organization/International society and federation of cardiology task force on the definition and classification of cardiomyopathies. *Circulation*, 93(5):841–842, March 1996. DOI: https://doi.org/10.1161/01.cir.93.5.841.

[Roar et al., 2016] Malte Roar, Zsolt Illes, and Tobias Sejbaek. Practice effect in symbol digit modalities test in multiple sclerosis patients treated with natalizumab. *Mult. Scler. Relat. Disord.*, 10:116–122, November 2016. DOI: https://doi.org/10.1016/j.msard.2016.09.009.

[Rotstein et al., 2012] D. Rotstein, P. O'Connor, L. Lee, and B. J. Murray. Multiple sclerosis fatigue is associated with reduced psychomotor vigilance. 39(2):180–184, 2012. DOI: https://doi.org/10.1017/s0317167100013196.

[Rudroff et al., 2016] Thorsten Rudroff, John H. Kindred, and Nathaniel B. Ketelhut. Fatigue in multiple sclerosis: Misconceptions and future research directions. *Frontiers in Neurology*, 7:122, 2016.

[Samn and Perelli, 1982] Sherwood W Samn and Layne P Perelli. Estimating aircrew fatigue: a technique with application to airlift operations. Technical report, School of Aerospace Medicine Brooks Afb tx, 1982.

[Scherer et al., 1997] P. Scherer, H. Bauer, and K. Baum. Alternate finger tapping test in patients with migraine. *Acta Neurologica Scandinavica*, 96(6):392–396, 1997. DOI: https://doi.org/10.1111/j.1600-0404.1997.tb00304.x.

[Schwid et al., 1999] Steven R. Schwid, Charles A. Thornton, Shree Pandya, Karishma L. Manzur, Mohammed Sanjak, Marie D. Petrie, Michael P. McDermott, and Andrew D. Goodman. Quantitative assessment of motor fatigue and strength in ms. *Neurology*, 53(4):743–743, 1999. DOI: https://doi.org/10.1212/WNL.53.4.743.

[Schwid et al., 2002] Steven R. Schwid, Melissa Covington, Benjamin M. Segal, and Andrew D. Goodman. Fatigue in multiple sclerosis: Current understanding and future directions. *Journal of Rehabilitation Research and Development*, 39(2):211–224, 2002.

[Schwid et al., 2003] Steven R Schwid, Carolyn M Tyler, Eileen A Scheid, Amy Weinstein, Andrew D Goodman, and Michael P McDermott. Cognitive fatigue during a test

requiring sustained attention: a pilot study. *Multiple Sclerosis Journal*, 9(5):503–508, October 2003.

[Sehle et al., 2011] Aida Sehle, Annegret Mündermann, Klaus Starrost, Simon Sailer, Inna Becher, Christian Dettmers, and Manfred Vieten. Objective assessment of motor fatigue in multiple sclerosis using kinematic gait analysis: a pilot study. *J Neuroeng Rehabil*, 8:59, 2011.

[Severijns et al., 2014] Deborah Severijns, Ilse Lamers, Lore Kerkhofs, and Peter Feys. Hand grip fatigability in persons with multiple sclerosis according to hand dominance and disease progression. *Journal of rehabilitation medicine*, 47, 09 2014. DOI: `https://doi.org/10.2340/16501977-1897`.

[Severijns et al., 2015] Deborah Severijns, Ilse Lamers, Lore Kerkhofs, and Peter Feys. Hand grip fatigability in persons with multiple sclerosis according to hand dominance and disease progression. *J Rehabil Med.*, pages 154–160, 2015. DOI: `https://doi.org/10.2340/16501977-1897`.

[Severijns et al., 2017] Deborah Severijns, Inge Zijdewind, Ulrik Dalgas, Ilse Lamers, Caroline Lismont, and Peter Feys. The assessment of motor fatigability in persons with multiple sclerosis: A systematic review. *Neurorehabilitation and Neural Repair*, 31(5):413–431, 2017.

[Shaffer and Ginsberg, 2017] Fred Shaffer and J P Ginsberg. An overview of heart rate variability metrics and norms. *Front Public Health*, 5:258, September 2017.

[Sheng et al., 2013] Ping Sheng, Lijun Hou, Xiang Wang, Xiaowen Wang, Chengguang Huang, Mingkun Yu, Xi Han, and Yan Dong. Efficacy of modafinil on fatigue and excessive daytime sleepiness associated with neurological disorders: a systematic review and meta-analysis. *PloS one*, 8(12):e81802–e81802, 12 2013.

[Shirani et al., 2017] Afsaneh Shirani, Braeden D. Newton, and Darin T. Okuda. Finger tapping impairments are highly sensitive for evaluating upper motor neuron lesions. *BMC Neurology*, 17(1):55, 2017. DOI: `https://doi.org/10.1186/s12883-017-0829-y`.

[Skurvydas et al., 2011] Albertas Skurvydas, Marius Brazaitis, Julija Andrejeva, Dalia Mickeviciene, and Vytautas Streckis. The effect of multiple sclerosis and gender on central and peripheral fatigue during 2-min mvc. *Clinical Neurophysiology*, 122(4):767 – 776, 2011. DOI: `https://doi.org/10.1016/j.clinph.2010.10.005`.

[Sletten et al., 2012] David M Sletten, Guillermo A Suarez, Phillip A Low, Jay Mandrekar, and Wolfgang Singer. COMPASS 31: a refined and abbreviated composite autonomic symptom score. *Mayo Clin. Proc.*, 87(12):1196–1201, December 2012.

[Smith, 1982] A Smith. Symbol digit modalities test (SDMT) manual (revised) western psychological services. *Los Angeles*, 1982.

[Society, 2018] National Multiple Sclerosis Society. Who gets ms? (epidemiology). Available at https://www.nationalmssociety.org/What-is-MS/Who-Gets-MS (2018/02/22), 2018.

[Spelten et al., 2003] E. R. Spelten, J. H. A. M. Verbeek, A. L. J. Uitterhoeve, A. C. Ansink, J. van der Lelie, T. M. de Reijke, M. Kammeijer, J. C. J. M. de Haes, and M. A. G. Sprangers. Cancer, fatigue and the return of patients to work—a prospective cohort study. *European Journal of Cancer*, 39(11):1562–1567, 2003. DOI: https://doi.org/10.1016/S0959-8049(03)00364-2.

[Stankoff et al., 2005] B. Stankoff, E. Waubant, C. Confavreux, G. Edan, M. Debouverie, L. Rumbach, T. Moreau, J. Pelletier, C. Lubetzki, and M. Clanet. Modafinil for fatigue in ms. *Neurology*, 64(7):1139–1143, 2005. DOI: https://doi.org/10.1212/01.WNL.0000158272.27070.6A.

[Steens et al., 2012] Anneke Steens, Astrid de Vries, Jolien Hemmen, Thea Heersema, Marco Heerings, Natasha Maurits, and Inge Zijdewind. Fatigue perceived by multiple sclerosis patients is associated with muscle fatigue. *Neurorehabilitation and Neural Repair*, 26(1):48–57, 2012. DOI: https://doi.org/10.1177/1545968311416991.

[Stroop, 1935] J R Stroop. Studies of interference in serial verbal reactions. *J. Exp. Psychol.*, 18(6):643–662, December 1935.

[Surakka et al., 2004] Jukka Surakka, Anders Romberg, Juhani Ruutiainen, Arja Virtanen, Sirkka Aunola, and Kari Mäentaka. Assessment of muscle strength and motor fatigue with a knee dynamometer in subjects with multiple sclerosis: a new fatigue index. *Clinical Rehabilitation*, 18(6):652–659, 2004. DOI: https://doi.org/10.1191/0269215504cr781oa.

[Tanigawa et al., 2017] Makoto Tanigawa, Jason Stein, John Park, Peter Kosa, Irene Cortese, and Bibiana Bielekova. Finger and foot tapping as alternative outcomes of upper and lower extremity function in multiple sclerosis. *Multiple sclerosis journal - experimental, translational and clinical*, 3(1), 2017. DOI: https://doi.org/10.1177/2055217316688930.

[Taylor Tavares et al., 2005] Ana Lisa Taylor Tavares, Gregory S.X.E Jefferis, Mandy Koop, Bruce C. Hill, Trevor Hastie, Gary Heit, and Helen M. Bronte-Stewart. Quantitative measurements of alternating finger tapping in parkinson's disease correlate with updrs motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. *Movement Disorders*, 20(10):1286–1298, 2005. DOI: https://doi.org/10.1002/mds.20556.

[Tombaugh, 2006] Tom N Tombaugh. A comprehensive review of the paced auditory serial addition test (PASAT). *Arch. Clin. Neuropsychol.*, 21(1):53–76, January 2006. DOI: https://doi.org/10.1016/j.acn.2005.07.006.

[Tong et al., 2019] Catherine Tong, Matthew Craner, Matthieu Vegreville, and Nicholas D. Lane. Tracking fatigue and health state in multiple sclerosis patients using connnected wellness devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), September 2019. DOI: https://doi.org/10.1145/3351264.

[Townsend et al., 2020] Liam Townsend, Adam H. Dyer, Karen Jones, Jean Dunne, Aoife Mooney, Fiona Gaffney, Laura O'Connor, Deirdre Leavy, Kate O'Brien, Joanne Dowds, Jamie A. Sugrue, David Hopkins, Ignacio Martin-Loeches, Cliona Ni Cheallaigh, Parthiban Nadarajan, Anne Marie McLaughlin, Nollaig M. Bourke, Colm Bergin, Cliona O'Farrelly, Ciaran Bannan, and Niall Conlon. Persistent fatigue following sars-cov-2 infection is common and independent of severity of initial infection. *PLOS ONE*, 15(11):1–12, 11 2020.

[Tran et al., 2009] Yvonne Tran, Nirupama Wijesuriya, Mika Tarvainen, Pasi Karjalainen, and Ashley Craig. The relationship between spectral changes in heart rate variability and fatigue. *J. Psychophysiol.*, 23(3):143–151, January 2009. DOI: https://doi.org/10.1027/0269-8803.23.3.143.

[Téllez et al., 2005] Nieves Téllez, Jordi Río, Mar Tintoré, Carlos Nos, Inés Galán, and Xavier Montalban. Does the modified fatigue impact scale offer a more comprehensive assessment of fatigue in ms? *Multiple Sclerosis Journal*, 11(2):198–202, 2005. DOI: https://doi.org/10.1191/1352458505ms1148oa.

[Valko et al., 2008] Philipp O Valko, Claudio L Bassetti, Konrad E Bloch, Ulrike Held, and Christian R Baumann. Validation of the fatigue severity scale in a swiss cohort. *Sleep*, 31(11):1601–1607, 2008. DOI: https://doi.org/10.1093/sleep/31.11.1601.

[van der Linden et al., 2003] Dimitri van der Linden, Michael Frese, and Theo F Meijman. Mental fatigue and the control of cognitive processes: effects on perseveration and planning. *Acta Psychol.*, 113(1):45–65, May 2003. DOI: https://doi.org/10.1016/s0001-6918(02)00150-6.

[Van Kessel et al., 2017] Kirsten Van Kessel, Duncan R Babbage, Nicholas Reay, Warren M Miner-Williams, and Paula Kersten. Mobile technology use by people experiencing multiple sclerosis fatigue: Survey methodology. *JMIR Mhealth Uhealth*, 5(2):e6, February 2017. DOI: https://doi.org/10.2196/mhealth.6192.

[van Oirschot et al., 2020] Pim van Oirschot, Marco Heerings, Karine Wendrich, Bram den Teuling, Marijn B Martens, and Peter J Jongen. Symbol digit modalities test variant in a smartphone app for persons with multiple sclerosis: Validation study. *JMIR Mhealth Uhealth*, 8(10):e18160, October 2020. DOI: https://doi.org/10.2196/18160.

[Vescio et al., 2018] Basilio Vescio, Maria Salsone, Antonio Gambardella, and Aldo Quattrone. Comparison between electrocardiographic and earlobe pulse photoplethysmographic detection for evaluating heart rate variability in healthy subjects in short-

and long-term recordings. *Sensors (Basel, Switzerland)*, 18(3), 2018. DOI: https://doi.org/10.3390/s18030844.

[Wahoo Fitness, 2018] Wahoo ticker fit, 2018.

[Walker et al., 2012a] L A S Walker, J A Berard, L I Berrigan, L M Rees, and M S Freedman. Detecting cognitive fatigue in multiple sclerosis: method matters. *J. Neurol. Sci.*, 316(1-2):86–92, May 2012. DOI: https://doi.org/10.1016/j.jns.2012.01.021.

[Walker et al., 2012b] Lisa A S Walker, Amy Cheng, Jason Berard, Lindsay I Berrigan, Laura M Rees, and Mark S Freedman. Tests of information processing speed: what do people with multiple sclerosis think about them? *Int. J. MS Care*, 14(2):92–99, 2012. DOI: https://doi.org/10.7224/1537-2073-14.2.92.

[Walker et al., 2019] Lisa A S Walker, Alyssa P Lindsay-Brown, and Jason A Berard. Cognitive fatigability interventions in neurological conditions: A systematic review. *Neurol Ther*, (2):251–271, December 2019. DOI: https://doi.org/10.1007/s40120-019-00158-3.

[Wallen et al., 2016] Matthew Wallen, Sjaan Gomersall, Shelley Keating, Ulrik Wisløff, and Jeff Coombes . Accuracy of heart rate watches: Implications for weight management. *PLOS ONE*, 11(5), 2016. DOI: https://doi.org/10.1371/journal.pone.0154420.

[Wang et al., 2014] Chao Wang, Mingzhou Ding, and Benzi M Kluger. Change in intraindividual variability over time as a key metric for defining performance-based cognitive fatigability. *Brain Cogn.*, 85:251–258, March 2014. DOI: https://doi.org/10.1016/j.bandc.2014.01.004.

[Wascher et al., 2014] Edmund Wascher, Björn Rasch, Jessica Sänger, Sven Hoffmann, Daniel Schneider, Gerhard Rinkenauer, Herbert Heuer, and Ingmar Gutberlet. Frontal theta activity reflects distinct aspects of mental fatigue. *Biol. Psychol.*, 96:57–65, February 2014. DOI: https://doi.org/10.1016/j.biopsycho.2013.11.010.

[Waxenbaum et al., 2019] J A Waxenbaum, V Reddy, and M Varacallo. Anatomy, autonomic nervous system. 2019. Accessed: 2023-3-2.

[Williamson et al., 2011] Ann Williamson, David A. Lombardi, Simon Folkard, Jane Stutts, Theodore K. Courtney, and Jennie L. Connor. The link between fatigue and safety. *Accident Analysis & Prevention*, 43(2):498–515, 2011. DOI: https://doi.org/10.1016/j.aap.2009.11.011.

[Wolkorte et al., 2015] Ria Wolkorte, Dorothea J. Heersema, and Inge Zijdewind. Muscle fatigability during a sustained index finger abduction and depression scores are

associated with perceived fatigue in patients with relapsing-remitting multiple sclerosis. *Neurorehabilitation and Neural Repair*, 29(8):796–802, 2015. DOI: https://doi.org/10.1177/1545968314567151.

[World Medical Association, 2002] World Medical Association. *World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. World Medical Association, 2002.

[Yu et al., 2013] Fei Yu, Arne Bilberg, Ulrik Dalgas, and Egon Stenager. Fatigued patients with multiple sclerosis can be discriminated from healthy controls by the recordings of a newly developed measurement system (famos): a pilot study. *Disabil Rehabil Assist Technol*, 8(1):77–83, 2013. DOI: https://doi.org/10.3109/17483107.2012.680941.

[Zhai et al., 2004] Shumin Zhai, Jing Kong, and Xiangshi Ren. Speed-accuracy tradeoff in fitts' law tasks: On the equivalency of actual and nominal pointing precision. *Int. J. Hum.-Comput. Stud.*, 61(6):823–856, 2004. DOI: https://doi.org/10.1016/j.ijhcs.2004.09.007.

[Zijdewind et al., 2016] Inge Zijdewind, Roeland Prak, and Ria Wolkorte. Fatigue and fatigability in persons with multiple sclerosis. *Exercise and Sport Sciences Reviews*, 80:123–128, 2016. DOI: https://doi.org/10.1249/JES.0000000000000088.

APPENDIX A

Fatigue Questionnaires

A.1. Fatigue Severity Scale (FSS)

Table A.1.: Fatigue Severity Scale

1.	My motivation is lower when I am fatigued.
2.	Exercise brings on my fatigue.
3.	I am easily fatigued.
4.	Fatigue interferes with my physical functioning.
5.	Fatigue causes frequent problems for me.
6.	My fatigue prevents sustained physical functioning.
7.	Fatigue interferes with carrying out certain duties and responsibilities.
8.	Fatigue is among my most disabling symptoms.
9.	Fatigue interferes with my work, family, or social life.

Notes: Fatigue Severity Scale (FSS). Patients rate their fatigue from "strongly disagree" (1) to "strongly agree" (7) [Krupp et al., 1989; Valko et al., 2008].

A.2. Fatigue Scale for Motor and Cognitive Functions (FSMC)

FSMC questionnaire by Penner et al. [2009].

FSMC

Fatigue Scale for Motor and Cognitive Functions

Date: _____

ID: _____

Initials: _____

Age: _____ Sex: m ◯ f ◯

Instructions

The following questionnaire is about problems in everyday life which are directly associated with an extreme form of tiredness (fatigue). This extreme form of tiredness refers to an overwhelming state of lethargy, exhaustion and lack of energy which comes on abruptly and is unrelated to any obvious external causes. It does not mean the sort of isolated episodes which everyone might experience in the course of the day, after exertion, or after a sleepless night!
Please read each statement carefully. Then decide to what extent each statement applies to you and your everyday life. Please try not to base your answers on the way you are feeling at the moment; instead try to give us a picture of the way you feel in normal day-to-day life. Please put a cross in the appropriate circle (only one cross per statement please!).

	Does not apply at all	Does not apply much	Slightly applies	Applies a lot	Applies completely
1. When I concentrate for a long time, I get exhausted sooner than other people of my age.	◯	◯	◯	◯	◯
2. When I am experiencing episodes of exhaustion, my movements become noticeably clumsier and less coordinated.	◯	◯	◯	◯	◯
3. Because of my episodes of exhaustion, I now need more frequent and/or longer rests during physical activity than I used to.	◯	◯	◯	◯	◯
4. When I am experiencing episodes of exhaustion, I am incapable of making decisions.	◯	◯	◯	◯	◯
5. When faced with stressful situations, I now find that I get physically exhausted quicker than I used to.	◯	◯	◯	◯	◯
6. Because of my episodes of exhaustion, I now have considerably less social contact than I used to.	◯	◯	◯	◯	◯
7. Because of my episodes of exhaustion, I now find it more difficult to learn new things than I used to.	◯	◯	◯	◯	◯

Please turn over

FSMC-cog = _____ FSMC-mot = _____ **FSMC total** = _____

FSMC

	Does not apply at all	Does not apply much	Slightly applies	Applies a lot	Applies completely
8. The demands of my work exhaust me mentally more quickly than they used to.	◯	◯	◯	◯	◯
9. I feel the episodes of exhaustion particularly strongly in my muscles.	◯	◯	◯	◯	◯
10. I no longer have the stamina for long periods of physical activity that I used to have.	◯	◯	◯	◯	◯
11. My powers of concentration decrease considerably when I'm under stress.	◯	◯	◯	◯	◯
12. When I am experiencing episodes of exhaustion, I am less motivated than others to start activities that involve physical effort.	◯	◯	◯	◯	◯
13. My thinking gets increasingly slow when it is hot.	◯	◯	◯	◯	◯
14. When I am experiencing an episode of exhaustion, my movements become noticeably slower.	◯	◯	◯	◯	◯
15. Because of my episodes of exhaustion, I now feel less like doing things which require concentration.	◯	◯	◯	◯	◯
16. When an episode of exhaustion comes on, I am simply no longer able to react quickly.	◯	◯	◯	◯	◯
17. When I am experiencing episodes of exhaustion, certain words simply escape me.	◯	◯	◯	◯	◯
18. When I am experiencing episodes of exhaustion, I lose concentration considerably quicker than I used to.	◯	◯	◯	◯	◯
19. When it is hot, my main feeling is one of extreme physical weakness and lack of energy.	◯	◯	◯	◯	◯
20. During episodes of exhaustion, I am noticeably more forgetful.	◯	◯	◯	◯	◯

Please make sure that you have written down your initials, age and sex on page 1 and have put a cross by each statement. Thank you.

Table A.2.: FSMC cut-off values. We focus our study in the motor aspect of fatigue and classify as motor fatigued participants with FSMC physical score ≥ 22; otherwise, we consider them non-fatigued.

	Cut-off	Classification
FSMC sum score	≥ 43	Mild fatigue
	≥ 53	Moderate fatigue
	≥ 63	Severe fatigue
FSMC cognitive score	≥ 22	Mild cognitive fatigue
	≥ 28	Moderate cognitive fatigue
	≥ 34	Severe cognitive fatigue
FSMC physical score	≥ 22	Mild motor fatigue
	≥ 27	Moderate motor fatigue
	≥ 32	Severe motor fatigue

Notes: Fatigue Scale for Motor and Cognitive Functions (FSMC) Penner et al. [2009].

APPENDIX B

Rapid Tapping Supplementary Materials

B.1. Mean Tapping Frequency vs. Mean Handgrip According to FSS

When grouping by gender (Figure B.1 center), there is a significant difference between non-fatigued and motor fatigued females as defined by the FSS questionnaire, with $H = 4.93$ ($p < 0.05$), a difference is also found in males, with $H = 4.82$ ($p < 0.05$). The handgrip shows no difference between non-fatigued and fatigued patients within the gender groups, but it shows a significant difference between genders, with $H = 10.58$ ($p < 0.01$) and $H = 6.35$ ($p < 0.01$) for non-fatigued and fatigued patients, respectively.

Figure B.1 (right) shows the boxplots when grouping by impairment as defined by the 9-HPT. There is a significant difference between the mean tapping frequency of non-fatigued and motor fatigued patients that are not hand impaired, with $H = 6.5$ ($p < 0.01$), while no significant difference is found in impaired participants, where we have a very small sample size. The mean handgrip strength shows no difference between and within the groups.

B.2. Maximum tapping frequency vs. maximum handgrip

As depicted in Figure B.2 (top left), there is a significant difference between the maximum tapping frequency of patients that do not have motor fatigue and those who are classified as motor fatigued using the FSMC questionnaire, with Kruskal-Wallis $H = 8.67$ ($p < 0.01$). However, there is no statistically significant difference between the same groups using the maximum handgrip strength (cf. Figure B.2 bottom left).

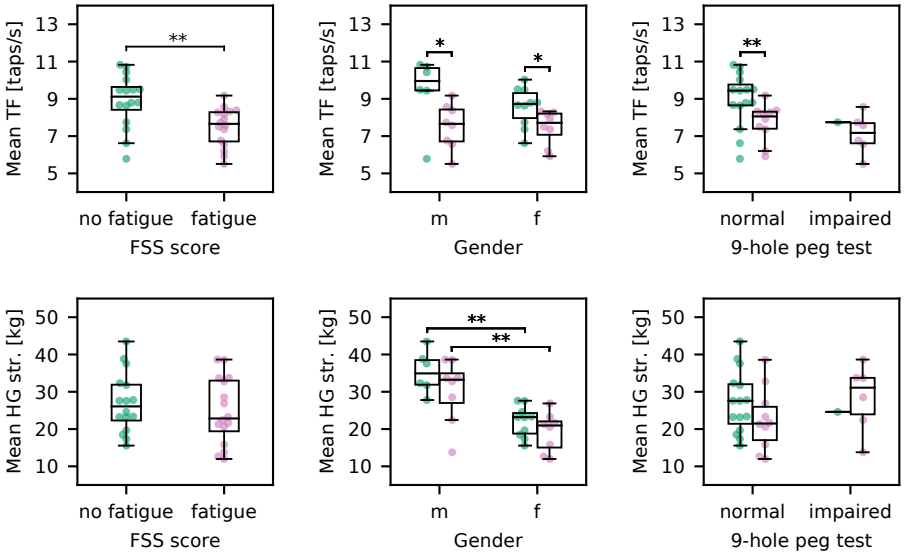

Figure B.1.: Mean tapping frequency (*top*) and mean handgrip strength (*bottom*) in function of FSS fatigue, gender, and impairment as defined by the 9-hole peg test.

When grouping by gender (Figure B.2 center), there is a significant difference between non-fatigued and motor fatigued females as defined by the FSMC questionnaire, with $H = 8.36$ ($p < 0.01$), while no significant difference is found in males, where we have a smaller sample size. The handgrip shows no difference between non-fatigued and fatigued patients within the gender groups, but it shows a significant difference between genders, with $H = 11.0$ ($p < 0.001$) and $H = 8.33$ ($p < 0.01$) for non-fatigued and fatigued patients, respectively.

Figure B.2 (right) shows the boxplots when grouping by impairment as defined by the 9-HPT. There is a significant difference between the maximum tapping frequency of non-fatigued and motor fatigued patients that are not hand impaired, with $H = 6.69$ ($p < 0.01$), while no significant difference is found in impaired participants, where we have a very small sample size. The max handgrip strength shows no difference between and within the groups.

B.3. Descriptive Statistics and Non-parametric Test Results

Table B.1.: FSMC motor fatigued vs. non-fatigued differences. Non-parametric hypotheses tests with dependent variable *Metric* (mean tapping frequency or mean handgrip strengths) and independent variable motor fatigue classification. Test conducted with IBM SPSS Statistics Version 27. Kruskal-Wallis H^*, Mann-Whitney U^*, Kolmogorov-Smirnov Z^*, Median Test $Median^*$.

Metric	Case	N	FSMC fatigued	FSMC Non-fatigued	Test	$p < 0.05$
mean TF [taps/sec]	all	34 M = 8.30 SD = 1.47	17 $M = 7.62$ $SD = 1.22$	17 $M = 8.99$ $SD = 1.40$	$H = 7.498$ $U = 65.000$ $Z = 1.543$ $Median$ = 8.35	✓$p = .006$ ✓$p = .006$ ✓$p = .017$ ✓$p = .016$
	male	15 M = 8.53 SD = 1.84	9 $M = 7.96$ $SD = 1.41$	6 $M = 9.41$ $SD = 2.19$	$H = 2.347$ $U = 14.000$ $Z = 1.265$ $Median$ = 8.56	$p = .126$ $p = .126$ $p = .082$ $p = .315$
	female	19 M = 8.13 SD = 1.12	8 $M = 7.25$ $SD = 0.91$	11 $M = 8.76$ $SD = 0.77$	$H = 8.84$ $U = 8.00$ $Z = 1.565$ $Median$ = 8.27	✓$p = .003$ ✓$p = .003$ ✓$p = .015$ ✓$p = .02$
	impaired	7 M = 7.21 SD = 1.00	5 $M = 7.03$ $SD = 1.17$	2 $M = 7.66$ $SD = 0.11$	$H = .600$ $U = 3.000$ $Z = .717$ $Median$ = 7.58	$p = .439$ $p = .439$ $p = .683$ $p = 1$
	non-impaired	27 M = 8.59 SD = 1.45	12 $M = 7.87$ $SD = 1.20$	15 $M = 9.17$ $SD = 1.41$	$H = 5.717$ $U = 41.000$ $Z = 1.42$ $Median$ = 8.67	✓$p = .017$ ✓$p = .017$ ✓$p = .035$ $p = .054$
mean HG [kg]	all	34 M = 25.97 SD = 8.30	17 $M = 24.88$ $SD = 9.07$	17 $M = 27.07$ $SD = 7.56$	$H = .406$ $U = 126.00$ $Z = .857$ $Median$ = 23.988	$p = .524$ $p = .540$ $p = .454$ $p = .732$
	male	15 M = 32.47 SD = 7.34	9 $M = 30.51$ $SD = 8.00$	6 $M = 35.41$ $SD = 5.59$	$H = .681$ $U = 20.00$ $Z = .527$ $Median$ = 33.27	$p = .409$ $p = .409$ $p = .944$ $p = 1.0$

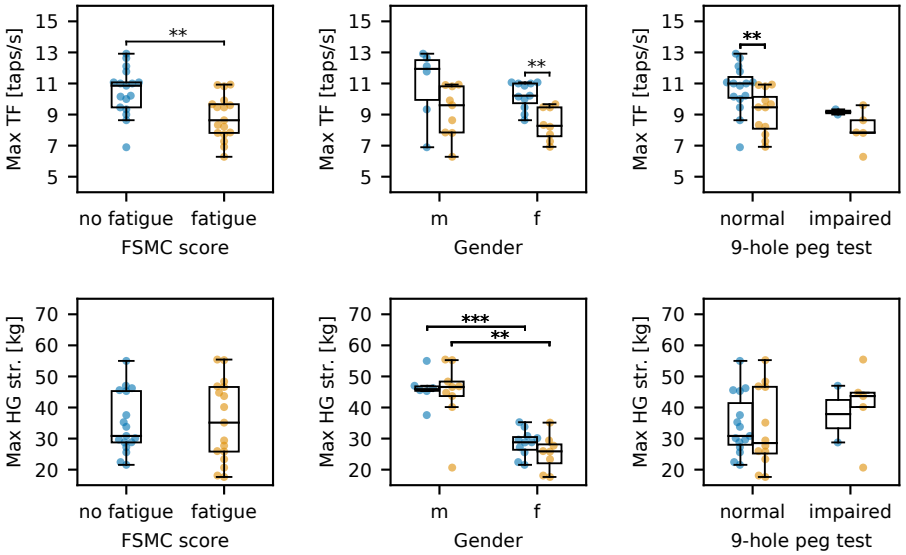

Figure B.2.: Maximum tapping frequency (*top*) and maximum handgrip strength (*bottom*) in function of FSMC motor fatigue, gender, and impairment as defined by the 9-hole peg test.

female	19 M = 20.84 SD = 4.62	8 $M = 18.55$ $SD = 5.34$	11 $M = 22.51$ $SD = 3.33$	$H = 2.727$ $U = 24.00$ $Z = 1.076$ *Median* = 21.32	$p = .099$ $p = .0993$ $p = .197$ $p = .650$
impaired	7 M = 27.91 SD = 8.40	5 $M = 26.43$ $SD = 8.47$	2 $M = 31.62$ $SD = 9.92$	$H = .600$ $U = 3.000$ $Z = .598$ *Median* = 28.521	$p = .439$ $p = .439$ $p = .867$ $p = 1$
non-impaired	27 M = 25.47 SD = 8.36	12 $M = 24.23$ $SD = 9.600$	15 $M = 26.46$ $SD = 7.41$	$H = 5.36$ $U = 75.000$ $Z = .861$ *Median* = 23.26	$p = .464$ $p = .464$ $p = .449$ $p = .704$

Table B.2.: FSS Fatigued vs. non-fatigued differences. Non-parametric hypotheses tests with dependent variable *Metric* (mean tapping frequency or mean handgrip strengths) and independent variable fatigue classification according to FSS. Test conducted with IBM SPSS Statistics Version 27. Kruskal-Wallis H^*, Mann-Whitney U^*, Kolmogorov-Smirnov Z^*, Median Test $Median^*$.

Metric	Case	N	FSS fatigued	FSS Non-fatigued	Test	$p < 0.05$
mean TF [taps/sec]	all	32 M = 8.19 SD = 1.42	16 $M = 7.50$ $SD = 1.04$	16 $M = 8.89$ $SD = 1.43$	$H = 9.091$ $U = 48.000$ $Z = 1.945$ $Median$ = 8.29	$\checkmark p = .003$ $\checkmark p = .003$ $\checkmark p = .001$ $\checkmark p = .005$
	male	14 M = 8.35 SD = 1.76	8 $M = 7.53$ $SD = 1.20$	6 $M = 9.45$ $SD = 1.89$	$H = 4.817$ $U = 7.000$ $Z = 1.543$ $Median$ = 8.47	$\checkmark p = .028$ $\checkmark p = .028$ $\checkmark p = .017$ $p = .103$
	female	18 M = 8.07 SD = 1.12	8 $M = 7.46$ $SD = 0.93$	10 $M = 8.56$ $SD = 1.04$	$H = 4.934$ $U = 15.00$ $Z = 1.467$ $Median$ = 8.22	$\checkmark p = .026$ $\checkmark p = .026$ $\checkmark p = .026$ $p = .152$
	impaired	7 M = 7.21 SD = 1.00	6 $M = 7.11$ $SD = 1.07$	1 $M = 7.746$	- - - -	- - - -
	non-impaired	25 M = 8.47 SD = 1.41	10 $M = 7.72$ $SD = 1.01$	15 $M = 8.97$ $SD = 1.44$	$H = 6.511$ $U = 29.000$ $Z = 1.715$ $Median$ = 8.632	$\checkmark p = .011$ $\checkmark p = .011$ $\checkmark p = .006$ $\checkmark p = .004$

mean HG [kg]	all	32 M = 25.89 SD = 8.41	16 $M = 24.77$ $SD = 8.88$	16 $M = 27.01$ $SD = 8.05$	$H = .513$ $U = 109.000$ $Z = .707$ *Median* = 23.988	$p = .474$ $p = .474$ $p = .699$ $p = .480$
	male	14 M = 32.41 SD = 7.61	8 $M = 30.26$ $SD = 8.49$	6 $M = 35.28$ $SD = 5.70$	$H = .600$ $U = 18.000$ $Z = .617$ *Median* = 33.22	$p = .439$ $p = .439$ $p = .841$ $p = 1$
	female	18 M = 20.81 SD = 4.75	8 $M = 19.27$ $SD = 5.27$	10 $M = 22.05$ $SD = 4.14$	$H = 1.334$ $U = 27.00$ $Z = .738$ *Median* = 21.48	$p = .248$ $p = .248$ $p = .648$ $p = .637$
	impaired	7 M = 27.91 SD = 8.40	6 $M = 28.46$ $SD = 9.06$	1 $M = 24.60$	- - - -	- - - -
	non-impaired	25 M = 25.33 SD = 8.50	10 $M = 22.55$ $SD = 8.44$	15 $M = 27.18$ $SD = 8.31$	$H = 1.772$ $U = 51.000$ $Z = .816$ *Median* = 23.26	$p = .183$ $p = .183$ $p = .518$ $p = .226$

Table B.3.: Female vs. male differences. Non-parametric hypotheses tests with dependent variable *Metric* (mean tapping frequency or mean handgrip strengths) and independent variable gender (male or female). Data set corresponding to FSMC motor fatigue classification. Test conducted with IBM SPSS Statistics Version 27. Kruskal-Wallis H^*, Mann-Whitney U^*, Kolmogorov-Smirnov Z^*, Median Test $Median^*$.

Metric	Case	N	Male	Female	Test	$p < 0.05$
mean TF [taps/sec]	all	34 $M = 8.31$ $SD = 1.47$	15 $M = 8.54$ $SD = 1.84$	19 $M = 8.13$ $SD = 1.12$	$H = .556$ $U = 121.000$ $Z = .894$ $Median = 8.35$	$p = .456$ $p = .456$ $p = .401$ $p = .300$
	fatigued	17 $M = 7.62$ $SD = 1.22$	9 $M = 7.96$ $SD = 1.41$	8 $M = 7.25$ $SD = 0.91$	$H = 1.815$ $U = 22.000$ $Z = 1.143$ $Median = 7.733$	$p = .178$ $p = .178$ $p = .146$ $p = .637$
	non-fatigued	17 $M = 8.99$ $SD = 1.41$	6 $M = 9.40$ $SD = 2.20$	11 $M = 8.76$ $SD = 0.78$	$H = 1.455$ $U = 21.000$ $Z = 1.31$ $Median = 1.455$	$p = .228$ $p = .228$ $p = .063$ $p = .228$
mean HG [kg]	all	34 $M = 25.97$ $SD = 8.30$	15 $M = 32.47$ $SD = 7.34$	19 $M = 20.84$ $SD = 4.61$	$H = 16.609$ $U = 25.000$ $Z = 2.509$ $Median = 23.98$	$\checkmark p = .000$ $\checkmark p = .000$ $\checkmark p = .000$ $\checkmark p = .000$
	fatigued	17 $M = 24.88$ $SD = 9.07$	9 $M = 30.50$ $SD = 8.00$	8 $M = 18.54$ $SD = 5.34$	$H = 7.259$ $U = 8.000$ $Z = 1.601$ $Median = 23.26$	$\checkmark p = .007$ $\checkmark p = .007$ $\checkmark p = .012$ $\checkmark p = .015$
	non-fatigued	17 $M = 27.06$ $SD = 7.56$	6 $M = 35.42$ $SD = 5.59$	11 $M = 22.51$ $SD = 3.33$	$H = 11.000$ $U = 0.000$ $Z = 1.970$ $Median = 24.602$	$\checkmark p = .001$ $\checkmark p = .001$ $\checkmark p = .001$ $\checkmark p = .002$

Table B.4.: Female vs. male differences. Non-parametric hypotheses tests with dependent variable *Metric* (mean tapping frequency or mean handgrip strengths) and independent variable gender (male or female). Data set corresponding to FSS fatigue classification. Test conducted with IBM SPSS Statistics Version 27. Kruskal-Wallis H^*, Mann-Whitney U^*, Kolmogorov-Smirnov Z^*, Median Test $Median^*$.

Metric	Case	N	Male	Female	Test	$p < 0.05$
mean TF [taps/sec]	all	32 $M = 8.19$ $SD = 1.42$	14 $M = 8.35$ $SD = 1.77$	18 $M = 8.07$ $SD = 1.12$	$H = .244$ $U = 113.000$ $Z = .735$ $Median = 8.29$	$p = .621$ $p = .6216$ $p = .653$ $p = .476$
	fatigued	16 $M = 7.50$ $SD = 1.04$	8 $M = 7.53$ $SD = 1.20$	8 $M = 7.46$ $SD = 0.93$	$H = .176$ $U = 28.000$ $Z = .750$ $Median = 7.65$	$p = .674$ $p = .674$ $p = .627$ $p = .1$
	non-fatigued	16 $M = 8.89$ $SD = 1.43$	6 $M = 9.45$ $SD = 1.89$	10 $M = 8.56$ $SD = 1.05$	$H = 2.647$ $U = 15.000$ $Z = 1.033$ $Median = 9.11$	$p = .104$ $p = .104$ $p = .236$ $p = .119$
mean HG [kg]	all	32 $M = 25.89$ $SD = 8.41$	14 $M = 32.41$ $SD = 7.62$	18 $M = 20.81$ $SD = 4.74$	$H = 15.013$ $U = 24.000$ $Z = 2.405$ $Median = 23.98$	$\checkmark p = .000$ $\checkmark p = .000$ $\checkmark p = .000$ $\checkmark p = .000$
	fatigued	16 $M = 24.77$ $SD = 8.89$	8 $M = 30.26$ $SD = 8.50$	8 $M = 19.27$ $SD = 5.27$	$H = 6.353$ $U = 8.000$ $Z = 1.5$ $Median = 22.84$	$\checkmark p = .012$ $\checkmark p = .012$ $\checkmark p = .022$ $p = .132$
	non-fatigued	16 $M = 27.02$ $SD = 8.05$	6 $M = 35.28$ $SD = 5.70$	10 $M = 22.06$ $SD = 4.14$	$H = 10.588$ $U = 0.000$ $Z = 1.936$ $Median = 26.078$	$\checkmark p = .001$ $\checkmark p = .001$ $\checkmark p = .001$ $\checkmark p = .007$

APPENDIX C

cFAST Supplementary Materials

C.1. ANCOVA Analysis

We conducted a one-way analysis of covariance (ANCOVA) to examine whether *correct* differs between the fatigue and non-fatigue groups when controlling for EDSS. First, we verified the test assumptions: The Shapiro-Wilk test indicates the data is normally distributed for the group with no fatigue $W(19)=.96$ ($P=.55$) and for the fatigued group $W(23)=.92$ ($P=.06$). Visual analysis with a scatter plot indicates similar regression slopes, and an F test shows no interaction between EDSS and fatigue groups (homogeneity of regression slopes) $F=(1,38)=.07$ ($P=.8$). Finally, Levene's Test confirms the homogeneity of variance with $F(1,40)=1.36$ ($P=.25$). ANCOVA analysis reveals that after controlling for EDSS (disability), there was no significant difference in fatigue on the *correct* score, $F(1,39)=2.36$ ($P=.13$). Estimated marginal means for no fatigue (M=104.48, SE=4.50) and fatigued (M=94.775, SE=4.06). EDSS is significantly related to correct ($F(1,39)=11.07$ $P=.002$). However, if we do not consider EDSS when analyzing fatigue then there is a difference in terms of correct between the groups $F(1,40)=7.84$, $P=.008$.

Table C.1.: Demographic Characteristics of Participants

		Not disabled	disabled	*P*
Number		23	29	
Age, mean (SD)		33.70 (8.93)	42.37 (13.77)	.03
Gender, n (%)				
	m	6 (26)	8 (42)	.44
	w	17 (74)	11 (58)	
MS type, n(%)				
	PMS	0 (5)	4 (21)	.04
	RRMS	23 (100)	15 (79)	
Disease duration, mean (SD)		9.61 (5.98)	13.16 (8.77)	0.14
DMT, n(%)				
	None	2 (9)	0 (0)	
	Interferon beta-1a	1 (4)	0 (0)	
	Dimethyl fumarate	3 (13)	0 (0)	
	Teriflunomide	1 (4)	1 (5)	
	Glatiramer acetate	1 (4)	1 (5)	
	Fingolimod	2 (9)	0 (0)	
	Natalizumab	7 (30)	7 (37)	
	Rituximab	0 (0)	4 (21)	
	Ocrelizumab	6 (26)	6 (32)	
Fatigue medication, n (%)				
	None	23 (100)	18 (95)	1.00
	Modafinil	0 (0)	1 (5)	
EDSS, mean (SD)		0.54 (0.67)	3.26 (1.54)	<.001
FSMC, mean (SD)				
	Total	39.78 (18.26)	60.53 (19.48)	.001
	cognitive	19.13 (9.55)	29.47 (10.39)	.002
	Motor	20.65 (9.22)	31.05 (10.28)	.002

Notes: Data are mean (SD) or n (%). PMS: progressive multiple sclerosis; RRMS: relapsing-remitting multiple sclerosis; Disease duration is measured in years since first manifestation; EDSS: expanded disability status scale; FSMC: Fatigue Score for motor functions and cognition; DMT: disease modifying therapy.

Table C.2.: Metrics comparison between fatigued and non fatigued patients with mean (SD), standard deviation, and Mann-Whitney U test (two-tailed) to assess whether there is a statistically significant difference between the groups.

	No fatigue ($n = 19$)	**Cognitive fatigue** ($n = 23$)	**U**	***P***
response time	2083.3 (358.31)	2586.88 (961.28)	316.0	.01
calibrated rate	3289.47 (1229.75)	3922.91 (1396.06)	298.0	.045
correct	109.11 (15.97)	90.96 (24.21)	119.5	.01
errors	7.58 (6.07)	8.04 (4.13)	243.5	.53
$\Delta correct$	3.51 (11.19)	-2.73 (9.95)	147.0	.07
$\Delta response\ time$	-0.96 (5.5)	2.69 (4.94)	298.0	.045
$\Delta errors$	-0.46 (2.05)	0.03 (1.86)	255.5	.35

Table C.3.: Metrics comparison between disabled and not disabled patients with mean (SD), standard deviation, and Mann-Whitney U test (two-tailed) to assess whether there is a statistically significant difference between the groups.

	Not disabled ($n = 23$)	**Disabled** ($n = 19$)	**U**	***P***
response time	2080.23 (317.39)	2696.61 (1030.37)	332.0	.004
calibrated rate	3211.22 (840.63)	4151.0 (1658.95)	312.0	.02
correct	108.3 (15.1)	88.11 (25.44)	105.5	.004
errors	7.96 (5.69)	7.68 (4.26)	220.0	.97
$\Delta correct$	0.55 (10.68)	-0.47 (11.35)	208	.80
$\Delta response\ time$	0.29 (5.55)	1.95 (5.33)	252.0	.40
$\Delta errors$	-0.03 (2.05)	-0.4 (1.83)	198.5	.61

C.2. User Interface Designs and Selection

We designed five user interfaces as prospects for our cognitive fatigability test. These are depicted in Figure C.1. We created two test modalities: single or grid modality. In the single modality, users have to map the middle symbol to its corresponding key, as displayed in the top mapping rule at the screen's top. While on the grid modality, users are presented with a four-by-three grid composed of six symbols and six keys. During each round, users have to map the elements within the grip following the mapping rule presented on top. The grid modality designs (Figure C.1 top row) were discarded after discussion with the neurologists and neuropsychologists. One of the main arguments against the grid design was that the added complexity would also result in more difficulties in making a fair comparison between the patients. The single modality design was further discussed with our specialist team and shown to patients attending the in-patient clinic. We presented printed and digital versions of each of the three single modality tests and asked for their preference in terms of style. After discussion with the specialist and informal feedback from the patients on the design, we opted for the SDMT style symbols as the other symbols were too similar, making the selection more complex and error-prone due to confusion.

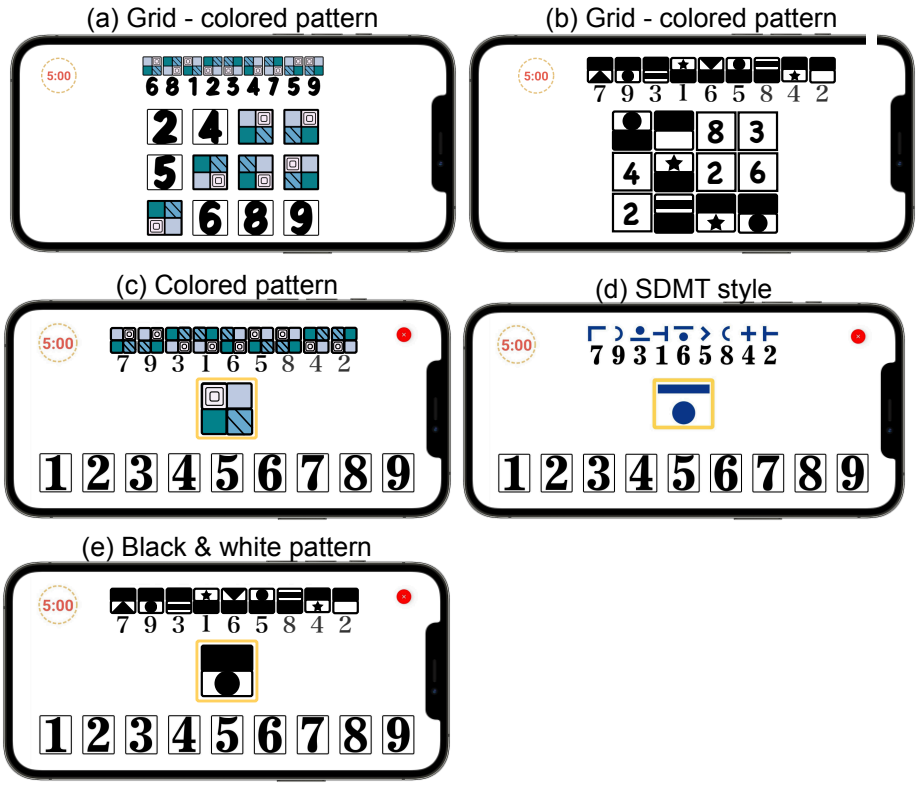

Figure C.1.: Different user interfaces were considered for cFAST. The top row depicts tests with a grid selection option, where users need to map each symbol within the grid with its corresponding key, as displayed in the mapping rule. The bottom rows depict the considered single selection interfaces.

APPENDIX D

PPG Sensor Validation Complete Tables

Table D.1.: Experiment II - Heart rate analysis per activity

Case	Activity	Size	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	Bias [95% LoA]
Everion Best (63592)	init	3808	83.09/82.83	14.03/13.94	.979 [+.978,+.981]	0.98	-0.26 [-5.83, +5.31]
	rest	21364	103.28/102.21	22.91/22.67	.988 [+.984,+.991]	0.99	-1.07 [-7.67,+5.53]
	bike (60 W)	9445	106.87/106.92	13.84/13.73	.989 [+.989,+.990]	0.99	+0.05 [-3.91, +4.01]
	bike (120 W)	8299	137.37/137.54	21.72/21.72	.995 [+.995,+.995]	1.00	+0.17 [-4.00, +4.34]
	walk	6722	104.00/103.90	16.57/16.54	.995 [+.995,+.995]	0.99	-0.10 [-3.36, +3.16]
	jog	8093	137.32/137.66	17.24/17.22	.973 [+.971,+.974]	0.97	+0.34 [-7.55, +8.23]
	run	5861	151.27/152.03	21.41/20.76	.974 [+.971,+.976]	0.97	+0.77 [-8.59, +10.12]
	avg		**117.60/117.59**	**18.25/18.08**	**0.985 [+.983,+.986]**	**0.99**	**-0.01 [-5.84,+5.81]**
Empatica Best (35705)	init	3852	81.63/81.64	11.56/12.83	.755 [+.741,+.769]	0.76	+0.01 [-16.74, +16.76]
	rest	20883	94.65/96.01	16.44/18.22	.834 [+.825,+.842]	0.84	+1.36 [-18.09, +20.82]
	bike (60 W)	4948	88.17/108.33	20.33/15.43	.031 [-.006,+.067]	0.05	+20.16 [-28.61, +68.94]
	bike (120 W)	4401	103.46/137.93	34.33/22.11	.118 [-.023,+.247]	0.22	+34.47 [-37.06, +106.00]
	walk	1287	104.02/106.43	16.46/18.03	.597 [+.558,+.634]	0.61	+2.41 [-27.74, +32.56]
	jog	233	107.95/140.09	22.33/22.17	.073 [-.044,+.195]	0.15	+32.14 [-24.77, +89.04]
	run	101	112.02/146.14	31.32/ 22.87	.120 [-.056,+.300]	0.22	+34.12 [-33.35, +101.59]
	avg		**98.84/116.65**	**21.83/18.81**	**0.361 [+.285,+.436]**	**0.41**	**17.81 [-26.62,+62.24]**

Table D.2.: Experiment I - Heart rate analysis per activity

Device	Activity	Size	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	Bias [95% LoA]
Empatica	init	6268	84.21/87.04	9.91/13.60	.464 [+.422,+.502]	0.50	+2.83 [-21.04, +26.69]
(79241)	rest	33676	92.94/100.15	15.87/18.61	.466 [+.349,+.558]	0.51	+7.21 [-26.46, +40.88]
	bike (60 W)	7856	90.66/108.61	21.23/12.88	.167 [-.002,+.314]	0.29	+17.95 [-24.07, +59.96]
	bike (120 W)	7968	104.54/137.18	36.99/20.56	.223 [-.027,+.422]	0.42	+32.64 [-33.95, +99.24]
	walk	8096	102.31/108.91	16.09/16.50	.433 [+.331,+.517]	0.47	+6.60 [-26.32, +39.52]
	jog	7690	102.98/138.73	26.28/18.29	.026 [-.015,+.068]	0.06	+35.76 [-25.11, +96.63]
	run	7687	102.85/152.94	30.08/19.74	.016 [-.014,+.046]	0.05	+50.08 [-18.77 ,+118.94]
	avg		**97.21/119.08**	**22.35/17.17**	**0.256 [+.149, +.346]**	**0.33**	**+21.87 [-25.10, +68.84]**
Everion	init	6268	87.71/87.04	14.26/13.60	.957 [+.953,+.960]	0.96	-0.67 [-8.59, +7.25]
(78821)	rest	33590	101.19/100.13	19.36/18.62	.972 [+.967,+.976]	0.97	-1.06 [-9.59, +7.47]
	bike (60 W)	7856	108.88/108.61	12.89/12.88	.981 [+.980,+.982]	0.98	-0.27 [-5.15, +4.61]
	bike (120 W)	7666	137.58/137.40	20.14/20.55	.988 [+.988,+.989]	0.99	-0.18 [-6.31, +5.95]
	walk	8096	109.14/108.91	16.61/16.50	.993 [+.993,+.993]	0.99	-0.23 [-4.05, +3.58]
	jog	7683	137.87/138.75	17.98/18.28	.965 [+.962,+.969]	0.97	+0.89 [-8.29, +10.07]
	run	7662	152.66/152.94	19.84/19.77	.952 [+.950,+.954]	0.95	+0.29 [-11.73, +12.30]
	avg		**119.29/119.11**	**17.30/17.17**	**0.972 [+.970,+.974]**	**0.97**	**-0.17 [-7.67, +7.32]**
Fitbit	init	3134	83.23/87.04	11.48/13.60	.767 [+.645,+.838]	0.82	+3.80 [-11.34, +18.95]
(39624)	rest	16812	98.71/100.14	19.81/18.62	.827 [+.821,+.832]	0.84	+1.43 [-20.21, +23.07]
	bike (60 W)	3928	99.02/108.61	16.56/12.88	.499 [+.170,+.682]	0.63	+9.59 [-16.21,+35.38]
	bike (120 W)	3984	118.37/137.18	27.97/20.56	.353 [+.079,+.542]	0.48	+18.81[-31.24, +68.87]
	walk	4048	103.48/108.90	12.73/16.50	.729 [+.541,+.824]	0.81	+5.43 [-13.62, +24.47]
	jog	3854	130.61/138.70	20.01/18.29	.703 [+.441,+.822]	0.78	+8.09 [-17.09, +33.27]
	run	3852	144.74/152.96	22.01/19.73	.782 [+.468,+.887]	0.86	+8.22 [-13.93, +30.38]
	avg		**111.17/119.08**	**18.63/17.17**	**0.665 [+.452, +.776]**	**0.74**	**+7.91 [-17.66, +33.48]**
Polar	init	3134	88.12/87.04	14.44/13.60	.959 [+.936,+.953]	0.95	-1.08 [-9.93, +7.77]
(39624)	rest	16812	101.60/100.14	19.45/18.62	.969 [+.959,+.976]	0.97	-1.46 [-10.27, +7.35]
	bike (60 W)	3928	108.73/108.61	13.30/12.88	.972 [+.970,+.973]	0.97	-0.13 [-6.24,+5.99]
	bike (120 W)	3984	136.90/137.18	21.23/20.56	.984 [+.983,+.985]	0.98	+0.28 [-7.10, +7.66]
	walk	4048	108.89/108.90	16.74/16.50	.989 [+.988,+.989]	0.99	+0.01 [-4.86, +4.89]
	jog	3854	137.51/138.70	19.17/18.29	.964 [+.957,+.969]	0.97	+1.19 [-8.41, +10.80]
	run	3852	152.27/152.96	21.09/19.73	.950 [+.947,+.953]	0.95	+0.69 [-11.89,+13.26]
	avg		**119.15/119.08**	**17.92/17.17**	**0.969 [+.963,+.971]**	**0.97**	**-0.07 [-8.38, +8.24]**
Wahoo	init	3094	84.91/86.97	13.77/13.51	.913 [+.875,+.936]	0.92	+2.06 [-8.41,+12.53]
(38492)	rest	16406	99.37/99.95	19.52/18.76	.965 [+.964,+.967]	0.97	+0.58 [-9.22,+10.38]
	bike (60 W)	3567	106.16/106.76	11.80/11.81	.971 [+.966,+.974]	0.97	+0.60 [-4.88, +6.08]
	bike (120 W)	3732	135.42/136.22	20.87/20.64	.984 [+.982,+.986]	0.99	+0.80 [-6.21, +7.81]
	walk	3975	107.80/108.94	17.00/16.63	.972 [+.964,+.977]	0.97	+1.14 [-6.36 ,+8.65]
	jog	3854	136.58/138.70	19.67/18.29	.939 [+.918,+.952]	0.95	+2.13 [-10.29, +14.54]
	run	3852	151.65/152.96	20.88/19.73	.937 [+.930,+.944]	0.94	+1.30 [-12.56, +15.18]
	avg		**117.41/118.64**	**17.64/17.05**	**0.954 [+.943,+.962]**	**0.96**	**+1.23 [-8.28, +10.74]**

Table D.3.: Experiment II - Heart rate variability analysis per activity for the Everion device.

Activity	Metric	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	R^2	Bias [95% LoA]
Init	RMSS	23.17/23.39	12.40/14.92	+.899 [+.742, +.967]	+0.91	+0.82	+0.22 [-12.22, +12.65]
μ len: 241 s	SDNN	58.49/60.62	16.71/17.54	+.876 [+.697, +.953]	+0.88	+0.75	+2.13 [-14.57, +18.84]
μ peaks: 318/314	PNN50	4.70/5.46	8.55/10.42	+.967 [+.912, +.988]	+0.99	+0.94	+0.76 [-3.99, +5.50]
# seg: 17	LF	1740.17/1801.81	1293.81/1265.94	+.982 [+.952, +.993]	+0.98	+0.96	+61.63 [-416.36, +539.63]
	HF	603.85/747.19	934.21/1333.10	+.918 [+.792, +.969]	+0.98	+0.87	+143.34 [-754.77, +1041.46]
	LF:HF	5.26/7.53	2.90/6.65	+.625 [+.215, +.846]	+0.92	+0.49	+2.27 [-5.85, +10.38]
	LFnu	80.86/81.10	8.90/14.16	+.745 [+.420, +.900]	+0.81	+0.63	+0.24 [-16.67, +17.15]
	HFnu	19.14/18.90	8.90/14.16	+.745 [+.420, +.900]	+0.81	+0.63	-0.24 [-17.15, +16.67]
Rest	RMSS	18.85/17.93	8.86/9.97	+.935 [+.904, +.956]	+0.95	+0.88	-0.91 [-7.35, +5.53]
μ len: 249 s	SDNN	58.11/58.21	20.89/21.95	+.982 [+.974, +.988]	+0.98	+0.97	+0.10 [-7.89, +8.09]
μ peaks: 363/357	PNN50	3.09/3.14	4.66/5.39	+.954 [+.933, +.968]	+0.96	+0.92	+0.04 [-2.97, +3.06]
# seg: 111	LF	1324.54/1361.74	1051.67/1177.61	+.946 [+.946, +.946]	+0.95	+0.89	+37.20 [-724.00, +798.40]
	HF	372.58/365.69	519.14/656.80	+.914 [+.877, +.940]	+0.94	+0.86	-6.90 [-489.31, +475.52]
	LF:HF	5.16/6.47	2.96/3.63	+.614 [+.413, +.744]	+0.67	+0.30	+1.31 [-4.05, +6.67]
	LFnu	79.79/83.05	10.32/8.96	+.701 [+.525, +.808]	+0.75	+0.26	+3.26 [-10.40, +16.93]
	HFnu	20.21/16.95	10.32/8.96	+.701 [+.525, +.808]	+0.75	+0.26	-3.26 [-16.93, +10.40]
Bike (60 W)	RMSS	10.82/8.13	4.28/2.78	-.018 [-.292, +.303]	-0.02	-3.42	-2.70 [-12.81, +7.41]
μ len: 294 s	SDNN	29.20/27.11	9.68/9.07	+.793 [+.596, +.899]	+0.81	+0.53	-2.10 [-13.50, +9.31]
μ peaks: 489/483	PNN50	0.28/0.14	0.56/0.33	+.394 [+.045, +.662]	+0.46	-1.44	-0.14 [-1.12, +0.85]
# seg: 28	LF	247.06/240.91	185.05/182.23	+.945 [+.886, +.974]	+0.94	+0.88	-6.15 [-126.78, +114.49]
	HF	93.06/52.20	101.05/37.16	+.053 [-.271, +.389]	+0.09	-8.16	-40.86 [-245.64, +163.93]
	LF:HF	3.15/5.73	1.15/2.93	-.012 [-.208, +.250]	-0.03	-0.98	+2.59 [-3.65, +8.82]
	LFnu	73.41/81.02	9.83/11.49	-.065 [-.345, +.266]	-0.08	-1.32	+7.61 [-23.17, +38.40]
	HFnu	26.59/18.98	9.83/11.49	-.065 [-.345, +.266]	-0.08	-1.32	-7.61 [-38.40, +23.17]
Bike (120 W)	RMSS	13.07/6.91	5.13/3.01	+.064 [-.161, +.427]	+0.15	-6.94	-6.15 [-17.04, +4.73]
μ len: 272 s	SDNN	54.11/49.63	20.39/16.76	+.835 [+.552, +.946]	+0.87	+0.55	-4.48 [-24.52, +15.57]
μ peaks: 541/533	PNN50	0.52/0.18	0.63/0.27	+.355 [-.112, +.728]	+0.59	-4.47	-0.34 [-1.36, +0.67]
# seg: 13	LF	100.62/42.73	96.74/39.70	+.266 [-.168, +.669]	+0.47	-5.93	-57.90 [-225.22, +109.43]
	HF	114.36/32.75	112.56/33.92	-.136 [-.464, +.347]	-0.34	-19.56	-81.61 [-332.89, +169.66]
	LF:HF	1.07/3.49	0.61/3.89	+.053 [-.316, +.507]	+0.23	-0.37	+2.43 [-5.03, +9.88]
	LFnu	48.23/62.76	12.94/25.16	+.031 [-.386, +.512]	+0.05	-0.58	+14.53 [-39.90, +68.96]
	HFnu	51.77/37.24	12.94/25.16	+.060 [-.386, +.512]	+0.05	-0.58	-14.53 [-68.96, +39.90]
Walk	RMSS	14.67/8.93	3.54/2.46	+.144 [-.091, +.478]	+0.42	-6.66	-5.74 [-12.31, +0.83]
μ len: 291 s	SDNN	33.36/29.83	9.17/8.28	+.755 [+.358, +.910]	+0.81	+0.38	-3.54 [-14.15, +7.07]
μ peaks: 444/446	PNN50	1.00/0.20	1.46/0.31	-.008 [-.361, +.412]	-0.03	-30.34	-0.79 [-3.74, +2.15]
# seg: 17	LF	414.73/337.55	195.72/115.77	+.417 [-.017, +732]	+0.51	-1.59	-77.18 [-407.37, +253.02]
	HF	208.90/50.43	181.55/24.16	+.046 [-.193, +.386]	+0.30	-97.74	-158.47 [-503.17, +186.23]
	LF:HF	3.17/8.13	2.09/3.90	+.238 [-.104, +.614]	+0.63	-1.32	+4.95 [-1.01, +10.92]
	LFnu	69.65/87.26	15.27/4.73	+.128 [-.111, +.455]	+0.49	-22.01	17.61 [-9.04, +44.26]
	HFnu	30.35/12.74	15.27/4.73	+.128 [-.111, +.455]	+0.49	-22.01	-17.61 [-44.26, +9.04]

APPENDIX E

Cronico User Interface

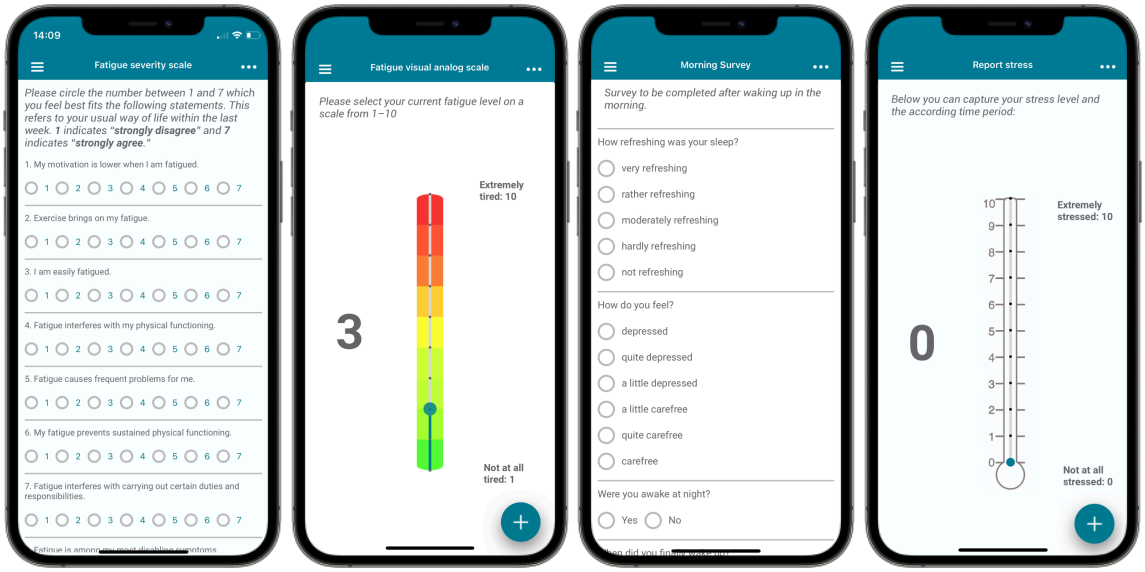

Figure E.1.: Some of cronicos' screenviews. From left to right fatigue severity scale, fatigue VAS, morning sleep protocol and stress VAS.