Enhancing the Accuracy of Disability Measurement Over Time in Multiple Sclerosis

Stanley Hum, B.Sc. (Biochemistry), M.Sc. (Biology)

School of Physical and Occupational Therapy
Faculty of Medicine
McGill University
Montreal, Quebec, Canada

September 2015

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Doctor of Philosophy (Rehabilitation Sciences)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADL	Activities of daily living
ARR	Annualized relapse rate
BIC	Bayesian information criteria
CNS	Central nervous system
CI	Confidence interval
CIS	Clinically isolated syndrome
ClinRO	Clinician reported outcome
CPC	Category probability curves
DASH	Disabilities of the Arm, Shoulder, and Hand
DIF	Differential Item Functioning
DMT	Disease modifying treatment
DSS	Disability Status Scale
EDSS	Expanded Disability Status Scale
EF	Environmental factor
ESS	Environmental Status Scale
FDA	The Food and Drug Administration
FMM	Finite mixture model
FS	Functional System
FSS	Functional System Score
GBTM	Group based trajectory modeling
Gd	Gadolinium
GEE	Generalized estimating equations
GHP	General health perception
HRQL	Health related quality of life
ICC	Item Characteristic Curve
ICF	International Classification and Functioning, Disability and Health
ICIDH	The International Classification of Impairments, Disabilities and Handicaps
ISS	Incapacity Status Scale
KM	Kaplan-Meier

LID	Local Item Dependency
MAR	Missing at random
MCAR	Missing completely at random
MCIC	Minimum clinically important change
ML	Maximum likelihood
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
MSFC	Multiple Sclerosis Functional Composite
MS-PDM	MS Physical Disability Measure
PASAT	Paced Auditory Serial Addition Test
PBMSI-V1	Preference Based MS Index-V1
PerfO	Performance-based outcome
PF	Personal factor
PP	Posterior probabilities
PPMS	Primary progressive multiple sclerosis
PRO	Patient reported outcome
PSI	Person Separation Index
QOL	Quality of life
RCT	Randomized control trial
RRMS	Relapsing remitting multiple sclerosis
SLCMSR	Sylvia Lawry Centre for MS Research
SPMS	Secondary progressive multiple sclerosis
SRO	Self reported outcomes
TPC	Threshold probability curves
WC	Wilson and Cleary model
WHO	World Health Organization

ABSTRACT

Multiple Sclerosis (MS) has a clinically heterogeneous disease course and, depending on the location of the lesions, different disabilities may manifest. This leads to challenges on how best to measure MS disability and track change over time. The Expanded Disability Status Scale (EDSS) is the standard measure of MS disability widely used to study the natural history of MS and to estimate the long-term effectiveness of disease modifying treatments (DMTs). The EDSS has been criticized for well known psychometric limitations, a narrow representation of disability, and an over emphasis on ambulation. There is a need for a mathematically sound measure of the full range of physical disability in MS.

The global aim of the research carried out for this doctoral thesis is to contribute evidence towards an optimal measurement approach to quantifying disability over time. The objectives were to: i) identify longitudinal patterns of MS progression using the current standard of disability assessment, the EDSS, and estimate the extent to which annualized relapse rate (ARR) contributes to disease course; ii) estimate the extent to which the EDSS is equivalently scored across experts; and iii) develop a prototype measure of MS physical disability.

Data for the longitudinal analysis of disease progression and the Rasch analysis of the FSS/EDSS are from the MS clinic database at the Montreal Neurological Institute/Hospital. To develop the prototype measure an existing dataset was used comprising of patient reported outcomes (PROs), performance based outcomes (PerfOs), and clinician reported outcomes (ClinROs) (271 items with questionnaires filled out at 2 time points) for 189 patients randomly selected from the three MS clinic in the Montreal area with disease onset post 1994.

Group Based Trajectory Modeling was used to identify trajectories of MS disease progression for women and men from three eras (manuscript 1). Three inception cohorts from distinct MS onset and treatment periods: (1) pre-magnetic resonance imaging (MRI) and DMTs (<1995); (2) MRI+1st generation DMTs (1995-2004); and (3) MRI+2nd generation DMTs (2005-present) were used to estimate the extent of heterogeneity in disease course over time. Results showed variability in disease course of MS patients over three different critical time periods. A majority of patients in the two post-1995 cohorts remain at their initial disability level within the

study observation periods. Having a higher ARR with reference to the lowest disability trajectory group increased the odds of being in a higher (worst) disability trajectory in the post-1995 cohorts.

Rasch analysis was used to test the extent to which the functional system scores (FSS) and EDSS items measure a single construct and fit an underlying theoretical hierarchy that forms a linear continuum (i.e. a "ruler") with interval-like units (manuscript 2). Rasch analysis was used as a statistical method to ensure data harmonization. Results show that the FSS support a unidimensional construct of MS disability measured on a linear interval scale. Several items had differential item functioning (DIF) by neurologist. Rasch analysis can identify and adjust for DIF to ensure data harmonization when pooling data from multiple sources.

A prototype measure was developed using Rasch analysis to integrate items from PROs, PerfOs, and ClinROs to form a single measure of physical disability in MS (manuscript 3). Results show that PROs, PerfOs and ClinROs health indices can co-exist in a linear hierarchical unidimensional measure of physical disability. However, these items were not evenly distributed along the linear continuum. Consistently, the less difficult items were PROs and located left of the continuum and PerfOs were more difficult items and located on the right of the continuum.

ABRÉGÉ

La sclérose en plaques (SP) est une maladie à évolution cliniquement hétérogène dans laquelle différentes incapacités (atteintes) peuvent survenir en fonction de la localisation des lésions. Cela engendre des défis quant à la meilleure façon d'évaluer les incapacités reliées à la SP et de mesurer le changement à travers le temps. L'échelle d'incapacités EDSS (Expanded Disability Status Scale) est la mesure couramment utilisée pour étudier l'histoire naturelle de la sclérose en plaques et pour évaluer l'efficacité à long-terme des médicaments modificateurs de l'évolution de la SP (MMÉSP). L'échelle EDSS a été critiquée pour ses limitations psychométriques bien connues, sa représentation étroite des incapacités, et parce qu'elle accorde trop d'importance à l'ambulation. Il existe donc un besoin pour une mesure mathématiquement saine de la totalité des incapacités dues à la SP.

Le but global de la recherche effectuée dans le cadre de cette thèse de doctorat est d'apporter des éléments de preuves vers une méthode optimale de mesure quantitative des incapacités dues à la SP dans le temps. Les objectifs étaient de i) identifier les configurations longitudinales de la progression de la maladie en utilisant la norme actuelle de la mesure de l'invalidité, l'EDSS, et d'estimer la mesure dans laquelle le taux annualisé de poussées contribue à évolution de la maladie; ii) évaluer dans quelle mesure l'EDSS est utilisée de manière équivalente par les experts; et iii) développer un prototype d'une mesure des incapacités dues à la SP.

Les données pour l'analyse longitudinale de la progression de la maladie et de l'analyse de Rasch des FSS/EDSS proviennent de la base de données de la clinique SP à l'Institut et hôpital neurologiques de Montréal. Afin de développer le prototype de la mesure, des données pré-existantes ont été utilisées. Il s'agissait de résultats rapportés par les patients (Patient Reported Outcomes – PROs), de résultats basés sur la performance (PerfOs), et des résultats fournis par les cliniciens (ClinROs) (271 items; les questionnaires remplis à 2 moments dans le temps) de 189 patients choisis au hasard parmi trois cliniques SP dans la région de Montréal et dont la maladie a débuté après 1994.

La technique de Group Based Trajectory Modeling a été utilisée afin d'identifier les trajectoires de progression de la maladie pour les femmes et les hommes sur trois époques (manuscrit 1). Trois cohortes de départ (inception cohorts) avec début de la maladie et périodes de traitement distinctes: (1) Pré-imagerie par résonance magnétique (IRM) et MMÉSP (<1995); (2) IRM + MMÉSP de 1ère génération (1995-2004); et (3) IRM + MMÉSP de 2ème génération (2005-aujourd'hui) ont été utilisés pour estimer le degré d'hétérogénéité dans l'évolution de la maladie au fil du temps. Les résultats ont démontré de la variabilité dans cette évolution chez les patients atteints de SEP sur trois différentes périodes critiques. La majorité des patients des deux cohortes post-1995 ont conservé leur niveau initial d'invalidité durant les périodes d'observation de l'étude. Les patients avec un taux annualisé de poussées (ARR) supérieur par rapport au groupe avec la trajectoire d'incapacité la plus faible avaient plus de chances de se retrouver dans un des groupes ayant la trajectoire d'incapacité la plus élevée dans les cohortes post-1995.

La technique d'analyse de Rasch a été utilisée pour évaluer la mesure dans laquelle les scores fonctionnels (FSS) et l'EDSS mesurent un concept unique et correspondent à une hiérarchie théorique sous-jacente qui forme un continuum linéaire (à savoir une "règle") avec des unités ressemblant à des intervalles (manuscrit 2). La technique d'analyse de Rasch a été utilisée en tant que méthode statistique pour assurer l'harmonisation des données. Les résultats montrent que les FSS soutiennent un concept unidimensionnel d'incapacité dues à la SP mesuré sur une échelle linéaire à intervalles. Plusieurs éléments avaient un fonctionnement différentiel (DIF) selon le neurologue. L'analyse de Rasch peut repérer et ajuster pour les items avec fonctionnement différentiel pour assurer l'harmonisation des données lors de la mise en commun des données provenant de plusieurs sources.

Un prototype de mesure a été développé en utilisant une analyse de Rasch pour intégrer les articles de PROs, PerfOs, et ClinROs afin d'obtenir une seule mesure d'incapacités physiques pour la SEP (manuscrit 3). Les résultats démontrent que les indices de santé PROs, PerfOs et ClinROs peuvent coexister dans une mesure unidimensionnelle hiérarchique linéaire d'incapacité. Cependant, ces éléments ne sont pas distribués uniformément le long du continuum linéaire. De

façon constante, les items les moins difficiles étaient les PROs, situés à gauche du continuum, tandis que les PerfOs étaient les articles les plus difficiles et se situaient à la droite du continuum.

ACKNOWLEDGEMENTS

First, many thanks to my supervisor, Dr. Nancy Mayo, for her guidance, advice, and support throughout my PhD. I am extremely grateful for her tireless enthusiasm and skilled tutelage. She shared her expertise on all aspects of this thesis from research methodology and analysis, to writing. Her curiosity in research and in life is infectious and made my PhD experience extremely rewarding.

I would like to thank the members of my committee, Drs. Pierre Duquette and Johanne Higgins for their assistance in developing the protocol for my thesis. I am grateful for their professional input and willingness to participate in my PhD.

I am grateful to Dr. Yves Lapierre for his generosity of spirit. He is always willing to share his clinical expertise, his limited time, and has provided me with tremendous support over the years. He is a rare breed, a dedicated leader in his field with the work ethic of a country doctor from a bygone era. I have enjoyed our conversations both related and unrelated to research over the years.

I would like to thank Dr. Lois Finch for her advice during my thesis. She provided invaluable knowledge on Rasch analysis and feedback on my protocol and to manuscript 2. I also thank Susan Scott for her statistical expertise and her willingness to share her knowledge and always answering my questions.

I thank Drs. Jack Antel and Pierre Duquette for their professional support, valuable feedback, and advice. I am grateful to Anne Marie Bismuth for her professional support, organizational skills, willingness to listen to my research ideas, and most importantly her friendship.

I would like to thank Dr. Lesley Fellows for her important contributions to manuscript 2 providing insight in research, interpretation, and context to the findings.

I thank Dr. Robert Brown for reviewing my work and his valuable comments and providing alternative perspectives to my research. I have enjoyed our debates on research with Drs Robert Brown and Dave Rudko on our many lunches and trips to Else's.

Thanks to Sabrina Figueiredo and Carolina Moriello for being great office mates. I appreciated your thoughtful feedback. I would like to thank the rest of the members of the lab, Vanessa Bouchard, Ayse Kuspinar, Marie-Eve Letellier, Christiane Lourenco, and Lyne Nadeau for always willing to help. To a previous member of my lab, Dr. Miho Asano, I thank her for encouraging me to continue with my research interests. I would also like to thank Pamela Ng for her help as a fellow student with Dr. Mayo and as a member of the MS clinic.

I would also like to thank Ami, Avi, Alison, Christine, Anna, and Vinod for their encouragement. I enjoyed my discussions with Uku concerning different approaches in research methodology.

I would like to acknowledge Christine Déry and Catherine Bigras for editing support. I would also like to acknowledge Catherine Bigras, Elaine Roger, and Dr. Yves Lapierre for translating my abstract to French. Thanks to Catherine Bigras for keeping things running.

Most importantly, I am sincerely grateful for the unconditional support from my Mom and Dad, and sister Carol throughout my life. They never questioned my reasons for returning to school but simply supported my decision.

Finally, I would like to acknowledge the financial assistance from the Canadian Institute of Health Research Neuroinflammation Training Program, The Edith Strauss Grant for Knowledge Translation, The Richard and Edith Strauss Doctoral Fellowship, and Fonds de recherche santé Québec.

One of the many famous quotes I've heard from Dr. Mayo during my PhD that resonates with me is“…when you see a fork in the road take it..." - Yogi Berra.

P.S. The NHL has the Stanley Cup; I have a one of a kind handcrafted mug by Mark and Nancy.

Thanks everyone

PREFACE

Statement of originality

The novel contributions from this thesis to the research knowledgebase of MS are the following: i) Quantifying the variation in MS disease course over time (Manuscript 1); ii) Variability in scoring the FSS/EDSS by neurologist impact pooling of data from multiple sources (Manuscript 2); and iii) Development of a prototype physical disability measure to comprehensively assess MS disability (Manuscript 3)

As a member of the MS clinic at the Montreal Neurological Institute and Hospital (MNI/H), my responsibilities included coordinating MS research projects and managing the longitudinal MS database. During that time working on Dr. Nancy Mayo's "Gender and Life Impact of Multiple Sclerosis Study", introduced me to methodologies and research approaches that were highly relevant to data recorded in the clinic's longitudinal database.

As a repository of demographic and clinical variables related to MS, the database has the potential to be a valuable tool to answer questions on the disease course of MS. There was a need for a method to better describe the variability in trajectories of disease course over time. It became evident to me that group based trajectory modeling (GBTM), a statistical approach not previously used in MS research, could be used to describe with increasing detail the heterogeneity in disease course of the MS subtype. Using the gold standard to assess MS disability, the Expanded Disability Status Scale (EDSS), from a longitudinal MS clinic database that had not been previously described, I was able to use GBTM to provide evidence of the variability in terms of stable and unstable trajectories of disease course across and *within* MS subtypes over time.

Like others, I questioned whether the EDSS should or could continue to be used as the gold standard to assess MS disability. Currently, the EDSS is included in the majority, if not all MS registries, including the database at the (MNI/H). It is currently assumed that neurologists trained to score the EDSS are doing so equivalently, and that data from multiple sources can be pooled. I tested this assumption using a modern psychometric methodology, Rasch analysis. By utilizing the expectations of unidimensionality and

invariance of the Rasch model I was able to apply Rasch analysis in an atypical manner as a data harmonization methodology to ensure comparability of existing data. I applied this methodology to test whether neurologists were using the EDSS in the same way, and in doing so I was able to identify and adjust for differences in scoring of the EDSS not previously quantified in this manner. If the research community is to continue to use the EDSS, studies must assess whether data generated from different sources are equivalent, or can be made equivalent, to ensure that data can be appropriately pooled for analysis.

Harmonizing EDSS data cannot overcome the inherent limitation of the EDSS, a clinician reported outcome, in only assessing a narrow representation of the spectrum of disability domains in MS. To assess disability beyond impairment from multiple perspectives a different approach is required. There exist many outcome measures assessing different domains of disability from the perspective of the patients, clinician, or from performance based tests. I applied Rasch analysis to an existing dataset of commonly used health indices used to assess a randomly selected sample of MS patients, to create a prototype measure of MS physical disability integrating performance based tasks, clinician reported outcomes, and patient reported outcomes. The new measure will have strong psychometric properties and able to comprehensively assess MS physical disability

Contribution of authors

Stanley Hum developed the design, analyzed, interpreted, and wrote all three manuscripts under the supervision of Dr. Nancy Mayo. Dr. Mayo provided expertise for all aspects of the research methodology and analysis of this thesis. Dr. Yves Lapierre is the Director of the MS clinic and provided full access to the database for manuscripts 1 and 2 and critically reviewed the manuscripts and provided important clinical interpretation of the study results. Susan C. Scott provided statistical analysis support for manuscript 1. Lois Finch provided analysis support, interpretation of results, and critically reviewed manuscript 2. Lesley Fellows provided extensive feedback, editing, interpretation of the results, and critically reviewed manuscript 2. Pierre Duquette critically reviewed critically reviewed manuscript 1 and provided invaluable interpretation of clinically relevance of the study results.

Organization of thesis

The thesis consists of three manuscripts. In accordance with the requirements of the Graduate and Postdoctoral Studies (GPS) additional chapters have been included in this thesis. An introduction and conclusion independent of the manuscripts have been included as required by the GPS. Due to the manuscript format it must be conceded that some duplication was inevitable in this thesis.

A brief outline of the thesis follows:

Chapter 1 is an introduction of an overview of MS including definitions of disease subtype and course. Current measures of disease activity, the consequence of relapses, and available treatments are summarized.

Chapter 2 provides an overview on how MS care and research are conceptualized from the perspective of the two predominant models related to MS care and research: The biomedical model and the biopsychosocial models (the International Classification of Functioning, Disability and Health (ICF) and Wilson-Cleary models).

Chapter 3 provides a review on how relapses have been measured over the decades in MS. Due to the large amount of literature on MS relapses the general approach of a scoping review was used to summarize the broad topics related to research in this area.

Chapter 4 presents an overview of MS disability from the perspective of the biomedical and biopsychosocial model. The biomedical model represents disability primary from the perspective of neurologists managing MS. The biopsychosocial model is the predominant model for rehabilitation.

Chapter 5 provides a brief overview of statistical methods applied in MS disability. A brief description of group based trajectory modeling is summarized.

Chapter 6 consists of manuscript 1 entitled " Trajectory of MS disease course for men and women over three eras". This study illustrates the use of group based trajectory modeling

to describe MS disease course over three time periods. This study was able to describe the variability in the trajectories of disease progression across and within MS subtypes.

Chapter 7 provides an overview of Rasch analysis. An algorithm of the typical steps of Rasch analysis using in manuscripts 2 and 3 is presented.

Chapter 8 Linking chapter 7 to (chapter 9) manuscript 2

Chapter 9 consists of manuscript 2 entitled "Rasch analysis as a method of data harmonization to resolve cross-rater variability in scoring the function system scores of the Expanded Disability Status Scale". This study provides evidence that the standard measure of MS disability is not scored equivalently across neurologist. A method of data harmonization using Rasch analysis was presented to identify scoring bias and adjusting for it.

Chapter 10 links (chapter 8) manuscript 2 to (chapter 9) manuscript 3

Chapter 11 consists of manuscript 3 entitled "Integrating patient reported outcomes, performance-based tasks, and clinician reported outcomes to produce a linear hierarchical unidimensional measure of physical disability in MS". This study presents the development of a prototype measure of physical disability using Rasch analysis to integrate patient reported outcomes, performance-based tasks, and clinician reported outcomes into a single measure.

Chapter 12 is the conclusion summarizing the findings and conclusions of the manuscripts and the overall thesis.

Corresponding figures and tables are presented at the end of each chapter/manuscript. Corresponding references and appendices were presented at the end of each manuscript. For the additional chapters, all references and appendices were presented at the end of the thesis.

Chapter 1

Introduction, rationale, and objectives

Introduction

Multiple Sclerosis (MS) is the most common neurological disease among young adults.[1,2] According to Statistics Canada's estimates from the years 2010/2011 there are over 90 000 Canadians with MS. [3] The overall Canadian MS prevalence is one of the highest in the world at 240/100 000 with a regional estimate of Quebec at 180/100000.[4] MS evolves over 30 to 40 years with an average age of onset in the mid to late 20s, the prime productive years, affecting women 3 times more than men.[5-8] Its etiology is unknown.[9]

Disease activity is presumed to be a T-cell mediated autoimmune inflammatory response targeting the central nervous system (CNS) causing inflammation and demyelination of axons and eventual neuronal damage.[10,11] This chronic disease has a long and variable clinical course with periods of quiescence and exacerbation.[1,12,13] In 1996, formal classifications were established. The MS phenotypes defined were relapsing-remitting (RRMS), secondary progressive (SPMS), primary progressive (PPMS), and progressive relapsing (PRMS). Disease is progressive from onset for the last two types.[13] The first relapse experienced by the patient defines disease onset. Most recently, the classifications were updated. Additional descriptors of more active and less active disease are included in classifying MS phenotype. PRMS is now included as part of PPMS as a more active form of MS. Clinically isolated syndrome (CIS) as a clinical descriptor of the first clinical presentation of signs that could lead to MS was also included. This is discussed further below. The most common and treatable type is RRMS affecting ~80% of patients.[6,9,13,14] An exacerbation or relapse is the most observable feature in RRMS patients. The resultant MS related neurological dysfunction is an indicator of disease activity. Symptom onset of relapses evolves over days, stabilizes, and often improves spontaneously or with corticosteroids.[6] Most relapses are followed by complete recovery but some people are left with residual deficits.[15] When neurological dysfunction persists after subsequent relapses and there is sustained worsening of disease, the patient is said to have transitioned from RRMS to SPMS.[16]

Definitions: Lacking a specific diagnostic test, guidelines for the diagnosis of MS are based predominantly on clinically observed disease activity (relapses), clinically determined lesion location and time. Poser et al. published the Poser diagnostic criteria for MS in 1983 that served as the standard for 2 decades.[17] They defined several diagnostic categories of MS, MS relapses, and clinical/paraclinical evidence for a CNS lesion[17]. Conclusive diagnosis still required a long clinical follow up.[18]

McDonald et al. developed a set of diagnostic criteria in 2001[19] to include new information and technology. It was revised in 2005[20] and in 2010[21] to allow magnetic resonance imaging (MRI) detected lesions to take a more prominent role in establishing a diagnosis. Included in these criteria is a new category, Clinically Isolated Syndrome (CIS) that can be assigned after the first demyelinating event (relapse). CIS patients meeting specific MRI criteria can obtain a diagnosis of MS. The McDonald criteria allow for a faster diagnosis leading to earlier treatment.[2,18]

Within all criteria are the definitions of a relapse. A relapse is defined generally as the development of new neurological symptom(s) or the worsening of existing symptom(s) lasting at least 24 hours in the absence of fever in those who had been neurologically stable or improving for the previous 30 days.[17,19-22]

Disease activity: Disease activity in MS is based on annualized relapse rate (ARR), MRI activity, and disability (most commonly measured by the Expanded Disability Status Scale [EDSS] which is performed by a neurologist).[2,23,24]

1) ARR estimate: It appears that estimates of ARR are variable.[25] Lhermitte et al., state in 1973 that ARR was 0.5 in the main reported studies of the time but did not provide a reference.[26] The estimated ARR in the placebo groups of the pivotal clinical trials of disease modifying therapies (DMTs) is approximately 1.0[27-30], whereas natural history data range from 0.4 to 1.0 relapse rate per year.[31] Others stated the natural history of relapse events averaged 1.1/year earlier in the disease and appears to decrease with advancing disease.[32]

2) MRI: MRI technology is the most promising biomarker for MS. Whereas MRI has been helpful in diagnosing MS, relapses are still defined clinically. MRI of the CNS allows for the

identification and measure of MS lesions in an objective and quantitative manner.[12] Conventional T2-weighted imaging (WI) can identify the number and volume of clinically silent lesions whereas a Gadolinium (Gd)-enhancing T1-WI lesion is believed to depict immune cells migrating across the blood-brain barrier to cause active MS inflammation.[33]

3) Disability (EDSS): The EDSS has 20 grades of impairment ranging from 0 (normal) to 10 (death due to MS) with 0.5 increments after 1.0. Scores from 0 to 4 are estimated using a rubric to compute results from 8 functional systems (FS) on patients that are fully ambulatory (able to walk 500m). FS scores ranges from 0 to 5 or 6. EDSS scores > 4.0 are based solely on ambulation status.[34]

Consequences of relapses: Relapses early in the course of the disease appear to be associated with earlier disease progression.[35-37] Confavreux et al., found that early relapse rate influenced disease progression but only until EDSS four.[38] Another study found that early relapses impacted disease progression in the short term but had no long-term impact (> 10 years or if already in secondary progressive phase),[39] whereas two other studies found no association between relapses and disease progression.[1,40] Despite the ambiguity in the literature, the impact at the patient level is very disruptive on the physical, social, financial and psychological wellbeing of people with MS.[41,42]

Disease management: Currently there are 10 disease modifying therapies (DMTs) approved in Canada. They have been shown to decrease the ARR in patients with RRMS or SPMS experiencing relapses.[43,44] For long-term management of MS, interferon beta-1a,[28,29] interferon beta-1b[27] and glatiramer acetate[30] are the first line drugs approved in Canada. It is generally believed that treating early can slow or prevent worsening of disability if initiated prior to the onset of more permanent CNS damage.[45] A summary of four clinical trials treating CIS patients showed that DMTs were able to delay onset of clinically definite MS.[46]

Natalizumab is a monthly intravenous infusion shown to reduce relapse rates by 68%.[47] However, it has been associated with a risk of a serious adverse event such as progressive multifocal leukoencephalopathy (PML).[48] Currently, there are three oral DMTs are approved in Canada; Fingolimod,[49,50] Teriflunomide,[51,52] and Dimethyl Fumarate.[53,54] Each of these DMT has a different associated safety profile with less long-term data on effectiveness and

safety. Most recently approved, Alemtuzumab is infused on five consecutive days and then a booster is given a year later.[55,56] One of the major concerns with Alemtuzumab is developing a secondary autoimmune disorder.[44] All DMTs were approved on their efficacy in reducing relapse frequency. The number of DMTs available is encouraging and will give patients more treatment options to meet their medical and personal needs. Unfortunately the newer DMTs are generally more costly.[57]

The multiple advances in MS treatment and disease management have led some to comment that MS patients can experience "remission" from disease activity.[58] Although encouraging, there is still a need for comprehensive rehabilitation intervention to minimize sequelae and symptoms from MS that can cause body function impairments, activity limitations, and participation restrictions.[59] A meta-analysis showed that the level of physical activity in MS populations were lower than in non-MS populations.[60]

The principal goal of rehabilitation is to maximize patient autonomy and quality of life. To achieve this exercise intervention is an important part of the rehabilitation process.[59] Exercise has been shown to have a positive impact on physical and psychosocial functioning and on quality of life but there is a lack of evidence on what should be prescribed to MS patients due to the broad range of exercise interventions.[61,62] In a recently published systematic literature search on exercise in MS, it was revealed that exercise studies using the EDSS as an outcome generally found no benefit in the intervention. The reviewer suggested that the EDSS is not an appropriate outcome for exercise interventions due to its poor responsiveness in combination with the short duration and small sample size of most exercise studies. The EDSS was also criticized for its psychometric properties and over emphasis on ambulation.[63] A Cochrane review suggests that there is an urgent need for a core set of outcomes for exercise studies.[64]

To guide the development of a new measure, the accepted conceptual framework for disability in rehabilitation, the World Health Organization's International Classification and Functioning (ICF), can be used to provide content validity for a disability measure.[65] Currently to comprehensive measure (physical) disability, multiple indices are needed to capture the spectrum from impairments and activity limitations.

To ensure items in a measure all belong to a single construct (such as global physical disability), Rasch analysis is a statistical methodology used to produce a unidimensional measure where items (difficulty) and people (ability) are organized hierarchically on the same linear logit scale.[66,67] Rasch analysis transforms ordinal scales (Likert scales) to interval-like scales using a logit transformation.[68] The ICF provides the conceptual framework to guide item selection whereas Rasch analysis is the methodology to develop new measures. It has also been used to combining items from different indices into a single measure.[69,70]

This brief introduction sets the stage for the importance of accurate measurement of MS disability and indicates that there are frameworks and statistical tools available to this end.

Rationale of the thesis

MS is a challenge to study longitudinally because it has a long and variable clinical course with periods of quiescence and exacerbation.[1,12,13] It causes a wide range of disability impacting physical, psychological, and social functions. The main measure used by neurologist to assess MS disability is the EDSS. The EDSS is the outcome used to study the natural history of MS and to estimate the long-term effectiveness of DMTs. The EDSS has well known psychometric limitations and has been criticized as to be heavily weighted on ambulation and not assessing important components of disability such as cognition and upper limb function.[71] There is a need for a more comprehensive measure of MS (physical) disability that includes the patient perspective. The new measure developed with modern psychometric methodology will measure a single construct organized in a hierarchy by item difficulty and person ability on the same linear scale. It will have an interval-like score that can be mathematically manipulated.

Objectives

The global aim of the research for this doctoral thesis is to contribute evidence towards an optimal measurement approach to quantifying disability over time. Three integrated analyses were carried out to address current challenges in measuring disability over time in MS. Each of these analyses is presented as manuscripts in this thesis.

The specific objectives were to:

I. identify the longitudinal patterns of MS progression using the EDSS, the current standard of disability assessment

Manuscript 1: Trajectory of MS disease course for men and women over three eras

II. estimate the extent to which the EDSS, the current standard disability assessment, is equivalently scored across raters

Manuscript 2: Rasch analysis as a method of data harmonization to resolve cross-rater variability in scoring the function system scores of the Expanded Disability Status Scale

III. Develop a prototype measure of MS physical disability

Manuscript 3: Integrating patient reported outcomes, performance-based tasks, and clinician reported outcomes to produce a linear hierarchical unidimensional measure of physical disability in MS

Chapter 2

Models for MS care and research: The biomedical model and the biopsychosocial model.

MS care and research can be viewed from the perspective of two different models, the biomedical model and biopsychosocial models. The concepts underlying each model point to different measurement and treatment approaches. This chapter will review the two predominant models: biomedical model and biopsychosocial models.

Biomedical model of medicine (disease):

Lhermitte's statement in 1933 is still apropos; "Medicine is not a contemplative science; its aim is to discover the immediate causes of diseases and to treat their manifestations".[72] Medicine is largely based on a biomedical model. Engel cites Ludwig's description of the medical model, with respect to psychiatry and medicine in general, as a philosophy that focuses on patients' signs and symptoms resulting in sufficient deviation from normal representing disease.[73] The disease is due to known or unknown natural causes and elimination of these causes will result in curing or improvement of the patient.[73,74] Within the context of this model, the physician's responsibility would pertain to differential diagnosis based on specific symptoms and signs, laboratory tests, knowledge of the natural history and prognosis of the condition, choice of therapeutic environment, and selection of therapy.[74]

Engel states that disease is viewed primarily within the confines of the biomedical model and is firmly based on biological sciences.[73] Wade et al., cites Virchow's cell theory that all diseases result from cellular abnormalities.[75] McCollum and Pincus, state that the causes, diagnosis, prognosis, treatment, and outcomes of diseases are determined largely by physical or somatic variables.[76] The underlying belief systems of the medical model are: 1) that all illnesses, symptoms and signs are from an underlying abnormality within the body (usually in the functioning or structure of specific organs) and referred to as a disease; 2) all disease gives rise to symptoms; 3) health is the absence of disease; 4) mental phenomena, such as emotional disturbance or delusions, are separate from and unrelated to other

disturbances of bodily function; 5) the patient is a victim of circumstance with little or no responsibility for the presence or cause of the illness; 6) the patient is a passive recipient of treatment although cooperation with treatment is expected.[75]

This model for medicine has been very successful when applied to acute conditions and/or in medical and surgical patient care where a "cause" is identified and "cure" is effective.[76-78] For example, an acute infectious disease is an example where a "cause" can be identified (in a microbial culture), leading to an available treatment (e.g. antibiotics) resulting in a "cure" with the host recovering to their normal condition.[76] The biomedical model has provided a foundation for understanding the underlying mechanisms in disease.[76,78]

The biopsychosocial models: (the ICF model and the Wilson and Cleary model)

In contrast to the biomedical model, the biopsychosocial models recognize that in addition to the pathological process and biological, physiological and clinical outcomes, psychological and social factors influence a patient's perceptions and actions as well as their experience of the illness. The biopsychosocial models uses all of these factors to guide diagnosis and treatment of the illness.[73,75] Two such models are the International Classification of Functioning, Disability and Health, also known as the ICF and the Wilson and Cleary (WC) model.[65,79] A discussion of each of these models follows.

1) ICF model: The goal of the ICF is to provide a unified and standard language and framework to describe health and health-related states *for all people*. The first part of the model deals with *functioning* and *disability*. Domains included are from the perspective of the body, individual, and society and listed in two sections: 1) body functions and structure; and 2) activities and participation. The term *functioning* is the umbrella term for all body functions, activities, and participation whereas the term *disability* includes all body function impairments, activity limitations, or participation restrictions have been classified. The second part listed within the ICF, is the contextual factors that include the environmental factors and personal factors that can interact with the domains. Although the ICF includes a list of environmental factors, personal factors are not classified in the ICF as yet. In order to assess the full ability of an individual, one would like to assess their functional capacity due to a health condition in a "*standardized environment*" (by applying

the EF classifications) to be able to separate the impact of the health condition from the varying impact due to different environments on the ability of the individual.[80] Personal factors are the particular background of an individual's life and living, and comprise features of the individual that are not part of a health condition or health states such as gender, race, age, education.[65] Despite the fact that personal factors can restrict functioning in a person's environment; these are not considered health-related restrictions of participation as classified in ICF.[65] Quality of life (QOL) is not included in the ICF but it is recognized as conceptually compatible with disability constructs. QOL, however concerns how people "feel" about their health condition or its consequences and is part of the construct of "subjective well-being".[65]

The importance of the ICF is that it puts "health" and "disability" in a new light. The emphasis is on health and function, rather than on disability. It dismisses the previous notion that disability begins when health ends. The ICF recognizes every human being can experience a decrement in health and thereby experience some degree of disability. Disability is not something that only happens to a minority but is a universal human experience. By focusing on the impact and not the cause of disability, all health conditions can be compared on a common metric of health and disability. As a biopsychosocial model, the ICF takes into account the social aspects of disability and does not see disability only as a 'medical' or 'biological' dysfunction. It provides a coherent view of different perspectives of health: biological, individual and social, as seen in Figure 2.1

As the figure indicates, in the ICF, the domains of functioning, activity and participation are viewed as outcomes of interactions between health conditions (disease, trauma, or disorder) and contextual factors.[80] The biomedical model on the other hand, defines health as the absence of disease, which essentially would preclude anyone with a chronic disease as being healthly.[75,81] Whereas, within the constitution of the WHO signed in 1946, it was argued that health is "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."[82]

As illustrated in Figure 2.1, the relationships in the ICF model are complex and multidimensional. It is important that when measuring these domains from a specific health condition (such as MS) that information is collected on each of the domains in order to explore

the associations and causal links between domains and provide evidence to complete the model relationships.

Within the ICF, there are over 1400 categories covering body structure and function, activities and participation, and environmental factors (EF) with personal factors (PF) not yet classified.[83] One of the important parts of the ICF is the recognition of the need to include EF and PF, where both contextual factors can facilitate or hinder a person's functioning.

2) The Wilson-Cleary model: The WC model is often cited when the outcome of interest is health related quality of life (HRQL).[79] A conceptual model of HRQL was described that categorizes measures of patient outcomes according to health domains and proposes specific causal relationship between domains. The five health domains are: biological and physiological factors, symptoms, functioning, general health perceptions (GHP), and overall quality of life (QOL).

The dominant causal associations are represented by arrows, but this does not mean that reciprocal relationships do not exist nor relationships between adjacent domains.[79] The model also includes characteristics of individuals and characteristics of the environment that can affect each health domain as depicted in Figure 2.2, although it was not discussed in the text.[79,84]

These health domains are thought to be on a continuum of increasing complexity and integration with biological measures being the most basic and overall QOL the most complex.[79] For example, GHP is more complex integrating all aspects of the domains that came earlier (left of it) resulting in this outcome being subjective in nature. The final domain is overall QOL (far right), it characterizes subjective well-being related to how happy or satisfied a person is with their whole life. Overall QOL is stated to be related to HRQL and to other salient life circumstances and experiences.[79] Subjective well-being is complex and does not represent a single construct. Therefore, overall QOL is subjective and influenced by patients' values and preferences. Unfortunately QOL has been used to mean a variety of different things so the term "health related quality of life" was intended to narrow the focus to the effects of health, illness, and treatment on QOL.[84]

The WC model provides a functional theoretical framework with a useful taxonomy of variables that have been commonly used to measure HRQL and proposes a specific causal relationships between them that link traditional clinical variables to measures of HRQL.[79,84] This model is different from other classification of health status measure where the goal is identify the dimensions of health that are necessary to comprehensively and validly describe health *(such as the ICF)*.[79] The goal of the biomedical model is to understand causal relationships, diagnose the patient and apply a medical treatment, whereas the social model focuses on functioning and overall well-being taking into account the social context including the environment, complementary systems devised by society to deal with the disruptive effects of illness and looking at how all these factor influence individuals.[73] The WC integrates both of these perspectives making it potentially useful to health care providers.[85] By emphasizing functioning, health, and quality of life, and not focusing solely on the pathophysiologic disturbance aspects of the disease, health is viewed as more than the absence of disease.[85] The WC model links *traditional* clinical variables to HRQL in other words, linking *objective* clinical measures to *subjective* patient health experiences.[86] An important feature of this model is its theoretical approach describing the causal relationships between health domains stating that the dominant relationship is unidirectional from the left (bio./physiol. variables) to the right (overall QOL).[85]

An atheoretical approach to conceptualize a multidimensional construct like HRQL would result in a list of variables with no hypotheses describing the relationship among the domains making assessment or interpretation of the domain relationship patterns difficult.[85] Terwee et al., states that without specifying a priori hypotheses (*such as the one proposed by the WC model*) "…the risk of bias is high because retrospectively it is tempting to think up alternative explanations for low correlations instead of concluding that the questionnaire may not be valid."[87] However, a validated HRQL model will help researchers understand the relationships among the domains, providers learn about different conditions impact on patients' lives, or evaluate different approaches to patient care.[85,88]

WC (HRQL) model provides the conceptual framework for understanding the associations among health and patient reported outcomes (PRO) and links *objective* clinical measures to *subjective* patient health experiences.[86] The WC model can help researchers design a

measurement plan and provide rationale to select health domain outcomes based on the hypothesized casual relationship and on where the expected efficacy of an intervention might occur based on the model. Patient centered outcomes such as HRQL measures take in to account the patient perspective and may be considered PROs.[89-91] The Food and Drug Administration (FDA) defines "A PRO is any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else."[92] There is evidence that HRQL measures are as reliable as most traditional clinical outcomes and have prognostic value.[91,93] The FDA supports the importance of PROs, such as measures of HRQL, as "true" outcomes for treatment benefit or risk and has developed guidelines for PRO measures to be used to support labeling claims.[90,92] PROs in general help to provide a better understanding of the impact of illness from the patient's viewpoint.[94] Having a conceptual model like the WC is an important requirement in PRO development and having it accepted as an endpoint for clinical trials when making efficacy label claims (such as the therapeutic drug under study improves HRQL).[90,92] It would seem appropriate to use outcome measures that include the patients' perspective and which take into account all aspects of the impact of the disease on the individual.

Summary of the biomedical and biopsychosocial models:

Under the biomedical model the goal is still focused on a better understanding of the pathophysiology of MS, developing more effective treatments to control disease activity, and overall symptom management. The ultimate goal is to find a cure. The viewpoint from people with MS appears to support this view. Results from a postal survey show a large proportion of participants with MS desired a cure or access to effective disease modifying therapies as their single greatest need.[95]

The ICF model provides a unified standard language and framework to describe health and health-related states. The terms *function* and *disability* are used as umbrella terms for body functioning, activity, participation and its impairment, limitation, restriction respectively. It has been translated into multiple languages.

The international effort to develop an ICF Core Set for MS by an evidence-based and formal

decision-making consensus process integrating research knowledge and expert opinion is important.[96] There is continual work to validate and improve the ICF Core Sets for MS. The methods used are focus groups and Delphi method using typical procedures for ICF mapping. Two recent studies used focus groups and Delphi method to describe functioning and disability from the points of view from patients and physicians respectively. This constant refinement of the ICF core sets for MS make it a useful tool.[97,98]

The usefulness of the ICF depends on whether researchers and/or clinicians globally embrace its use as a tool and put it into practice.[99] Recently, a systematic review on the use, implementation and operationalization of the ICF seem to suggest that more researchers in a variety of fields are adopting the ICF *concept* and that a "cultural change and a new conceptualization of functioning and disability is happening."[100] If the ICF is adopted internationally, it may be possible to compare results across conditions.[101,102]

Rehabilitation, a multidimensional discipline dedicated to optimizing patient functioning and health, has adopted the use of ICF as the model of choice in instrument development, to assess needs of the MS population in order to provide appropriate services in an interdisciplinary setting, or policy guideline that result in patient centered approach to care.[102]

The strength of the ICF is the standard language and framework for the description of health and health-related states, the ability to map existing measures to ICF, and having developed the infrastructure to allow the ICF model to be a tool in research.[102] The neutral terms used in the ICF to classify the components will hopefully put the notion of "health" and "disability" in a new light.[80,102] It can have a role in education and communication among different health care professionals and has been recommended by the National Institute for Clinical Excellence in 2003 as the model and vocabulary for clinicians, professional groups, and organization involved in the care of those with MS.[103] It may improve communication between patients and health care professional.[102] With a common language it will be easier for patients to understand their functioning and health, rehabilitation goals, and intervention plan.[102]

WC is relevant when conducting research on HRQL. The strength of the WC model is its

simplicity (unlike the ICF) and its theory of the causal relationship among health domains. WC provides a more global perspective of impacts on health domains to include biological and physiological factors most *objective* to the more *subjective* domains of GHP and overall QOL. These domains are not covered by the ICF. WC model's inclusion of subjective outcomes (GHP and HRQL) each assimilating the outcomes *downstream* of it based on theory makes it important in HRQL PRO development. There is no internationally agreed upon gold standard for HRQL measurement in MS.[104] As such, development and validity cannot be established by typical criterion validity.[105] In this circumstance, the conceptual framework (of the WC model) has a natural importance in the development of HRQL patient reported outcomes to help establish construct validity of the measure.[85,87,105,106] This view was reflected in the FDA requirement for a conceptual framework for PRO development as an endpoint for labeling claims.[90,92] Patient centered outcomes such as HRQL measures take in to account the patient perspective and may be considered PROs.[89-91]

There is a need for continual refinement and validation of the WC model's proposed causal relationships among the domains in MS as mentioned above. There has not been a concerted effort to systematically validate the WC model. It appears to be an *ad hoc* process of researchers interested in HRQL and attempt to validate the models causal relationship in health conditions they are studying. As discussed above, different methods have been used to validate this model with the most useful appearing to be structural equation modeling although it has not been done in MS. In 2003, a task force on PRO development did not endorse the WC model stating that there was insufficient evidence to support the hypothesized relationships.[107]

With the many different HRQL measures available for use in MS,[108] selecting the appropriate HRQL measures can be aided by a valid WC model as it has been described in the examples of cancer and orthopedic patient care.[106,109] The WC model can be used to tailor patient assessment and help clinicians with patient counseling or for referring patients to others in a multidisciplinary care environment using the example of cancer and orthopedic patient care.[106,109] Unfortunately, there appears to be little effort in HRQL assessment in routine MS care.[104]

Combining the ICF and WC models:

In a recent article, the authors were able to illustrate the compatibility of the two biopsychosocial models, the ICF and the WC model. Both models were developed independently but share commonalities.[110] The ICF considers disability as a functioning continuum.[110,111] The WC model extends its model beyond ICF functioning and includes two subjective domains of patient health outcomes; GHP and HRQL (overall QOL in the original model).[79,84] QOL is not included in the ICF but it is recognized as conceptually compatible with disability constructs.[65] Both models have health-related variables and contextual factors that are divided into environmental and individual characteristics (Figure 2.3). The Biological & physiological and symptom status variables in the WC model correspond to the body function & structure component of the ICF, WC functional variables correspond to activity and participation of the ICF.[110]

Both biopsychosocial models view the patient as a "whole person" and include personal and environmental factors. The knowledge gained from research based on the two models has been used to better understand the person with MS as an individual and the needs of the MS population as a whole in their living context. This knowledge will enhance the ability to develop interventions to improve the person's function in their personal, family, and civic life and will provide better patient centered care. This may have a resulting benefit in enhancing communication between patient and doctor.

The differences and similarities in the models are summarized in Table 2.1. Although the biomedical model and biopsychosocial models are different, their appropriate application in MS care and research can *ultimately* lead to the same goal, to improve health in people with MS.

Chapter 2 presented the two predominant models for MS, the biomedical model and the biopsychosocial model. To paraphrase Lord Kevin, "You can't measure what you don't understand and you can't fix what you can't measure". Models lead to measurement frameworks as they indicate the targets of concern. The models contribute understanding of the concerns, which lead to a measurement framework.

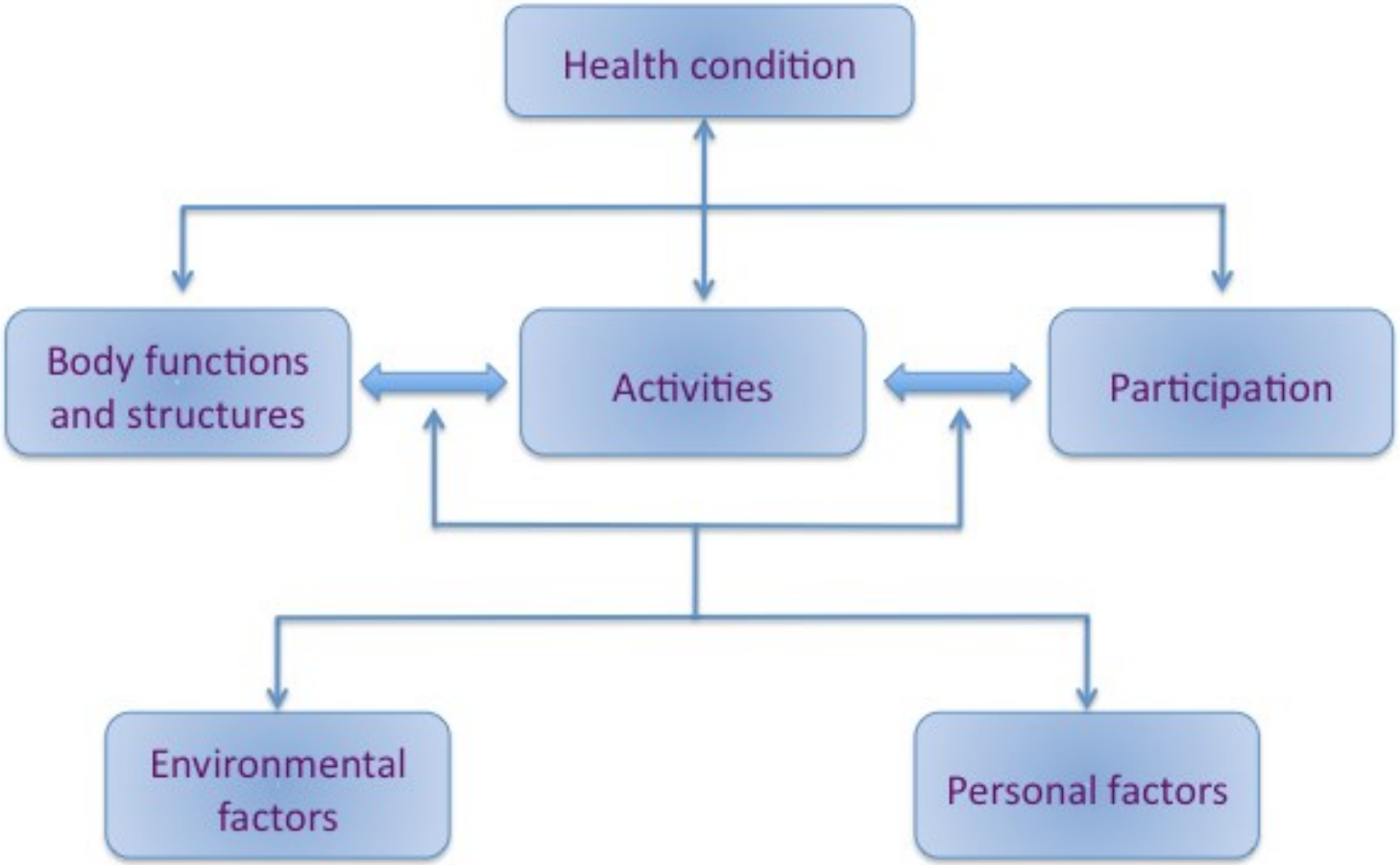


Figure 2.1: The current conceptual framework of disability-The World Health Organization International Classification of Functioning, Disability and Health (ICF)

World Health Organization (2001). International Classification of Functioning, Disability and Health: ICF. Geneva, World Health Organization

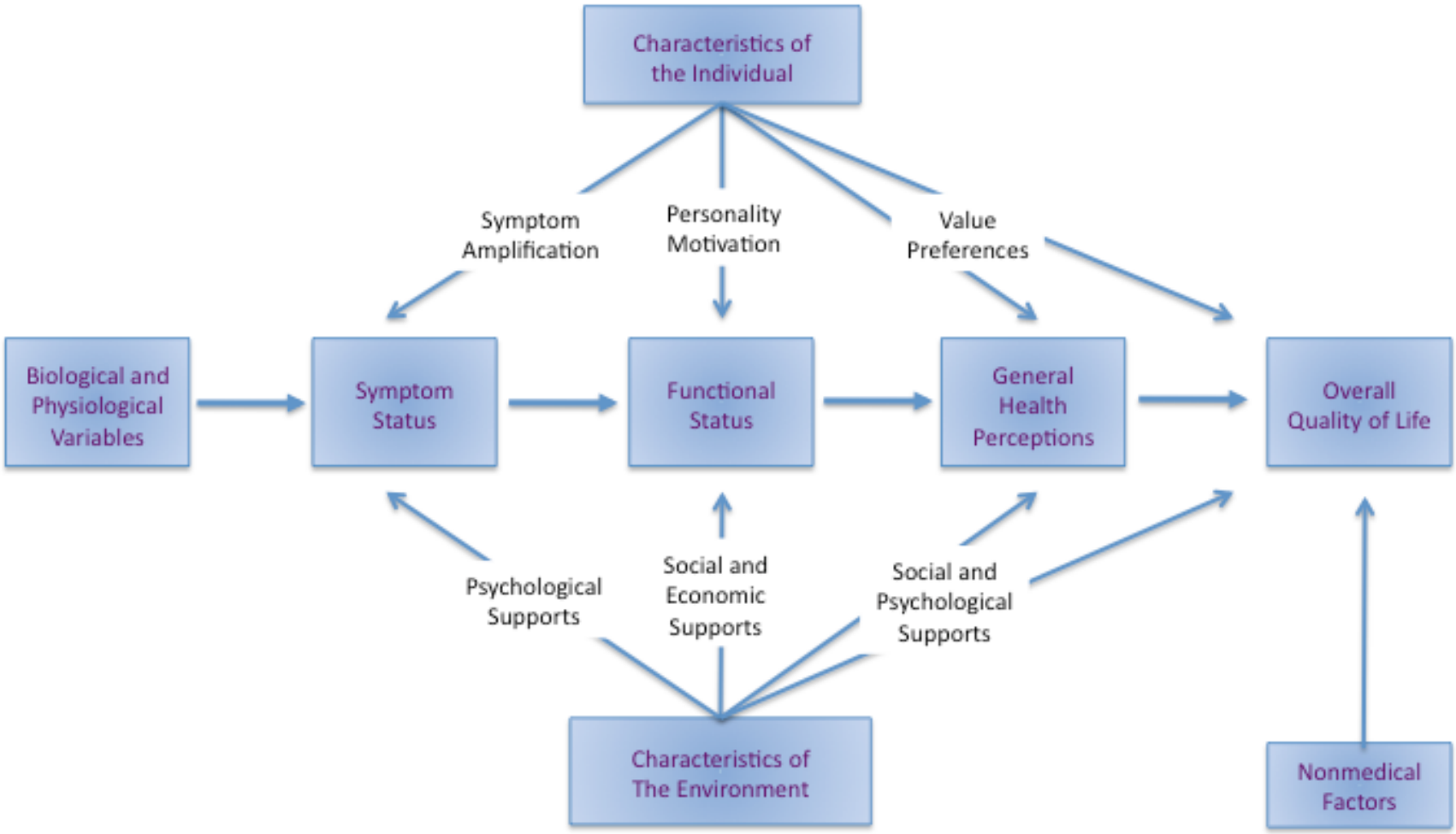


Figure 2.2: Wilson and Cleary Model: Relationships among measures of patient outcome in a health-related quality of life conceptual model

Wilson, I.B. and Cleary, P. D. (1995). Linking Clinical Variables with Health-Related Quality of Life. JAMA: The Journal of the American Medical Association 273(1): 59-65.

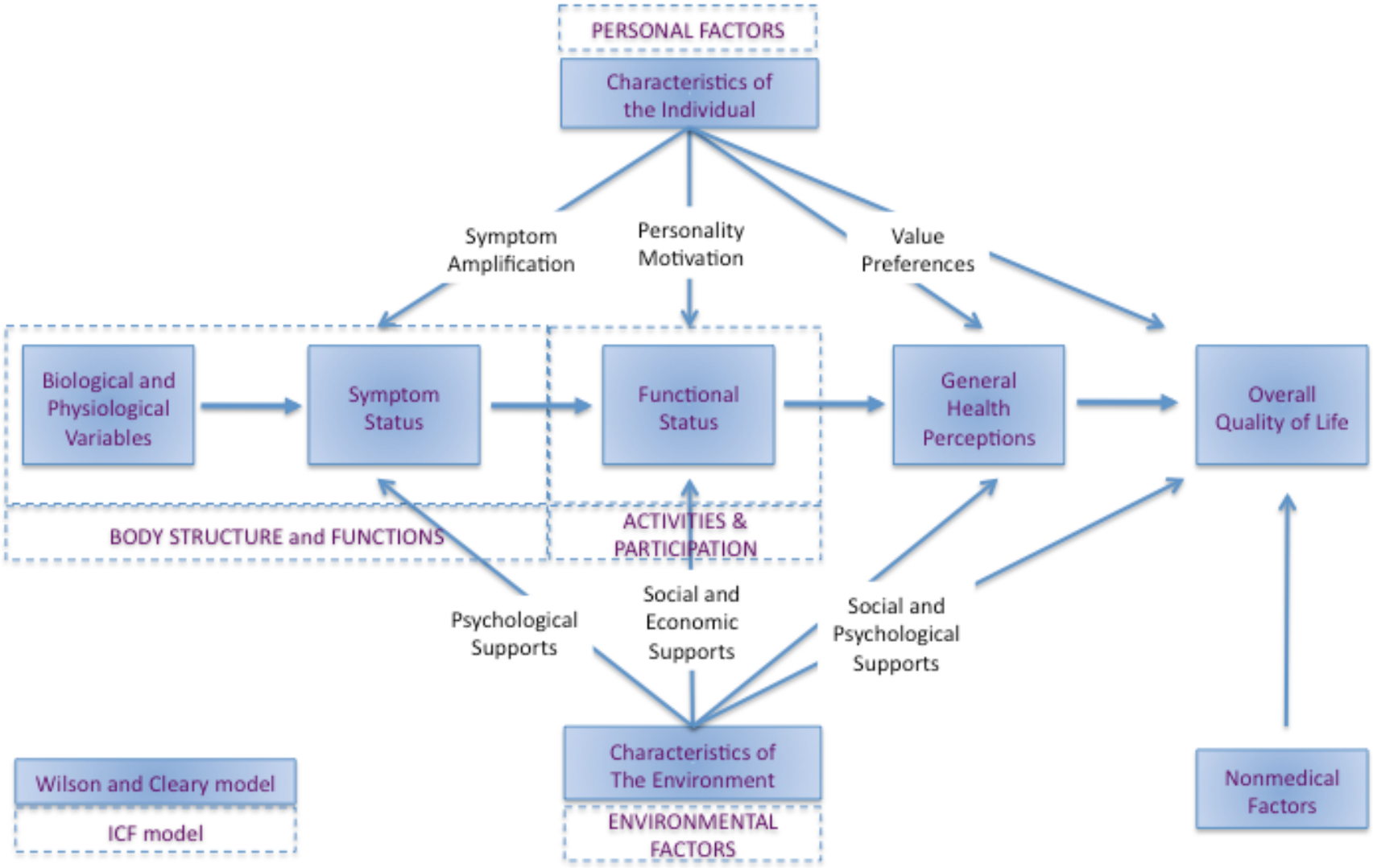


Figure 2.3: An integrated model for health outcomes (modified)

Valderas, J. and Alonso, J. (2008). Patient reported outcome measures: a model-based classification system for research and clinical practice. Quality of Life Research 17(9): 1125-1135

Table 2.1: Comparison of the Biomedical model, ICF model, and Wilson-Cleary models for MS

Multiple sclerosis	Biomedical model	ICF model	Wilson Cleary model
Cause	Cause unknown; search for a single cause	Multifactorial	Multifactorial
Treatment decision	Mainly by the neurologist, patient is a passive participant with minimal input	Patient is involved, patient perspective is important	Patient is involved, patient perspective is important
Treatment goal	Control or stop disease activity; find a cure	Improve body function, activity and participation	Improves all health domains including GHP and QOL
Types of outcomes	Measured directly	Measured directly & self-reported	Measured directly & self-reported
How is health affected?	Loss of health when you have a disease	Functioning is a continuum	Impact of health condition has a linear impact among health outcome domains affecting QOL
Measuring health domains	Body part: Cell, tissue, organ, organ system	Whole body: tissue, organ, organ system, whole body and person level	Whole body: cell, tissue, organ, organ system, whole body and person level
Biological/ physiological	**Yes**, explicitly described: genetic markers, biomarkers of disease activity; MRI, evoke potentials, immune system and neuronal health	**Body structure:** structure of the nervous system, structures of the immune system **Body function:** mental, sensory, voice and speech, immune system, genitourinary, reproductive, neuromusculoskeletal and movement.	**Yes** not explicitly specified
Symptoms	(Prefer neurological signs on examination) Pain, fatigue, sensory, muscle weakness, spasticity, vision, bowel & bladder, depression, etc.		**Yes** not explicitly specified
Functioning	Ambulation, upper limb (range of motion), muscle function, executive function, etc.	**Activity and participation:** learning and applying knowledge, general tasks and demands, communication, mobility, self-care , domestic life, interpersonal interactions and relationships, major life areas, community, social and civic life	**Yes** not explicitly specified
GHP	No	No	**Yes** not explicitly specified
QOL	No	No	**Yes** not explicitly specified

Chapter 3

Measurement of a relapse in MS

Under the biomedical model, measures of disease activity are of primary interest. These measures are relapses, presence of new lesions as seen on brain MR imaging, and findings from neurological examination using the EDSS. This chapter presence a review on how relapses have been measured over the decades in MS. A full review of MRI as indicators of disease activity is beyond the scope of this thesis and a review of the EDSS and its measurement challenges is the subject of future chapters.

The disability a person with MS experiences during a relapse is the most observable clinical feature of disease activity. It is presumed to be caused by lesions in corresponding locations in the CNS. The diagnosis of definite MS can be obtained when a patient experiences two relapse at different times with different lesions in the CNS.[17] The definition on what constitutes a relapse has evolved over the years.[17,19-22]

Aim: A scoping review was used to map the major areas of research related to MS relapses and to identify any area (gap in knowledge) to focus a more structured review on measurement challenges in relapses.

Due to the large amount of literature on MS relapses the general approach of a scoping review was used to summarize the broad topics related to research in this area (to take stock of what has been learned).[112] A concept map is provided to summarize (descriptively) the key concepts in MS relapses research and to help identify the measurement challenges of relapses. The focus was on RRMS patients as they are most likely to experience relapses, are relatively earlier in the disease process and the most treatable group that will respond to current approved therapies; therefore this is the largest group in relapse related studies.

A literature search was completed using Medical Subject Heading (Mesh) terms and a keyword from three online electronic databases, Cinahl, Medline (1948 to July Week 1 2011) and Embase (1947 to 2011 July 12). Mesh terms “multiple sclerosis” and “recurrence” were used for Cinahl and Medline. The keyword, “relapse” and Mesh term

"multiple sclerosis" was also used in Cinahl. Mesh terms for Embase were "multiple sclerosis" and "disease exacerbation". Duplicate articles were deleted. Only English and French journals were part of the review. This initial search yielded over 2000 articles (Cinahl=227;Medline=1145; Embase=822). The title and abstract of each article were reviewed. Article text was also reviewed in reference to information related to challenges in measuring relapses. All articles unrelated to remitting-relapsing MS or relapses were removed. Secondary and primary progressive MS articles were removed. Articles related to pediatric MS, animal studies, complications in MS not directly related to relapses, case studies and differential diagnosis were not part of the analysis. The working number of articles for the scoping review was 1414.

An update of the scoping review included articles from 2011 to March 8th 2015. The same databases were searched with the same criteria. After duplicates were deleted from each database search, the update yielded 912 articles (Cinahl=410; Medline=170; Embase=332). An additional 63 duplicate articles were deleted leaving 849 to merge with the original 1414 articles. From the 2263 articles, 31 duplicate articles were deleted. An additional 346 unrelated articles using the same criteria as from the original review were deleted. The final dataset contained 474 updated articles and 1414 articles from the original dataset. Major categories of research topics related to relapses are depicted in a concept map. Although these groups are related to each other, only major relationships related to relapses were depicted with lines for the sake of readability. Many articles may have been only tangentially related to measurement challenges of relapses but are part of the search results because relapse is synonymous with disease activity and contain section(s) concerning relapses. These articles were retained in the concept map but coded orange. Articles that were specifically related to relapses were coded blue or have an asterisk (*) in front of the number on the map.

The types of studies found from the scoping review are given below, in percentages of all studies up to 2012 is given first and the value including the update to 2015 is given in square brackets.

The types of studies found are as follows (Figure 3.1 include studies up to 2012 [Figure 3.1a include studies updated to 2015]):

- 33.3% [35.6%] **Drug related** (472 [671]): This group contains articles on commentaries and studies on all drugs, their mechanism of action and adverse events, and treatment reviews and treatment guidelines.
- 22.4% [20.7%] **Relapse related** (318 [390]): Commentaries and studies on relapse treatments (steroids), optic neuritis, relapse management, MRI and immunology of relapses (predictors and biomarkers), and outcome measures
- 23.9% [21.9%]**MS in general** (324 [413]): MRI and immunology/molecular studies, miscellaneous paraclinical measures, genetics related, statistics and outcome articles, articles on all aspects of MS and the disease process, symptom and disease management
- 10.9% [11.2%] **External factors** (18+*136** [211]): Infections, vaccines, pregnancy and fertility, environmental factors (seasons weather), stress trauma (not related), smoking, air pollution, month of birth, and high sodium diet
- 4.9% [4.6%] **Natural History & Cohort Studies** (55+*14** [70+17*]): Natural History of MS, disease course, natural history related to relapses, long-term studies, and database studies
- 3.5% [3.8%] **Quality of life** (50 [72]): Quality adjusted life years, cost, quality of life, and qualitative studies
- 1.9% [2.3%] **Non-drug interventions** (23+*4** [39+5*]): Exercise and rehabilitation, diet and supplements, alternative medicines, and alternative interventions.
- There were four phenomenological studies identified but none concerned MS relapses specifically.[113-116]

The scoping review of relapses in MS provided a comprehensive overview of the research interests in this topic.

From the major groups of studies identified, some general statements can be made. Drug related research articles predominated with the largest group at 33.3%, [35.6%]. In summary," research related to drugs use to control disease activity" and "relapse related research" represents about two thirds of the research activity in MS. All DMTs target disease activity by reducing relapse frequency. The EDSS is the main tool used by neurologist to assess MS disability including relapses. Relapse frequency is an outcome for many studies. ARR and EDSS are important in cohort studies and a variable tracked in longitudinal database registries. There is a definition for a relapse. There are a few articles

on relapse severity using indirect measures such as need for steroids, hospitalization, or emergency room visits.[117-119] There are two articles in the original review that reported having residual deficits after suffering a relapse. Lublin et al., reported residual deficit as measured by EDSS 2 months after the relapse.[15] Vercellino et al, used operational definitions of a severe relapse such as a 2-point or more increase in EDSS from baseline.[120] They reported that incomplete recovery at 1 month was a predictor of long-term persistent residual deficits.[120] Although there was no discussion on relapse severity, one might make the assumption that a relapse with residual deficits is more severe than a relapse with full recovery.[45] One article was found in the update also used a two point EDSS change as a severe relapse. The authors found that of the 226 relapses experienced by 144 patients, 32% of relapses were severe with 11% failing to recover. The study estimated that the majority of improvement in physical disability after a relapse occurs by 2 months but some patients took up to 12 months to recover.[121] There is little research on estimating relapse severity or duration.

In reviewing these articles it is apparent that there is no specific relapse measurement tool used to decide when a patient is experiencing a relapse or to estimate its severity. There were no reports of relapse duration from any of the articles in the original scoping review and only one in the update. There appears to be less focus on research concerning the duration or severity of a relapse. It appears that the impact of relapse frequency on the long-term disease progression is the main interest.

Secondary structured review: measuring relapses in randomized control trials

A gap in research identified is that there is no specific relapse measurement tool. A secondary structured review was conducted to provide a more focused and detailed examination of the importance of relapse as an outcome and the measurement challenges of frequency, duration, and severity of a relapse. Randomized control trials (RCT) are the most methodologically vigorous study designs and thus should have the best standardized methodology to assess relapse frequency, severity, and duration.[122] RCTs for MS treatments currently target the inflammatory process to reduce disease activity (reduce relapses). These studies have relapses as an outcome and often as the primary outcome. Cohort/database studies with longitudinal data are the best method of understanding the

natural history of the disease process and importance of relapses on MS.[123-126] Online library database search of Embase and Medline only identified a partial list of relevant studies with many non-relevant articles. Multiple combinations of mesh terms (multiple sclerosis, multiple sclerosis/ dt [Drug Therapy], randomized controlled trials, investigational drugs, recurrence, relapse, double-blind method, immunosuppressive agents,) to find RCTs and mesh terms (database, registries, multiple sclerosis, recurrence, relapse rate [keyword], cohort studies) cohort/databases studies only identified a partial list of relevant studies with many non-relevant articles. To augment this list, experts in the field, neurologists from the MS Clinic at the Montreal Neurological Institute that are involved in fundamental MS research and clinical research were asked to provide information to search for relevant studies. Article bibliographies were reviewed to obtain relevant articles. RCTs and cohort/database were also identified from Inusah et al.,[127] Flachenecker et al.,[128] and Hurwitz, et al.,.[43,129] Pivotal RCTs for MS disease modifying drugs with a placebo group of approximately 100 subjects or more and large cohort and/or database registries with information on MS relapses were identified. The majority of the cohort/ database studies were post-1980s. Prior to these studies, the MS diagnostic criteria (and relapses) used were different and represent measurement challenges that may no longer be relevant. NARCOMS was not included since it is designed for patient self-report measures that are not compatible for the purposes of this paper.[128]

Results:(RCT)

RCTs with placebo groups (13±7) were identified and summarized in Tables 3.1a-c in the appendix. RCTs that led to currently approved MS drugs in Canada were included.[27-30,47,49-54,130,131] Additionally RCTs were included; studies on Laquinimod [132-134], Cladribine[135], Daclizumab[136] and Pegylated interferon beta-1a[137]. One additional study, the oral Copaxone (CORAL) study[138] was added because it had a large placebo group even though it was an efficacy failure.

Defining a relapse: ARR was the primary outcome[27,29,30,47,49,51-54,133-138] in 15 of the 20 RCTs and included as an outcome in all the studies. Relapse definitions varied in this set of clinical trials. Twelve studies used 24 hours[27,29,47,50-54,130,135-137] as the minimum amount of time to experience symptoms in order to qualify as a relapse, 7 used 48 hours[28,30,131-134,138] and 1

study[49] did not specify a time in the relapse definition. Twelve studies required changes in EDSS or FSS to be labeled a relapse.[28,30,49-52,131-135,138] Ten studies excluded changes to bowel and bladder or cognitive function.[28,30,49-52,131,132,135,138] Interestingly, one of these studies also excluded fatigue and sensory symptoms.[28]

The time allowed to evaluate a suspected onset of a relapse also varied. One study required an evaluation of a suspected relapse within 24 hours[28], 1 within 72 hours[47], 2 within 5 days,[53,54] 11 within 7 days,[29,30,130 49,51,131-135,138] and unknown in 5 studies.[27,50,52,136,137] All studies required relapses to be confirmed by a neurologist

Severity: Only 5 studies reported an outcome for relapse severity.[27,29,130,131,138] The studies used "decreased use of steroids" and/or "decrease hospitalization" as indirect measures of relapse severity. Of these, 3 used the Scripps outcome measure to classify the relapse as mild, moderate or severe.[27,29,130] The EDSS was used in all the RCTs. Twelve RCTs had formal definitions of FSS and EDSS change to define a relapse but it is assumed that the EDSS was used to assess relapses in all the RCTs, since all relapses had to be confirmed by a neurologist with objective signs of worsening neurological dysfunction (disability). The MS Functional Composite (MSFC) was used in 4 RCTs to assess overall neurological function[49,50,133,134] however data was not reported in the 2006 study on FTY720.[50]

Duration: Only 1 study reported measuring relapse duration but did not report the data.[27] Interestingly, 15 studies recorded the time to first relapse.[27-30,49,50,52-54,130,133-138]

Results: (Cohort/Database)

Cohort/database studies (17) were identified and summarized in table 3.2a-b in the appendix. In this sample, more recent registries benefit from more standardized diagnostic and relapse criteria.

Defining a relapse: The relapse definition was unclear in 2 studies.[139,140] The Sylvia Lawry Centre for MS Research (SLCMSR) dataset is a pool of 31 RCTs and does not have a single relapse definition.[8] Two cohort studies use the McAlpine criteria to define relapses.[26,141] However, Confavreux et al.[141] modified the McAlpine criteria by adding the requirement for symptoms to last at least 24 hours making the criteria resemble Schumacher's according to

Lublin et al,.[13] This longitudinal database (EDMUS-Lyon) changed the criteria to the Poser's criteria in the 2003 study.[142] Diagnostic criteria also changed in the EDMUS-LORSEP database, where the Poser criteria was used before 2002 but switched to McDonald criteria after 2002.[143] This group along with the group from EDMUS-Burgundy also added a statement that fatigue alone as a symptom does not constitute a relapse.[143,144] Trojano et al., did not provide complete information for their 2 studies[145,146] but are using iMed database software which is coded for Poser and McDonald diagnostic criteria[146]. Four studies used Schumacher's definition.[40,144,147,148] Goodkin et al, used a proprieties definition for a relapse requiring a 0.5 point EDSS or 1 point Ambulation Index change for more than 5 days and less than 60 days to be a relapse.[149] This may skew recorded relapses to be of longer duration since the typical definition of a relapse require symptoms to only persist for $\geq$ 24 hours. Tremlett et al., did not specify a criteria but provided a definition.[125]

Thirteen studies used 24 hours as the minimum duration of a relapse.[36,40,125,139,141-148,150] One study used 5 days as the minimum relapse duration.[149] A relapse definition was not applicable to the SLCMSR database since the data is a composite of 31 RCTs.[8] Two studies did not state a minimum relapse duration.[26,140] Only 3 studies stated that the onset relapse was deleted to correct for an overestimation of the first year relapse rate.[26,125,141] Studies presented ARR results differently: 3 studies reported ARR for 5 year periods,[26,125,143] 1 study had an ARR but no period reported, 2 studies reported ARR for 1 year before start of a drug,[142,148] 4 studies reported ARR for 2 years before start of a drug,[8,139,142,148] 2 studies reported ARR for an average period[145,150] of 7.4 years and 14.2 years respectively, 4 studies reported ARR for 2 year from onset.[36,144,149,151] Tremlett et al, had average ARR for every 5 years for up to 30 years.[125] Two studies did not report the ARR.[140,147] Some of the cohort data was gathered retrospectively, a few had a combination of historical and prospective data and others generally stated that data was prospective although it was generally difficult to judge how well data was collected (refer to Table 3.2a in the appendix).

Severity: Two studies indicated that relapse severity was measured but did not indicate how it was defined.[141,151] All studies, except two, used versions of the EDSS as the standard MS measure of disability.[26,141]

Duration: The two studies that reported measuring relapse severity also reported measuring duration but did not present results.[141,151]

Summary:

RCT definition of a relapse: RCTs reviewed above generally had well defined definitions of what constituted a relapse. However, the definitions were often modified from published guidelines to better assess the more *objective (directly measured)* neurological signs of a relapse possibly excluding milder relapses with symptoms/signs that are more difficult to assess. As seen in the results of the review, 10 studies excluded bowel and bladder and cognitive function as relapse symptoms.[28,30,49-52,131,132,135,138]

In more recent trials, study investigators were all trained on the neurostatus©, a training program designed to increase intra and inter-rater reliability of the EDSS. Also, the neurostatus©, rescored bowel and bladder and visual FS scores by decreasing the likert scale by 1 point thus decreasing the contribution of these two functional systems to the overall EDSS scores[152] further changing the assessment of a relapse within the RCT. The RCT solution to measurement challenges of a relapse is to measure only what can be done relatively *objectively (directly)* on a neurological disability scale using the EDSS. This solution may be used for the purpose of trying to have better reproducible measures of a relapse but is a poor solution outside the RCT. By excluding symptoms and decreasing their impact on the EDSS scale, researchers are ignoring the full impact of a relapse on the patient.

Defining a relapse outside trial protocols: Schumacher et al., stresses that symptomatic worsening should be counted as a relapse only by appropriate change in "*objective"* neurologic function as determined by examination.[22] This essentially ignores patient reports whereas Poser et al., allows for "…completely subjective and anamnestic…" patient reports that are consistent with MS.[17] Interestingly Poser et al., does not mention fever as an indicator of a pseudoattack in the article.[17] In 2001, McDonald et al., also endorsed the description of Poser et al. that an attack can be defined by either a "…subjective report or by objective observation…" lasting at least 24 hours.[19] However, a point of ambiguity exists since it "…assumes that there is expert clinical assessment that the event is not a pseudoattack…" leaving the reader wondering if a *subjective* patient report would be acceptable on its own as a relapse.[19] In 2005, the relapse definition was clarified to state "subjective reports (backed up by objective findings) or objective observation" should be used.[20] In 2010, "…patient-reported symptoms or *objectively* observed signs typical of an

acute inflammatory demyelinating event in the CNS, current or historical, with a duration of at least 24…" could constitutes a relapse. It recommends a timely neurological exam but does recognize that historical events without "*objective*" neurological finding but consistent with MS can provide reasonable evidence of a prior demyelinating event.[21] The version of the criteria applied to define a relapse and if one decides to include patient reported symptoms may affect the doctor's decision in making a diagnosis of (measuring) a relapse.

There are three studies on the understanding of MS diagnostic criteria and relapses. In two surveys, one with primary care physicians (PCP) and another with neurologists indicated there are still problems interpreting what is a relapse and/or knowing when to treat and/or how to treat relapses. The survey with neurologists illustrated that there was some confusion on interpreting the definition of a relapse and diagnosing MS using the McDonald criteria.[153] The survey with PCPs showed that there was confusion in the definition of relapse and understanding that steroids were the standard treatment for acute relapses.[154] At a general neurology practice, patient charts were reassessed and revealed that over 50 percent of patients with a diagnosis of MS did not satisfy the 2001 McDonald criteria for definite MS.[155] The authors felt that while a majority of patients would eventually fulfill the criteria it still leaves the question whether there will be patients that have been misdiagnosed with MS. Although there is a definition for a MS relapse, the details of the definition and how to apply it have been evolving over the years and appear to still pose a challenge for doctors in *measuring* a relapse. These issues would be very relevant to the longitudinal cohort/database studies. The majority of the current registries use published guidelines for MS diagnosis and to identify a MS relapse (Table3.2b in the appendix). If neurologists are to use relapses as an outcome for treatment failure and/or assess patient disease stability, there is a need for an agreed upon relapse definition and measures for the appearance and severity of a relapse.[25]

ARR measurement and interpretation issues

There is little research on the duration and severity of MS. There is evidence that ARR in clinical trials is declining.[127,156] Some of the possible reasons cited are changes in diagnostic criteria or the recruitment milder patients since the more active patients may already be on treatment.[127] In RCTs, patients selected were relatively early in the disease process with

relatively high pre-study relapse rates. The study durations were relatively short compared to the course of the disease. This does not negate the results of the RCTs but the ARRs measured are not generalizable to the RRMS population. For example, relapse rate before entry to RCT was a predictor of on-study relapse rate.[157] Due to the change in the ARR observed in clinical trials, the efficacy results from more recent RCTs, such as Natalizumab should not be compared to the pivotal trials of the interferon betas or glatiramer acetate.[158]

Variation in ARR was also reported in the cohort/database studies. Unfortunately, in the review, cohort/database studies did not report ARR results in a similar manner to allow for easy comparison. Descriptively the ARR was quite variable (Table 2b). Including the onset attack could artificially inflate the ARR estimate of the initial year. Several researchers include the practice of deleting the onset relapse to avoid overestimating the ARR in the first year.[26,125,141] A recent retrospective longitudinal cohort study examined the relapse rate and provided more evidence of the decline in ARR over a period of 30 years. The relapse rate decreased by 17% every 5 years.[125] A possible reason cited for the declining ARR is regression to the mean.[26,31] There is also evidence that prospective data yield higher estimates of ARR than retrospective data.[25,36] The variation in the definition of a relapse, and retrospective or prospective data gathering process may impact the ARR estimate.[31] Systematically measuring ARR is a measurement challenge making variation in ARR difficult to interpret.

IN SUMMARY

This chapter reviewed one of the key measures of disease activity, relapses. Alternatively, the other common measure of disease activity is observed deterioration in neurological status as measured by the EDSS. The EDSS is considered a measure of disability comprised of 8 functional system scores (FSS) to score EDSS <= 4.0 whereas score > 4.0 is based solely on ambulation. The next chapter will discuss disability from the biomedical and biopsychosocial perspective. Disability is represented more comprehensively by the biopsychosocial models.

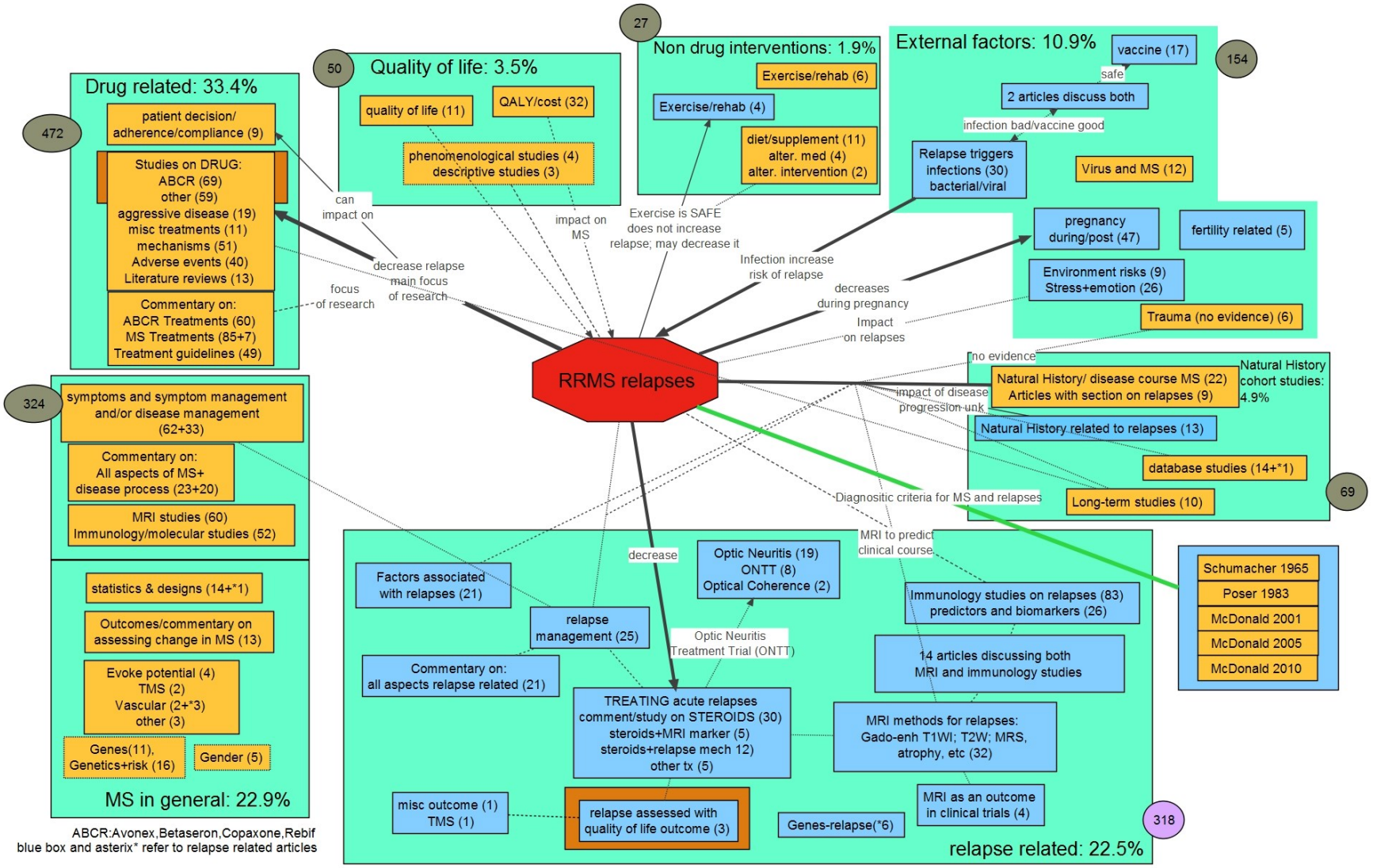


Figure 3.1: Concept map of major research areas in relapses up to 2012

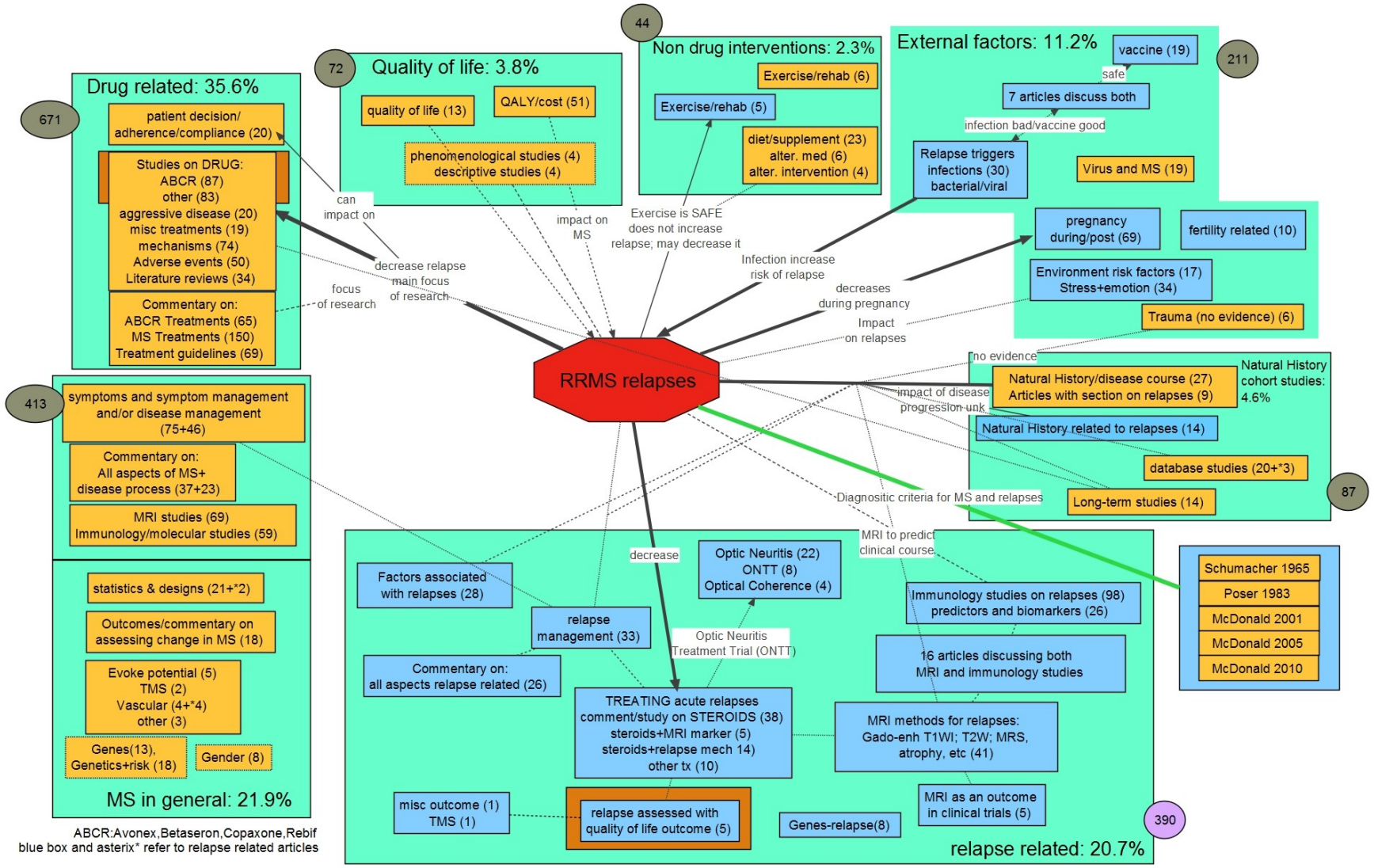


Figure 3.1a: Concept map of major research areas in relapses updated to 2015

Chapter 4

MS disability from the perspective of the biomedical and biopsychosocial model

The biomedical model represents disability primary from the perspective of neurologists managing MS. The biopsychosocial model is the predominant model for rehabilitation.

The biomedical view of MS disability

Disability measures:

Under the biomedical model the focus is on understanding the pathophysiology of MS, removing the cause(s) and to return the patient to normal function (or as close as possible). The role of the patient is passive with the expectation to cooperate in receiving treatment.[75] Disease activity from MS causes damage to the CNS and is purported to result in the observed disability experienced by the patient. MS disability within this framework measures damage related to body function impairment. Disease activity in MS is based on ARR, MRI activity, and disability (using the EDSS).[2,23,24]

In this model, the neurologist will be assessing MS disability. The EDSS is the standard tool used by neurologist to judge the patient's level of disability.[34] The EDSS has a long history. Kurtzke and Berlin first described the use of the original Disability Status Scale (DSS) as a single item rank ordered scale with categories from 0(normal) to 10(death due to MS) with 1-point increments. This scale was "intended to measure the maximal function of each patient as limited by his neurologic deficits".[160] The FSS were developed later as a means of collapsing the neurological exam into categories and were intended to complement and be a useful check of the scoring of the DSS.[161] Kurtzke divided what he felt were all neurological deficits into 8 FSS ((i) pyramidal (motor), (ii) cerebellar (ataxia, co-ordination), (iii) brainstem (cranial nerve function including speech, swallowing), (iv) sensory, (v) bowel and bladder function, (vi) vision, (vii) "mentation" covering mood alteration and cognitive impairment, and (viii) other) based on central nervous system functions affected by MS, and not by brain anatomy. [161] Each functional system (FS) has a mutually exclusive numerical rating per category where "a higher number would reflect a

greater level of dysfunction".[161] A unique scoring rubric was required to translate the FSS into a "total overall score" of the DSS since attempts to simply add all the FSS resulted in a plateau long before the theoretical maximum score.[162] The EDSS was modified to have 20 grades of impairment with scores still ranging from 0 (normal) to 10 (death due to MS) but with 0.5 increments after 1.0. EDSS scores between 0 and 4.0 are based on neurological exams and FSS, whereas scores above 4.0 are based solely on ambulation. MS disability as measured by the EDSS is narrowly defined with an emphasis on walking disability. Cognitive and upper limb function are not measured.[163]

Although based on sound clinical knowledge, the EDSS and FSS were developed without psychometric input and this limits their usefulness as an evaluative outcome measure in MS.[164,165] EDSS is not an actual "measure", as the grades represent ranks (ordinal) and not numerical values; the distance between ordinal categories is unknown. A 2014 systematic review of the psychometric properties of the EDSS[71] summarized the measurement limitations for interpretation and mathematical manipulation, owing to the ordinal scoring system, the typical bimodal distribution[164,166], and the non-linear response for each scoring step.[167,168] The EDSS can only validly be reported as a median affecting interpretation of change. This review also highlighted issues with standardization, sensitivity, and reliability.[168] A key criterion for an evaluative measure is the minimum clinically important change (MCIC). The established MCIC of a 1.0 for people with EDSS ≤ 5.5 and 0.5 for people with EDSS ≥ 6[169], has been challenged by Kragt et al.[170] Under the biomedical model disability the EDSS is based exclusively on assessing body function impairment.

The MSFC was an attempt to provide an alternative to the EDSS. The MSFC, a multi-component performance outcome measure, comprised of 3 parts, (Paced Auditory Serial Addition Test (PASAT), 9-Hole peg test, 25-foot walk) measuring cognition, and upper and lower limb function respectively[168]. It has been used mostly as a research tool, as administration requires equipment, training, and the need to use *z*-scores in order to generate a total score.[23] The disadvantage is that it takes approximately 15 minutes to administer, there are practice effects, and the interpretation of *z*-scores across trials using different reference populations is problematic.[23,171] The MSFC is not a tool used by

neurologist. This may be an additional reason why it has not been widely adopted at MS centers. Again the MSFC only measures body function impairment.

There was an early attempt by the International Federation of Multiple Sclerosis Societies to establish a Uniform Minimum Record of Disability to characterize all patients at MS centers using a common vocabulary defined by The International Classification of Impairments, Disabilities and Handicaps (ICIDH)[172] from the World Health Organization (WHO), an early version of the ICF. These measures would include domains beyond simply body function impairment. The three proposed measures were the DSS, the Incapacity Status Scale (ISS), and the Environmental Status Scale (ESS) to measure impairment (now body function impairment), disabilities (now activity limitations), and handicaps (now participation restrictions) respectively. Dr. Kurtzke did not fully endorse all the proposed measures partly because he felt the ESS constructs were considered too distant from the illness and potentially influenced by too many extraneous factors to be of any practical value as a measurement of treatment effects in clinical trials.[162] He also felt that the ISS was too nonspecific, redundant to the neurological assessment and had a poor scoring system.[162] Kurtzke only endorsed the disability scale based on the neurological exam he developed, the DSS.[34,162] He does state that measuring the degree of involvement in their patients does not come naturally to neurologist. They are generally more concerned with establishing the presence of particular neurologic deficits and obtaining a diagnosis.[162] These statements suggest an adherence of neurologists to the biomedical model at the time and *possibly* not their full understanding of the role of the biopsychosocial model. Other possible issues were that the definitions and the classification of impairments in this early version were difficult or ambiguous to apply.[173] One of the criticisms was that the ICIDH appeared to offer a unidirectional causal model linking the health conditions and impairments, and then to disabilities and handicaps.[65,111] Importantly, contextual factors were not part of this early model ICF.[172]

Relapses:

As discussed earlier, relapses are the clinical indicators of disease activity. Relapse severity is assessed with the EDSS. Patients that experience frequent and/or severe relapses are candidates for DMT to control disease activity to decrease the risk of long-term disease progression. Acute symptoms can be treated in the short with corticosteroids. Chapter 3 also summarizes, in detail, the definition of a relapse and the preference for neurological signs determined by a neurologist examination over patient reported symptoms. The importance of relapses in disease progression is briefly discussed in the introduction and in more detail in chapter 6 (manuscript 1).

MRI:

MS disability observed from a relapse is considered disease activity. MRI has allowed the detection of disease activity without the patient experiencing any symptoms.[33] MRI of the brain and spinal cord is now used to establish a diagnosis of MS and can be used to monitor disease activity.[174] This technology has led to improvements in diagnosis time from years to months.[18] MRI of the CNS allows for the identification and measurement of MS lesions in an objective and quantitative manner.[12] The inflammatory activity and myelin damage caused by MS to the CNS has been estimated by lesion counts (using T2-weighted scans) or disease burden (using T1-weighted scans).[174] Gadolinium (Gd)-enhancing T1-weighted lesions are believed to depict immune cells migrating across the blood-brain barrier to cause active MS inflammation.[33]

MRI has been used as endpoints in phase II and III studies of disease-modifying therapies. However MRI metrics have correlated poorly with clinical status as measured by the EDSS or as a predictor of disease progression.[175,176] Gd-enhancing T1-weighted scans appears to detects 5-10 times more activity than clinical observation (relapses).[33] Ninety percent of Gd-enhancing lesions are not associated with identifiable signs or symptoms.[32] Zivadinov et al., concluded that conventional MRI such as Gd-enhancing T1 and T2 lesions have only limited value for predicating clinical status in MS due to their poor sensitivity and

specificity for the underlying pathophysiologic process and feel that newer techniques may be better.[177] These newer techniques have only been used in a research setting.[177]

While the EDSS and MRI are very important tools to monitor disease activity for the treating neurologist, they may not represent the patients' interests.[178] Researchers have suggested perhaps it is time to redefine "function" by including the patient perspective and domains of MS disability beyond the narrow representation by the EDSS.[178] Perhaps it is not surprising that EDSS or MRI often poorly correlate with quality of life measures.[178]

The biopsychosocial view of MS disability

Common symptoms in MS are sensory disturbance, limb weakness, pain, bladder & bowel dysfunction, vision problems, fatigue, muscle spasticity and cognitive dysfunction.[179] However, depending on the extent and location of damage to the CNS, any number of imaginable disabilities may occur. Fatigue is the most common symptom experienced by people with MS.[180] It is more severe than fatigue experienced by healthy individuals and impacts activities of daily living (ADL).[181] Difficulties or dependence in self-care, mobility, and domestic life were also predicted by impairment in fine hand motor task (dominant hand), balance, gait speed, and walking distance.[182] Rao et al., found that people with MS having cognitive dysfunction were less likely to be working and have fewer social and activities or hobbies.[183] Severe MS disability can impact on the standard of living and psychological well-being of people with MS and their families.[184]

The goal of rehabilitation is to reduce the burden due to the health condition and maintain optimal functioning of people with MS.[185] To achieve this, a comprehensive assessment of the impact of MS on the individual must be completed before a plan of action can be developed.[186] As part of the rehabilitation process there is planned follow ups to evaluate the effectiveness of the treatment plan.[186] To fully describe MS disability, its impact on the body, individual, and society needs to be included.

Most recently, there has been work to develop an ICF Core Sets for MS. A core set is a list of agreed upon ICF categories most relevant to patients with a particular health condition (MS) to specify functioning.[83,187] A MS Core Set would be a minimal dataset of ICF coded

items necessary to adequately describe MS and sufficient enough to account for the majority of the disability associated with the disease.[188] The process leading to the approval of the ICF Core Sets for MS is considered an evidence-based and formal decision-making consensus process integrating research knowledge and expert opinion.[96] There are a total of seven publications on MS ICF core sets and checklists[96-98,187,189-191] identified categories to reflect the different perspectives of patients, physiotherapist, clinicians, and other health care professionals. A summary of all identified ICF categories for MS are 139 body function impairments, 21 body structures, 111 activity limitations and participation restrictions, and 77 environmental factors (Table 1a in chapter 11 (manuscript 3) of the appendix).

This work provides a globally agreed upon framework and system for comprehensively classifying the typical spectrum of functioning and disability of persons with MS placed in the environmental context in which they live. Under the ICF model, MS disability extends beyond body structure and body function impairment to include activity (execution of a task or action by an individual) limitations and participation (involvement in a life situation) restrictions.[80] It is intended as an international standard to aid in decisions on what to measure and report (not how to measure it) and to help in the assessment, interpretation and grouping of data for any health information in any setting.[96] The ICF Core Sets for MS can serve as a valuable and practical tool based on a universal language understood by health providers, researcher, and patients alike.[96] This tool may be important given that it has been argued that a comprehensive perspective on functioning and disability is important in MS care and research especially in the multidisciplinary area of rehabilitation.[178,192,193]

Future of ICF model in MS research

The ICF and especially having a ICF Core Sets for MS can be used for content validity and will benefit future measurement development serving as a guide to help select measurement items from existing measures or creating new measures that match the appropriate ICF domains and categories.[83]

Once content validity has been established for a measurement tool, the next step is to test for construct validity. Using a modern psychometric method such as Rasch analysis, which requires the assumption of unidimensionality can aid in establishing construct validity.[194,195] The ICF appears to compliment this method. For example, the ICF core sets for a musculoskeletal condition was used to selected items for mobility of upper and lower extremity and Rasch analysis was used to create a measure for that single construct.[195] Johansson et al., suggested in their study assessing a fatigue measure for MS, that has content validity, should undergo Rasch analysis to further develop the psychometric properties of the measure.[196] The goal is to develop a psychometrically sound measure of a construct within the ICF framework specific to a health condition. Rasch analysis will be discussed further in chapter 7.

Rehabilitation clinicians and researchers have a large inventory of measurement tools covering all ICF domains. These tools, also called rating scales as outcome measures are used to measure latent (unobservable) constructs such as disability.[197] Recently, a taskforce from the American Physical Therapy Association identified 120 outcome measures (generic and disease-specific) that have been used to assess disability in people with MS.[198] The taskforce recognized that the sheer number of indices available would be a barrier to clinicians and researchers selecting the appropriate and the number of outcome measures to comprehensively assess the diverse disabilities people with MS can experience.

These measures are comprised of patient reported outcomes (PROs), performance-based outcomes (PerfOs), and clinician reported outcomes (ClinROs). Each of these assessments are defined by the U.S. Food and Drug Administration[199] as the following:

Patient-reported outcome (PRO)— A PRO is a measurement based on a report that comes from the patient (i.e., study subject) about the status of a patient's health condition without amendment or interpretation of the patient's report by a clinician or anyone else. A PRO can be measured by self-report or by interview, provided that the interviewer records only the patient's response. Symptoms or other unobservable concepts known only to the patient (e.g., pain severity or nausea) can only be measured by PRO measures. PROs can also assess the patient perspective on functioning or activities that may also be observable by others.

Clinician-reported outcome (ClinRO) — A ClinRO is based on a report that comes from a trained health-care professional after observation of a patient's health condition. A ClinRO measure involves a clinical judgment or interpretation of the observable signs, behaviors, or other physical manifestations thought to be related to a disease or condition. ClinRO measures cannot directly assess symptoms that are known only to the patient (e.g., pain intensity).

Performance outcome (PerfO) — A PerfO is a measurement based on a task(s) performed by a patient according to instructions that is administered by a health care professional. Performance outcomes require patient cooperation and motivation. These include measures of gait speed (e.g., timed 25 foot walk test), memory recall, or other cognitive testing (e.g., digit symbol substitution test).

Disability at the level of body function impairment includes symptoms and can be measured from the perspective of the patient using PROs. Body function impairments such as signs (disabilities) related to the CNS and muscle are usually assess by a trained clinician using ClinROs.

However, there may be differences between PRO and ClinRO assessments of body function impairment. When a PRO (the Multiple Sclerosis Impact Scale-physical component) and ClinRO, the EDSS were used to assess MS disease progression, 33% of patients identified as worsened only in the PRO.[200] When a self-reported questionnaire was compared to a physician's assessment of the patients disease course, people with MS were more likely to classify themselves as progressive compared to physician evaluators.[201] Others have shown that the perspective of people with MS differ to those of physicians on the relative importance of the eight domains of the SF36.[202] Clinicians were more concerned about the physical aspects of the disease whereas patients were more concerned about mental health and vitality.[202]

At the level of activities limitations and participation restrictions assessment can be made by using PROs or PerfOs. PROs measures are easy to administer and capable of evaluating several aspect of disability in a single test. However, responses can be influenced by cognitive function[203,204] and the willingness of the patient to answer the question accurately (social desirability bias).[205] For self-reports of ADL for example, there has been two main criticisms. First, people have trouble judging their own competency accurately and second assessing what a person think they can do (self-report) does not provide any

information on what they can actually do (task).[206] In terms of PerfOs, they do provide an actual measure of the person's capacity. PerfOs are less influenced by cognitive function, culture, language, or education. However, each task will only assess a single attribute of the domain (ADL in this case)[204] thus several PerfOs might be needed to more comprehensively assess a patient. For example, a typical attribute for a PerfO to assess lower extremity is to record time. Other attributes that may need to be assessed are distance, speed, endurance, and strength thus requiring additional tasks.[204] Logistically PerfOs cannot always be done due to time, safety considerations, space restrictions, or the person's physical and/or medical condition.[206] More interestingly, it appears that self-reports and performance based ADLs are only weakly correlated thus indicating that these two methods of assessment of ADL are complimentary and assessing different attributes of the ADL domain.[182,206]

In rehabilitation medicine patients are comprehensively assessed in multiple domains of disability. Currently, this requires multiple outcome measures with the choice of PROs, PerfOs, and ClinROs. The drawback is response burden for the patient and time commitment from the clinician.

Chapter 5

Statistical methods applied in MS disability

From the literature review it is evident that a number of changes have occurred in MS diagnosis and management. There has been new diagnostic criteria incorporating the use of MRI decreasing time to the diagnosis of MS and the development of several approved drugs all targeting disease activity, specifically relapses. These secular changes are likely to have impacted on MS disease course over time. The statistical methods to estimate secular changes are reviewed in the next chapter. The subsequent chapter presents a manuscript describing secular changes in MS disease course for men and women over three eras using a relatively new approach to model longitudinal change, not previously used in MS, Group Based Trajectory Model (GBTM).[207]

The choice of method for any analysis is the measurement scale of the outcome variable. Most medical research uses *events* such as mortality, disease occurrences, or change in health state as outcomes. Statistically these can be quantified as binary (present/absent) or as time-to-event. When outcomes are binary some form of logistic regression can be used. When time-to-event is the outcome survival analysis is used.

In fact, the New England Journal of Medicine, the most highly ranked of the biomedical journals, survival analysis is the most commonly reported method for analyzing longitudinal data.[208] So much so that in many studies outcomes that are not naturally binary are converted to time-to-event. This is the case in MS research using disability as an outcome. Disability is a construct measured on a continuous scale but often converted to time to reach a specific disability milestone.

Survival analysis using Kaplan-Meier (KM) survival curves and/or Cox regression has been well established in clinical trials.[209] The RCTs that led to currently approved MS drugs in Canada all included time to event analysis.[27-30,47,49-54,130,131] Survival analysis was used to estimate whether the new drug delayed the onset of a new relapse and/or delayed the time to a certain disability level compared to a placebo.[210-213]

Longitudinal studies using registries are motivated by the need to better understand MS disease course over time. Survival analyses are used in MS cohort/database registry studies to estimate time to disability milestones such as EDSS 4 (limit of fully ambulatory), 6 (need for a cane), or 7 (need of a wheelchair).[8,36,141,144-146,150,214-218] Some have recommended that survival analysis should be the standard analysis for MS registries to generate comparable (time to specific disability milestones) results.[129]

The optimal use of survival analysis

Survival analysis is optimized for endpoints that are "absorbing states," such as death, where a transition to the state can only occur once, and for situations where the state prior to reaching the endpoint is uninformative. The accuracy of the time-to-event estimates is affected by measurement error on the time of crossing the threshold, and also on the supposition that events are irreversible when they may not be. If for example, when the event is time to the use of a cane or wheelchair (as mentioned above) then the survival time estimate will naturally have an additional level of uncertainty. There are similar uncertainties when using endpoints such as confirmed or sustained progression (which can mean a 0.5 or 1 point EDSS change over 3 or 6 months) in shorter duration studies. It has been estimated that a large proportion of patients (~40-50%) have been shown to regress back to baseline EDSS when followed for a longer period of time.[219,220] Another issue is that the EDSS is an ordinal outcome such that a one point change at different EDSS levels do not necessarily represent similar change.

A limitation when using survival analysis is that all subject must complete the study follow up time to have an accurate estimate of mean survival time, if subjects are censored then information is lost and will effect the accuracy of the result.[221,222] Additionally, as more subjects reach the event or are censored fewer subjects remain at the tail end of the curve. This makes survival estimates at the beginning of the curve more reliable.[222]

Outside the framework of time to event analysis other methods are available to analyze longitudinal change. In MS or any chronic disease where people are measured repeatedly over time other methods are needed to analyze longitudinal change.

Clearly to model disability it would be best to use statistical methods that do not depend on transforming the data to fit a specific statistical model but rather use the data as they come. The challenge with using non-binary data over time is when there are multiple time points as the "growth" in outcome needs to be modeled rather than absolute change, recognizing that growth can be monotonically increasing or decreasing or more rarely following a non-monotonic pattern. A critical feature of longitudinal data is that values at one time point are correlated with values at other time points and this correlation structure needs to be considered in the analysis. In addition, data can be missing at one or more time points.

The details of these methods are beyond the scope of this thesis. Two popular regression models that deal with correlated data structure and can handle missing data[223] [224,225] are generalized estimating equations (GEE)[226,227] and mixed effects model[228]. GEE and mixed effects model can model dependency and can be used for longitudinal and clustered data.[226,227,229] GEE and mixed effects model methods can include time-varying predictors and time-invariant predictors.[226,229,230] GEE and mixed effects model use all available data and can handle data missing completely at random (MCAR).[223,225,231] GEE estimates with missing data at random (MAR) did not perform as well as methods using maximum likelihood (ML) estimation of mixed effects model.[232] A major difference between GEE and mixed effects model is the way missing data are treated.[225] ML estimation can treat MCAR and MAR as *ignorable response mechanism* where as data MAR in GEE is not since a quasi-likelihood estimation method is used.[233] GEE's robustness of not needing to know the correlation structure becomes a problem with MAR. In this situation the working correlation structure needs to be the true correlation structure.[233]

When growth over time is the parameter of interest mixed effects models are the optimal model. There are several advantages in using mixed effects models to analysis longitudinal data; they are very flexible, allow for a tailored structure of the correlation over time and across person to correct for dependency,[230] they uses the ML estimation method,[234] and regression coefficients can vary between individuals. Data are truly modeled at the individual level, allowing one to exam the individual variability of the intercepts and slopes.[235] GEE and mixed effects model addresses many of the challenges of modeling longitudinal data that

might occur in any chronic disease or from longitudinal registry such as those for MS.

Mixed effects models are an appropriate choice to analyze longitudinal change when data can be represented by an average trajectory and individual differences can be captured by estimating a random coefficient (random intercept and/or random slope) to represent the variability surrounding the average intercept and average slope.[236] However, if one suspects that one average trajectory cannot be used to represent all individuals in the sample then GBTM may provide a better method to model longitudinal change.[236]

Group-Based Trajectory Modeling

GBTM is designed to identity clusters of individuals, called trajectory groups, who follow a similar developmental trajectory on an outcome of interest.[236] GBTM are based on finite mixture models (FMM)[237] a class of statistical models designed to analyze data composed of a mixture of two or more groups whose outcome are generated by distinct statistical processes.[236] FMM are an extension of the ML model. The likelihood function is flexible enough to accommodate different forms of data such as censored normal, count, and binary data. Thus, GBTM can handle normal, and Poisson and binary & logit distributions.[207] The specific form of the GBTM depends on the type of data being analyzed. The shape of each trajectory group depends on the distribution of the data type and the parameters of the polynomial function of age (or time) associated with it. A separate set of parameters is estimated for each group so that the shape can be different for each. As a result the model allows the trajectory shapes to vary across groups. This is a key feature of GBTM.[207,238] An important use of FMM is to analyze data from a population that is thought to be composed of subpopulations that are not identifiable from measured characteristics *ex-ante.*[236] If two groups were distinguishable based on measured characteristics, they would simply be analyzed separately using a mixed effect model.[236]

GBTM assumes that the population is composed of a mixture of distinct groups defined by their developmental trajectories.[207,236] GBTMs use a non-parametric ML estimator for the distribution of unobserved individual differences by approximating the distribution with FMM. The idea is that a finite number of groups will be used to approximate a continuous

distribution.[207,236,238] Rather than assuming that the population distribution of trajectories varies continuously across individuals and in a fashion that can ultimately be explained by a multivariate normal distribution of population parameters, it assumes that there may be clusters or groupings of distinctive developmental trajectories that themselves may reflect distinctive etiologies and these clusters or groups of distinctive developmental trajectories can approximate the actual continuous distribution by using FMM.[207,236,238] GBTM does not allow for variation within the latent class by not including the random effects in each group's trajectory model.

Advantages of GBTM: GBTM is capable of identifying qualitatively distinct trajectories that are not identifiable using classification rules *a priori*. It is able to distinguish real differences across individuals from those due to chance by using a formal statistical structure.[236] GBTM does not make any assumptions on the population distribution of trajectories and instead uses the trajectory groups as a statistical device to approximate the unknown distribution of trajectories across the population members.[236]

One of the key decisions in GBTM is to determine the number of groups that should be used to represent the different developmental trajectories. In contrast to GBTM, developmental trajectories are often identified using assignment rules based on subjective categorization to construct the categories.[236] One of the limitations of such a process is that the existence of distinct development trajectories must be assumed *a priori*.[207,236] It is also difficult to know how well these categories (classifications) are represented. One does not know how well an individual's trajectory actually fits into the group classification.

Researchers must decide on the number of groups to be extracted from the data, several statistical methods can be employed as indices for goodness of fit. These fit indices will test how right (or less wrong) the selection of the number of groups "fit" compared to more and to less number of groups. These fit indices are the Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC), Lo-Mendell-Rubin Likelihood Ratio Test (LMR-LRT), or entropy score.[207] However, caution must be taken in making a model selection based only on goodness of fit indices. Depending on the question being asked and available data, substantive knowledge of the subject matter should always be part of the decision in

selecting the appropriate model. One must not simply rely on a test statistic.[207,236] One does not want to end up with a best fitting model statistically that is inadequate in answering the research question.

The strength of GBTM, is that it provides several statistical criteria for assessing model adequacy. GBTM uses a set of probabilities calculations known as the "posterior probabilities of group membership" (PPGM). Based on the model coefficient estimates of the individual's longitudinal pattern on the outcome, each individual's probability of membership in each group is calculated.[239] They are called *posterior* probabilities (PP) because they are calculated after the model estimation using the model's estimated coefficient. These probabilities measure a specific individual's likelihood of belonging to each of the models trajectory groups and assess the quality of the model's fit to the data. PPGM is different from probability of group membership (PGM). PGM measures the proportion of the population that belongs to a specific group. It can be thought of, as the probability of a randomly selected individual will follow a specific group trajectory. In contrast, PPGM measures the probability that an individual with a specific measured profile belongs to a specific trajectory group, providing a valuable source of information. Individuals assigned to each group should have an average PPGM of greater than a minimum threshold of 0.7. Group assignment is probabilistic not deterministic.

When using GBTM, one must keep in mind that "group membership" is a convenient statistical approximation and individuals do not actually belong to trajectory groups. The group trajectory is intended to capture a long-term behavioral pattern (weighted average), not short-term individual variability about that pattern. The number of groups and the shape of each group's trajectory are not fixed realities. A trajectory group is a cluster of individuals following a similar trajectory, with more waves of data the cluster may split into more groups. Sample sizes also influence the number of trajectories. Groups are not immutable.[238]

The manuscript in the following chapter is an illustration of using GBTM to model disease course over three time periods.

Chapter 6 (Manuscript 1)

Trajectory of MS disease course for men and women over three eras

Stanley Hum[1], Yves Lapierre[2], Susan C. Scott[3], Pierre. Duquette[4], Nancy E. Mayo[1,3]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]The Montreal Neurological Institute, McGill University Health Center, Montreal, QC, Canada

[3]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

[4]Neurologie, Centre hospitalier de l'Université de Montréal, Montréal, QC, Canada

In preparation for submission to *Multiple Sclerosis Journal*

Communication addressed to:

Stanley Hum, M.Sc., PhD. Candidate
School of Physical and Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
514-398-5981
Email: stanley.hum@mcgill.ca

ABSTRACT

Multiple Sclerosis (MS) disease progression is often measured with time-to-event endpoints. Group Based Trajectory Model (GBTM) is a relatively new statistical approach not previously used in MS that is available to model longitudinal change. It provides a means to describe and explain variability in disease progression.

The objective is to estimate disease course heterogeneity over three distinct MS onset periods: (1) pre-magnetic resonance imaging (MRI) and disease modifying drugs (DMTs) (<1995); (2) MRI+1st generation DMTs (1995-2004); and (3) MRI+2nd generation DMTs (2005-present). A secondary objective is to estimate the extent to which annualized relapse rate (ARR) contributes to disease course.

The data are from the Montreal Neurological Institute MS Clinic longitudinal database established during the 1980s. GBTM, a specialized mixture model, estimates clusters of individuals following similar developmental trajectories on an outcome of interest within the population. A GBTM described disease progression for three inception cohorts: pre-1995 onset; onset between 1995-2004; and 2005 onward. Secular and gender contrasts were made on the proportion of patients with stable and unstable disability trajectories. Stable trajectories were defined as ≤ one EDSS point change and having an EDSS ≤ 3.0 over the study period.

Percentage of women in each cohort was 73.0%, 71.1% and 71.0%, respectively. Cohort sizes were 237, 648, 567 respectively. Average onset age ranged from 32 to 36. For women, the number of trajectories were 4, 7 and 6 for the three cohorts, respectively; for men these numbers were 4, 6, and 5. The proportion of women classified as stable was 0% pre-1995, 69.0% (CI: 61.3-76.8) for 1995-2004, and 83.9% (CI: 74.2-93.6) post-2005; for men, these proportions were 18.4% (CI: 5.2-31.8), 41.4% (CI: 31.63-51.2), and 53.8% (CI: 43.1-65.4) respectively. The proportion of men with stable disease was significantly lower than women only in both post-1995 cohorts (Chi-square tests: p<0.0001).

Odds ratios with 95% confidence intervals (95% CI) were calculated for each trajectory with reference to the "best" lowest trajectory. For women in the pre-1995 cohort, there

were no associations between ARR and trajectory; this null association was also true for men. For both post-1995 cohorts, all odds ratios were > 1.0 with 95% CI above 1.0 except for one trajectory group for men in the 1995-2004 and in the post-2005 groups. These groups had small sample sizes (12 and 13) and a large proportion of progressive patients (75% and 54%) respectively.

GBTM is a valuable tool to describe longitudinal data showing the variability in disease course of people with MS under real-life management strategies over three distinct onset periods. It is encouraging to observe large proportions of patients remaining stable at their initial disability level for at least 15 years. Higher ARR within the first five years of disease increases the odds of a patient being in a higher disability trajectory. Progressive MS patients are more likely to have a poorer prognosis. Women have milder disease course than men.

INTRODUCTION

Disease course in MS is characterized, for the most part, by periods of exacerbation and quiescence, so much so that early observers thought that each patient followed a unique disease course.[1] The different MS subtypes were formally classified in 1996.[2] Recognizing that there is additional heterogeneity, these classifications have recently been refined based on new clinical, imaging, and biomarker advances.[3] Although the core descriptions of relapsing remitting MS (RRMS), secondary progressive MS (SPMS), and primary progressive MS (PPMS) are retained as the principal descriptors of different MS disease course phenotypes, important distinctions are now made for more active and less active disease. However, specific criteria for "active" disease are lacking.[3]

More precise estimates of disease course heterogeneity are informative for clinicians and desirable for patients anxious about their long-term prognosis. Information derived from historical natural history studies that involve modeling what happens to people over time in the absence of treatment is the current standard to projecting MS progression in newly diagnosed people.[4] However, with the introduction of disease modifying therapies (DMTs), the "natural history of MS" is unknowable, and the existing knowledgebase has essentially been relegated to the role of a historical reference.

These historical natural history studies generally have revealed that male sex, older age at onset, shorter time to SPMS, shorter inter-attack interval, high relapse rate in first years, short time to an Expanded Disability Status Scale (EDSS) score of 3, and progressive disease course are predictors of poorer long-term prognosis.[5-7] Historically MS has been viewed as a rapidly progressing disease with patients eventually requiring a cane to ambulate, or a wheelchair. The best data to understand disease progression comes from well-defined cohorts that have been assembled and followed systematically. Figure 1 is a graphical illustration of key cohorts that provide data on time to progression, in this case, time to EDSS 6 (need for a cane to ambulate). The cohorts are presented chronologically and separated into prevalent and inception cohorts. Prevalent cohorts register all people in view regardless of date of onset and follow from date of registration; inception cohorts include all people from a common onset time, which can be an event, such as a stroke, or in

the case of MS, the reported date of symptom onset, recorded at the first (usually neurological) visit.

The two inception cohorts provided estimates of 9.4 years (London, Ontario; ending 1984) and 14 years (Florence, Italy; ending 1996) until time to EDSS 6. For prevalent cohorts, there is a trend towards longer time to EDSS 6 with later cohorts. The earliest cohorts estimated median time as 14.4 years with time frame 1980-1998,[8] 15 years with time 1979-1984[9], and 20 years with time frame 1976-1997,[10] later cohorts were generally greater than this, ranging from 18.6 years in Nova Scotia with time frame 1998-2004,[8], 20 years with time frame 1996-2008, 24 years with time frame 1991-2000,[11] and 28 years in British Columbia with a time frame spanning 1988-2003.[8,11-13]

Although much has been learned about disease progression from these historical natural history cohorts, they typically employ time-to-event analysis. With the current use of database registries to record patients' clinic visit data over time, other statistical methods that can use all the data on hand would provide more information about disability course than simply when a person crosses a specific disability threshold. Survival analysis is optimized for endpoints that are "absorbing states" such as death and when the state prior to reaching the endpoint is uninformative. Typical endpoints or disability thresholds are at EDSS 4 (limit of fully ambulatory), 6 (need for a cane), and 7 (need for wheelchair). Accuracy of time-to-event estimates is affected by measurement error on the time of crossing the threshold and also on the supposition that events are irreversible when they may not be. For example, there is additional uncertainty when using endpoints such as confirmed or sustained progression (which can mean a 0.5 or 1 point EDSS change over 3 or 6 months) in shorter duration studies. It has been estimated that a large proportion of patients (~40-50%) have been shown to regress back to baseline EDSS when followed for a longer period of time.[14,15]

Over time, there have been advances in diagnosis and treatment of MS. Pre-1995, without any effective DMTs, clinical care was limited to managing acute relapses and persistent symptoms.[16] The diagnosis of MS was based on a patient experiencing two attacks with clinical evidence of lesions in different locations at least one month apart.[17] Between 1995-

2004, four DMTs with partial efficacy were approved to treat MS and (magnetic resonance imaging) MRI was used to support the diagnosis of MS. Currently, 2005-2014, there are now ten different DMTs approved based on efficacy in decreasing relapse frequency and MRI activity. The McDonald criteria were developed in 2001 to formally include MRI results in the diagnosis of MS. These were updated in 2005[18] and in 2010[19]. Based on these changes to the definition and treatment of MS over time, we would expect to see corresponding changes in the disease course.

Just as there have been advances over time in diagnosis and treatment, there have also been advances in statistical methods to better describe longitudinal change. This study takes a new approach to describing longitudinal change in MS using Group Based Trajectory Modeling (GBTM). GBTM is designed to identify clusters of individuals, called trajectory groups, who follow a similar developmental trajectory on an outcome of interest.[20] A key feature of this method is that groups do not need to be identified *a priori*. This is important when subpopulations are thought to exist but are not identifiable from measured characteristics *ex-ante*.[20] Also, the trajectory shapes can vary across groups.[21] This method works by assigning individuals to groups such that variance is minimized within groups and maximized between them, thereby identifying distinct subpopulations.[22]

Additionally, optimal methodology to quantify the natural course of a health condition is to assemble a representative population of people with the target condition at time of onset, and follow this cohort ascertaining the key outcomes on all persons from inception to a time when the cohort has either died out or reached a plateau in progression, if plausible. In the case of MS, the past has revealed what happens to people over time in the absence of treatment. There is now sufficient evidence for short-term efficacy of disease modifying drugs such that current guidelines are recommending treatment of patients experiencing relapses.[23,24] Currently, there is a need to follow well-characterized inception cohorts to quantify the course of MS regardless of the treatment approach.

OBJECTIVE

The purpose of this study is to estimate the extent of heterogeneity in disease course over three distinct MS onset and treatment periods: (1) pre-MRI and DMTs (<1995); (2) MRI+1st generation DMTs (1995-2004); and (3) MRI+2nd generation DMTs (2005-present). A secondary objective is to estimate the extent to which annualized relapse rate (ARR), a common target of DMTs, contributes to disease course.

METHODS

The database at the MS Clinic of the Montreal Neurological Institute was started in the late 1980's to collect socio-demographic and neurological disability data on MS patients. Currently there are over 5000 patients registered, and 3000 actively followed by seven neurologists. The EDSS, relapse history, and DMT status are recorded in the database for each patient visit (every 6 to 8 months). All neurologists have all been trained to perform the EDSS according to the Neurostatus© guidelines which were developed to increase intra and inter rater reliability of the measure.[25]

Patients included in this study were required to have a recorded initial visit at the MS clinic within two years of disease onset in order to minimize information bias, particularly when estimating the date of disease onset, start of any DMTs, and the number of relapse experienced. Based on this criterion, three inception cohorts were defined to represent clinical practice during eras before and after the advent of DMTs and the use of MRI for the diagnosis of MS: pre-1995 onset; onset between 1995-2004; and 2005 onward. The latest cohort represents the era where diagnostic criteria were refined and 2nd generation DMTs were available. All MS types were included in the analysis. EDSS score was the outcome variable. Disease duration was used as the time variable and was calculated from recorded MS onset date.

Data Analysis

Descriptive statistics were calculated to summarize demographic and clinical characteristics. All trajectory analyses were conducted using SAS® 9.3TS1M2 32bit version

(SAS Institute, Gary, N.C., USA) and the PROC TRAJ macro. GBTM was used to identify groups of individuals with similar longitudinal disease course. The EDSS was as the outcome and disease duration was the time variable. Choosing the optimal number of trajectories is an iterative process and here we used the forward classification approach, starting with three groups and adding additional groups until the best fit to the data was achieved.[22] Each trajectory is plotted with 95% confidence interval (CI) bounds.

GBTM provides several fit statistics to select the best model. The most appropriate shape of each trajectory is estimated by the posterior probability (PP) of group membership fit statistic and the final number of trajectories is selected based on the Bayesian Information Criterion (BIC) and clinical knowledge. A minimum PP of 0.70 is recommended for each trajectory and is interpreted as a 70% probability that the patients selected in the group belong to the trajectory. [20,21]

Each trajectory was then described by the patients' disease characteristics (ARR, age of onset, MS subtype, and proportion treated with DMTs). The proportion of stable and unstable disease progression trajectories for each cohort was summarized for men and women separately. Stable (S) MS trajectories (with relatively low levels of disability) were defined as having a change of no more than one EDSS point and having an EDSS ≤ 3.0 over the study period. Trajectories that were stable but with moderate or high levels of disability were labeled S_{MD} or S_{HD} respectively. Otherwise, trajectories were considered to be progressing (P) or improving (I) depending on their slope. Individual trajectories were then sorted and grouped by similar disease progression patterns. Based on the results observed in the proportion of stable trajectories for men and women, post-hoc chi-square analysis or Fisher's exact test, where indicated, were performed to test the difference in proportions between men and women.

To identify the role that ARR plays in defining a trajectory, ARR for the first 5 years after disease onset was calculated and the GBTM was refit with this as a fixed covariate. By convention, the onset relapse was not included in the calculation of ARR to avoid artificially inflating the magnitude of relapse rate estimates overtime.[26] Odds ratios (OR) for ARR were

estimated with respect to trajectory #1 (MS patients with the lowest level of disability) as the reference group. Confidence intervals (95%) were calculated for all odds ratios.

Sample size

In GBTM, sample size has two dimensions, the number of cases/individuals and the number of repeated measures of the same outcome (waves of data). Nagin estimated that the model was stable when the sample size was 300-500 people.[21] Van Dulmen et al, contributed evidence that sample size fewer than 250 were adequate for modeling relevant trajectories.[27]

RESULTS

Table 1 summarizes the demographic and clinical characteristics of 1452 patients for the three inceptions cohorts, overall and according to sex. The proportion of women in each inception cohort was 73.0%, 71.1% and 71.0%, respectively. Average age at recorded onset ranged from 32 to 36. Clinically Isolated Syndrome and RRMS patients were 62.5% of the pre-1995 inception cohort and > 80% in the two later cohorts. The mean ARR for women ranged from 0.3 to 0.36, across cohorts; for men, the ARR ranged from a high of 0.47 (SD: 0.83) in the pre-1995 cohort to a low of 0.21 (SD: 0.38) in the post-2005 cohort.

The trajectories produced for women and men are shown in Figure 2a-2f. Trajectories with varying starting EDSS and slope were revealed. For women, the number of trajectories were 4, 7 and 6 for the three inception cohorts, respectively; for men these numbers were 4, 6, and 5. Each trajectory had a somewhat different number of years in view. The proportion of people classified as belonging to stable trajectories, as defined by EDSS ≤ 3.0 and a change of no more than one point on EDSS classification over the study period, was calculated and summarized in Table 1. The proportion of women classified as stable was 0% pre-1995, 69.0% (CI: 61.3-76.8) for 1995-2004, and 83.9% (CI: 74.2-93.6) post-2005; for men, these proportions were 18.4% (CI: 5.2-31.8), 41.4% (CI: 31.63-51.2), and 53.8% (CI: 43.1-65.4) respectively. The proportion of men with stable disease was significantly lower than women only in both post-1995 cohorts (Chi-square tests: $p<0.0001$).

Figure 3 presents a descriptive classification for the different trajectory shapes for women and men, within each cohort and salient model parameters associated with each trajectory. The BIC indicated good fit for all models. As shown in figure 3, another indication of fit is the closeness of the actual percentage of patients assigned to a trajectory to the predicted percentage of patients, and the magnitude of the PP of trajectory assignment. The PP is optimally >70% and for the models here PP ranged from 81% to 100%. For women in the pre-1995 cohort, all trajectories were classified as progressive (P) although with differing rates of progression (slope) as previously shown in Figure 2a. For later cohorts, more stable trajectories (S) are evident although the degree of disability and cohort entry varied from stable low (S: EDSS ≤3) to stable moderate (S_{MD}: EDSS 3-3.5) to stable high (S_{HD}: EDSS: 5 or 6); see Figure 2c and 2e. For men, a similar pattern can be seen although with different proportions in the specific trajectories; see Figure 2d and 2f.

Table 2 summarizes the clinical characteristics of women and men assigned to the different trajectories which are labeled both by number to refer to Figure 2a to 2f and by shape to refer to Figure 3. For people in the pre-1995 cohort, the number of people assigned to each trajectory ranged from 18 to 78 for women and 7 to 30 for men. Also shown is the proportion of MS type assigned to each trajectory. ARR for women ranged from 0.07 to 0.43 and for men, the range of ARR was 0.38 to 0.58, with increasing ARR generally associated with assignment to a more disabled trajectory except for the highest disability trajectories, which had the lowest proportion of RRMS. On average people in this cohort were approximately 31 years of age at reported onset. For the two later cohorts, the proportion of people on DMTs is also included.

Table 3 presents the results of analyses linking ARR to trajectory. Each trajectory is treated as a categorical variable and the comparison is to the "best" trajectory, logistic regression was used and the parameter of association is the odds ratio (OR). For women in the pre-1995 cohort (n=173), there were no associations between ARR and trajectory; this null association was also true for men (n=64). For women in the 1995-2004 cohort, each trajectory describing high disability was associated with a higher ARR. For example, the OR for women in trajectory #2 was 1.16 (95% CI: 1.01-1.34) indicating that there is a 16%

increase in the odds of being assigned to trajectory #2 with every 0.1 increase in ARR. For men, there was also an increase in the odds of being assigned to a more disabled trajectory with higher ARR. Some of the ORs were greater 1.4. For the latter two cohorts, all but one of the ORs had 95% CI that excluded the null value of 1.0.

DISCUSSION

The results from this study show variability in the disease course of people with MS under real-life management strategies over three different critical time periods where improvements in diagnostic criteria and new treatment regimes were implemented to minimize damage caused by MS disease activity. The number of distinct trajectories varied from 4 in the pre-1995 era to between 5 and 7 in later eras (see Figure 2a-2f). In the pre-1995 cohort, among the 173 women, none had a stable trajectory over a 20-year period; for the 64 men, few (n=9) showed stable (or slightly improving) trajectory over a 15-year period. In later cohorts (1995-2004 and post-2005), the proportions of patients with stable trajectories were substantial: women [men], 69%[41%] and 84% [54%], respectively. The unexpected result in the pre-1995 cohort may be due to the small numbers representing men for this cohort. There was also a 5-year difference in the trajectories between men and women.

It is encouraging to observe for our inception cohort (MNI cohort) that a substantial proportion of patients in the two post-1995 cohorts remained stable at their initial disability level with proportionally more women showing stable disease course than men (see Figure 2c-2f and Table 1). Women had a milder disease course than men, as has been reported elsewhere[4,6].

For the latest period, post-2005, there is insufficient follow-up to conclude accurately about duration of stability. However, in the 1995-2004 cohort, 69% of women and 41% of men met a typical definition of "benign" MS. Historically, the term "benign" MS was used to describe people remaining with low disability. This term should always be estimated historically and used with caution since it has often been misunderstood and misused.[3] It does however, provide an indication of disease severity[3] and allows us to compare our

historical results on proportions with stable/low disability trajectories to published estimates.[28]

A review of studies reporting estimates of benign MS are summarized in Table 4, according to the definition of "benign" used. For comparison purposes, the estimates from this study were recalculated to match each published definition of benign MS and the information added to the table. These published estimates are derived from data collected pre-1995, except for one study. For the most common definition (EDSS ≤ 3.0 for 10 years), the estimates varied, noting that those from small studies (n~60) are imprecise.[29,30] For the four larger studies, only one estimate exceeded 20%. In our 1995-2004 onset cohort, 69% of women and 41% of men fit this definition, with women having consistently higher than published estimates and a majority of men with higher estimates.[28] The post-2005 cohort has not reached 10 years but appears to be on track for similar disease progression patterns.

Our approach is to use all persons in each era and let the observed data define the trajectories irrespective of what the future held for assignment of MS type, respecting the rigor of the inception cohort approach. The differences in disease course between RRMS and progressive MS have been reported and indicate that RRMS has a less aggressive trajectory.[6] After the best fitting trajectory model was developed, the distribution of MS type as defined at the end of the cohort period was examined within each trajectory.

Within the RRMS subtype there is important variation in the long-term prognosis. Persons assigned to stable/low disability trajectories (see Table 2) were almost exclusively RRMS but RRMS can be found in all but one of the other trajectories, in all cohorts. This suggests that having a stable or low disability trajectory is a near sufficient criterion for RRMS but it is not necessary.[31]

Progressive MS patients are more likely to have poorer prognosis. PPMS in our study cluster only into progressing (unstable) trajectories. Similarly, the majority of SPMS patients were in unstable trajectories. The few SPMS patients found in the stable trajectories may represent patients recently transitioned to a more progressive phase. We

suspect that with longer follow up these people will migrate to a more progressive trajectory. The study results are consistent with the RRMS subtype as having a better prognosis.[7]

Until a cure becomes available, keeping patients in the lowest possible disability trajectories is the role of MS therapies. It is encouraging to observe that a majority of patients in both post-1995 cohorts remained stable at their initial disability level. DMTs were introduced in this era, and have been shown to have clinical efficacy on decreasing relapse rate. Relapses early in the course of the disease appear to be associated with earlier disease progression.[32-34] Confavreux et al., found that early relapse rate influenced disease progression but only until EDSS four.[10] Another study found that early relapses impacted disease progression in the short term but had no long-term impact (> 10 years or if already in secondary progressive phase).[35] In this study, we estimated clinical disease activity early in the disease course by calculating the ARR during the first five years after MS onset. ARR was then linked to the trajectory providing an estimate of the impact of early disease activity on long-term disease progression.

The effect of DMT on trajectory was not directly estimated due to confounding by indication. This arises because patients doing well with the mildest and most stable disease, with relatively low ARR, are not immediately started on a DMT. This is shown in Table 2 where > 50% of people in trajectory #1 are treatment naïve. This result is compatible with a philosophy of having individualized treatment goals and interventions for people with MS. Patient treatment plans are often a negotiated balance between the advice and recommendation from their MS specialist and the achievable goals set by the patient and/or partner and/or family members.[36] Patients are likely to refuse treatment if they are doing well and may not be willing to undertake a treatment regimen that may be intrusive in their life. In support of confounding by indication is that we found that higher disability trajectories had higher proportions of patients on DMTs. (see Table 2)

Instead of modeling the impact of DMTs, we chose to model the impact of ARR as an indicator of the level of disease activity experienced by each patient. As all current drugs have efficacy in decreasing relapses, ARR incorporates the effect of DMTs. ARR for

treatment naïve patients would follow the natural history ARR and disease progression whereas patients on DMTs would presumably follow an altered ARR and disease course, reflecting the patient's response to therapy. ARR as a variable integrates the consequences of individualized patient treatment decisions and the resulting disease activity experienced over time and, therefore, would be an indicator of the "real-world" process behind the patient's observed disability trajectory.

The results shown in Table 3, indicate that having a higher ARR within the first five years from MS onset increases the odds of a patient being in a higher disability trajectory, relative to the odds of being in the best trajectory. Even patients with stable and mild disease course at EDSS~1 or 2 had higher ARR than the reference group (EDSS~0). In fact, ARR was associated with all higher disability trajectories (whether stable or not) as shown by an OR>1 and the 95%CI excluding the null value of 1.0. These results support previous reports of early relapses as a predictor of disease progression.[6,7]

Limitations

This analysis used historical data, the quality of which is affected by era. The application of the EDSS has changed during the period of the historical dataset, as modifications occurred up to 1983[37] and additional training programs were recommended after 1997.[25] The EDSS is known to be unreliable across observers, particularly outside of a trial protocol. In later eras, more trial protocols were available. Of course the EDSS does not measure all aspects of disability that are important to patients as it has a strong focus on neurological signs and walking.

The definition of a MS relapse has not changed substantively during the study periods; however, the interpretation and application of the definition may be more consistent in the more recent cohorts.[38] Recording practices with respect to a relapse could also have varied across clinician and era. For example, once the possibility of drug therapy emerged, the accurate documentation of relapses became more important. Documentation accuracy could be one reason why our ARR did not differ by era despite information from other sources indicating a decrease in ARR over time.[38]

We included patients with first clinic visit falling within one of the eras. However, as is typical, the date of disease onset is estimated by from the patient's self report of their first neurological symptoms attributable to MS, all medical history, and when available paraclinical test results. We selected patients with this onset date within 2 years of initial visit to minimize length bias sampling.[39] The uncertainty associated with date of onset is an inherent limitation in establishing the start of disease onset. There is evidence that the disease process occurs even before the first symptoms experienced by the patient.[40]

A limitation in comparing across eras, particularly including the early pre-1995 cohort, is lead-time bias[41], arising from change in diagnostic technology. The diagnosis of MS in the pre-1995 cohort was exclusively obtained using the Poser criteria[17], requiring two relapses each at different times and involving impairment in different neurological domains. The introduction of the McDonald criteria[18,19,42], which integrated the results of MRI into the diagnostic process, advanced the date of diagnosis with respect to earlier eras. In our 1995-2004 cohort, patients would have been diagnosed with the Poser criteria until 2001 and then a combination of Poser and McDonald criteria were used at the discretion of the neurologist afterwards. Estimates from United States based NARCOMS's registry showed that the time to diagnosis advanced from an average ~7 years in the 1980s to ~7.6 months after 2000.[43]

In our cohorts the mean time from onset to confirmed diagnosis did not decrease with the introduction of the McDonald criteria (data not shown). A possible reason is the availability of MRI machines estimated from pre-2005 show Canada (4.6 MRI units/million population) had less machines than in the USA (25.3 MRI units/million population) per capita.[44] The specific MRI wait times in 2007 specific to our institution was ~9 months for non-contrast MRI scans and ~12 months for contrast MRI scans.[45] However we cannot definitively discount the presence of lead-time bias in our estimates and purposefully did not use the pre-1995 cohort for comparison to the later cohorts.

Observing that a proportion of patients had very little disease progression over time even though they were not on DMTs is particularly interesting. Whether these patients will continue to do well in the long-term without treatment is still uncertain. The fact that

patients with high ARR can be stable at EDSS~1 or ~2 begs the question whether treating patients earlier (as has been advocated) with the potential of preventing one or more relapses can migrate patients to a lower disability trajectory.

The data presented here along with the availability of new and more potent therapies has brought about the discussion of possibly entering a era for the "New MS"[46] where patients can experience "NEDA" an acronym for "No Evidence of Disease Activity" as defined as free from: relapses, 3-month confirmed disability progression, gadolinium enhancing T1 lesions, and new or newly enlarged T2 lesions.[47] This may signal hope for patients newly diagnosed with MS, as a large proportion of patients could remain with low disability for many years as shown for the 1995-2004 cohort.

CONCLUSION

MS course is highly variable even within MS sub-types, but there is strong evidence that a large proportion of people with MS will not progress on the EDSS from their initial disability level over a period of 15 years. The results from this analysis support the newest MS classification to include descriptors of "more" and "less" active disease[3]. The trajectories indicate that people who remained at the entry EDSS level for 2 years were unlikely to progress. Future validation of this prognostic indicator would be warranted.

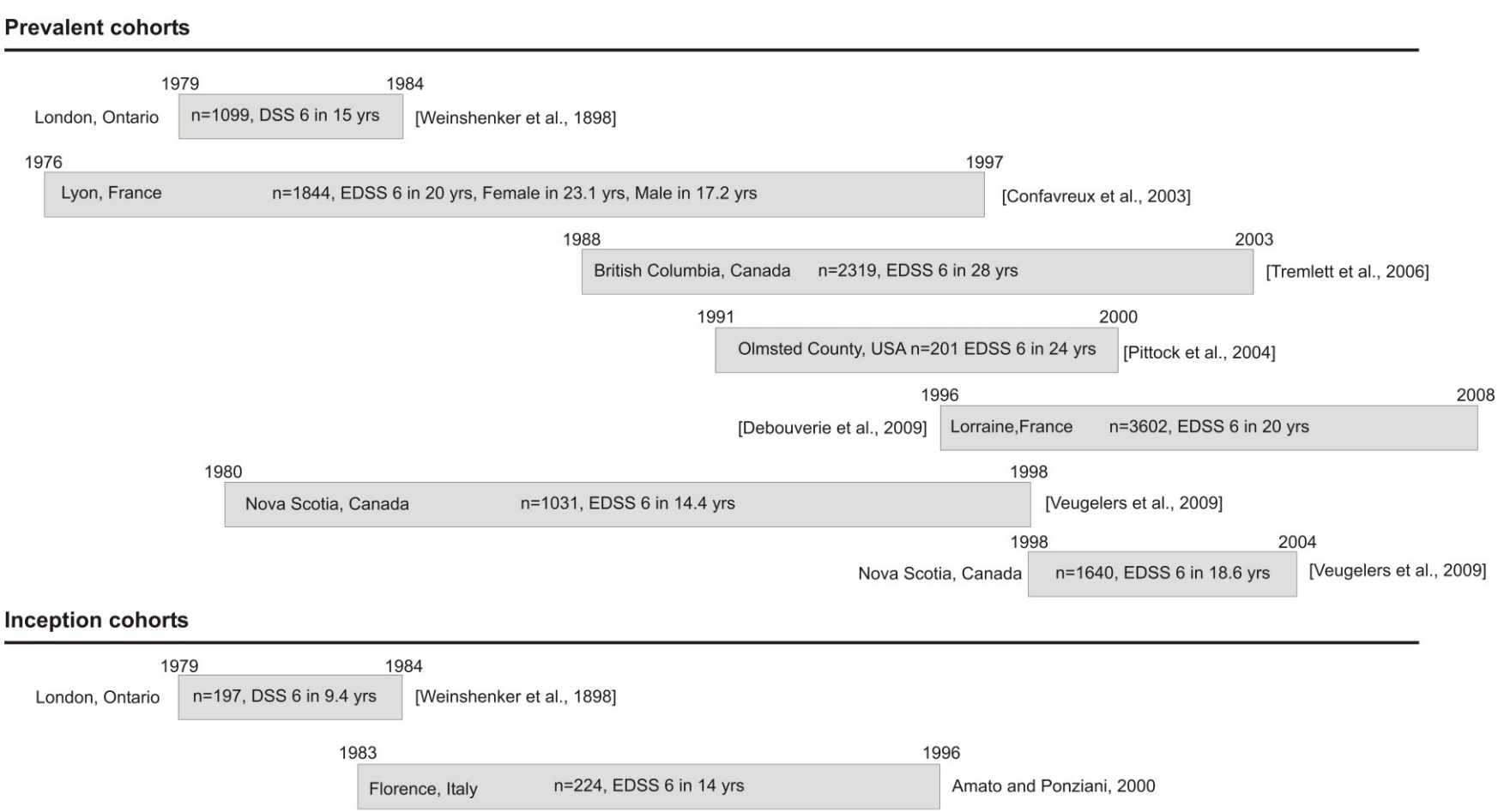


Figure 1. Median time to (E)DSS 6 for prevalent and inception MS cohorts

Table 1. Demographic and clinical characteristics of each inception cohort, overall and for women and men separately.

	Full sample			Women			Men		
Inception cohorts	< 1995	1995-2004	≥ 2005	< 1995	1995-2004	≥ 2005	< 1995	1995-2004	≥ 2005
N	237	648	567	173	461	404	64	187	163
Age at onset (SD)	31.8 (9.5)	35.0 (10.3)	36.1 (10.3)	31.6 (9.9)	35.0 (10.1)	35.7 (9.9)	32.1 (8.6)	34.8 (10.8)	37.1 (11.0)
MS type n (%)									
C.I.S	N/A	9 (1.4)	95 (16.8)	N/A	5 (1.1)	68 (16.8)	N/A	4 (2.1)	27 (16.6)
RR	148 (62.5)	513 (79.2)	438 (77.2)	111 (64.2)	389 (84.4)	318 (78.7)	37 (57.8)	124 (66.3)	120 (73.6)
SP	72 (30.4)	85 (13.1)	16 (2.8)	48 (27.7)	48 (10.4)	11 (2.7)	24 (37.5)	37 (19.8)	5 (3.1)
PP	11 (4.6)	32 (4.9)	14 (2.5)	8 (4.6)	12 (2.7)	4 (0.9)	3 (4.7)	20 (10.7)	10 (6.1)
Unknown	6 (2.5)	9 (1.4)	4 (0.07)	6 (3.5)	7 (1.5)	3 (0.7)	N/A	2 (1.1)	1 (0.6)
ARR for first 5 yrs (SD)	0.34 (0.61)	0.33 (0.45)	0.27 (0.66)	0.30 (0.50)	0.36 (0.45)	0.30 (0.74)	0.47 (0.83)	0.27 (0.43)	0.21 (0.38)
Summary of trajectory analysis									
Number of trajectories	---	---	---	4	7	6	4	6	5
Time period modeled (yrs)	---	---	---	14.1 to 20.5	13.9 to 15.3	8.2	15.3	13.2 to 14.4	7.1 to 9.2
Proportion of patients in stable trajectories (95% CI)	---	---	---	0 (0)	69 (61.3-76.8)	83.9 (74.2-93.6)	18.4 (5.2-31.8)	41.4 (31.6-51.2)	53.8 (43.1-65.4)

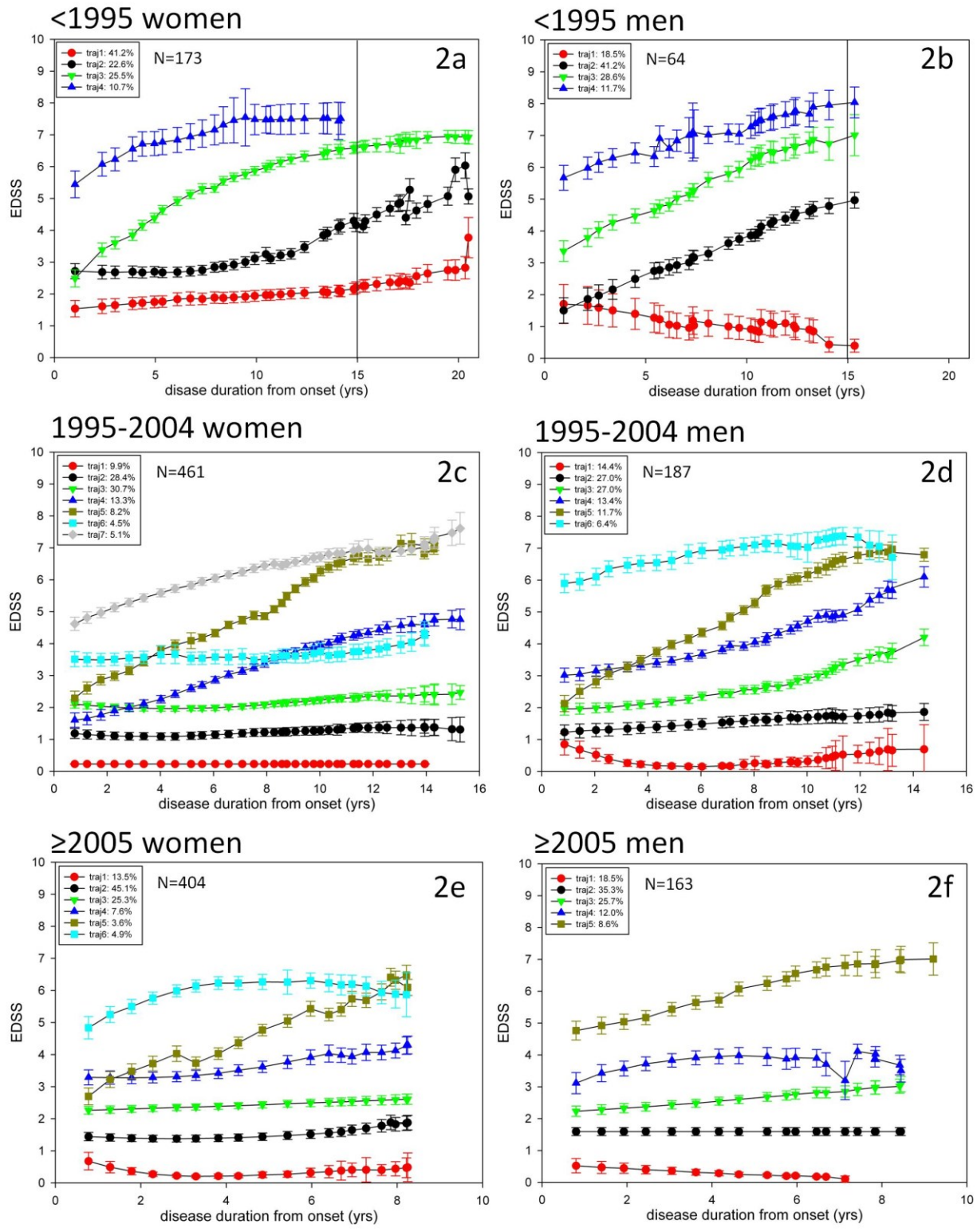


Figure 2a-2f. Group Based Trajectory Model plots by sex for each inception cohort

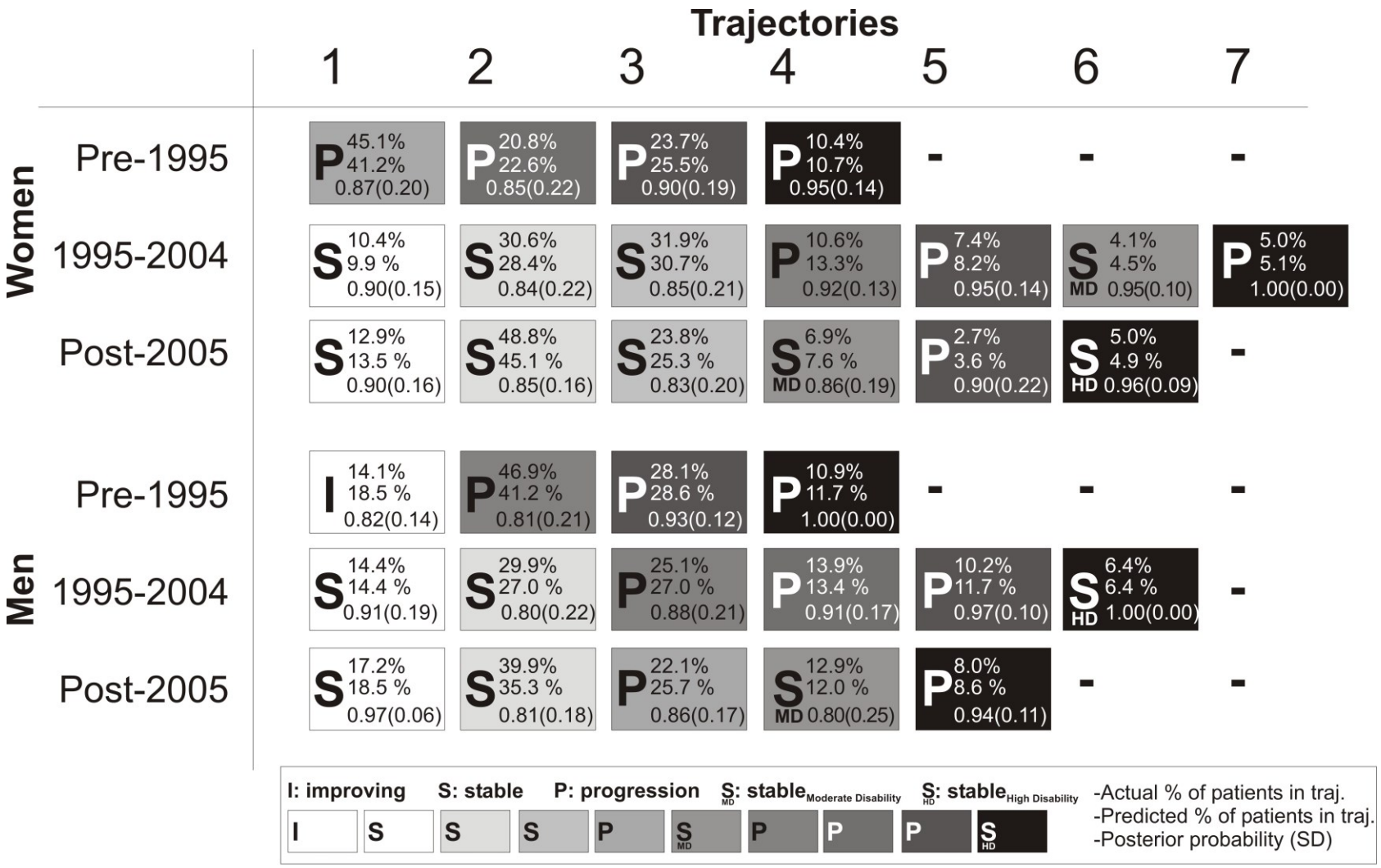


Figure 3. Classification of trajectories by shape and fit statistics

Table 2. Characteristics of MS patients assigned to trajectories by inception cohort and sex

		Women				Men			
		traj1 **(P)**	traj2 **(P)**	traj3 **(P)**	traj4 **(P)**	traj1 **(I)**	traj2 **(P)**	traj3 **(P)**	traj4 **(P)**
<1995	N	78	36	41	18	9	30	18	7
	MS type								
	CIS (%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	RR (%)	89.7	69.4	24.4	33.3	100.0	70.0	38.9	0
	SP (%)	5.1	22.2	63.4	55.6	0	30.0	50.0	85.7
	PP (%)	0	5.6	9.8	11.1	0	0	11.1	14.3
	Unknown (%)	5.1	2.8	2.4	0	0	0	0	0
	ARR (SD)	0.25 (0.06)	0.43 (0.09)	0.35 (0.08)	0.07 (0.05)	0.38 (0.17)	0.46 (0.15)	0.51 (0.23)	0.58 (0.31)
	Age onset (SD)	31.72 (8.59)	31.91 (10.56)	31.87 (11.56)	30.17 (10.37)	30.88 (7.47)	32.31 (7.87)	31.76 (7.58)	33.55 (15.09)

	95-04	Women							Men					
		traj1 **(S)**	traj2 **(S)**	traj3 **(S)**	traj4 **(P)**	tra5 **(P)**	traj6 **(S_{MD})**	traj7 **(P)**	traj1 **(S)**	traj2 **(S)**	traj3 **(P)**	traj4 **(P)**	tra5 **(P)**	traj6 **(S_{HD})**
1995-2004	N	48	141	147	49	34	19	23	27	56	47	26	19	12
	MS type													
	CIS (%)	2.1	1.4	1.4	0	0	0	0	7.4	1.8	2.1	0	0	0
	RR (%)	91.7	95.7	95.2	77.6	38.2	73.7	21.7	88.9	94.6	70.2	26.9	21.1	25.0
	SP (%)	0	1.4	2.0	20.4	47.1	15.8	60.9	0	3.6	21.3	46.2	52.6	25.0
	PP (%)	0	0	0	2.0	14.7	10.5	17.4	0	0.0	4.3	26.9	26.3	50.0
	Unknown (%)	6.3	1.4	1.4	0	0	0	0	3.7	0.0	2.1	0	0	0
	ARR (SD)	0.14 (0.04)	0.28 (0.03)	0.41 (0.04)	0.40 (0.06)	0.54 (0.10)	0.62 (0.10)	0.43 (0.09)	0.11 (0.05)	0.17 (0.03)	0.39 (0.08)	0.20 (0.07)	0.60 (0.15)	0.20 (0.08)
	Age onset (SD)	34.19 (9.28)	32.59 (9.72)	34.72 (10.04)	35.68 (9.14)	38.09 (10.99)	39.81 (1.34)	44.02 (8.20)	29.56 (8.36)	33.13 (10.33)	35.37 (11.43)	39.48 (9.53)	34.86 (12.64)	41.63 (8.73)
	0 DMT (%)	72.9	43.3	42.9	18.4	20.6	36.8	30.4	63.0	51.8	34.0	34.6	15.8	41.7
	1 DMT (%)	16.7	31.2	23.1	32.7	26.5	15.8	34.8	33.3	33.9	19.1	38.5	36.8	41.7
	≥ 2 DMTs (%)	10.4	25.5	34.0	49.0	52.9	47.4	34.8	3.7	14.3	46.8	26.9	47.4	16.7

		Women						Men				
		traj1 **(S)**	traj2 **(S)**	traj3 **(S)**	traj4 **(S_{MD})**	tra5 **(P)**	traj6 **(S_{HD})**	traj1 **(S)**	traj2 **(S)**	traj3 **(P)**	traj4 **(S_{MD})**	tra5 **(P)**
≥2005	N	52	197	96	28	11	20	28	65	36	21	13
	MS type											
	CIS (%)	19.2	22.8	9.4	3.6	0.0	15.0	17.9	23.1	8.3	9.5	15.4
	RR (%)	76.9	76.1	88.5	89.3	90.9	40.0	78.6	76.9	77.8	76.2	30.8
	SP (%)	0	0.5	2.1	3.6	9.1	30.0	0	0	5.6	4.8	15.4
	PP (%)	0	0	0	3.6	0	15.0	0	0	8.3	9.5	38.5
	Unknown (%)	3.8	0.5	0	0	0	0	3.6	0	0	0	0
	ARR (SD)	0.08 (0.02)	0.28 (0.07)	0.37 (0.05)	0.42 (0.08)	0.71 (0.18)	0.33 (0.11)	0.03 (0.01)	0.17 (0.04)	0.40 (0.09)	0.32 (0.09)	0.12 (0.07)
	Age onset (SD)	33.58 (8.87)	33.98 (9.16)	37.94 (10.78)	39.01 (9.22)	34.77 (8.99)	43.74 4.63)	34.58 (12.17)	34.71 (9.14)	37.56 (10.73)	41.07 (13.23)	46.57 (8.40)
	0 DMT (%)	53.8	39.1	26.0	7.1	9.1	35.0	53.6	40.0	19.4	38.1	61.5
	1 DMT (%)	34.6	39.6	41.7	39.3	27.3	40.0	39.3	36.9	44.4	28.6	7.7
	≥ 2 DMTs (%)	11.5	21.3	32.3	53.6	63.6	25.0	7.1	23.1	36.1	33.3	30.8

n: sample size; act %: actual percentage of patients; pred %: predicted percentage of patients; PP: posterior probability

Table 3. Association between ARR and trajectory by inception cohort and sex

	Women N=173		Men N=64	
	traj	OR (CI 95%)	traj	OR (CI 95%)
<1995	1	---	1	---
	2	1.07 (0.98-1.16)	2	1.06 (0.93-121)
	3	1.03 (0.95-1.12)	3	1.05 (0.93-1.20)
	4	0.56 (0.29-1.08)	4	1.06 (0.92-1.22)

	Women N=461		Men N=187	
	traj	OR (CI 95%)	traj	OR (CI 95%)
1995-2004	1	---	1	---
	2	1.16 (1.01-1.34)	2	1.34 (1.03-1.74)
	3	1.22 (1.07-1.40)	3	1.42 (1.08-1.85)
	4	1.22 (1.06-1.41)	4	1.30 (0.98-1.74)
	5	1.34 (1.15-1.55)	5	1.50 (1.15-1.96)
	6	1.36 (1.17-1.58)	6	1.28 (0.94-1.73)
	7	1.26 (1.08-148)	---	---

	Women N=404		Men N=163	
	traj	OR (CI 95%)	traj	OR (CI 95%)
≥2005	1	---	1	---
	2	1.30 (1.08-1.57)	2	1.62 (1.01-2.62)
	3	1.44 (1.19-174)	3	1.92 (1.19-3.09)
	4	1.50 (1.23-1.83)	4	1.90 (1.18-3.08)
	5	1.62 (1.30-2.02)	5	1.44 (0.80-2.60)
	6	1.43 (1.16-1.76)	---	---

S: significant; NS: not significant

Table 4. Estimates of the proportion of benign MS

Author	study or recruitment period	n	Study	Country	Benign MS (%)
Definition 1: Kurtzke grade 3 or less after 10 years of illness					
Hutchinson, 1986	1970s	60	Community study	Ireland	54
Thompson et al., 1986	1980-1984	400	Hospital based	Ireland	42
McDonnell and Hawkins, 1998	pre-july 1, 1996	280	Prevalence	Ireland	20
Cabre et al., 2001	Nov 1997-Oct 1999	62	Population-based survey	Martinique	19.4
Bencsik et al., 2001	data from pre-1996	248	Prevalence	Hungary	15
Kalanie et al., 2003 1625	1996-2001	265	Natural history	Iran	14
Hum et al.	1995-2004	648	Clinical database	Canada	69(F); 41(M)
Definition 2: DSS 0-2 after more than 10 years duration					
Kurtzke et al., 1977	diagnosed 1942-1951 followed 1959-1963	234	Natural history	USA	20.1
Lauer and Firnhaber, 1987	1980s	363	Retrospective	Germany	19
Hum et al.	1995-2004	648	Clinical database	Canada	38(F); 41(M)
Definition 3: EDSS score ≤3 with normal neurophysiological examination in a period of 15 or more years after clinical onset of the disease.					
Perini et al., 2001	mean disease duration at study entry 20 ± 6.6 years	500	Genetic	Italy	6
Hum et al.	1995-2004	648	Clinical database	Canada	59(F)

table modified from Ramsaransing and De Keyser, 2006

1. Minderhoud JM, van der Hoeven JH, Prange AJ. Course and prognosis of chronic progressive multiple sclerosis. Results of an epidemiological study. Acta Neurol Scand 1988;78:10-5.
2. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: Results of an international survey. Neurology 1996;46:907-11.
3. Lublin FD, Reingold SC, Cohen JA, et al. Defining the clinical course of multiple sclerosis: The 2013 revisions. Neurology 2014;83:278-86.
4. Confavreux C, Compston A. Chapter 4 - The natural history of multiple sclerosis. In: Wekerle ACCLMMNS, ed. McAlpine's Multiple Sclerosis (Fourth Edition). Edinburgh: Churchill Livingstone; 2006:183-272.
5. Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. Neurology 2010;74:2004-15.
6. Richards RG, Sampson FC, Beard SM, Tappenden P. A review of the natural history and epidemiology of multiple sclerosis: implications for resource allocation and health economic models. Health Technology Assessment 2002;6:1-73.
7. Kantarci O, Wingerchuk D. Epidemiology and natural history of multiple sclerosis: new insights. Current Opinion in Neurology 2006;19:248-54.
8. Veugelers P, Fisk J, Brown M, et al. Disease progression among multiple sclerosis patients before and during a disease-modifying drug program: a longitudinal population-based evaluation. Multiple Sclerosis 2009;15:1286-94.
9. Weinshenker BG, Bass B, Rice GPA, et al. The Natural History of Multiple Sclerosis: A Geographically Based Study 1. Clinical Course and Disability. Brain 1989;112:133-46.
10. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003;126:770-82.
11. Pittock SJ, Mayr WT, McClelland RL, et al. Disability profile of MS did not change over 10 years in a population-based prevalence cohort. Neurology 2004;62:601-6.
12. Tremlett H, Paty D, Devonshire v. Disability progression in multiple sclerosis is slower than previously reported. Neurology 2006;66:172-7.
13. Debouverie M, Laforest L, Van Ganse E, Guillemin F, Group ftL. Earlier disability of the patients followed in Multiple Sclerosis centers compared to outpatients. Multiple Sclerosis 2009;15:251-7.
14. Liu C, Blumhardt LD. Disability outcome measures in therapeutic trials of relapsing-remitting multiple sclerosis: effects of heterogeneity of disease course in placebo cohorts. Journal of Neurology, Neurosurgery & Psychiatry 2000;68:450-7.
15. Gray O, Butzkueven H. Measurement of disability in multiple sclerosis. Neurology Asia 2008;13:153-6.
16. Compston DA. The management of multiple sclerosis. Quarterly Journal of Medicine 1989;70:93-101.
17. Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: Guidelines for research protocols. Annals of Neurology 1983;13:227-31.
18. Polman CH, Reingold SC, Edan G, et al. Diagnostic Criteria for Multiple Sclerosis: 2005 Revisions to the "McDonald Criteria". Annals of Neurology 2005;58:840-6.
19. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. Annals of Neurology 2011;69:292-302.
20. Nagin DS, Odgers CL. Group-Based Trajectory Modeling in Clinical Research. Annual Review of Clinical Psychology 2010;6:109-38.

21. Nagin D. Group-Based Modeling of Development. Cambridge: Harvard University Press; 2005.
22. Twisk J, Hoekstra T. Classifying developmental trajectories over time should be done with great caution: a comparison between methods. Journal of Clinical Epidemiology 2012;65:1078-87.
23. Goodin DS, Frohman EM, Garmany GP, Jr., et al. Disease modifying therapies in multiple sclerosis: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and the MS Council for Clinical Practice Guidelines. Neurology 2002;58:169-78.
24. Freedman MS, Selchen D, Arnold DL, et al. Treatment Optimization in MS: Canadian MS Working Group Updated Recommendations. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 2013;40:307-23.
25. Kappos L, Lechner-Scott J, Lienert C. Neurostatus.net. 2007.
26. Tremlett H, Zhao Y, Joseph J, Devonshire V, the UCN. Relapses in multiple sclerosis are age- and time-dependent. J Neurol Neurosurg Psychiatry 2008;79:1368-74.
27. van Dulmen MHM, Goncy EA, Vest A, Flannery DJ. Group-Based Trajectory Modeling of Externalizing Behavior Problems from Childhood through Adulthood: Exploring Discrepancies in the Empirical Findings. In: Savage J, ed. The Development of Persistent Criminality. 198 Madison Avenue, New York, New York 10016: Oxford University Press, Inc.; 2009:289-314.
28. Ramsaransing GSM, De Keyser J. Benign course in multiple sclerosis: a review. Acta Neurologica Scandinavica 2006;113:359-69.
29. Hutchinson M. Disability due to multiple sclerosis: a community-based study of an Irish county. Irish Medical Journal 1986;79:48-50.
30. Cabre P, Heinzlef O, Merle H, et al. MS and neuromyelitis optica in Martinique (French West Indies). Neurology 2001;56:507-14.
31. Mayo N, Bronstein D, Scott S, Finch L, Miller S. Necessary and sufficient causes of participation post-stroke: practical and philosophical perspectives. Quality of Life Research 2014;23:39-47.
32. Ebers GC. The natural history of multiple sclerosis. Neurol Sci 2000;21:S815-7.
33. Weinshenker BG, Bass B, Rice GPA, et al. The Natural History of Multiple Sclerosis: A Geographically Based Study 2. Predictive Value of the Early Clinical Course. Brain 1989;112:1419-28.
34. Kantarci O, Siva A, Eraksoy M, et al. Survival and predictors of disability in Turkish MS patients. Neurology 1998;51:765-72.
35. Tremlett H, Yousefi M, Devonshire V, Rieckmann P, Zhao Y, Neurologists UBC. Impact of multiple sclerosis relapses on progression diminishes with time. Neurology 2009;73:1616-23.
36. Giovannoni G, Rhoades RW. Individualizing treatment goals and interventions for people with MS. Current Opinion in Neurology 2012;25:S20-S7 10.1097/01.wco.0000413321.32834.aa.
37. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 1983;33:1444-52.
38. Uitdehaag BMJ, Barkhof F, Coyle PK, Gardner JD, Jeffery DR, Mikol DD. The changing face of multiple sclerosis clinical trial populations. Current Medical Research & Opinion 2011;27:1529-37.

39. Porta M. Dictionary of Epidemiology. Fifth Edition ed. New York, New York: Oxford University Press; 2008.
40. Wolfson C, Wolfson DB. The Latent Period of Multiple Sclerosis: A Critical Review. Epidemiology 1993;4:464-70.
41. Everitt BS. The Cambridge Dictionary of Statistics. Cambridge University Press2006.
42. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. Annals of Neurology 2001;50:121-7.
43. Marrie RA, Cutter G, Tyry T, Hadjimichael O, Campagnolo D, Vollmer T. Changes in the ascertainment of multiple sclerosis. Neurology 2005;65:1066-70.
44. Emery DJ, Forster AJ, Shojania KG, Magnan S. Management of MRI Wait Lists in Canada. Healthcare Policy 2009;4:76-86.
45. Atwood V, McGregor M. Wait times at the MUHC. No. 4 Diagnostic Imaging Revisted Adult Hospitals of the MUHC Has there been progress? Where are the bottlenecks? How can they be removed? Montreal: McGill University Health Centre; 2008 Feb 29, 2008. Report No.: 32.
46. Mayo N. Setting the agenda for multiple sclerosis rehabilitation research. Multiple Sclerosis 2008;14:1154-6.
47. Nixon R, Bergvall N, Tomic D, Sfikas N, Cutter G, Giovannoni G. No Evidence of Disease Activity: Indirect Comparisons of Oral Therapies for the Treatment of Relapsing–Remitting Multiple Sclerosis. Advances in Therapy 2014;31:1134-54.

Chapter 7

An overview of Rasch analysis

Linking chapter for manuscript 2 and 3

The previous chapter demonstrated a method of modeling disability (EDSS) over time using GBTM. This analysis assumed the EDSS is equivalently used across all neurologists who contributed EDSS values to the database over time. This is an assumption that needs to be verified as increasingly large pooled dataset are being used in MS research. As a result there is an increasing need for techniques for data harmonization. Rasch analysis is a method of estimating the extent to which the items of the EDSS (the FSS) are consistently used by neurologists over time. The next chapter provides an overview of Rasch analysis, which will serve two purposes: 1) to explain the methods for data harmonization for manuscript 2; and 2) support the development of a more comprehensive measure of disability, which will be presented in a subsequent manuscript 3.

The Rasch model was named for Georg Rasch, a Danish mathematician, working in the 1960's formulated a model on how peoples' responses to items on a questionnaire can be used to measure ability on the construct being queried. In statistical terms this is referred as the latent trait. The past decade has seen a major methodological development in the area of measurement, the Rasch Measurement Model or simply the Rasch model.[240,241] Applying the model through Rasch analysis has revolutionized the development and quantification of health outcome measures. Since 2000, the number of publications has increased exponentially. Publications indexed in PubMed with "Rasch analysis" in the title or abstract has risen from 20 to now over 150 per year.

As defined in the "Dictionary of Quality of Life and Health Outcomes Measurement", the Rasch model, widely used in health outcome measurement, transforms ordinal response categories into a linear scale with interval-like properties. It is based on a logit transformation of the probability of response to a particular item; an item that 50% of respondents pass or endorse has a logit of 0. A scale that defines the full spectrum of a construct will range from -4 to +4 logits, corresponding to ± 4 standard deviations defining the full range of a standard normal distribution. People at the low end of the logit scale have less ability whereas people at the high end have more ability. The Rasch model is a probabilistic model used to specify an observed rating of a person on a variable of interest as a function of the ability of the person and the difficulty of the items used to derive the rating, where both are defined by their location on continuum from least (easiest) to most (hardest). Items that fit a Rasch model would form a measure with a total score that is sufficient to determine that person's ability on the underlying construct.[242]

Additional key features of the Rasch model are the expectation of unidimensionality and invariance.

Unidimensionality

Besides the common sense that in order to know what you are measuring, it is best to measure a single construct at a time. Following the assumptions of the Rasch model, all items must measure a single construct. The probability of responding to the item (either

correctly or incorrectly) is a function of the ability of the person and the difficult of the task of the single construct being measured.[243] Additionally, the items should function independently of each other. A correct or incorrect response to one item should not predict the response of another item. If this assumption is violated it can biases parameter estimates and affect the unidimensionality of the test.[244]

Invariance

Measurement invariance means the item locations (item estimating level of disability) and person locations (person with estimated level of disability) could be estimated independent of each other. Stated by Hobart et al., when the data fit the Rasch model, the relative location (level of disability) of any two patients does not depend on the items they took, and the relative location (estimate of disability) of any two items does not depend on the patients from which the estimate were made.[245] Item difficulty is not dependent on the sample and person's ability is not dependent of the items.[245-247]

Rasch analysis as defined in the "Dictionary of Quality of Life and Health Outcomes Measurement":

> A method of analyzing data according to the Rasch model, to identify whether or not adding the scores from a collection of items is justified in the data. This is called the test of fit between the data and the model. If the invariance of responses across different groups of people does not hold, then taking the total score to characterize a person is not justified.[242]

Described below are the analyses required to test the data against the expectations of the Rasch model: 1) model fit; 2) threshold order; 3) local item dependency (response dependency and unidimensionality); 4) differential item functioning; and 5) scale targeting.

1) Model fit will be assessed using three summary statistics:

To produce a unidimensional hierarchical continuous linear measure of disability *data must fit the Rasch model*. The "fit" is determined in several ways using: chi-square goodness of fit, item and person standardized fit residuals, and F-statistic probability values.[248] A good overall fit to the Rasch model is indicated by a non-significant ($p > 0.05$ after Bonferroni adjustment) chi-square goodness of fit test such that the differences between the observed and expected responses were

due to chance alone. This is a formal test for invariance.[249] This shows that the items have a hierarchical ordering and is consistent over all levels of the construct (disability).

The average item and average person fit are indicated by two item-person interaction standardized fit residuals that should have a mean of zero and a standard deviation of one indicating an ideal fit to the Rasch model. The individual person or individual item fit residuals are interpreted as *z* scores and should be with in ± 2.5.[249,250] Fit residual > 2.5 indicate misfit of the item to the construct. Fit residual < 2.5 indicate a redundant item. The chi-square probability for the item is based on each person's observed and expected scores based on the model. The F-statistic probability is based on an analysis of variance of groups with different levels of "ability" (class intervals). Significant $p < 0.05$ (after Bonferroni adjustment) indicate the item doe not meet the expectations of the Rasch model.[249,250]

Internal consistency of the scale is estimated by the Person Separation Index (PSI) and is interpreted the same as a Cronbach's α coefficient.[251] A measurement instrument with a minimum PSI value of 0.7[68] is required for group use and 0.85[68] or 0.90[252] for individual use.

2) Ordered thresholds:

Another requirement is ordered thresholds. Each item's response options are expected to increase monotonically. A patient with higher "ability" (less disability) is expected to endorse (select) a corresponding higher response option and a patient with low "ability" (more disability) is expected to endorse the corresponding lower response options.[68,250] A threshold is a point between two response options where there is an equal probability of selecting either response option. Disordered thresholds occur when patients inconsistently endorse the response options. This can occur with poorly labeled or by having too many response options. This impacts item reliability. Category probability curves can be used to visual identify disordered thresholds.[68,250] Collapsing adjacent response options can resolve disordered thresholds.[251]

3) Local item dependency (response dependency and unidimensionality):

a) Response dependency: Local item independence is an assumption in Rasch model. Each item in a Rasch analysis is expected to be independent of each other.[244] Response dependency is

deemed to occur when the response to one item determines the response to another item.[253] Once the Rasch factor is extracted in the Rasch analysis there should be no pattern remaining in the residuals.[68] This is examined by looking at the residual correlation matrix between pairs of items with a correlation greater than 0.3.[251]

b) Unidimensionality: One of the assumptions of the Rasch model is that any measure developed is only measuring one construct. To ensure that this disability scale is unidimensional, a principal component analysis (PCA) of the fit residuals will be performed within the Rasch analysis software (RUMM2020). In the first component items with residual correlation > +0.4 and < -0.4 are used to form two subset of items that are the most different (most negative and most positive). The two item subsets are used to make separate person estimates for each person which can be compared by the application of series of independent *t*-tests. Less than 5% of *t* values outside ±1.96 would support unidimensionality.[254,255] When the value is greater than five percent a binomial test of proportions can be used to calculate the 95% confidence interval around the t-test estimate. Evidence of unidimensionality is still supported if the 5 percent value falls within the 95% confidence interval of the *t*-test estimates.[256]

NB
Item reduction: After rescoring items with disordered thresholds, items with the worst (highest) fit residuals > 2.5 were deleted iteratively. Items with most negative fit residual < 2.5 were also examined for deletion. Items with response dependency were either combined into testlets or deleted depending on model fit.[257] Items that were deemed not to fit the model were deleted iteratively, one at a time until the best model is obtained. After each deletion, item and person fit statistics will be re-examined to look for improvements to the model.[258]

4) Differential item functioning (DIF):

DIF is said to occur when different groups (gender, age, or disease type) response differently to an individual item despite have the same level of ability (disability).[68,253] Once the data is deemed to fit the Rasch model, a two-way analysis of variance will examine if each item's location was stable across different groups. The significance level was adjusted for multiple comparisons by

using the Bonferroni adjustment.[251,258] Items with DIF will affect both unidimensionality[259] and measurement invariance[260].

5) Structure of the measure (targeting):

How well the items (level of difficulty) match the level of ability of the sample is termed targeting. The distribution of persons and item across the construct is depicted on a person-item location distribution plot. A well targeted scale should include a set of items that span the full range of person estimates.[251] The average item location should have a mean logit of zero and a standard deviation of one matching the average person location also with a mean logit of zero and a standard deviation of one.[253] Ideally, to cover 99% of the construct would require the items and person to be normally distributed over ± 4 logits. Poorly targeted measures often result in floor or ceiling effects and thus not provide reliable information for that population.[250]

Rasch analysis can be used as a method for data harmonization or measurement development.

Data harmonization:

When pooling data from multiple sources the properties of the Rasch model can be exploited to facilitate data harmonization. The presence of DIF when pooling data from different sources shows there is heterogeneity in the data and will impact unidimenionality and invariance of the measure. Rasch analysis provides a method to identify and adjust for DIF.[259]

Additional steps aside from the typical Rasch analysis may be required. As an initial step a decision must be made whether to "rack" or "stack" the data. Typically when subjects are accessed at two time-points the goal is to study change. For this situation each of the subject's two assessments are entered as "stacks" in rows. This has been explained for example, as measuring each patient's level of function when entering and leaving rehabilitation. The change in the level of function would be the difference at entry and leaving rehabilitation. Each patient will have two sets of observations. Each set of observations is entered so as to have the same frame of reference for the analysis. In this case the dataset will have twice as many observations as patients. When person change is under study data is stacked. In contrast, to provide evidence that items were consistently being

used a "racked" data format is required where time points are entered in columns. Using the same example as above, between the two time-points if there was an intervention or some change occurred, the items should reflect this change from the same person. This data format allows the focus to be on how items vary, rather than examining how subjects change.[261]

For the purposes of data harmonization, a racked dataset and DIF analysis by rater would provide evidence on whether the items are being consistently used by different raters in order to pool data.[259,260] Adjustments can then be made to account for DIF.

Measurement development:

Rasch analysis has been used to develop new measures or reevaluate the psychometric properties of existing legacy measure. Rasch analysis has been used to combine existing indices into a single measure.[69,70] Redundant items and "poorly" functioning items that do not meet the expectations of the Rasch model are removed. This results in an optimal number of items being included. It also provides evidence that the full spectrum of the construct is or is not represented in the measure. Rasch analysis enables shorter forms of the full measure to be used to assess patients without compromising the comparability of results. The properties of a Rasch measure can decrease the response burden for patients and administration for clinicians. More recently, PerfOs and patient-reported outcomes (PROs) have been combined with the use of a Rasch analysis to form a single, unidimensional measure that is used in traumatic brain injury and stroke.[70,262]

Rating scale vs partial credit scale:

An additional decision needs to be made for items with polytomous response options as to whether the rating scale model or partial credit scale model should be used for the analysis.[68] With the rating scale model[263], each item will have the same number of response options and a common threshold pattern is imposed. In this context, the number of thresholds is equal to one minus the number of response options. The partial credit scale model,[264] an unrestricted model allows a different number of response options across items and variable threshold pattern. If items have the same number of response options, either model can be used; RUMM software provides a Fisher's log likelihood ratio test and

rationale to help select the appropriate model in their manual.[265] A non significant result of the Fisher's log likelihood ratio comparing the rating scale against the unrestricted partial credit model suggests that re-parameterizing the model by varying threshold values did not improve the model and the simpler rating scale thresholds can be used.[265]

NB. For the purposes of this thesis where the selection of response options was not an option, the partial credit model was used for both analyses.

Rasch Analysis Decision Flowchart:

As described above, Rasch analysis can be used in two ways, as a tool for data harmonization or to development a measure. Rasch analysis is complex and requires a stepwise process. The order of the procedures and specific decision points were formulated from user experience and interpretation of general guidelines from several published articles with comprehensive descriptions of Rasch analysis.[68,250] There are common steps required for any Rasch analysis that are iterative and require multiple decision rules. Hence, they lend themselves well to a diagrammatic approach to summarized steps undertaken in a Rasch analysis and are illustrated here using a standard flowchart algorithm. The *Rasch Analysis Decision Flowchart* with brief explanations of key decision steps was developed to represent the Rasch analysis by this author. Fit statistics and output available are based on the RUMM2020 software used for this thesis.[266]

The order of the steps outlined in the Figure 7.1 is a graphical representation derived from our experience with Rasch analysis and steps that are typically reported in Rasch analysis articles.

Figure 7.1 presents the "Rasch Analysis Decision Flowchart". As it is typical in computer flowcharting, a oval representing a start or end of a process, a hexagon represents a preparation process, a rectangle represent a process, a rhombus represents a decision point, and the arrows indicate flow. The numbers in circles correspond to Figure 7.1, which provides additional information of each process and specific criteria for each decision point. The principle steps outlined in the flowchart (the grey rhombus) should be generalizable to other platforms although fit statistics are not directly comparable.[68]

Figures 7.2a and b are examples of category probability curves for an item with 5 response categories. Figure 7.2a displays an item with disordered categories; Figure 7.2b displayed an item with ordered categories. Along the y-axis is the probability that a category is endorsed and along the x-axis is the value of the latent construct as defined by the all the items included. On the disordered category figure (Figure 7.2a) the probability curve for category#2 should cross category#3 before category #4 indicating the item was not functioning in a monotonic fashion. For Figure 7.2b showing ordered categories, each category has its own unique probability space (a row of hills) indicating the item was functioning as expected.

Figure 7.3 depicts threshold probability curves, remembering that the number of thresholds equal the number of categories minus one. The axes are the same as in Figure 7.2a. Each ogive shaped line represents the probability of passing a threshold according to the person's ability as represented by the location along the x-axis. To illustrate, imagine a vertical line drawn at the point, -2 logit along the x-axis; this point represents people with relatively low ability (low level of the construct). The vertical line on this figure shows the probability a person at -2 logit will have of passing each of the thresholds shown on the graph. For example, they will have almost no probability of passing threshold #3 and #4, ~25% probability for threshold #2, and ~75% probability of passing threshold #1. So a person at this level of the latent trait will likely receive a score of 1 on the item. Similar response probabilities can be predicted for each ability. The dots represent the actual observed response of each different class intervals and should fall on the ogive curves, the expected responses as determined by the model.[248] People of all class intervals have a high probability of passing threshold #1 and that item level would be considered very easy. Persons of all class intervals have a much lower probability of passing threshold #4 and that would represent a harder level to achieve.

Figure 7.4 is an example of item characteristic curves (ICC) showing differential item functioning DIF by the personal factor, gender. Along the y-axis is the expected value for the item. Along the x-axis is person location as defined by the latent construct. The observation that the two gender-specific lines are not superimposed on the expected line

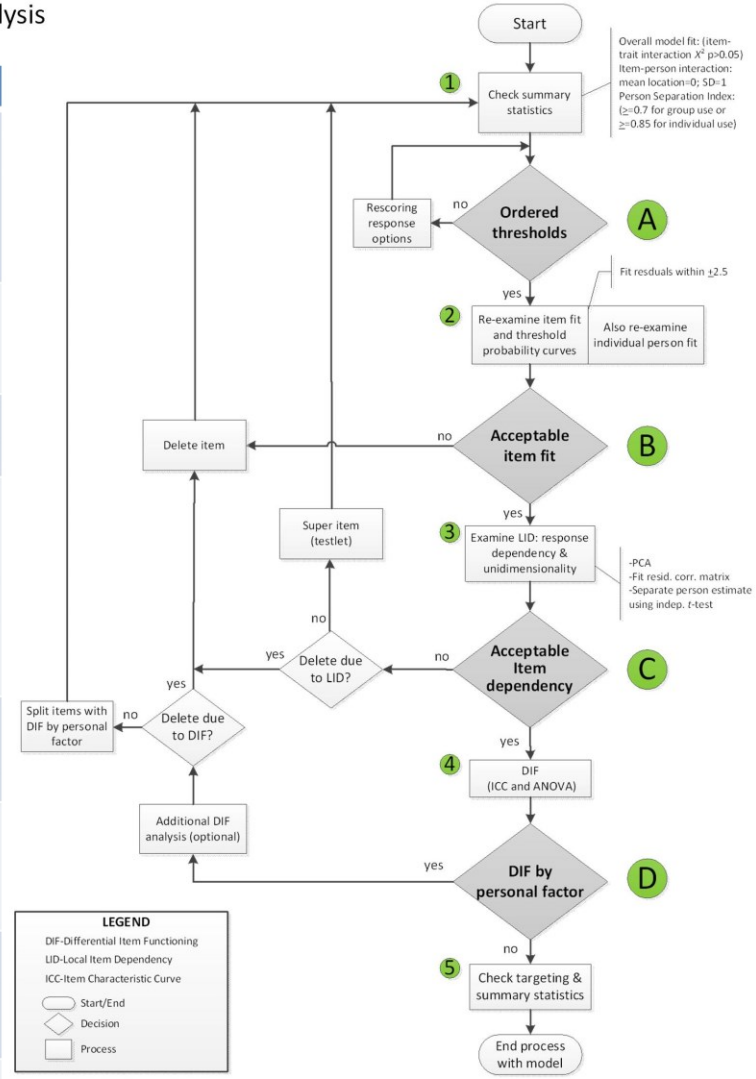
indicates DIF. The observed response pattern for men is greater than expected for all class intervals. For women, the response pattern is above expectation at the lower end (easier end) and below expectation at the higher end (harder end) of ability on the x-axis. A two-way ANOVA provides evidence of possible DIF. Evidence of DIF can be tested using a variety of other standard statistical methods.[267,268]

Checking for item DIF is examined as one of the final steps of a Rasch analysis. At this point the summary statistics can fit the expectations of the Rasch model but item bias still needs to be assessed. Common personal factors such as gender, education, or language are tested against the item to assess whether the personal factor impacts the items response. In the example provided, it appears that gender may impact how the item is answered.

DIF occurs when two groups with equal levels of ability do not response similarly on the item and the reason for this difference is not part of the construct understudy. DIF may impact on measurement unidimensionality.[68,259]

Figure 1. Flow diagram of the process underlying Rasch analysis

	Test/procedure and criteria
Step 1: Overall model fit	• Item-trait interaction χ^2 probability > 0.05[250, 255] (Bonferroni adjusted for the number of items)[269] • Item-person interaction: mean fit residual=0; SD=1[250] • Reliability estimate: Person Separation Index ≥ 0.7[68] for group use; ≥ 0.85[68] or 0.90[252] for individual use • Start with examining item and person misfit • Check data for data entry errors
Step 2: Item response category - threshold order	• Visually inspect category probability curves (CPC) and threshold probability curves (TPC) – Figures 2 and 3 • Identify disordered threshold by examining table of threshold estimates and graphic indicator such as the CPC[250]
Step 2: Item and person fit to model	• Re-examine fit residuals should be within range of -2.5 to 2.5[250] • Re-examine observed data fit to expect curves of the Item Characteristic Curves (ICC)[270]
Step 3: Local item dependency (LID): Response dependency and unidimensionality	• Fit residuals between two items with r > 0.3 indicate response dependency[251] • Examine the correlation of the residuals between two subset of items (the most positively and negatively correlated items) of the 1st component of the residuals from the principal component analysis (first residual factor). Separate person estimates for the two sets of items are performed using the independent *t*-test. The percentage of such tests outside the range -1.96 to 1.96 should not exceed 5 percent[256]
Step 4: Differential item functioning (DIF)	• Examine ICC by personal factor • Examine ANOVA (by personal factor, Class interval, or Class Interval X personal factor interaction)[68, 253] • Additional DIF analysis (optional)[259, 267, 268]
Step 5: Targeting	• Mean location of items and persons=0; SD=1[253] • Mean fit residual of items and persons=0; SD=1[253] • Examine Person-Item Threshold Distribution Map • A range of ± 4 logits will theoretically cover the full range of the latent variable being measured.
Step 5:	• Item-trait interaction χ^2 probability > 0.05[255] (Bonferroni adjusted)[269] • Item-person interaction: mean fit residual=0; SD=1[250] • Person Separation Index ≥ 0.7[68] for group use; ≥ 0.85[68] or ≥ 0.90[252] for individual use



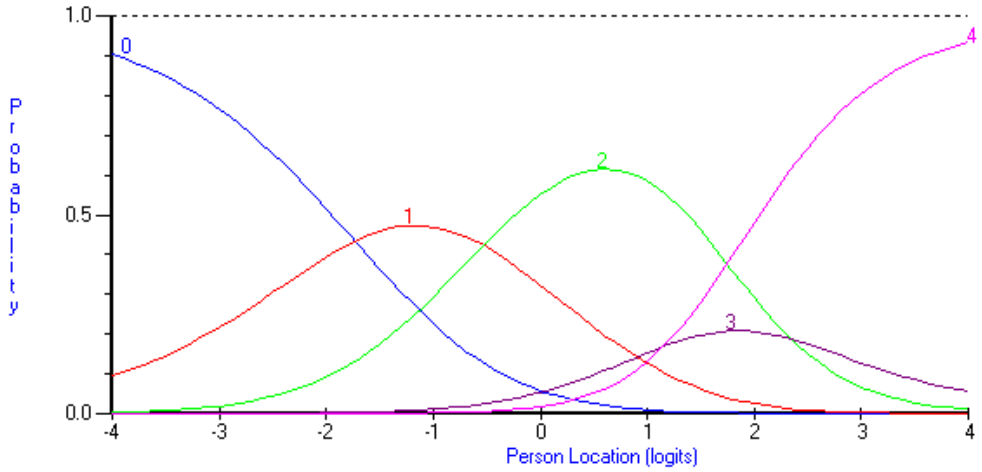


Figure 7.2a: Typical Category Probability Curves with disordered thresholds

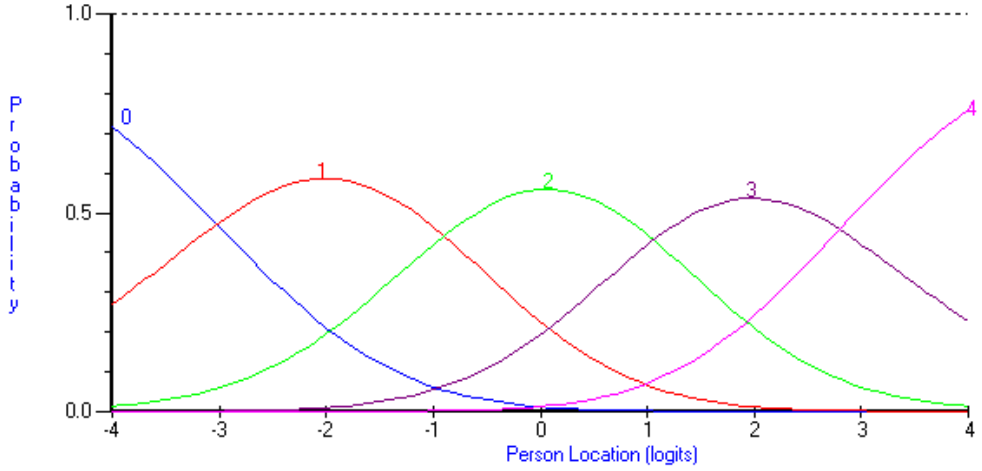


Figure 7.2b Typical Category Probability Curves with ordered thresholds

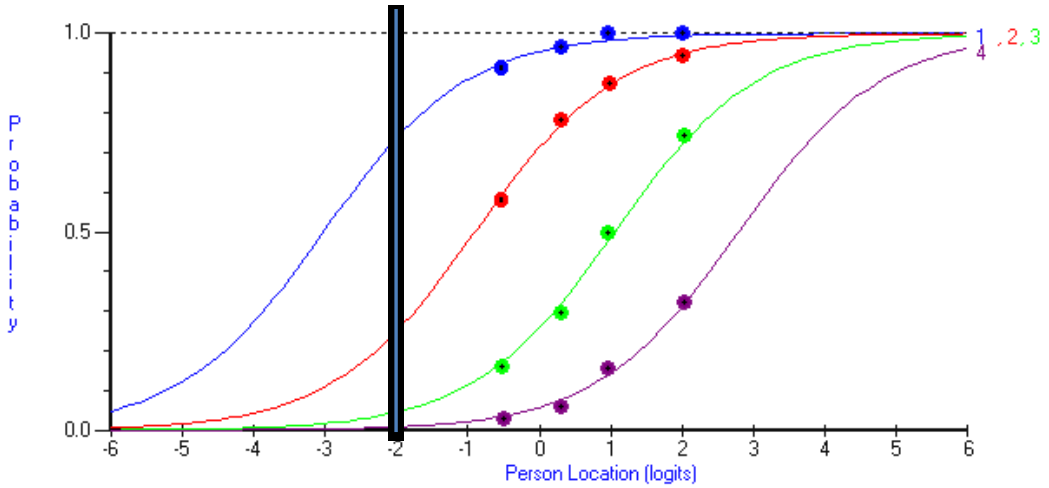


Figure 7.3: Typical Threshold Probability Curves with ordered thresholds

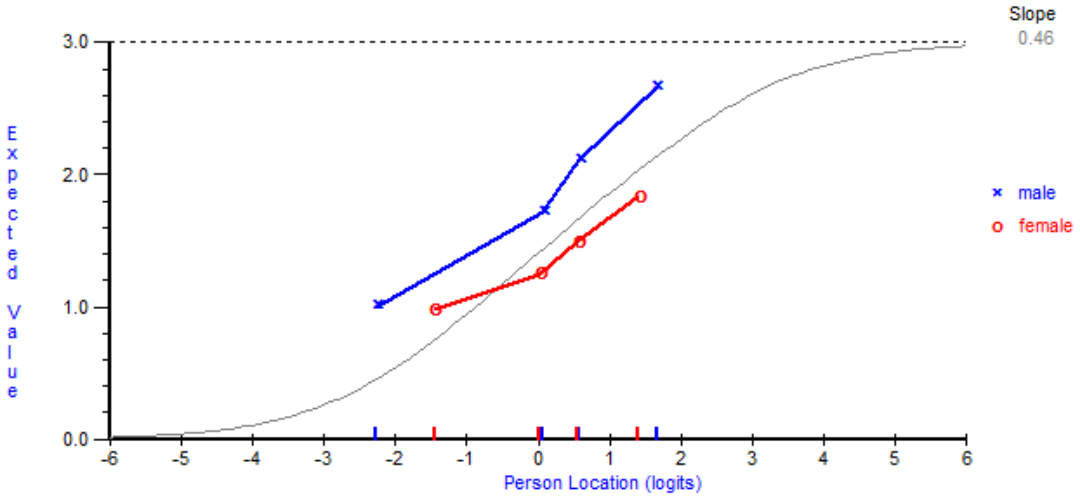


Figure 7.4: Example of an Item Characteristic Curve with DIF by personal factor, gender

Chapter 8

Linking chapter 7 to (chapter 9) manuscript 2

The last chapter provided an overview of Rasch analysis and describes the two ways it can be used. For Rasch analysis, data are typically entered in rows, called a stacked format when the interest is in how people change. When the interest is in how items change, data are entered in columns, called a racked format.

The next manuscript is an example how Rasch analysis uses a racked data format to identify whether data can be harmonized across raters and across time. The next manuscript entitled “Rasch analysis as a method of data harmonization to resolve cross-rater variability in scoring the function system scores of the Expanded Disability Status Scale” gives an example of this procedure using the EDSS across five neurologists and two time points to illustrate this.

Chapter 9 (Manuscript 2)

Rasch analysis as a method of data harmonization to resolve cross-rater variability in scoring the functional system scores of the Expanded Disability Status Scale

Stanley Hum[1], Yves Lapierre[2], Lois Finch [3], Lesley Fellows[4], Nancy E. Mayo[1,3]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, QC, Canada

[2]The Montreal Neurological Institute, McGill University Health Center, Montreal, QC, Canada

[3]Division of Clinical Epidemiology, McGill University Health Center, Montreal, QC, Canada

[4]Cognitive Neuroscience, McGill University, Montreal, QC, Canada

In preparation for submission to *Multiple Sclerosis Journal*

Communication addressed to:

Stanley Hum, M.Sc., PhD. Candidate
School of Physical and Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
514-398-5981
Email: stanley.hum@mcgill.ca

ABSTRACT

The wide spread adoption of the Expanded Disability Status Scale (EDSS) as a measure of Multiple Sclerosis (MS) disability allow pooling of datasets for observational studies. However, strong statistical methodology is essential to deal with data heterogeneity from different sources. The objective of this study is to estimate the extent to which items of the Functional System Scores (FSS) and EDSS are equivalently scored by neurologists ensuring data from different sources can be pooled.

Rasch analysis of the EDSS and its FSS data from five neurologists at the Montreal Neurological Institute MS Clinic were used to estimate differential item functioning (DIF) of raters. Randomly queried from the clinic database were 100 patients/neurologist with two assessments at a 1-year interval. There were no age or MS-type restrictions. Personal factors were age, gender, and neurologist. Estimated were parameters for threshold order, model fit, unidimensionality, DIF, targeting, and response dependency using a *racked* dataset.

Sample age was 47.5 years with 68% women. MS-types were clinical isolated syndrome and relapsing-remitting=65.4%, secondary-progressive=24.2% and primary-progressive=10.4%. Disordered thresholds occurred in 11/16 items. Rescored FSS fit the model. EDSSs were redundant and deleted. Items were well distributed along the continuum (mean logit=0; SD=1.2), but there were few items for high functioning people (mean logit=3.52; SD=1.9).

Only DIF by neurologist was detected for five items. To correct for this: three item pyramidal2, mental2, and brainstem1 were split by raters. Bowel-bladder1, mental1, mental2 for neurologist 3 & 5 were deleted. Reliability (person separation index=0.84) was excellent. The global model fit with Bonferroni corrected p values = 0.006.

The FSS support a unidimensional construct of MS disability measured on a linear interval scale. Rasch analysis can detect and adjust for DIF to ensure data harmonization when pooling data from multiple sources. The FSS items form a measure primarily targeting people with high disability.

Introduction

The Expanded Disability Status Scale (EDSS) has been universally adopted as the measure of disability in multiple sclerosis (MS).[1] The use of a common measure facilitates evaluation of treatments and has the potential to allow pooling of datasets for research purposes. The EDSS is included in all European registries surveyed in 2014[2] and as part of the international MSBASE registry.[3] Increasingly, these pooled datasets are being used to answer questions related to epidemiology, long-term drug effectiveness, and health outcomes in MS.[2,4,5] However, combining data from different registries, or even across local cohorts, requires strong statistical methodology to deal with data heterogeneity.[6] EDSS scores from different sources require validation to ensure the data are directly comparable. The method ensuring that information being combined data collected from different sources is compatible is called data harmonization.[7]

Although based on sound clinical knowledge, the EDSS and its Functional System Scores (FSS) were developed without psychometric input, compromising their mathematical properties and limiting their usefulness as evaluative outcome measures.[8,9] The limitations of the EDSS have been described by traditional psychometric analyses. In a 2014 systematic review of the measurement properties of the EDSS, validity was supported, but inter-rater reliability was sub-optimal, ranging from slight to moderate (Kappa=0.32 to 0.76) for the EDSS, and slight to substantial (Kappa=0.23-0.58) for the FSS.[10] Intra-rater agreement was slightly better than inter-rater agreement. Agreement was more variable in the lower EDSS range[10], where more precision is needed to detect mild changes. A key criterion for an evaluative measure is a valid value for minimum clinically important change (MCIC). The established MCIC of a 1.0 for people with EDSS ≤ 5.5 and 0.5 for people with EDSS ≥ 6[11] has been challenged by Kragt et al.[12]

Variability in scores was simply adjusted by grouping scores until there was 100% agreement among raters. Noseworthy et al. estimated typical variation among raters were 0.5 point for the EDSS and 1 point for the FSS and suggested a 2-step change for EDSS and FSS as an indicator of real change.[13] Goodkin et al., found that to obtain 100% intra-rater agreement a variation within 1.0 EDSS was typical where as to obtain 100% inter-rater

agreement a variation was estimated at within 1.5 EDSS.[14] There have been attempts to increase FSS/EDSS reliability by providing training material (Neurostatus©)[15] however there has been no information on improvements on its scoring.

Rasch analysis is a modern statistical method that can be applied to ensure harmonized data across multiple sources. Named for the Danish statistician George Rasch,[16] Rasch analysis has been typically used to develop new health outcome measures or to reexamine existing 'legacy' measures. This method can provide additional evidence of the scientific merit of a measure for research purposes and reassess reliability and validity within a modern psychometric framework.[17] Rasch analysis is able to elucidate additional properties of a scale such as the ordering of item response options, differential item functioning (DIF) or response bias (i.e. by rater or by gender), unidimensionality, and the appropriate targeting (matching) of items to a person's level of ability (or disability).

Rasch analysis tests the extent to which the FSS/EDSS items measure a single construct and fit an underlying theoretical hierarchy that forms a linear continuum (i.e. a "ruler") with interval-like units.[18] Tangibly, the FSS/EDSS from different sources needs to be mapped to a standard metric to be "inferentially equivalent".[19] Here the standard "metric" is provided by the Rasch model.

When observed data fulfill the expectations of the Rasch model, the measure would also have the important property of "measurement invariance".[18] In a Rasch model, measurement invariance means the item locations (item estimating level of disability) and person locations (the person with the estimated level of disability) could be estimated independently of each other. In other words, stated by Hobart et al., when data fit the Rasch model, the relative location (level of disability) of any two patients does not depend on the items that were used to assess them, and the relative location (estimate of disability) of any two items does not depend on the patients from which the estimate was made.[20] Others have called this measurement stability.[20] By establishing invariance of the FSS/EDSS we also establish it is "inferentially equivalent" across different data sources in order to harmonize data.[20] Of particular interest, as part of the Rasch analysis, item DIF by a personal factor (such as by neurologists) is estimated. The EDSS is known to lack reliability,

especially outside of a clinical trial setting but it is assumed that the neurologists scoring the FSS/EDSS are interpreting the items in a similar manner. DIF analyses will provide evidence on whether neurologists are applying the FSS/EDSS items in a similar manner. It can also identify how different each neurologist is in scoring the FSS/EDSS and provide a method to adjust for DIF by splitting the item by neurologist.[21]

For this study, we assess the reliability of this scale in a clinical setting, typical of those contributing data to large-scale registries. By applying Rasch analysis as the data harmonization method, we address the expected variability in scoring the FSS/EDSS and provide additional information about the scale's measurement properties crucial for optimal interpretability and direct comparability of data.

Objective

The purpose of this study is to estimate the extent to which the FSS/EDSS are equivalently scored by neurologists, allowing data from multiple sources to be pooled.

Methods

Subjects were selected from the MS Clinic of the Montreal Neurological Institute's (MNI) longitudinal database. This database was established in the 1980's to characterize the population seen at the MNI clinic in terms of socio-demographic and neurological status. As of 2013, it contains information for 5000 registered patients, of which 3,000 attend regularly. At each clinic visit, the patient's EDSS and FSS are recorded. We imposed no restrictions on age or MS type for this analysis. The EDSS and seven FSS items (not "other") from five neurologists were included in the Rasch analysis. Each half point of the EDSS was rounded down to the corresponding whole unit to resemble the original Disability Status Scale.[22] All of the neurologists (men) had extensive clinical experience as MS specialists (ranging from 10 to 40 years), had participated in clinical drug trials as principal investigators and had obtained the Neurostatus© certification.[23]

Each neurologist had registered varying numbers of patients in the database and so a random sample of 100 patients was selected for each of the five neurologists, for the period

2008 to 2013. From each patient's record, we chose for analysis two neurological exams (Index visit 1 and 2) one year apart (± 1 month).

The general methods for Rasch analysis have been described in detail elsewhere.[24,25] For the analysis of cross rater validity by differential item functioning (DIF) analysis, the methods were adopted from published guidelines.[21,25] This method provides evidence on whether physicians are scoring the FSS/EDSS items in a similar manner and can make adjustments when differences are detected.

Typically in a Rasch analysis, data from multiple time points (here 2) are "stacked" as rows allowing for the focus to be on how the subjects change. For this analysis, time points are "racked" as columns allowing for the focus to be on how items vary.[26] Index visits 1 and 2 were entered in columns, resulting in an analysis with a sample size of 500 (100 subjects X 5 neurologists) with 16 items (7 FSS plus EDSS for each of 2 index visits). A sample size of 250 patients can typically estimate an item difficulty to within ± 0.5 logits with a 99% confidence level; sample sizes of at least 500 are recommended for critical applications.[27]

The RUMM2020 software (version 4.1) was used to fit a partial credit Rasch model.[28] The steps of the Rasch analysis are depicted in Figure 1. Parameter estimates were tested for threshold order, model fit, response dependency, unidimensionality, DIF, and item/person targeting. All p values < 0.05 after Bonferroni adjustment were considered to be significant.

Briefly, data were fit to the Rasch model and evaluated by testing item-trait interaction using a χ^2 goodness of fit test; two item-person interaction statistics, and the person separation index (PSI) were also estimated.[25] Disordered thresholds were rescored by collapsing adjacent response options and then rerunning the analysis. A response option *threshold* is a transition point where a respondent is equally likely to choose either of two adjacent responses. Ordered thresholds are expected to occur such that a person with a higher level of "ability" would choose a correspondingly more "difficult" response option. The number of response thresholds are one less than the number of actual response options.

Fit residuals for items should be within ± 2.5 indicating adequate fit to the model.[24] Response dependency among items was estimated by examining the residual correlation matrix for pairs of items, and inferred when the correlation value was greater than 0.3.[29] To assess unidimensionality, the principal component analysis (PCA) of the fit residuals will be performed within the Rasch software. In the first component, items with residual correlation > +0.4 and < -0.4 are used to form two subset of items that are the most different (most negative and most positive). The two item subsets are used to make separate person estimates for each person which can be compared by the application of a series of independent *t*-tests. Less than 5% of *t* values outside ± 1.96 would support unidimensionality. When this value is greater than 5% a binomial test of proportions can be used to calculate the 95% confidence interval around the *t*-test estimate. Evidence of unidimensionality is still supported if the 5% value falls within the 95% confidence interval of the *t*-test estimate.[30] DIF, or item bias, was assessed for each of the personal factors of age, gender, and rater. DIF was detected by two-way ANOVA and adjusted by splitting by personal factor or deleting the item; see Figure 1. DIF between raters (i.e. the 5 neurologists) was used to estimate cross-rater validity and tested using *post-hoc* Tukey tests.[21]

Results

The characteristics of the patients assessed by each neurologist (labeled 1 to 5) are presented in Table 1. The sample consisted of 68% women; 65.4% with relapsing remitting MS, 24.2% with secondary progressive MS and 10.4% with primary progressive MS. Sample characteristics were similar across neurologists, except that neurologist 3 had more patients with a high level of disability (median EDSS of 6.0), more men, and more people with progressive MS.

Table 2 shows the results of the first analysis, attempting to fit all items and neurologists to the Rasch model. Disordered thresholds were observed in 11 of the 16 items (69%). The remaining items had observed threshold probability curves that deviated from the expected values. The analysis of the sample for each neurologist demonstrated disordered thresholds that ranged from 56-100% of the items.

In Table 3, the model fit of the original (unmodified) scale, described by the item-trait interaction statistics, was significant (χ^2=179.750; df=48; p<0.000001), indicating deviation from the expectation of the Rasch model. After rescoring, all functional systems, item fit residuals were within ± 2.5. The EDSS items (#8 and #16) consistently had fit residuals < -2.5, indicating redundancy, and were deleted. The overall model statistics now fit the model expectations (χ^2=71.462; df=42; p<0.0031) with a mean item fit residual of -0.588 (SD 0.998), a mean person residual of -0.337 (SD 0.698) and a mean person location of 3.8 (SD 1.7). PSI was 0.85.

Following the steps illustrated in Figure 1, response dependencies were found between item #1 (Pyramid2) and item #7 (Mental2) with residual correlations of -0.343 and between item #1 (Pyramidal2) and item #15 (Mental1) with residual correlations of -0.371. The independent *t*-test showed that 28 of the 465 patients (6.02%) of the person estimates derived from the two most different subsets of items differed from estimates derived from all items, however the 95% confidence interval (3.9% to 8.2%) included the 5% value indicating that unidimensionality was still supported.

There was no DIF by patient factors (gender, age). There was DIF by neurologist for 5 of the remaining 14 items (7 items from FSS X 2 visits). Table 4 shows the results of the DIF analysis; the items with DIF were: #1 (pyramidal2), #7(mental2), #11(brainstem1), #13(bowel and bladder1), and #15(mental1). *Post hoc* Tukey tests identified DIF between specific pairs of neurologists and these are summarized in Figure 2. On item 1, the first block in Figure 1, neurologist 1 (the first row) had DIF (marked by X) with all other neurologists as did neurologist 5. The others, only had DIF with 1 and 5, but not amongst themselves. For the other items, neurologist 5 consistently showed DIF with the other neurologists.

Table 5 summarizes the steps used to adjust for DIF. Item #1, #7, #11, #13 and #15 were split by neurologist. Splitting item #15 did not improve the model or correct DIF and was deleted. Item #13 for neurologist 1 and 5 (I0013A: BB1A), item #7 for neurologist 3 and 4 (I0007A Mental2A) and item #13 for neurologist 2, 3, and 4 (I0013B: BB1B) were deleted, in that order, due to DIF or statistical misfit. Item # 1 (I0001A: Pyramidal2A) for

neurologist 1 and 5 has a location of 0.669 logit and for neurologist 2,3 and 4 (I0001B: Pyramidal2B) a location of 2.653 logit. Item #7 for neurologist 1 and 2 (I0007B: Mental2B) has a location of 0.725 logit and for neurologist 5 (I0007C: Mental2C) a location of 1.478 logit. Item #11 for neurologist 5 (I00011A: Brainstem1A) has a location of -0.169 logit and for neurologist 1, 2, 3, and 4 (I00011B: Brainstem1B) a location of -0.099 logit. Also shown in the item map of Figure 4, each of the items split by neurologists are situated at different threshold locations.

In the final model, the item-trait interaction was not significant after Bonferroni adjustment indicating fit to the Rasch model (χ^2: 72.554; df=45; p=0.0057). Figure 3 shows the relationship between items, the people, and the Rasch model. Item location (threshold) spread along the continuum was quite wide, ranging from -6.4 to 6.1 logits as shown in the bars below the horizontal axis, where each bar represents the number of thresholds. The locations of the people on the continuum are shown in the bars above the horizontal axis, where each bar represents the frequency (%). Targeting was poor; i.e. patients were mainly in the high functioning (low disability) ranging from -1.1 to 7.4 logits; mean location was 3.52 (SD 1.9). None of the patients were at the low functioning (high disability) range of the scale, and no items were difficult enough to properly test high functioning patients. There were also gaps in coverage of the items so patients at those locations will not be well measured. The final PSI was 0.84.

Discussion:

The aim of this study is to use Rasch analysis to confirm whether existing FSS/EDSS data commonly found in MS clinic databases can be pooled across neurologists for observational research purposes. Having the FSS/EDSS fit the Rasch model requirements allow certain claims to be made concerning the psychometric properties of the items. Rasch analysis of the FSS/EDSS revealed some previously unknown characteristics. Items exhibited disordered thresholds indicating that neurologist were not able to score an item with an increasing value to match an increasing level of disability. One possible reason is that neurologists applied the response options inconsistently, leading to sources of misfit to the Rasch model. Decisions on transition to the next level of impairment for each response

option may be ambiguous, or overly complicated, leading to disordered thresholds. This can be due to confusing labels or too many response options. The authors of the recent Rasch analysis of the Medical Research Council (MRC) grading system for muscle strength, used in part for the standard neurological examination, found similar difficulties in the raters' ability to consistently applying the response options of the scale.[31] The authors suggest that the disordered thresholds observed in 75% of the items might be improved with fewer (4) response options. They also felt that the descriptions provided for each response option may have the undesirable effect of making item scoring more complex.[31] However, there is no consensus on the optimal number of response options. Given the difficulties shown for the MRC scale, is not surprising that neurologists using the MRC grading system to score the FSS may also be having problems generating consistent scores. Inconsistent scale application has an impact on item reliability, and will contribute to noise in pooled data. Solutions can include more rigorous rater training, fewer response options, or both.

An item with DIF by rater indicates it item is not being scored in a similar manner by every neurologist providing evidence the original data may not be inferentially equivalent. Four FSS items had DIF by neurologist, two of which (pyramidal and brainstem) were adjusted by splitting the item by neurologist. For the brainstem item, the DIF was attributed to a single neurologist, 5. He consistently had the most occurrences of DIF compared to the four other neurologists in the items that exhibited DIF (Figure 2). For the pyramidal item, neurologist 5 was implicated again, along with neurologist #1. Both grouped differently from the remaining neurologists (2, 3 and 4). DIF in these cases may reflect the wide range of clinical experience, level of training, and expertise of the raters. Items exhibiting DIF by neurologist provide strong evidence that the FSS are not being scored the same way amongst the five neurologists and, by extension, show that EDSS scores ≤ 4.0 cannot be directly compared with certainty in this sample.

DIF in the remaining items (mental and bowel & bladder) are not based on the neurological examination. It is not surprising that the mental FSS showed DIF by rater. There is no standard assessment tool adopted to assess cognitive function in MS. The neurologist is left

to his or her own devices to score this functional system. Bowel & bladder assessment is typically based on the patient's self report during their clinic visit as a presenting complaint or as part of their past medical history in conversation with their neurologist. As with cognition, the assessment method is at the discretion of each neurologist, and can also depend on the openness of the patient to discussing this issue and the communication skills of the physician. The discretionary interpretation of these two FSS among the raters would impact scoring consistency (reliability) and meaning of the FSS (validity), increasing data heterogeneity. As shown in Table 2, all five neurologists scoring these two items resulted in disordered thresholds indicating variability in interpreting the response options. If the scoring of these two items is influenced by factors other than the expected change in the domains that these scales are intended to measure, as seems to be the case here, this will have an impact on the unidimensionality of the construct being measured and will increase data noise.

Table 5 shows the results of splitting items by rater to adjust for DIF. Items I0001 (Pyramid2), I0007 (Mental2), and I0011 (Brainstem1) were split by rater resulting in different location estimates for the item depending on the neurologist. Figure 4 shows items split by neurologist are situated at different thresholds. For example, patients with the same score on the Pyramid2 item are assigned to a lesser degree of disability by neurologist 2, 3, and 4 (Pyr2B) than neurologist 1 and 5 (Pyr2A). This provided a method to account for different scoring behavior of individual neurologists; adjusting the scores so that the data can be considered inferentially equivalent before being pooled.

The Person-Item Threshold Distribution Map in Figure 3 shows the distribution of estimated person disability levels did not overlap completely with the distribution of FSS item thresholds. By convention the Rasch analysis always centers the scale on zero to represent the average item difficulty (here disability). For a well-targeted measure, the estimated average location of the patients' level of disability should also be around zero logits to match the items.[25,29] Our results indicate the patients are relatively high functioning (with low disability) compared to the items (too easy) used to measure them.

Poorly targeted measures are often result in a floor or ceiling effects and result in poor precision when measuring the mistargeted patients.[24]

Whether this is an important problem depends on how the measure is intended to be used. The gaps in the high disability region with no patients reflect the fact that the high disability end of the score for each item of the function systems is a theoretical range that is not actually attained by any patients in the sample. For example, the sensory and bowel and bladder FSS ranges from 0 (normal) to 6 (sensation essentially lost below the head or loss of bowel and bladder function) but there are rarely any MS patients scoring 6 for either system. The estimates the item thresholds in the low functioning range (high disability) observed in Figure 3 are extrapolated values.

Patients with extreme scores (at the ceiling) with the lowest levels of disability estimates were located at 7.4 logit and represented 7.2% of the sample. There are also no items at the low disability (high functioning) end of the scale, leaving gaps in the scale where many persons are located in this sample, which we expect is typical of many tertiary MS clinics in academic medical settings. This is an important problem: the scale provides inadequate measurement at the (low disability/high functioning) end of the spectrum that is the focus of all current interventions. The ability to detect small (subtle) changes in this range of ability may be crucial, since the goal is to maintain a patient's level of function early in the disease. The gaps in measurement precision here will have an impact on the reliability of the measure and on data quality in this clinically important range of disability. The item gaps means that the rescored items do not measure this patient group well and results are extrapolated estimates of the patient's level of disability. A poorly targeted measure with gaps in measurement precision here may not have a direct effect on data harmonization, but will have an impact on the reliability of the measure and on data quality in this clinically important range of disability.

Training programs to improve consistent scoring and interpretation of the FSS and increase the intra and inter rater reliability of EDSS scoring have been developed in recent years.[23] The study sample was designed to reflect clinicians with many years of experience who had received FSS/EDSS training during the study period. Despite this training and

experience, we found disordered thresholds and DIF by rater for several FSS. Although all raters are highly trained clinicians, variability in scoring may result from their individual style and thoroughness in the performance of the neurological exam, influenced by their general clinical training and years of experience, which ranged from ~10 to 40 years. Alternatively, the scoring of the FSS items may be too complex to reliably score consistently.[32]

Conclusion

Rasch analysis provides a strong statistical methodology to assess the psychometric properties of the FSS/EDSS, a technique to detect data heterogeneity, and a method to harmonize data from different sources when heterogeneity is found. Identifying and eliminating sources that impact on threshold order improves item reliability and decreases data noise. The DIF analysis offers a sophisticated method to identify inter-rater item bias among the neurologist and can under certain circumstances adjust for it. Estimating and accounting for item bias ensures that pooled data are inferentially equivalent across all sources.

Collapsing response options (to adjust for disordered thresholds) to those where there is agreement between raters on transitions from one level to the next will improve item reliability, but can impact the validity of the scale. There is evidence that validity tends to increase and residual error tends to decrease as the number of response options increases.[33]

DIF by rater in this study indicates the neurologists are conceptualizing the scoring of the functional systems differently even though they are all trained to perform the FSS/EDSS. If response options cannot be found that are both reliably rated and clinically meaningful, alternative measurement approaches should be considered. Standardized instruments to assess domains exhibiting disordered thresholds or DIF to aid the neurologist in scoring the affected functional systems can be included. Including patient reported outcomes as part of the patient assessment should be considered. The impact of some disabilities important in MS such as pain, fatigue, bladder, and bowel can readily be assessed from the

patient perspective. Alternatively a performance-based outcome can be considered to provide an objective score, such as neuropsychological tests for cognition or urodynamic tests for bladder function.

There are plans to merge MS databases in Europe to create a continent-wide registry under the project called "EUReMS".[2] The benefits can be large, with the ability to answer questions concerning MS epidemiological and clinical surveillance or long-term drug effectiveness and safety.[2] However considering the issues with disordered thresholds and DIF by rater, effort should be made to ensuring data comparability from different sources. We provide evidence that "inferential equivalence" needs to be tested even when the same measure is being pooled, and demonstrate a method to address issues of inter-rater variability when these are identified.

Figure 1. Flow diagram of the process underlying Rasch analysis

	Test/procedure and criteria
Step 1: Overall model fit	• Item-trait interaction χ^2 probability > 0.05[24, 34] (Bonferroni adjusted for the number of items)[35] • Item-person interaction: mean fit residual=0; SD=1[24] • Reliability estimate: Person Separation Index ≥ 0.7[25] for group use; ≥ 0.85[25] or 0.90[31] for individual use • Start with examining item and person misfit • Check data for data entry errors
Step 2: Item response category - threshold order	• Visually inspect category probability curves (CPC) and threshold probability curves (TPC) – Figures 2 and 3 • Identify disordered threshold by examining table of threshold estimates and graphic indicator such as the CPC[24]
Step 2: Item and person fit to model	• Re-examine fit residuals should be within range of -2.5 to 2.5[24] • Re-examine observed data fit to expect curves of the Item Characteristic Curves (ICC)[36]
Step 3: Local item dependency (LID): Response dependency and unidimensionality	• Fit residuals between two items with r > 0.3 indicate response dependency[29] • Examine the correlation of the residuals between two subset of items (the most positively and negatively correlated items) of the 1st component of the residuals from the principal component analysis (first residual factor). Separate person estimates for the two sets of items are performed using the independent *t*-test. The percentage of such tests outside the range -1.96 to 1.96 should not exceed 5 percent[30]
Step 4: Differential item functioning (DIF)	• Examine ICC by personal factor • Examine ANOVA (by personal factor, Class interval, or Class Interval X personal factor interaction)[25, 37] • Additional DIF analysis (optional)[21, 38, 39]
Step 5: Targeting	• Mean location of items and persons=0; SD=1[37] • Mean fit residual of items and persons=0; SD=1[37] • Examine Person-Item Threshold Distribution Map • A range of ± 4 logits will theoretically cover the full range of the latent variable being measured.
Step 5:	• Item-trait interaction χ^2 probability > 0.05[34] (Bonferroni adjusted)[35] • Item-person interaction: mean fit residual=0; SD=1[24] • Person Separation Index ≥ 0.7[25] for group use; ≥ 0.85[25] or ≥ 0.90[31] for individual use.

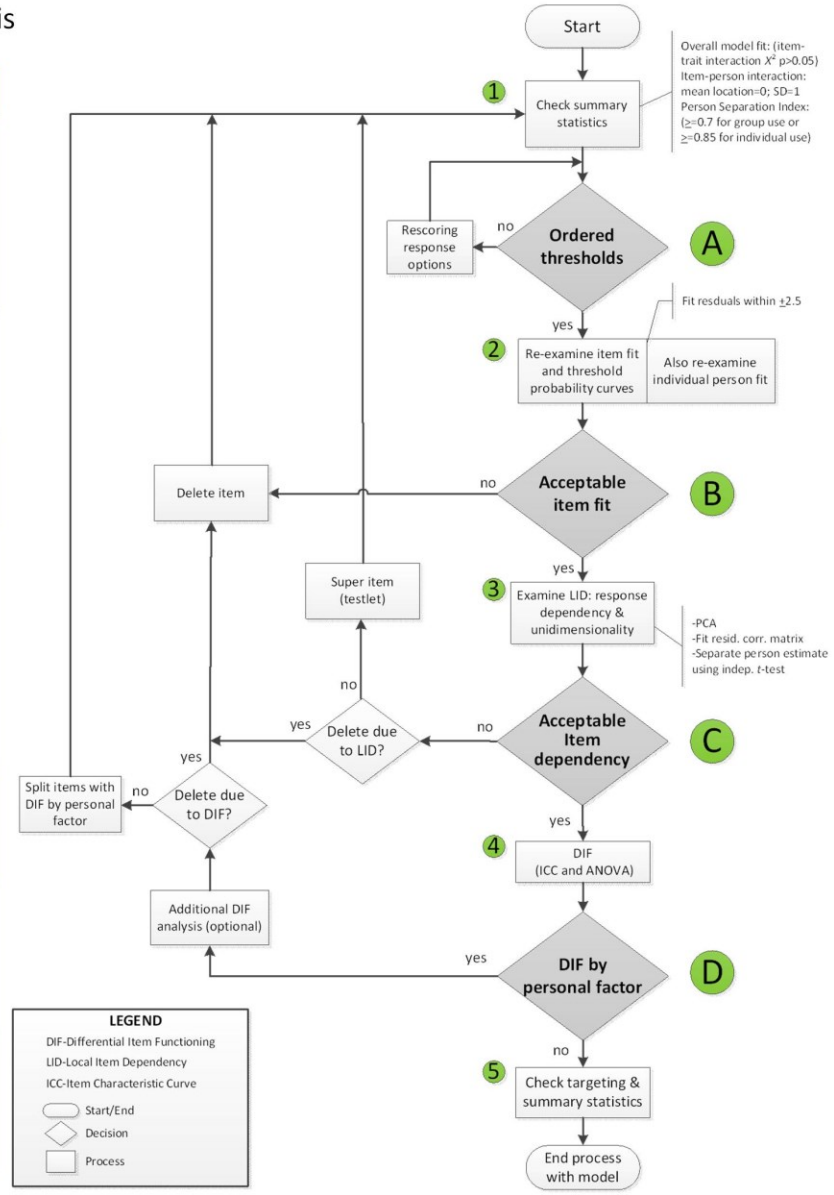


Table 1. Demographic and clinical characteristics of study sample

Sample characteristics	Entire sample (n=500)	Neurologist #1 (n=100)	#2 (n=100)	#3 (n=100)	#4 (n=100)	#5 (n=100)
Age at: mean (y) (SD)						
onset	32.7 (10.2)	32.3 (10.0)	32.2 (10.3)	33.4 (10.4)	32.2 (10.4)	33.7 (10.2)
Index visit 1	47.5 (12.9)	51.5 (11.4)	46.4 (12.4)	54.3 (13.0)	44.3 (12.1)	41.1 (11.3)
Index visit 2	48.5 (12.9)	52.6 (11.4)	47.4 (12.4)	55.3 (13.0)	45.3 (12.1)	42.1 (11.3)
Women n(%)	340 (68)	76	65	59	69	71
MS type, %						
RR/CIS	65.4	56.0	68.0	37.0	80.0	86.0
SP	24.2	33.0	21.0	48.0	14.0	5.0
PP	10.4	11.0	11.0	15.0	6.0	9.0
Median EDSS for visit 1 median (IQR 25%; 75%)	3.0 (1.5; 6.0)	3.5 (2.0; 6.5)	2.0 (1.0; 6.0)	6.0 (2.0; 7.0)	2.5 (1.5; 3.5)	2.5 (1.5; 4.0)
Median EDSS for visit 2 median (IQR 25%; 75%)	3.0 (1.5; 6.0)	3.5 (2.0; 6.125)	2.0 (1.0; 6.125)	6.0 (2.5; 7.5)	2.5 (2.0; 3.5)	2.5 (1.5; 4.0)

Table 2. Threshold ordering of the FSS and EDSS

		Entire sample racked (n=500)	Neurologist #1 (n=100)	Neurologist #2 (n=100)	Neurologist #3 (n=100)	Neurologist #4 (n=100)	Neurologist #5 (n=100)
Item	description	# thresholds	order	order	order	order	order
I0001	Pyramidal2	6	**	**	✓	✓	**
I0002	Cerebellar2	5**	**	**	**	✓	✓
I0003	Brainstem2	5	**	**	✓	✓	✓
I0004	Sensory2	6**	✓	**	✓	**	**
I0005	Bowel & Bladder2	6**	**	**	**	**	**
I0006	Visual2	6**	**	**	✓	**	**
I0007	Mental2	5**	**	**	**	**	**
I0008	EDSS2	9**	**	**	✓	**	✓
I0009	Pyramidal1	6	**	**	✓	✓	✓
I0010	Cerebellar1	5**	**	**	**	✓	**
I0011	Brainstem1	5	**	**	✓	**	✓
I0012	Sensory1	6**	✓	**	**	✓	**
I0013	Bowel & Bladder1	6**	**	**	**	**	**
I0014	Visual1	6	**	**	**	**	✓
I0015	Mental1	5**	**	**	**	**	**
I0016	EDSS1	9**	**	**	**	**	**
	Disordered n(%)	11(69)	14 (88)	16 (100)	9(56)	10(63)	10(63)

** disordered threshold

No consistent pattern of threshold order of these functional system assessments performed by the 5 neurologists

Table 3. Summary fit statistics for the Rasch analysis of the FSS and EDSS

Action	Overall model fit	Item Location	Item fit Mean	Person Location	Person fit Mean	PSI
original scale	X^2=179.750, df=48, p=0.000000	0.0 (0.649)	-0.739 (3.339)	2.152 (1.599)	-0.365 (1.238)	0.95
rescoring scale	X^2=71.462, df=42, p=0.0031	0.0 (1.012)	-0.588 (0.998)	3.800 (1.769)	-0.337 (0.698)	0.85
final model	X^2=72.554, df=45, p=0.0057	0.0 (1.161)	-0.680 (1.071)	3.521 (1.914)	-0.380 (0.641)	0.84

Table 4. ICC DIF summary statistics

		UNIFORM DIF			
Item	description	MS	F	DF	Prob
I0001	Pyramidal2	22.58	33.05	4	0.000000
I0002	Cerebellar2	0.67	1.12	4	0.347325
I0003	Brainstem2	0.36	0.57	4	0.687520
I0004	Sensory2	2.27	2.59	4	0.036480
I0005	Bowel & Bladder2	2.08	3.38	4	0.009620
I0006	Visual2	1.95	2.15	4	0.073540
I0007	Mental2	23.73	38.47	4	0.000000
I0009	Pyramidal1	2.54	4.31	4	0.002000
I0010	Cerebellar1	2.52	3.02	4	0.017740
I0011	Brainstem1	12.97	19.15	4	0.000000
I0012	Sensory1	1.45	1.78	4	0.132684
I0013	Bowel & Bladder1	10.04	16.02	4	0.000000
I0014	Visual1	0.98	0.98	4	0.417623
I0015	Mental1	22.28	27.20	4	0.000004

DIF summary statistics - highlighting Bonferroni probability adjustment at the 0.0001667 significance level

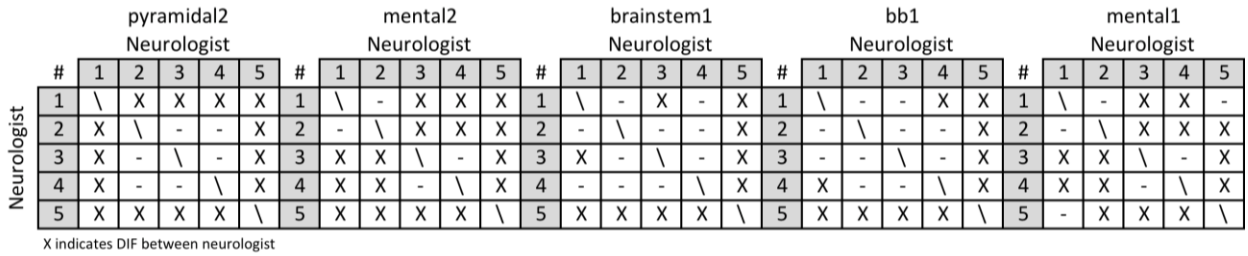


Figure 2. Results of Tukey test post hoc analysis to identify DIF between neurologist

Table 5. Final model after splitting items by neurologist and deleting misfit items

Item	Type	# of thresholds	Location	SE	FitResid	Prob
I0001A	Pyramidal2A-#1#5	3	0.669	0.142	-1.590	0.74
I0001B	Pyramidal2B-#2#3#4	3	2.653	0.118	-2.219	0.57
I0002	Cerebellar2	1	1.039	0.141	-1.605	0.06
I0003	Brainstem2	2	-1.713	0.134	-1.364	0.10
I0004	Sensory2	2	-0.747	0.114	0.211	0.44
I0005	Bowel & Bladder2	1	0.188	0.172	-0.539	0.21
I0006	Visual2	2	-1.255	0.123	0.329	0.07
I0007A	Mental2A-#3#4	DELETE	STEP4			
I0007B	Mental2B-#1#2	2	0.725	0.182	0.120	0.46
I0007C	Mental2C-#5	2	1.478	0.238	-0.233	0.01
I0008	EDSS2	DELETE	STEP1			
I0009	Pyramidal1	2	-0.149	0.163	-1.689	0.37
I0010	Cerebellar1	2	-0.617	0.115	-0.014	0.05
I0011A	Brainstem1A-#5	2	-0.169	0.263	-0.581	0.86
I0011B	Brainstem1B-#1#2#3#4	2	-0.099	0.131	-1.879	0.19
I0012	Sensory1	2	-0.835	0.115	-0.897	0.26
I0013A	Bowel & Bladder1-#1#5	DELETE	STEP3			
I0013B	Bowel & Bladder1-#2#3#4	DELETE	STEP5			
I0014	Visual1	2	-1.168	0.120	1.754	0.11
I0015	Mental1	DELETE	STEP2			
I0016	EDSS1	DELETE	STEP1			

#n refers indicate rating neurologist; letter suffix indicate split items

PROCEDURE	ITEM DELETED	DESCRIPTION & REASON
STEP1:	8; 16	DELETE EDSSs (redundant; fit residuals < -2.5)
STEP2:	15	observed data deviating from expected threshold probability curves
STEP3:	13A	BB1-#1#5; deleted-DIF was still present
STEP4:	7A	data deviating from expected threshold probability curves
STEP5:	13B	BB1-#2#3#4; overall model did not fit; observed data deviating from expected threshold probability curves

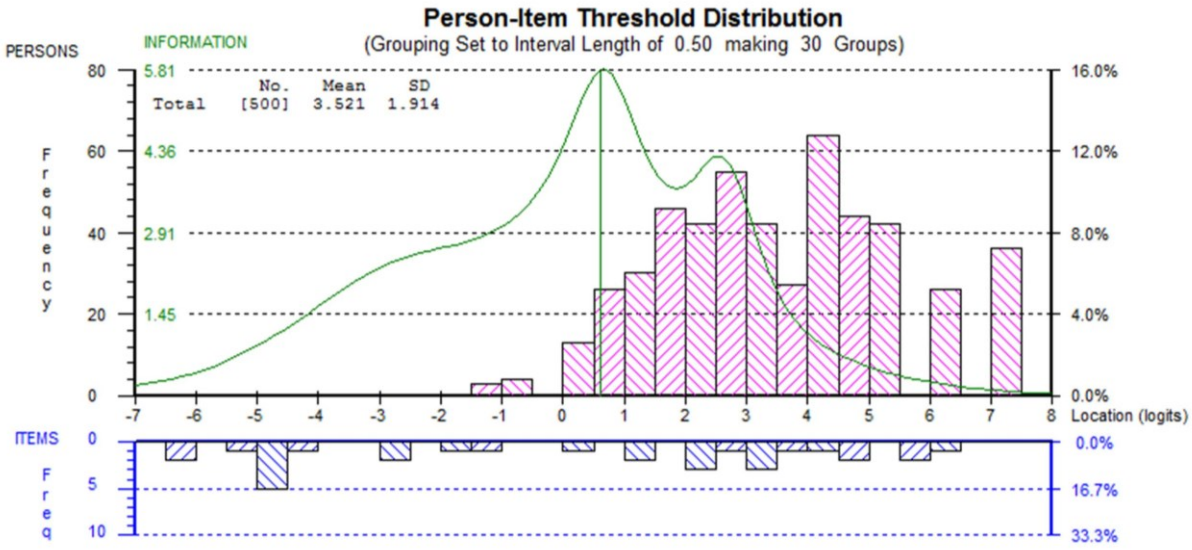


Figure 3. Person-Item Threshold Distribution

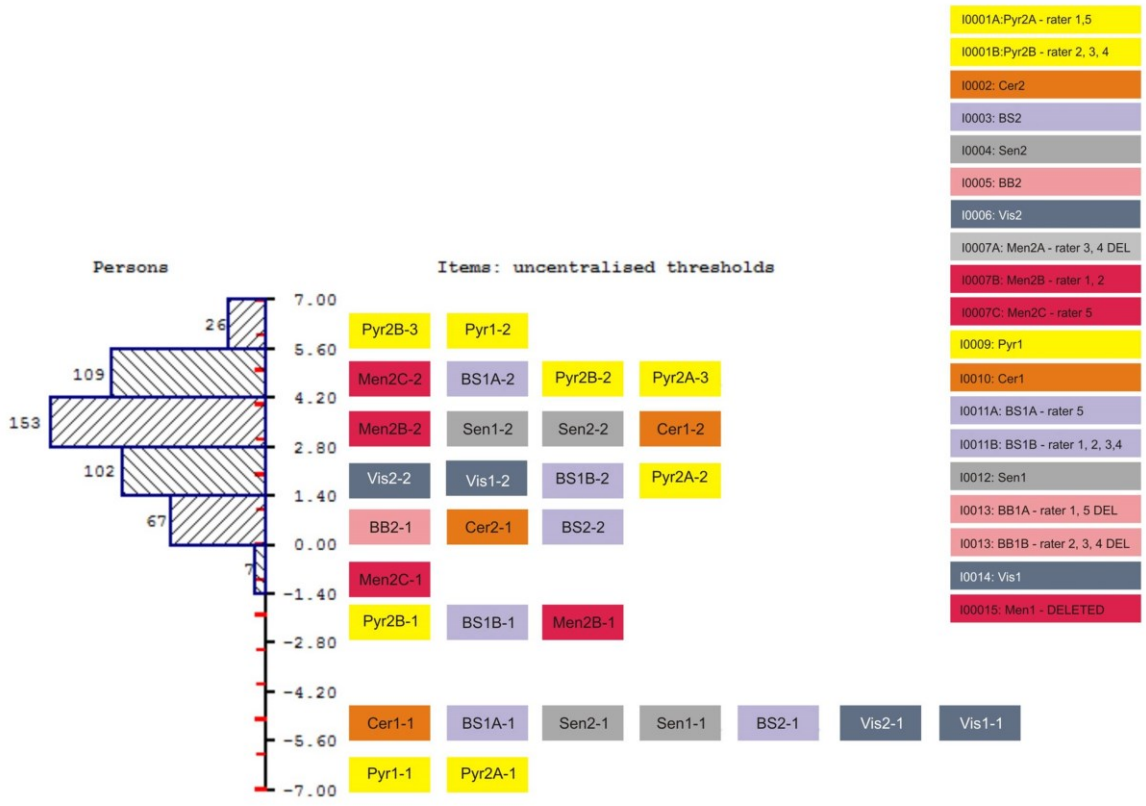


Figure 4. Item map

1. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 1983;33:1444-52.
2. Flachenecker P, Buckow K, Pugliatti M, et al. Multiple sclerosis registries in Europe – results of a systematic survey. Multiple Sclerosis Journal 2014.
3. Butzkueven H, Chapman J, Cristiano E, et al. MSBase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. Multiple Sclerosis 2006;12:769-74.
4. Hurwitz BJ. Analysis of current multiple sclerosis registries Neurology 2011;76:S7-S13.
5. Hurwitz BJ. Registry studies of long-term multiple sclerosis outcomes. Neurology 2011;76:S3-S6.
6. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. International Journal of Epidemiology 2010;39:1383-93.
7. Magalhaes S, Wolfson C. Harmonization: a methodology for advancing research in multiple sclerosis. Acta Neurologica Scandinavica 2012;126:31-5.
8. Sharrack B, Hughes RAC, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999;122:141-59.
9. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000;123:1027-40.
10. Meyer-Moock S, Feng Y-S, Maeurer M, Dippel F-W, Kohlmann T. Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. BMC Neurology 2014;14:58.
11. Goodkin DE. EDSS reliability. Neurology 1991;41:332.
12. Kragt JJ, Nielsen JM, van der Linden FA, Uitdehaag BM, Polman CH. How similar are commonly combined criteria for EDSS progression in multiple sclerosis? Multiple Sclerosis 2006;12:782-6.
13. Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC. Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. Neurology 1990;40:971-5.
14. Goodkin DE, Cookfair D, Wende K, et al. Inter‐ and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS). Neurology 1992;42:859.
15. Lechner-Scott J, Huber S, Kappos L. Expanded Disability Status Scale (EDSS) training for MS-multicenter trials. Journal of Neurology 1997;244:S25.
16. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institution for Educational Research; 1960.
17. Cano SJ, Mayhew A, Glanzman AM, et al. Rasch analysis of clinical outcome measures in spinal muscular atrophy. Muscle & Nerve 2014;49:422-30.
18. Bond TG, Fox CM. Applying the Rasch Model: Lawrence Erlbaum Associated; 2001.
19. Fortier I, Doiron D, Burton P, Raina P. Invited Commentary: Consolidating Data Harmonization—How to Obtain Quality and Applicability? American Journal of Epidemiology 2011;174:261-4.

20. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technology Assessment 2009;13:1-177.
21. Tennant A, Penta M, Tesio L, et al. Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales through Differential Item Functioning within the Framework of the Rasch Model: The PRO-ESOR Project. Medical Care 2004;42:I37-I48.
22. Kurtzke JF. On the evaluation of disability in multiple sclerosis. Neurology 1961;11:686-94.
23. Kappos L, Lechner-Scott J, Lienert C. Neurostatus.net. 2007.
24. Pallant JF, Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology 2007;46:1-18.
25. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care & Research 2007;57:1358-62.
26. Wright BD. Rack and Stack: Time 1 vs. Time 2. Rasch Measurement Transactions 2003;17:905-6.
27. Linacre JM. Sample Size and Item Calibration [or Person Measure] Stability. Rasch Measurement Transactions 1994;7:328.
28. Andrich D, Lyne A, Sheridan B, Luo G. Rasch Unidimensional Measurement Models (RUMM2020 Version 4.1). Duncraig, Western Australia: Rumm Laboratory Pty Ltd; 2003.
29. Ramp M, Khan F, Misajon R, Pallant J. Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). Health and Quality of Life Outcomes 2009;7:58.
30. Smith EV. Detecting and Evaluating the Impact of Multidimensionallity using Item Fit Statistics and Principal Component Analysis of Residuals. Journal of Applied Measurement 2002;3:205-31.
31. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. Brain 2012;135:1639-49.
32. Cohen JA, Reingold SC, Polman CH, Wolinsky JS. Disability outcome measures in multiple sclerosis clinical trials: current status and future prospects. The Lancet Neurology 2012;11:467-76.
33. Andrews FM. Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. Public Opinion Quarterly 1984;48:409-42.
34. Mills RJ, Young CA, Nicholas RS, Pallant JF, Tennant A. Rasch analysis of the Fatigue Severity Scale in multiple sclerosis. Multiple Sclerosis 2009;15:81-7.
35. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310.
36. Marais I. Local Dependence. In: Christensen KB, Kreiner S, Mesbah M, eds. Rasch Models in Health. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2013:111-30.
37. Muller S, Roddy E. A Rasch Analysis of the Manchester Foot Pain and Disability Index. Journal of Foot and Ankle Research 2009;2:29.
38. Scott N, Fayers P, Aaronson N, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health and Quality of Life Outcomes 2010;8:81.

39. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Medical Care 2006;44:S115-S23 10.1097/01.mlr.0000245183.28384.ed.

Chapter 10

Linking chapter 9 (manuscript 2) and chapter 11 (manuscript 3)

The previous chapter illustrated an application of Rasch analysis in the context of data harmonization as no one neurologist used the scoring structure of the FSS in the same way. In order to pool data across multiple sources the methodology of using Rasch analysis with a racked data format provides a way of adjusting FSS scores to be equivalent across neurologist. This method used existing data available on the FSS, but neither the FSS nor the EDSS can be considered a comprehensive measure of physical disability as the domains covered relate to neurological impairment and ambulation. In addition, it would be preferable to have a measure that was not based on interpretation but would contain items that are directly measured or the information is obtained directly from the patient. The EDSS, as a ClinRO, requires interpretation by the neurologist and as shown in the previous manuscript, neurologists do not use the scoring system in the same way. An optimal measurement disability measure would cover domains beyond ambulation, and have mathematical properties to permit accurate estimation over time. The next chapter shows the development of such as measure.

Chapter 11 (Manuscript 3)

Integrating patient reported outcomes, performance-based tasks, and clinician reported outcomes to produce a linear hierarchical unidimensional measure of physical disability in MS

Stanley Hum[1] and Nancy E. Mayo[1,2]

[1]School of Physical and Occupational Therapy, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.

[2]Division of Clinical Epidemiology, McGill University Health Center, Montreal, Quebec, Canada

For Submission to the Journal of Rehabilitation Medicine

Communication addressed to:

Stanley Hum, MSc, Ph.D. Candidate
School of Physical & Occupational Therapy
Faculty of Medicine, McGill University
3654 Prom Sir William Osler
Montreal, Quebec, H3G 1Y5
Canada
Tel: 514-398-5981
Email: Stanley.Hum@mcgill.ca

ABSTRACT

Multiple Sclerosis (MS) disability can only be measured indirectly by how it is manifested. The goal is to develop a comprehensive linear hierarchical unidimensional measure of physical disability that includes the patient perspective and is relevant to a multidisciplinary team.

The conceptual framework for this study was the International Classification of Functioning, Disability, and Health (ICF). Several ICF MS core sets provided content validity for this analysis. Rasch analysis was used to develop the disability measure and provided evidence of construct validity.

This is a secondary analysis of information from 189 MS patients comprehensively assessed on health indices. Performance outcomes (PerfOs), clinician reported outcomes (ClinROs), and patient reported measures (PROs) were assessed at baseline and questionnaires were assessed twice (6 months apart). A total of 136 items were selected from domains of body function impairment and activity limitation as a starting point.

RUMM2020 software was used to fit a partial credit Rasch model. Items with continuous scores were categorized by their frequency distribution. Measurement fit to the Rasch model was tested for threshold order, model fit, response dependency, unidimensionality, DIF, and item/person targeting. All p values < 0.05 after Bonferroni adjustment were considered to be significant.

Disordered thresholds were observed in 55.9% of the items. Residual correlation ≥ 0.3 was observed in 266 item pairs. The final 22-item physical disability measure fit the expectations of the Rasch measurement model. There was some mistargeting with mean person location of 1.595 (SD=2.188). Person separation index was excellent at 0.90. Three PRO measures contributed items to the final measure: the RAND36-PFI (5 item), the Preference Based MS Index–V1 (PBMSI-V1; 1 item), and the Disabilities of the Arm, Shoulder, and Hand (DASH; 7 items). A total of 8 PerfOs were also included in the measure: 6 minute walk, 9hole peg test dominant hand, push-ups, partial-curl-ups, gait speed (comfortable and fast), VO_2 max, and vertical jump. The Equi-balance contributed one item,

a ClinRO. PROs were consistently located on the left (less difficult) end of the continuum and PerfOs were located on the right (more difficult).

PROs, PerfOs and ClinROs health indices can co-exist in a linear hierarchical unidimensional measure of physical disability. However, these items were not evenly distributed along the linear continuum. PROs were easier than PerfOs.

Introduction

In MS, the neurological damage can be diverse, and depending on the location of the lesions, different disabilities may manifest. Disability is a latent construct that cannot be measured directly, but can be inferred by observing how a person performs on tasks or reports on symptoms within a disability framework.

The internationally accepted conceptual model for disability is the World Health Organization's International Classification of Functioning, Disability, and Health, which is known also as the ICF. The ICF provides a standard framework and common language with which to identify the presence and severity of disability.[1] Disability is the umbrella term that refers to all of the negative aspects of function, namely impairments, activity limitations and participation restrictions. The ICF is organized by domain, category, and classification with an increasing level of detail used to describe disability.[1] MS disability can be comprehensively described as shown in Figure 1, which lists the relevant ICF domains. ICF Core sets of MS disabilities have been identified[2-8] at the category levels that reflect the different perspectives of patients, clinicians, and researchers in a common language.

The ICF framework is intended to provide assistance in deciding what to measure and report, but does not indicate how to measure disability.[7] At the impairment level, the disabilities include symptoms that can be measured from the perspective of the patient with no interpretation from any other observer, termed a patient reported outcome (PRO). Body function impairments, such as signs related to the central nervous system and muscle disabilities usually assessed by a trained clinician using outcomes termed clinician reported outcomes (ClinRO).[9]

At the activity and participation level, the ICF recognizes the difference between "capacity", what the person can do in a testing situation, and "performance", what the person actually does in their own environment, usually identified through self-report on activities and participation termed "self-reported outcomes" (SRO). Currently, the literature does not distinguish between PRO and SRO although clearly they are different, and pose unique measurement challenges.

Tests of capacity have "bio-physical" units but assess only narrow aspects of disability and many tests have to be administered to fully represent the spectrum of disability in MS. A feature of bio-physical units, such as height or weight, is that they are on an interval scale with a natural zero. In the measurement literature, tests of "capacity" have traditionally been called "performance-based" outcomes (PerfO), adding to the confusion about measuring disability. We will use PerfO to refer to these tests of capacity. How the person functions in their own environment is usually captured by asking the patient a series of questions about the degree to which they are limited or not in activities that are normal for a person to do in everyday life. The majority of these self-report indices use rating scales that have different numbers of ordinal categories to represent the underlying (latent) construct. The use of ordinal scales provide information useful only to rank people according to different levels of disability;[10-12] in order to provide information on their actual level of disability interval or ratio scales are more optimal to measure change.[11,12] A ratio scale is an interval scale with a true zero value such as degrees Kelvin or a test of capacity such as gait speed or distance walked. This allows ability to be measured from none to complete, and also to situate people relative to each other. Only interval scales can legitimately be used in mathematical operations. To meet this requirement, Rasch analysis has been used to transform ordinal rating scales to interval-like scales using a logit transformation.[13]

The main MS disability measure is the Expanded Disability Status Scale (EDSS)[14], referred to as the gold standard.[15] The EDSS has 20 grades of impairment, with scores ranging from 0 (normal) to 10 (death due to MS) with 0.5 increments above 1.0. The scale is based in part on the neurological examination with eight functional system scores (FSS). The FSS were developed to help score the grades of the EDSS from 0 to 4.0; above EDSS 4.0, categories are based solely on ambulation. While useful as a "shorthand" for MS specialist[16] it is limited as a measure as something as complex as disability.

The Multiple Sclerosis Functional Composite (MSFC) was an early attempt to provide an alternative to the EDSS. The MSFC, a multi-component PerfO, comprised 3 parts: the Paced Auditory Serial Addition Test (PASAT), 9-Hole peg test and 25-foot walk, measuring

cognition, upper and lower limb function, respectively. A total score is derived from summing *z*-transformed scores on the subtest.[17] The disadvantage is that it takes approximately 15 minutes to administer, there are practice effects, and the interpretation of *z*-scores across trials using different reference populations is problematic.[15,18]

The use of the EDSS in the research community is waning. There is a need to identify or develop new psychometrically sound measures of MS disability. The FDA organized a workshop bringing together industry, regulatory agencies, and academia to discuss the need for psychometrically *strong* clinical outcome measures to help drug development with the specific goal of improving drug development in MS.[19] The organizer's approach is to leverage existing data from MS clinical trials representing an aggregate of over 20 000 patients assessed on existing legacy outcome measures. Their goal was to produce a composite multi-dimensional measure from existing ClinROs to meet the specific need for an efficacy primary endpoint in clinical trials aimed at reducing, stopping, or reversing MS disability progression.[20]

Currently there is no "core set" of outcome measures used to assess MS disability.[21] Even for interventions that are specifically designed to target disability, a Cochrane review identified an urgent need for a core set of outcomes.[22] The American Physical Therapy Association organized a taskforce to address the barriers in using and selecting outcome measures to assess MS disability. The taskforce identified 120 outcome measures (PerfO, ClinRO, and PROs) that have been used to assess MS disability and recognized that the sheer number of indices available would be a barrier to clinicians and researchers selecting appropriate outcome measure(s) in MS. Additionally, the taskforce understood that MS encompasses a wide range of disabilities experienced by patients over the course of the disease, making the selection of outcome measures challenging. The taskforce selected 63 outcome measures to review and made recommendations on when and how to use each measure.[23] With the large number of outcome measures available to assess different aspects of disability, there is a need for a method to test how they work together to comprehensively measure people with MS.

Rasch Model

Rasch analysis uses an experimental paradigm to develop new measures or to re-evaluate the psychometric properties of existing "legacy measures" that have been developed by more traditional methods.[12,24,25] Rasch analysis can be used to verify whether the common practice of adding ordinal scores from rating scales is appropriate. When the data "fits" the expectations of the underlying Rasch model, ordinal scores are transformed to interval-like scores.[13] In the context of measuring MS disability, Rasch analysis allows the ability of patient to be quantified along a hierarchical linear continuum from less ability to more ability.

Rasch analysis has been used to combine existing indices into a single measure.[26] Redundant items and "poorly" functioning items that do not meet the expectations of the Rasch model could be removed and still retain the measurement properties. Ability to create shorter forms of the measure can dramatically reduce response burden for patients and administration for clinicians.

As disability is a complex construct and cannot be measured from one perspective only (performance on tests, patient's experiences, clinician assessment), an optimal measure would integrate all these perspective into a common latent variable. Rasch analysis is ideal for this integration and only requires creating meaningful ordinal cut points for performance tests that are measured on a continuous scale. PerfOs and PROs are beginning to be combined using Rasch analysis to form single unidimensional measures, examples are from traumatic brain injury and stroke.[27,28]

Objectives

The overall goal of this analysis is to develop a comprehensive unidimensional measure of physical disability. The specific objective is to contribute evidence for the extent to which commonly used disability tests and questionnaires fit a unidimensional hierarchical linear continuum reflecting the spectrum of MS physical disability.

Methods

This is a secondary analysis of an existing dataset from a CIHR funded study titled "Gender Differences in the Life Impact of Multiple Sclerosis" conducted in 2007.[29] Patients with disease onset after 1994 were randomly selected from databases at the three largest tertiary care outpatient MS Clinics (Montreal Neurological Institute (MNI), Centre Hospitalier de l'Université de Montréal (CHUM), and Neuro Rive–Sud (NRS)) from the Greater Montreal area, Quebec, Canada. Information on 189 MS patients who were comprehensively assessed in person (including interview assisted completion of questionnaires as needed) on health indices of body function impairments, activity limitations, participation restrictions, illness intrusiveness, and health-related quality of life were included in the dataset. PerfOs and PROs were assessed at baseline with questionnaires reassessed at a 2nd time point (6 month). A computerized database was used to record all items from tests and questionnaires. A dataset of 271 items were potentially available for this analysis. Sex ratio was 2.86 females to 1 male. Median EDSS was 2.0. Mean age of disease onset was 34.2 years (SD 9.7) and mean age at study baseline was 43.0 years (SD 10.2). The dataset comprised patients from MNI=65, CHUM=64 and NRS=60.

Measures

The general ICF model was the conceptual framework for function (disability) in the original study to provide content validity in selecting a set of items that reflect the impact of MS on patients. This process included a multi-disciplinary team of MS specialists, epidemiologists, physical and occupational therapists, psychologists, and psychiatrists in reviewing and supplementing items deemed important in assessing the impact of MS on all domains of function and health related quality of life.

To identify the items for inclusion in the Rasch analysis the available items were mapped to domains from the several MS core sets.[2-8] Details of the ICF core sets and checklists are summarized in Table 1a, included in the appendix. This permits the identification of items

related to body function impairments and activity limitations as the basis of the MS physical disability measure (MS-PDM).

Although cognition, mood, and emotions are important components of functioning, there is evidence to suggest that these domains do not fit the construct of physical disability. [30] Fatigue items were included in the measure because there are components presumably related to MS pathology[31] and there is evidence that fatigue predicts change in physical activity.[32] Although Bakshi et al., showed that fatigue was not associated with physical disability.[33] and Chamot et al., showed that fatigue groups with mental disability.[34] From the 271 items available, 136 items were selected for the measure development and summarized in Table 1. Items related to participation were not included in the measure as participation is affected directly by impairments and activity limitations.[35]

Statistical methods

The steps for the Rasch analysis were adopted from published guidelines[13,36] and are described in Figure 2. RUMM2020 version 4.1 software was used to fit a partial credit Rasch model.[37] PerfO's with continuous scores were categorized using their distributions and known clinically meaningful values. The greatest number of clinically relevant cut points supported by the data was used to allow for flexibility in identifying the optimal cut points to represent the latent construct. The increment units used for the following PerfOs are summarized in Table 2a in the appendix.

Construct validity is supported when the selected items fit the expectations of the Rasch model.[38] Convergent and discriminant validity were estimated with correlation analysis: a correlation from 0.4 to 0.8[39] between the MS-PDM and similar constructs was used as supporting evidence for convergent validity; correlations <0.4 between MS-PDM and dissimilar constructs was used as supporting evidence for discriminant validity. Constructs for the validity analysis were those that did not contribute items to the MS-PDM. Additionally, measures contributing item(s) to the final measure should also correlate with it. This will provide evidence that the construct of the new measure is similar to the

construct of the measure contributing items. Also, in the final model, the EDSS was reintroduced to the MS-PDM to situate the measurement items to a well know reference.

Sample size: To have a stable item calibration (precision) within ± 0.5 logits with a 99% confidence level between person and items estimates, an average sample size of 150 is needed. Sample size requirement can vary between 108 to 243 depending how well the population is targeted.[40] This dataset was sufficient to achieve this level of precision.

Results

Demographic and clinical characteristics of the sample are summarized in Table 2. The preliminary set of 136 items summarized in Table 1 consists of PerfOs, PROs, and ClinROs assessing MS ICF domains of impairment or activity limitations. Additional details of each item are included in the appendix in Table 3a. In the initial set of items, 76 items had disordered thresholds and were rescored. For example, the EDSS had disordered thresholds (see Figure 3) and did not fit the final model of the latent construct of disability. The chi-square probability at this point was significant after Bonferroni correction, thus failing to support the model and summarized in Table 3. Figure 4 summarizes the individual steps to systematically remove items with misfit, local item dependency (LID), or redundant to obtain a working model of the MS-PDM.

A total of 28 items fit the criteria of the Rasch analysis and was included in the working model of the MS-PDM. All items had fit residuals within -2.5 and +2.5 as shown in Table 4. In Table 3, the model fit described by the item-trait interaction statistics, was not significant (χ^2=95.099; df=84; p=0.191), indicating appropriate overall fit to the expectations of the Rasch model. Mean item fit residual was -0.420 (SD 1.000), mean person residual was -0.314 (SD 0.599), and mean person location was 2.279 (SD 2.128). PSI was 0.91.

The first component of the residuals in the principal component analysis was 9.29% and the proportion of t values outside ± 1.96 from the two most different subsets of items was only 3.23% indicating unidimensionality. There was no DIF by time however several items had DIF by gender. There was DIF for Item #266 (VO_2 max), #252 (vertical jump) #161 (Carry a heavy object over 10 lbs) and #155 (Open a heavy door), depicted in Figures 5 to 8

respectively. These items were adjusted by splitting the items by gender. Additional LID was then detected between 13 item pairs. To adjust for LID, items (155Male»154»161Male»153»47»34»152»161Female) were deleted individually in that order and summarized; see Figure 4.

A final 22-item MS-PDM model fit is summarized in Table 3. The item-trait interaction statistics were not significant (χ^2=45.334; df=48; p=0.583), indicating appropriate fit to the expectations of the Rasch model. Mean item fit residual was -0.371 (SD 0.867), mean person residual was -0.255 (SD 0.469) and a mean person location was 1.595 (SD 2.188). The final PSI was excellent at 0.90. Details of each item location and fit statistics are listed in Table 5. Three PRO measures contributed items to the final measure: the RAND36-PFI (5 items), the PBMSI-V1 (1 item), and the DASH (7 items). Also included in the measure are 8 PerfOs (6 minute walk, 9hole peg test dominant hand, push-ups, partial-curl-ups, gait speed (comfortable and fast), VO_2 max, and vertical jump) and 1 item from the Equi-balance, a ClinRO.

Figure 9 shows the Person-Item Threshold Distribution map person histogram was skewed to the right with the mean person location at 1.595 logit, indicating that the sample had an average higher functioning (lower disability) than the average difficulty of the items available. Person location ranged from -7.0 to 7.0 logits. The inverted histogram for item thresholds ranged from -7.2 to 8.4 logits but shows gaps in item coverage at the low and high end of the disability spectrum. The proportion of MS patients that obtained the highest (extreme) scores in the MS-PDM was 16.9% indicating a ceiling effect. There were no patients with the worst (extreme) score, indicating no floor effect.

Figure 10 shows the threshold location of each item along the MS disability scale. PROs had item difficulty ranging from -7.3 to 2.3 logits whereas PerfOs had item difficulty ranging from -2.1 to 8.5 logits. PROs and PerfOs overlap between -2 to +2 logits.

Reintroducing the EDSS in the final model provided a crude estimate of the location of the items of the MS-PDM with reference to the EDSS. After collapsing the response categories to adjust for disordered threshold the EDSS had two thresholds; the first between EDSS 0 and (1-3) and the second between EDSS (1-3) and ≥ 4. The thresholds between EDSS 0 and (1-3)

were 5.160 and 3.290 logits and the thresholds between EDSS (1-3) and EDSS ≥ 4 were 1.476 and 0.629 logits for men and women respectively as shown in Figure 10. The EDSS impact on all the items was a minimal shift in all items of between 0.1 and 0.2 logits higher (harder). Although the EDSS has some issues of LID with the items of the MS-PDM we simply wanted to situate these items along the continuum of our measure.

Summarized in Table 6, are the results of the correlation analysis to estimate convergent and discriminant validity. The MS-GDPM correlation was moderate to high, with measures accessing similar domains with *r* ranging from 0.58 to 0.76 and r_s between 0.59 to 0.69. All *p* values were < 0.01. There was low correlation between MS-PDM and several measures accessing domains distinct from physical disability. The correlation of the MS-PDM with cognition was <0.4 for the PDQ. Measures of mood had low correlation with the MS-PDM having *r* values between 0.169 and 0.256. Interestingly, fatigue items consistently did not fit the latent construct of our Rasch analysis, but correlated with the fatigue questionnaire at r=-0.592 and r_s=-0.616 with the RAND36-Vitality subscale at r=0.438 and r_s=0.466. All *p* values were < 0.01. The correlations of the new MS-PDM and each of the individual measures that contributed items(s), "feeder measures" were calculated and summarized in Table 6. The average correlation of the feeder measures with the MS-PDM was r=0.77 (95% CI: 0.68-0.84) and r_s=0.75 (95%CI (0.68-0.82)

Discussion

A 22-item prototype measure was developed consisting of 8 PerfOs and items from 3 PROs and 1 ClinRO health indices commonly used to assess physical disability in MS patients. PerfOs, PROs, and ClinROs were successfully integrated as hypothesized indicating a measure of disability can reflect multiple perspectives. The advantage is that the measure will have meaning and relevance for all stakeholders in the MS experience. Reliability of the measure was excellent.

To achieve this, typical issues of disordered thresholds, item misfit, LID, and DIF were continuously tested. Disordered threshold were observed in 55.9% (76/136) of the items in the Rasch analysis of the MS-PDM and required rescoring by collapsing the response

categories. Disordered threshold in the PROs provide evidence that the measurement item is not functioning monotonically as expected. The rater cannot reliably select a score to reflect the level of disability. Rescoring by collapsing response categories was used to adjust for disordered thresholds.

Item reductions in Rasch analysis were achieved by removing items that misfit or were redundant. This study was able to identify items related to fatigue, pain, and functions associated with spinal lesions such as bladder and sexual function that consistently did not fit the construct of physical disability. Although these disabilities are important to the MS patient, they appear not to have a direct impact on physical disability. Fatigue not fitting within the MS-PDM may be explained by a lack of an accepted definition of MS fatigue. It is difficult to distinguish between "mental" and "physical" MS fatigue. Additionally there is also "primary" fatigue that is related to MS pathology and "secondary" fatigue that is indirectly related to MS. All of which may impact what is actually being measured.[31,41]

A large proportion of items had LID suggesting the presence of redundant items. Of the 90 items remaining, 93 item pairs had residual correlations indicating LID. The RAND36-PFI, ABC confidence scale, Ashworth scale, and Equi-balance had LID with their respective original scales. The Equi-balance had 20 items pairs with residual correlation above 0.3. Some of the LID may be explained in that the original measure was developed with a relatively small number of MS patients (n=55) and the authors of the Rasch version of the Equi-balance had only eight items, having removed two items suspected of being redundant.[42] LID was similarly observed in the RAND36-PFI subscale, the DASH, and ABC confidence scale. For the RAND36-PFI subscale[43] and DASH[44] previous Rasch analysis showed LID with items within the original measure.

PerfOs

Vertical jump and VO_2max are the most difficult items in the measure respectively. These two items may be good indicators of milder changes in the patient's physical status. Failing either item may indicate deviation from the person's optimal physical function. Many of the PerfOs have different expected scores based on gender. As expected, vertical jump and

VO_2max had DIF by gender (Figure 10). There was no DIF for push-ups as women and men did different forms of the test. For partial curl-ups, the final response category was binary 0 (0 to 24 partial curl-ups) and 1 (25 partial curl-ups). There were few patients between the two extremes with similar results for men and women resulting in no DIF.

MS can affect components of muscle control, balance, and coordination impacting gait speed.[45] Comfortable gait speed and fast gait speed items co-exist in the same Rasch measure with the former item being more difficult. This suggests that these two items are measuring different components of MS physical disability. Gait speed (comfortable) has been used to predict risk of mortality or other health conditions and is recommended as a "vital sign" in different medical conditions[46] whereas gait speed (fast) is a possible indicator of independent community walking.[47]

How well the items assessing physical disability match the level of function (disability), or targeting, can be shown by the Person-Item Threshold Distribution Map, Figure 9. Overall, the sample of MS patients had lower levels of disability than the average difficulty of the items indicating some mistargeting. The mean person location was 1.595 (SD=2.188). There were some gaps in item coverage at the low disability end of the scale as shown in Figure 9. All patients with extreme scores were at the high physical functioning end of the measure indicating a ceiling effect.

Depending on the measure's purpose, the impact of the gaps and ceiling effect can be minimal. The vertical jump is the hardest item; one might judge that a patient has little physical disability if they can jump and have not deteriorated if they maintain this ability. On the other hand even with minimal disability, if the patient loses the ability to achieve the highest item (e.g. jump), that may be an important time to intervene, and not wait until the person has accumulated irreversible disability. If the interest is in identifying patients at risk of falling or unable to live independently in the community one might only be interested in items below a certain cut-off score. If one is interested in detecting patients at risk of falling or assessing independent living in the community, one might only be interested in patients below a certain cut-off score.

There was overlap between the PROs and PerfOs between -2 and 2 logits. MS patients with disability within this middle range of the MS-PDM would be assessed with both PROs and PerfOs. Theoretically it would be possible to the use of PROs to assess patients in this range when it was not practical to use PerfOs and still obtain results on the same measurement scale.

However, PROs and PerfOs did not integrate through out the entire range of the measure; many of the PRO items were only located at the easier end of the scale (more disability) and PerfO items at the harder end (low disability). This supports previous research that PerfOs may detect functional limitations in ADL before PROs.[48] This may be because what people think they can do (self-report) provides different information than what they can actually do (task).[49] By combining PROs and PerfOs in a single measure using Rasch analysis, we can estimate which PRO items have similar difficulty locations on the "ruler" with PerfOs and these could be interchanged where they overlap. The fact that PROs and PerfOs are well distributed along the continuum suggests that these items provide different and useful information on the single construct of physical disability.

ClinRO

The only ClinRO that fit the disability construct was one item from the Equi-balance scale. Notably the EDSS did not fit. There was DIF for gender for the EDSS when forced back into the model such that men were assigned an EDSS level of 4.0 with a lesser degree of disability than women. As in standard clinical practice, the ability to walk 500 metres without a walking aid, the criterion to remain at EDSS 4.0 is not usually measured but inferred from observation and patient interview.

To provide some context for the level of disability measured by the MS-PDM, the EDSS was reintroduced into the measure as a variable with two thresholds, one to distinguish between EDSS 0 and EDSS (1-3) and one for EDSS ≥ 4. These were the only thresholds that showed a monotonic increase with disability. Most of the PerfOs overlapped the two EDSS thresholds indicating adequate coverage of the important lower range between EDSS 0- 4. At the higher disability end of the EDSS threshold (≥4), assessment can be made with PROs.

Limitations

Presented here is only a prototype measure. It requires refinement and retesting on a different sample of MS patients. Not all items were assessed at a second time point. A larger sample over time and with more men will provide better estimates of DIF by time and gender.

Ideally PerfOs would be integrated with PROs so that an alternative short form test could be selected with an even distribution of items with varying difficulty. Although we were able to merge both types of measures, PROs were located at the lower end of the continuum and PerfOs at the higher end, as seen in Figure 10.

Conclusion

We were able to select measurement items from health indices commonly used to assess MS disability to produce a 22-item prototype measure. We provided a method to reduce the assessment burden to both the clinician and patient. Although no single measure can serve all purposes and all populations, the MS-PDM covered as wide range of disability levels and has good reliability. We were able to include performance based and patient reported outcomes into a single measure of MS physical disability.

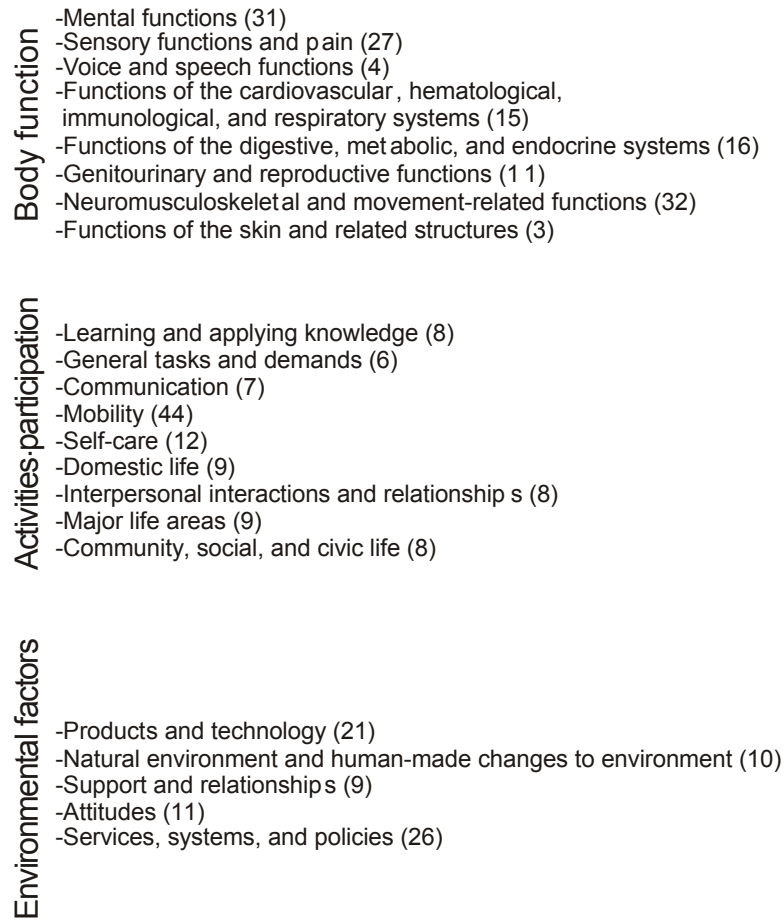


Figure 1. ICF domains related to MS disability

Table 1. 136 items included in the analysis (number of items included)

Performance based measures	Patient reported outcome	Clinical reported outcome
Vertical jump	RAND36-PFI (10)	Ashworth scale (14)
2 minute walk test	RAND36-Pain (2)	Equi-balance (10)
6 minute walk test	RAND36-Vitality (4)	EDSS
9 hole peg test (LR)	Fatigue (19)	
Push ups (full for men; modified for women)	ABC confidence scale (16)	
Partial curl ups	Disabilities of the Arm, Shoulder, and Hand (DASH) (21)	
VO_2 max	Perference Based MS Index Version 1 (9)	
Gait speed:confortable & fast (2)	EQ-5D (4)	
Grip strength	Symptom checklist (15)	

Figure 2.Flow diagram of the process underlying Rasch analysis

	Test/procedure and criteria
Step 1: Overall model fit	• Item-trait interaction χ^2 probability > 0.05[36, 50] (Bonferroni adjusted for the number of items)[51] • Item-person interaction: mean fit residual=0; SD=1[36] • Reliability estimate: Person Separation Index ≥ 0.7[13] for group use; ≥ 0.85[13] or 0.90[52] for individual use • Start with examining item and person misfit • Check data for data entry errors
Step 2: Item response category - threshold order	• Visually inspect category probability curves (CPC) and threshold probability curves (TPC) – Figures 2 and 3 • Identify disordered threshold by examining table of threshold estimates and graphic indicator such as the CPC[36]
Step 2: Item and person fit to model	• Re-examine fit residuals should be within range of -2.5 to 2.5[36] • Re-examine observed data fit to expect curves of the Item Characteristic Curves (ICC)[53]
Step 3: Local item dependency (LID): Response dependency and unidimensionality	• Fit residuals between two items with r > 0.3 indicate response dependency[54] • Examine the correlation of the residuals between two subset of items (the most positively and negatively correlated items) of the 1st component of the residuals from the principal component analysis (first residual factor). Separate person estimates for the two sets of items are performed using the independent *t*-test. The percentage of such tests outside the range -1.96 to 1.96 should not exceed 5 percent[55]
Step 4: Differential item functioning (DIF)	• Examine ICC by personal factor • Examine ANOVA (by personal factor, Class interval, or Class Interval X personal factor interaction)[13, 56] • Additional DIF analysis (optional)[57-59]
Step 5: Targeting	• Mean location of items and persons=0; SD=1[56] • Mean fit residual of items and persons=0; SD=1[56] • Examine Person-Item Threshold Distribution Map • A range of ± 4 logits will theoretically cover the full range of the latent variable being measured.
Step 5:	• Item-trait interaction χ^2 probability > 0.05[50] (Bonferroni adjusted)[51] • Item-person interaction: mean fit residual=0; SD=1[36] • Person Separation Index ≥ 0.7[13] for group use; ≥ 0.85[13] or ≥ 0.90[52] for individual use

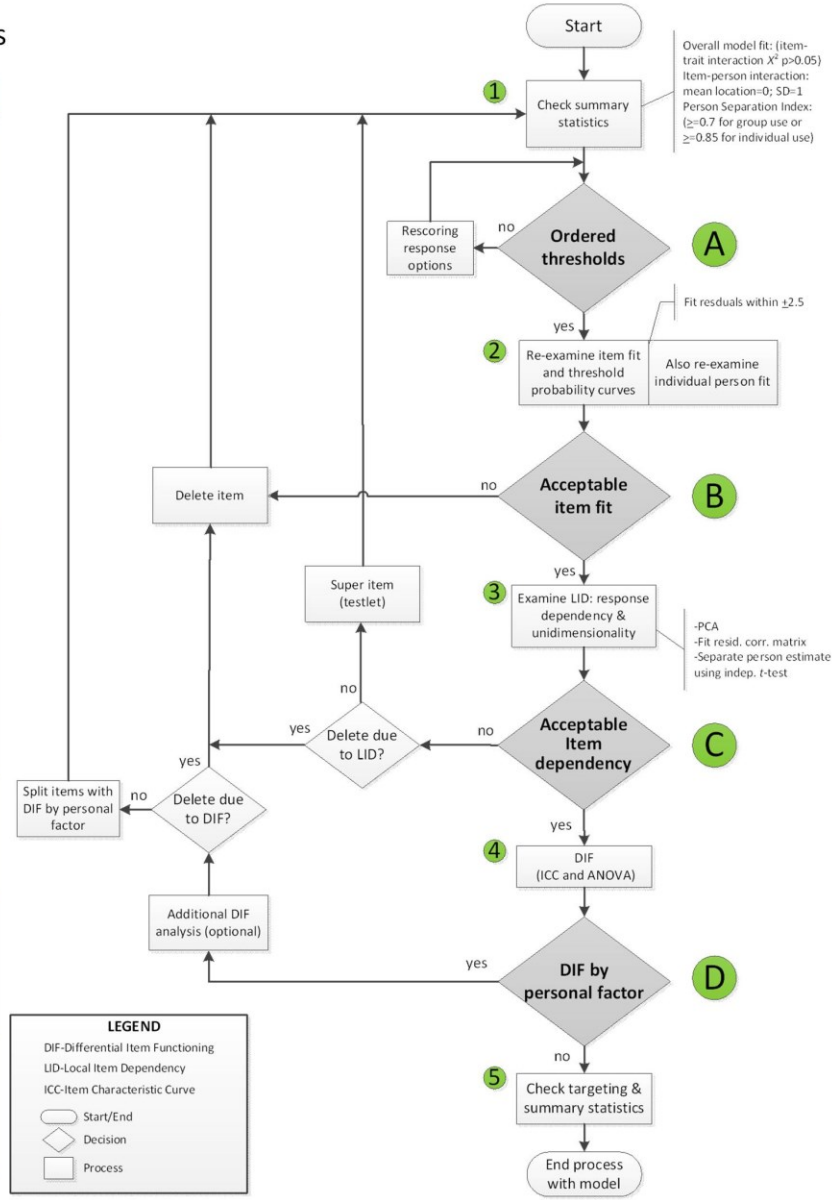


Table 2. Demographic and clinical characteristics of the sample (n=189)

Characteristics	
Age at onset: mean (y) (SD)	34.2 (9.7)
Age at baseline: mean (y) (SD)	43.0 (10.2)
F/M	140/49
MS type, n(%)	
RR/CIS	158 (83.6)
SP	17 (9.0)
PP	14 (7.4)
EDSS, median (IQR)	2.0 (1.0-3.5)
Drug, n(%)	
On DMT	97 (51.3)
None	85 (45.0)
Unknown	7 (3.7)

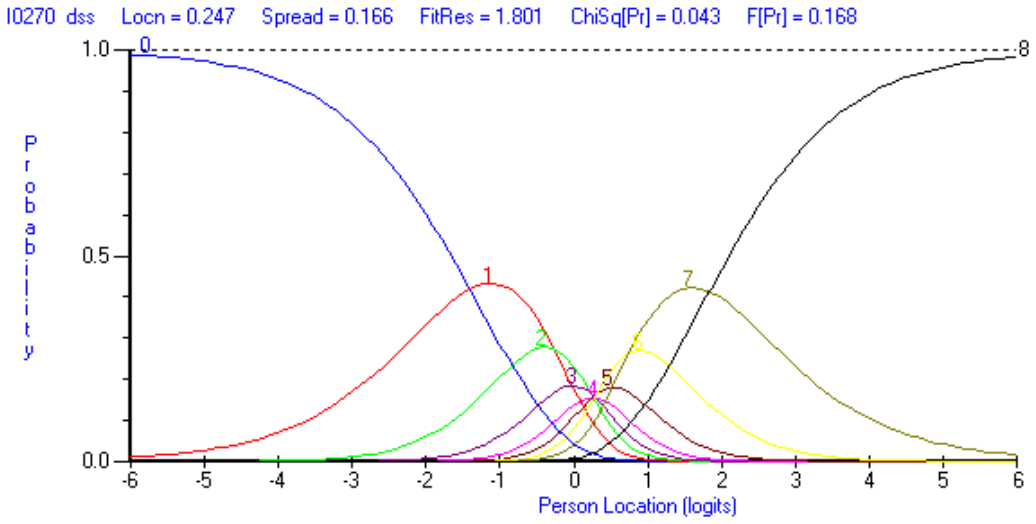


Figure 3. Disordered threshold of the EDSS

Table 3. Summary fit statistics for the physical disability measure in MS

Model	Overall model fit	Item Location	Item fit Mean	Person Location	Person fit Mean	PSI
136 items	X^2=1467.053, df=408, p=0.000000	0.00 (1.728)	-0.516 (2.161)	1.826 (1.613)	-0.347 (1.149)	0.94
28 items	X^2=95.099, df=84, p=0.191449	0.00 (2.453)	-0.420 (1.000)	2.279 (2.128)	-0.314 (0.599)	0.91
22 items	X^2=45.334, df=48, p=0..582712	0.00 (2.933)	-0.371 (0.867)	1.595 (2.188)	-0.255 (0.469)	0.90

Figure 4. Results of item reduction (because of misfit, LID, item redundency or DIF)

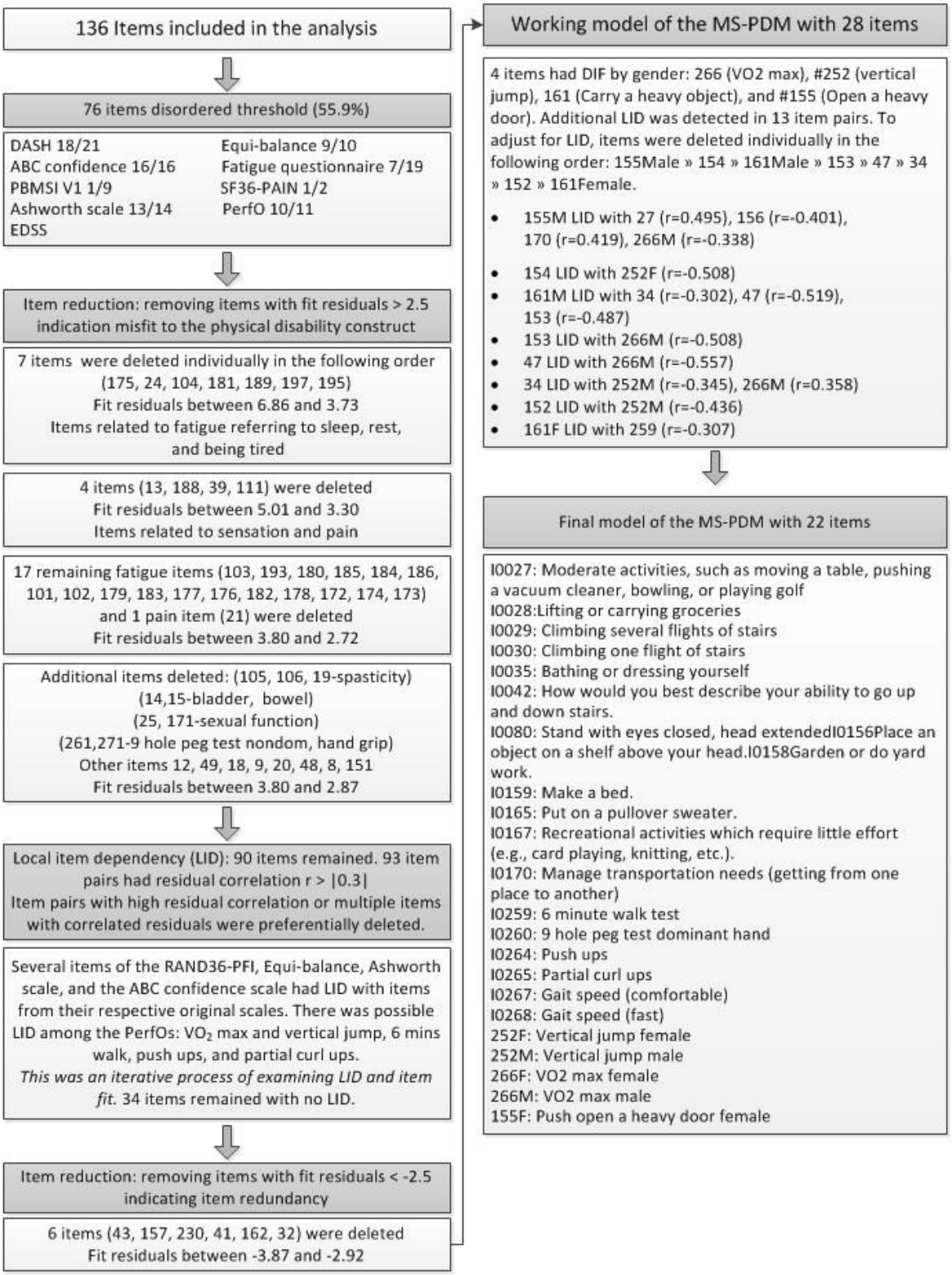


Table 4. Working model of the physical disability measure consisting of 28 items

Item	Statement	Location	SE	FitResid	ChiSq	Prob
I0027	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	2.294	0.153	-0.487	13.288	0.001302
I0028	Lifting or carrying groceries	0.452	0.120	-2.048	2.872	0.237924
I0029	Climbing several flights of stairs	1.768	0.111	-1.855	4.208	0.121948
I0030	Climbing one flight of stairs	-1.957	0.241	-1.090	1.369	0.504415
I0034	Walking one block	-0.100	0.126	-1.364	2.508	0.285324
I0035	Bathing or dressing yourself	-1.139	0.146	-1.001	0.468	0.791242
I0042	How would you best describe your ability to go up and down stairs.	-3.139	0.326	0.171	0.682	0.711231
I0047	How would you best describe your ability to speak.	-2.609	0.281	1.754	6.940	0.031124
I0080	Stand with eyes closed, head extended	0.385	0.224	-0.898	1.458	0.482383
I0152	Write.	-0.939	0.142	1.139	2.254	0.324021
I0153	Turn a key.	-0.702	0.193	0.209	1.789	0.408739
I0154	Prepare a meal.	-3.109	0.322	-0.548	0.228	0.892177
I0155	Push open a heavy door.	-2.218	0.154	-0.050	2.526	0.282753
I0156	Place an object on a shelf above your head.	-0.555	0.134	-0.871	1.531	0.465061
I0158	Garden or do yard work.	-1.442	0.222	-0.533	0.657	0.720146
I0159	Make a bed.	-4.764	0.543	-0.661	1.469	0.479732
I0161	Carry a heavy object (over 10 lbs).	-0.054	0.131	-1.529	6.526	0.038281
I0165	Put on a pullover sweater.	-0.996	0.202	0.576	0.563	0.754801
I0167	Recreational activities which require little effort (e.g., card playing, knitting, etc.).	-0.831	0.198	1.431	2.300	0.316582
I0170	Manage transportation needs (getting from one place to another)	-3.803	0.395	-0.284	0.701	0.704374
I0252	Vertical jump	4.708	0.138	-0.468	1.107	0.574803
I0259	6 minute walk test	2.391	0.131	-1.688	0.104	0.949134
I0260	9hole peg test dominant hand	1.971	0.199	0.697	1.622	0.444374
I0264	Push ups	4.205	0.141	-0.188	2.375	0.305000
I0265	Partial curl ups	3.385	0.171	-0.120	2.100	0.349982
I0266	VO2 max	3.402	0.195	0.578	3.107	0.211550
I0267	Gait speed (comfortable)	2.191	0.142	-0.659	2.840	0.241658
I0268	Gait speed (fast)	1.207	0.160	-1.920	0.271	0.873358

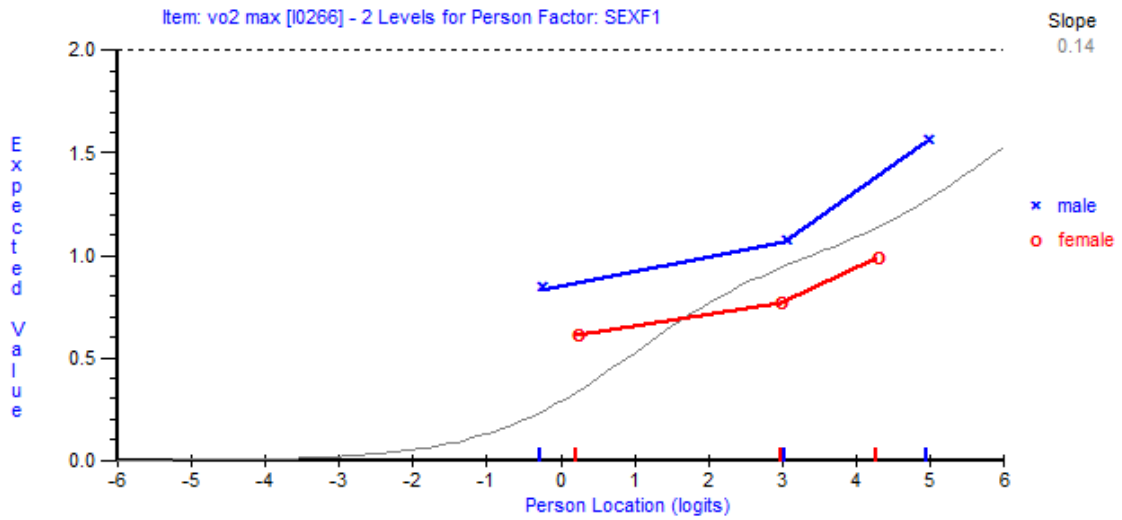


Figure 5. Item 266-VO_2max: DIF by sex

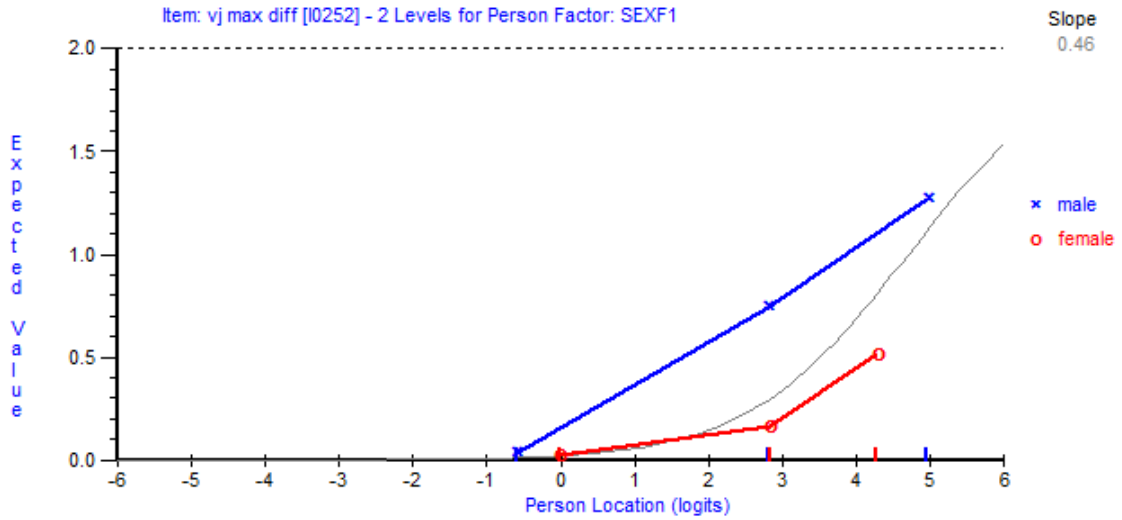


Figure 6. Item 252-Vertical jump: DIF by sex

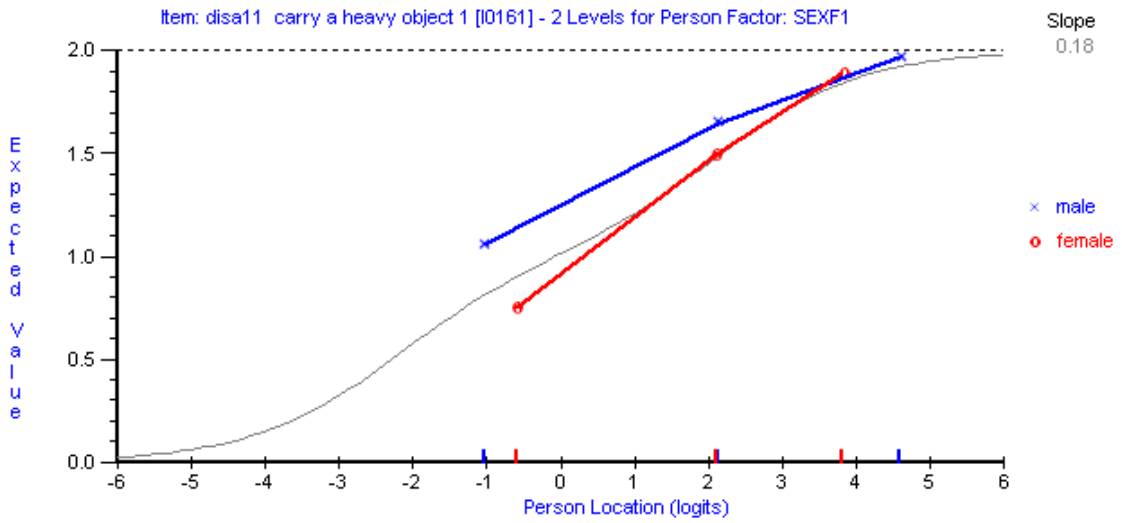


Figure 7. Item 161-DASH Carry a heavy object: DIF by sex

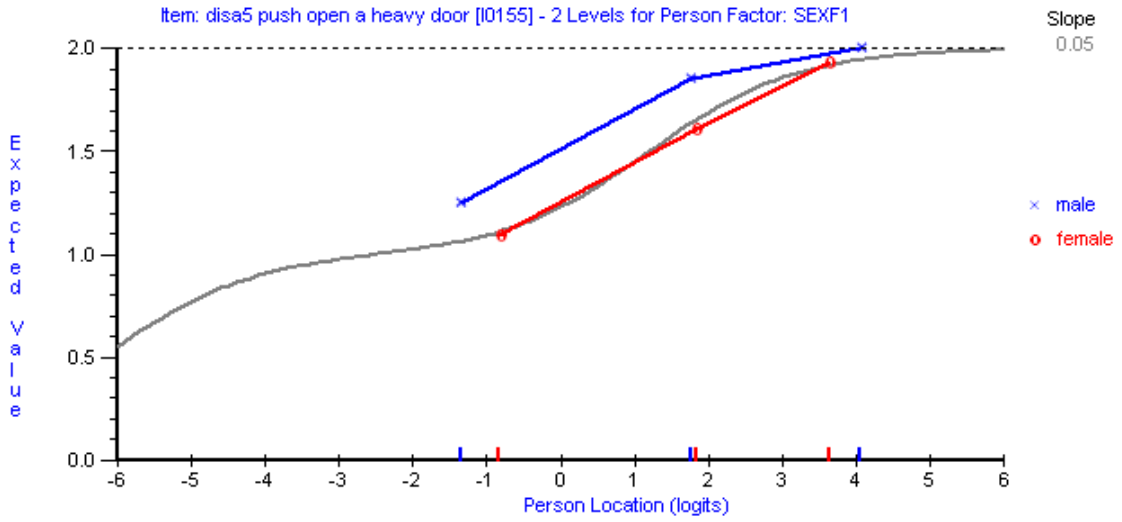


Figure 8. Item 155-DASH Open a heavy door: DIF by sex

Table 5. 22 items from the MS-PDM - item fit statistics

Item	Description	Location	SE	FitResid	ChiSq	Prob
I0027	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1.680	0.155	-0.064	12.253	0.002185
I0028	Lifting or carrying groceries	-0.185	0.121	-1.191	1.604	0.448430
I0029	Climbing several flights of stairs	1.145	0.113	-1.518	2.252	0.324272
I0030	Climbing one flight of stairs	-2.711	0.251	-1.098	1.257	0.533381
I0035	Bathing or dressing yourself	-1.842	0.150	-1.037	1.182	0.553904
I0042	How would you best describe your ability to go up and down stairs.	-3.887	0.335	0.320	1.213	0.545145
I0080	Stand with eyes closed, head extended	-0.280	0.229	-0.457	0.232	0.890666
I0156	Place an object on a shelf above your head.	-1.203	0.136	-0.345	1.323	0.516007
I0158	Garden or do yard work.	-2.175	0.230	-0.528	0.574	0.750615
I0159	Make a bed.	-5.874	0.594	-0.567	1.742	0.418515
I0165	Put on a pullover sweater.	-1.642	0.206	0.801	0.749	0.687539
I0167	Recreational activities which require little effort (e.g., card playing, knitting, etc.).	-1.454	0.201	1.880	3.180	0.203911
I0170	Manage transportation needs (getting from one place to another)	-4.543	0.400	-0.216	3.233	0.198625
I0259	6 minute walk test	1.767	0.133	-1.675	0.403	0.817547
I0260	9hole peg test dominant hand	1.331	0.201	0.693	1.896	0.387556
I0264	Push ups	3.631	0.142	-0.170	1.386	0.500031
I0265	Partial curl ups	2.821	0.171	-0.036	1.921	0.382795
I0267	Gait speed (comfortable)	1.532	0.144	-0.646	2.731	0.255195
I0268	Gait speed (fast)	0.546	0.163	-1.847	1.518	0.468059
252fe	Vertical jump female	6.049	0.215	-0.214	0.530	0.767279
252ma	Vertical jump male	3.009	0.245	-0.478	0.034	0.983239
266fe	VO2 max female	4.085	0.255	0.902	0.796	0.671770
266mal	VO2 max male	1.219	0.417	-0.473	2.163	0.339073
155fe	Push open a heavy door female	-3.019	0.171	-0.939	1.163	0.559025

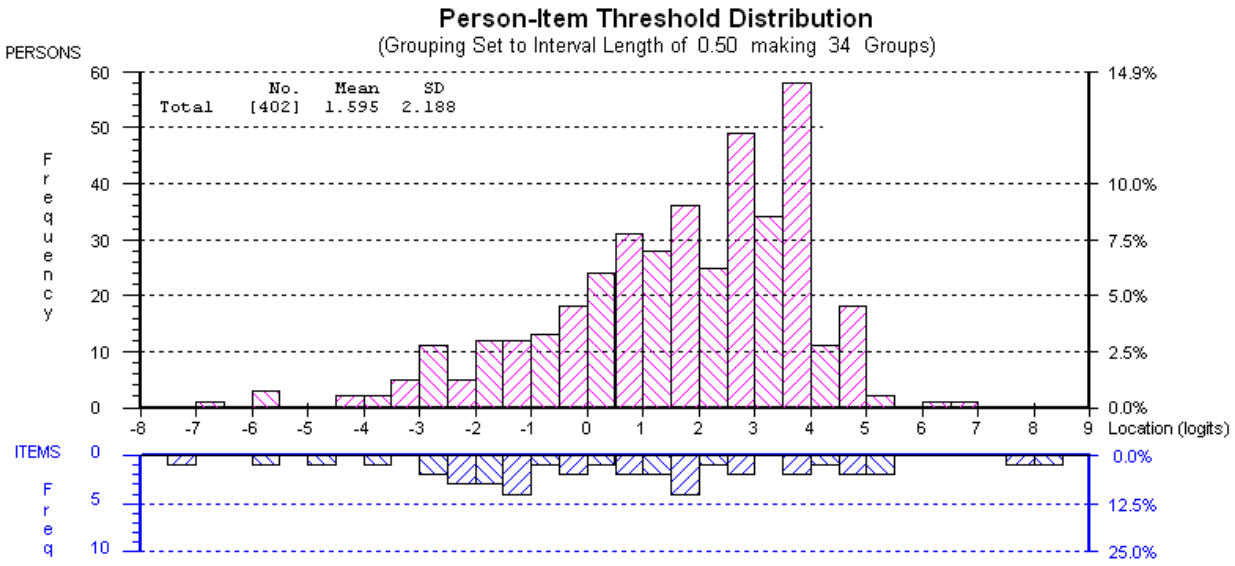


Figure 9. MS-PDM – Person-Item Thresholds. The patient distribution is represented by a histogram and the item threshold distribution by an inverted histogram

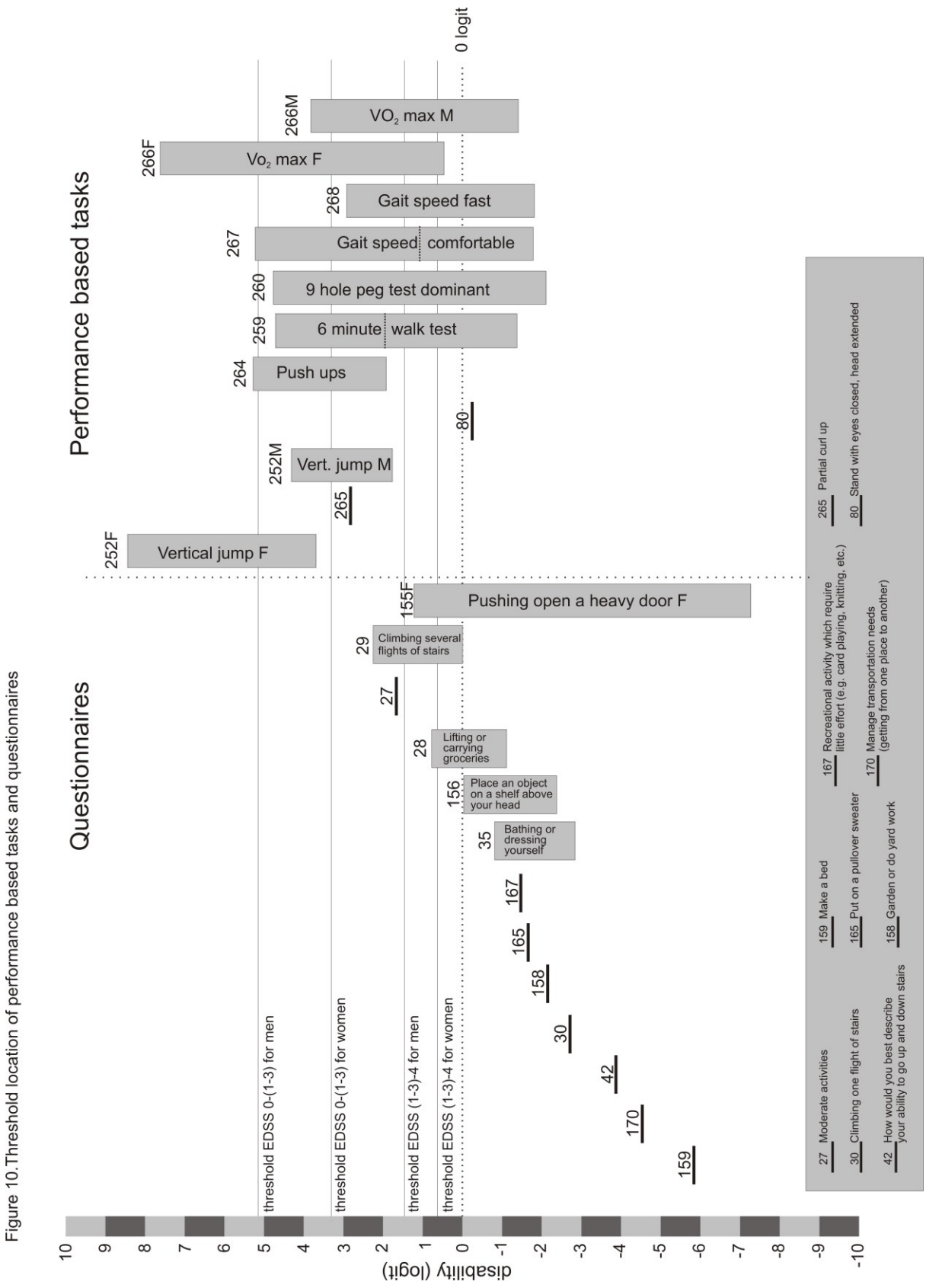


Figure 10. Threshold location of performance based tasks and questionnaires

Table 6 Convergent and discriminant validity of the MS-PDM

Measures to test convergent validity	n	Pearson correlation	Spearman's rho
EDSS	335	-0.764**	-0.685**
Illness Instrusiveness Scale (13 items)	341	-0.582**	-0.589**
SF36-Role Physical	342	0.586**	0.605**
Measures to test discriminant validity			
Cognition and Mood			
PDQ	342	-0.281**	-0.321**
SF36-Role emotional	342	0.203**	0.199**
SF36-Mental health index	342	0.171**	0.169**
Mental component-SF36	342	0.237**	0.256**
Correlation between measures that contributed items to the MS-GPDM			
SF36-Physical function index	342	0.920**	0.900**
DASH	341	-0.892**	-0.853**
Gait speed fast	186	0.887**	0.861**
6 minute walk test	186	0.884**	0.845**
Gait speed comfortable	186	0.819**	0.719**
Equi-balance	185	0.802**	0.790**
PBMSI-V1	333	0.788*	0.771**
Vertical jump	186	0.792**	0.790**
9 hole peg test dominant	182	-0.619**	-0.587**
VO2 max	138	0.616**	0.585**
Partial curl up	186	0.615**	0.668**
Push up	186	0.563**	0.658**
Overall correlation		0.77	0.75
(95% CI)		(0.68-0.84)	(0.68-0.82)

** is significant at the 0.01 level (2-tailed); * is significant at the 0.05 level (2-tailed)

1. World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva: World Health Organization; 2001.
2. Conrad A, Coenen M, Schmalz H, Kesselring J, Cieza A. Validation of the Comprehensive ICF Core Set for Multiple Sclerosis From the Perspective of Physical Therapists. Physical Therapy 2012;92:799-820.
3. Berno S, Coenen M, Leib A, Cieza A, Kesselring J. Validation of the Comprehensive International Classification of Functioning, Disability, and Health Core Set for multiple sclerosis from the perspective of physicians. Journal of Neurology 2012:1-14.
4. Coenen M, Basedow-Rajwich B, König N, Kesselring J, Cieza A. Functioning and disability in multiple sclerosis from the patient perspective. Chronic Illness 2011;7:291-310.
5. Khan F, Pallant JF. Use of International Classification of Functioning, Disability and Health (ICF) to describe patient-reported disability in multiple sclerosis and identification of relevant environmental factors. Journal of Rehabilitation Medicine 2007;39:63-70.
6. Karhula ME, Kanelisto KJ, Ruutiainen J, Hämäläinen PI, Salminen A-L. The activities and participation categories of the ICF Core Sets for multiple sclerosis from the patient perspective. Disability and Rehabilitation 2013;35:492-7.
7. Coenen M, Cieza A, Freeman J, et al. The development of ICF Core Sets for multiple sclerosis: results of the International Consensus Conference. Journal of Neurology 2011;258:1477-88.
8. Holper L, Coenen M, Weise A, Stucki G, Cieza A, Kesselring J. Characterization of functioning in multiple sclerosis using the ICF. Journal of Neurology 2010;257:103-13.
9. Drug Development Tools Qualification Programs. (Accessed June 18th, 2015, 2015, at http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284395.htm.)
10. Cutter GR, Baier M, Balcer L. Measures of neurological impairment and disability in multiple sclerosis. In: Cohen JA, Rudick RA, eds. Multiple Sclerosis Therapeutics. United Kingdom: Informa Healthcare; 2007.
11. Merbitz C, Morris J, Grip JC. Ordinal Scales and Foundations of Misinference. Arch Phys Med Rehabil 1989;70:308-12.
12. Hobart J. Rating scales for neurologists. Journal of Neurology, Neurosurgery & Psychiatry 2003;74:iv22-iv6.
13. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care & Research 2007;57:1358-62.
14. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 1983;33:1444-52.
15. Goldman MD, Motl RW, Rudick RA. Possible clinical outcome measures for clinical trials in patients with multiple sclerosis. Therapeutic Advances in Neurological Disorders 2010;3:229-39.

16. Thompson AJ, Hobart JC. Multiple sclerosis: assessment of disability and disability scales. Journal of Neurology 1998;245:189-96.
17. Cutter GR, Baier ML, Rudick RA, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999;122:871-82.
18. Fischer JS, Rudick RA, Cutter GR, Reingold SC, Force NMSCOAT. The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. Multiple Sclerosis 1999;5:244-50.
19. Regnault A, Acquadro C. FDA Workshop - Measurement in Clinical Trials: Review and Qualification of Clinical Outcome. PRO Newsletter 2012;47:12-7.
20. Rudick RA, LaRocca N, Hudson LD, MSOAC. Multiple Sclerosis Outcome Assessments Consortium: Genesis and initial project plan. Multiple Sclerosis Journal 2014;20:12-7.
21. Lamers I, Kelchtermans S, Baert I, Feys P. Upper Limb Assessment in Multiple Sclerosis: A Systematic Review of Outcome Measures and their Psychometric Properties. Archives of Physical Medicine and Rehabilitation 2014;95:1184-200.
22. Rietberg MB, Brooks D, Uitdehaag BMJ, Kwakkel G. Exercise therapy for multiple sclerosis. The Cochrane Database of Systematic Reviews 2004:CD003980.
23. Potter K, Cohen ET, Allen DD, et al. Outcome Measures for Individuals With Multiple Sclerosis: Recommendations From the American Physical Therapy Association Neurology Section Task Force. Physical Therapy 2013.
24. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institution for Educational Research; 1960.
25. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine 2003;35:105-15.
26. Hsueh I-P, Wang W-C, Sheu C-F, Hsieh C-L. Rasch Analysis of Combining Two Indices to Assess Comprehensive ADL Function in Stroke Patients. Stroke 2004;35:721-6.
27. Johnston MV, Shawaryn MA, Malec J, Kreutzer J, Hammond FM. The structure of functional and community outcomes following traumatic brain injury. Brain Injury 2006;20:391-407.
28. Finch LE, Higgins J, Wood-Dauphinee SL, Mayo NE. A Measure of Physical Functioning to Define Stroke Recovery at 3 Months: Preliminary Results. Archives of physical medicine and rehabilitation 2009;90:1584-95.
29. Kuspinar A, Andersen RE, Teng SY, Asano M, Mayo NE. Predicting Exercise Capacity Through Submaximal Fitness Tests in Persons With Multiple Sclerosis. Archives of physical medicine and rehabilitation 2010;91:1410-7.
30. Finch LE, Higgins J, Wood-Dauphinee S, Mayo NE. A measure of early physical functioning (EPF) post-stroke. J Rehabil Med 2008;40:508-17.
31. Lapierre Y, Hum S. Treating Fatigue. The International MS Journal 2007;14:64-71.
32. Motl RW, Suh Y, Weikert M, Dlugonski D, Balantrapu S, Sandroff B. Fatigue, depression, and physical activity in relapsing-remitting multiple sclerosis: Results from a prospective, 18-month study. Multiple Sclerosis and Related Disorders 2012;1:43-8.
33. Bakshi R, Shaikh ZA, Miletich RS, et al. Fatigue in multiple sclerosis and its relationship to depression and neurologic disability. Multiple Sclerosis 2000;6:181-5.
34. Chamot E, Kister I, Cutter G. Item response theory-based measure of global disability in multiple sclerosis derived from the Performance Scales and related items. BMC Neurology 2014;14:1-12.

35. Motl RW, McAuley E. Longitudinal analysis of physical activity and symptoms as predictors of change in functional limitations and disability in multiple sclerosis. Rehabilitation Psychology 2009;54:204-10.
36. Pallant JF, Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology 2007;46:1-18.
37. Andrich D, Lyne A, Sheridan B, Luo G. Rasch Unidimensional Measurement Models (RUMM2020 Version 4.1). Duncraig, Western Australia: Rumm Laboratory Pty Ltd; 2003.
38. Baghaei P. The Rasch Model as a Construct Validation Tool. Rasch Measurement Transactions 2008;22:1145-6.
39. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. second edition ed. Oxford: Oxford University Press; 1995.
40. Linacre JM. Sample Size and Item Calibration [or Person Measure] Stability. Rasch Measurement Transactions 1994;7:328.
41. Kluger BM, Krupp LB, Enoka RM. Fatigue and fatigability in neurologic illnesses: Proposal for a unified taxonomy. Neurology 2013;80:409-16.
42. Tesio L, Perucca L, Franchignoni FP, Battaglia MA. A short measure of balance in multiple sclerosis: Validation through Rasch analysis. Functional Neurology 1997;12:255-65.
43. Horton M, Tennant A. Applying Rasch analysis to the SF-36 physical function scale: effect of dependent items. Trials 2011;12:A75.
44. Cano S, Barrett L, Zajicek J, Hobart J. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. Multiple Sclerosis Journal 2011;17:214-22.
45. Remelius JG, Jones SL, House JD, et al. Gait Impairments in Persons With Multiple Sclerosis Across Preferred and Fixed Walking Speeds. Archives of Physical Medicine and Rehabilitation 2012;93:1637-42.
46. Bohannon RW, Glenney SS. Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. Journal of Evaluation in Clinical Practice 2014;20:295-300.
47. Rosa MC, Marques A, Demain S, Metcalf CD. Fast gait speed and self-perceived balance as valid predictors and discriminators of independent community walking at 6 months post-stroke – a preliminary study. Disability and Rehabilitation 2015;37:129-34.
48. Rozzini r, Frisoni GB, Ferrucci L, Barbisoni P, Bertozzi B, Trabucchi M. The effect of chronic diseases on physical function. Comparison between activities of daily living scales and the Physical Performance Test. Age and Ageing 1997;26:281-7.
49. Goverover Y, Kalmar J, Gaudino-Goering E, et al. The Relation Between Subjective and Objective Measures of Everyday Life Activities in Persons With Multiple Sclerosis. Archives of Physical Medicine and Rehabilitation 2005;86:2303-8.
50. Mills RJ, Young CA, Nicholas RS, Pallant JF, Tennant A. Rasch analysis of the Fatigue Severity Scale in multiple sclerosis. Multiple Sclerosis 2009;15:81-7.
51. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310.
52. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. Brain 2012;135:1639-49.

53. Marais I. Local Dependence. In: Christensen KB, Kreiner S, Mesbah M, eds. Rasch Models in Health. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2013:111-30.
54. Ramp M, Khan F, Misajon R, Pallant J. Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). Health and Quality of Life Outcomes 2009;7:58.
55. Smith EV. Detecting and Evaluating the Impact of Multidimensionallity using Item Fit Statistics and Principal Component Analysis of Residuals. Journal of Applied Measurement 2002;3:205-31.
56. Muller S, Roddy E. A Rasch Analysis of the Manchester Foot Pain and Disability Index. Journal of Foot and Ankle Research 2009;2:29.
57. Scott N, Fayers P, Aaronson N, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health and Quality of Life Outcomes 2010;8:81.
58. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Medical Care 2006;44:S115-S23 10.1097/01.mlr.0000245183.28384.ed.
59. Tennant A, Penta M, Tesio L, et al. Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales through Differential Item Functioning within the Framework of the Rasch Model: The PRO-ESOR Project. Medical Care 2004;42:I37-I48.
60. Mayo N, Bayley M, Duquette P, Lapierre Y, Anderson R, Bartlett S. The role of exercise in modifying outcomes for people with multiple sclerosis: a randomized trial. BMC Neurology 2013;13:69.
61. Paltamaa J, Sarasoja T, Leskinen E, Wikström J, Mälkiä E. Measuring Deterioration in International Classification of Functioning Domains of People With Multiple Sclerosis Who Are Ambulatory. Physical Therapy 2008;88:176-90.
62. Schwid SR, Goodman AD, McDermott MP, Bever CF, Cook SD. Quantitative functional measures in MS: What is reliable change? Neurology 2002;58:1294-6.

Appendix

Table 1a. Summarized MS ICF core set from the perspective of patients and health care professionals

code	item	Conrad 2012	Berno 2012	Coenen 2011	Khan 2007	Karhula 2013	Coenen 2011a	Holper 2010
Body function								
Chapter 1: mental functions								
b1101	Continuity of consciousness		x					
b114	Orientation functions	x	x	x				x
b117	Intellectual functions							x
b126	Temperament and personality functions	x	x	x			x	x
b1263	Psychic stability		x					
b1265	Optimism		x					
b1266	Confidence	x						
b130	Energy and drive functions		x		x		x	x
b1300	Energy level	x	x	x				
b1301	Motivation	x	x	x				
b1308	Energy and drive functions, other specified (fatigue)	x	x	x				
b134	Sleep functions	x	x	x	x			x
b1342	Maintenance of sleep	x						
b140	Attention functions	x	x	x	x		x	x
b1402	Dividing attention	x						
b144	Memory functions	x	x	x	x		x	x
b1440	Short-term memory		x					
b147	Psychomotor functions						x	x
b152	Emotional functions	x	x	x	x		x	x
b1522	Range of emotion	x	x					
b156	Perceptual functions	x	x	x			x	x
b1563	Gustatory perception	x						
b160	Thought functions						x	x
b1600	Pace of thought		x					
b164	Higher-level cognitive functions	x	x	x				x
b1641	Organization and planning	x	x					
b1644	Insight	x						
b1646	Problem solving	x						
b167	Mental functions of language							x
b180	Experience of self and time functions	x					x	
b1801	Body image	x						
Chapter 2: sensory functions and pain								
b210	Seeing functions	x	x	x	x		x	x
b215	Functions of structures adjoining the eye	x	x					
b2150	Functions of internal muscles of the eye	x						

b220	Sensations associated with the eye and adjoining structures						x	
b230	Hearing functions		x				x	x
b235	Vestibular functions	x	x	x	x		x	x
b240	Sensations associated with hearing and vestibular function	x	x					
b2401	Dizziness	x	x					
b250	Taste function		x					
b255	Smell function		x					
b260	Proprioceptive function	x	x	x				
b265	Touch function	x	x	x	x		x	
b270	Sensory functions related to temperature and other stimuli	x	x	x			x	
b2700	Sensitivity to temperature	x	x					
b2702	Sensitivity to pressure	x	x					
b280	Sensation of pain	x	x	x	x		x	x
b2800	Generalized pain	x	x					
b2801	Pain in body part	x						
b28010	Pain in head and neck	x	x					
b28012	Pain in stomach or abdomen		x					
b28013	Pain in back	x	x					
b28014	Pain in upper limb	x	x					
b28015	Pain in lower limb	x	x					
b28016	Pain in joints	x	x					
b2802	Pain in multiple body parts	x						
b2803	Radiating pain in a dermatome	x						
b2804	Radiating pain in a segment or region	x						
Chapter 3: voice and speech functions								
b310	Voice functions	x	x	x				x
b3100	Production of voice	x						
b320	Articulation functions	x	x	x			x	x
b330	Fluency and rhythm of speech functions	x	x	x			x	x
Chapter 4: functions of the cardiovascular, hematological, immunological, and respiratory systems								
b410	Heart functions	x						
b415	Blood vessel functions	x						
b4152	Functions of veins	x						
b435	Immunological system functions		x					
b43501	Non-specific immune response		x					
b4352	Functions of lymphatic vessels	x						
b440	Respiration functions	x	x				x	
b4402	Depth of respiration	x						
b445	Respiratory muscle functions	x	x	x				x
b450	Additional respiratory functions	x						
b455	Exercise tolerance functions	x	x	x	x		x	
b4550	General physical endurance	x						
b4551	Aerobic capacity	x						
b4552	Fatiguability		x					

b460	Sensations associated with cardiovascular and respiratory functions						x	
Chapter 5: functions of the digestive, metabolic, and endocrine systems								
b510	Ingestion functions	x					x	
b5101	Biting		x					
b5102	Chewing		x					
b5104	Salivation	x	x	x				
b5105	Swallowing	x	x	x				
b515	Digestive functions	x						x
b525	Defecation functions	x	x	x	x		x	x
b5252	Frequency of defecation		x					
b5253	Fecal continence	x	x					
b530	Weight maintenance functions	x						x
b535	Sensations associated with with the digestive system						x	x
b550	Thermoregulatory functions		x					x
b5500	Body temperature	x	x	x				
b5501	Maintenance of body temperature	x						
b5508	Thermoregulatory functions, other specified (sensitivity to heat)	x	x	x				
b5508	Thermoregulatory functions, other specified (sensitivity to cold)	x	x	x				
Chapter 6: genitourinary and reproductive functions								
b610	Urinary excretory functions	x						
b620	Urination functions	x	x	x	x		x	x
b6200	Urination		x					
b6201	Frequency of urination		x					
b6202	Urinary continence	x	x					
b630	Sensations associated with urinary functions						x	
b640	Sexual functions	x	x	x	x		x	x
b6400	Functions of sexual arousal phase		x					
b6403	Functions of sexual resolution phase		x					
b660	Procreation functions		x					
b6700	Discomfort associated with sexual intercourse		x					
Chapter 7: neuromusculoskeletal and movement-related functions								
b710	Mobility of joint functions	x	x	x				x
b7101	Mobility of several joints	x						
b715	Stability of joint functions	x						
b7150	Stability of a single joint	x						
b7151	Stability of several joints	x						
b7152	Stability of joints generalized	x						
b730	Muscle power functions	x	x	x	x		x	x
b7300	Power of isolated muscles and muscle groups	x						
b7301	Power of muscles of 1 limb	x						

b7303	Power of muscles in lower half of the body	x	x					
b7304	Power of muscles of all limbs	x						
b7305	Power of muscles of the trunk	x	x					
b735	Muscle tone functions	x	x	x	x		x	x
b7350	Tone of isolated muscles and muscle groups	x	x					
b7353	Tone of muscles of lower half of body		x					
b7354	Tone of muscles of all limbs		x					
b7355	Tone of muscles of trunk		x					
b7356	Tone of all muscles of the body		x					
b740	Muscle endurance functions	x	x	x	x			
b7401	Endurance of muscle groups		x					
b750	Motor reflex functions	x	x	x				
b755	Involuntary movement reaction functions	x						
b760	Control of voluntary movement functions	x	x	x	x		x	
b7600	Control of simple voluntary movements	x						
b7602	Coordination of voluntary movements	x						
b765	Involuntary movement functions	x	x				x	x
b7650	Involuntary contractions of muscles	x	x	x				
b7651	Tremor	x	x	x				
b770	Gait pattern functions	x	x	x	x		x	x
b780	Sensations related to muscles and movement functions	x	x	x			x	x
b7800	Sensation of muscle stiffness	x	x					
b7801	Sensation of muscle spasm	x	x					
Chapter 8: functions of the skin and related structures								
b810	Protective functions of the skin	x	x					
b840	Sensation related to the skin		x				x	
b850	Functions of hair						x	
Body structures								
Chapter 1: structures of the nervous system								
s110	Structure of brain	x	x	x	x			x
s1104	Structure of cerebellum	x						
s1106	Structure of cranial nerves		x					
s120	Spinal cord and related structures	x	x	x				x
Chapter 2: structures related to eye, ear, and related structures								
s2	Structures of eye, ear and related structures							x
Chapter 3: structures involved in voice and speech								
s3	Structures invovled in voice and speech							x
Chapter 5: structures related to the digestive, metabolic, and endocrine systems								
s5	Structures related to the digestive, metabolic, and endocrine systems							
s560	Structure of liver	x						
Chapter 6: structures related to the genitourinary and reproductive systems								

s610	Structure of urinary system	x	x	x	x			x
s630	Structure of reproductive system							x
s6100	Kidney	x						
Chapter 7: structures related to movement								
s710	Structure of head and neck region							x
s7104	Muscles of head and neck region	x						
s720	Structure of shoulder region							x
s730	Structure of upper extremity	x	x	x	x			x
s740	Structures of pelvic region							x
s750	Structure of lower extremity	x	x	x	x			x
s760	Structure of trunk	x	x	x	x			x
s7701	Joints	x						
s7702	Muscles	x						
Chapter 8: skin and related structures								
s810	structure of areas of skin	x	x	x				
Actvities and participation								
Chapter 1: learning and applying knowledge								
d110	Watching	x	x	x				x
d155	Acquiring skills	x	x	x		x		x
d160	Focusing attention	x	x	x	x			
d163	Thinking	x	x	x				x
d166	Reading	x	x	x		x	x	x
d170	Writing	x	x	x		x	x	x
d175	Solving problems	x	x	x	x			x
d177	Making decisions	x	x	x	x		x	x
Chapter 2: general tasks and demands								
d210	undertaking a single task	x	x	x		x	x	x
d220	undertaking multiple tasks	x	x	x	x		x	x
d230	carrying out daily routine	x	x	x	x	x	x	x
d2303	Managing one's own activity level		x					
d240	handling stress and other psychological demands	x	x	x	x	x		
d2401	Handling stress		x					
Chapter 3: communication								
d310	Communicating with/receiving spoken messages	x						
d315	Communicating with/receiving nonverbal messages	x						
d330	Speaking	x	x	x			x	x
d335	Producing nonverbal messages		x					
d350	Conversation	x	x	x		x		x
d360	Using communication devices and techniques	x	x	x				
d3600	Using telecommuncication devices	x						
Chapter 4; mobility								
d410	Changing basic body position	x	x	x		x	x	
d4100	Lying down	x						
d4101	Squatting	x						
d4102	Kneeling	x						

d4103	Sitting	x	x					
d4104	Standing	x	x					
d4105	Bending	x						
d4106	Shifting the body's center of gravity	x						
d415	Maintaining a body position	x	x	x		x	x	
d4150	Maintaining a lying position	x						
d4152	Maintaining a kneeling position	x						
d4153	Maintaining a sitting position	x						
d4154	Maintaining a standing position	x	x					
d420	Transferring oneself	x	x	x		x		x
d4200	Transferring oneself while sitting	x						
d4201	Transferring oneself while lying	x						
d430	Lifting and carrying objects	x	x	x	x	x	x	x
d4300	Lifting	x						
d435	Moving objects with lower extremities	x						
d440	Fine hand use	x	x	x	x	x	x	x
d4402	Manipulating	x						
d445	Hand and arm use	x	x	x	x	x	x	x
d450	Walking	x	x	x	x	x	x	x
d4500	Walking short distances	x	x					
d4501	Walking long distances	x	x					
d4502	Walking on different surfaces	x						
d4503	Walking around obstacles	x						
d455	Moving around	x	x	x	x	x	x	x
d4551	Climbing	x	x					
d4552	Running	x	x					
d4553	Jumping	x						
d4554	Swimming	x						
d460	Moving around in different locations	x	x	x		x	x	x
d4600	Moving around within the home	x	x					
d4601	Moving around within building other than the home	x	x					
d4602	Moving around outside the home and other buildings	x	x					
d465	Moving around using equipment	x	x	x	x	x	x	x
d470	Using transportation	x	x	x	x	x	x	x
d4700	Using human-powered vehicles	x						
d4701	Using private motorized transportation	x						
d4702	Using public motorized transportation	x						
d475	Driving	x	x	x	x	x	x	x
d4570	Driving human-powered transportation	x						
d4751	Driving motorized vehicles		x					
Chapter 5: self-care								
d510	Washing oneself	x	x	x	x	x	x	x
d5101	Washing whole body	x	x					
d520	Caring for body parts	x	x	x	x	x		x
d530	Toileting	x	x	x		x	x	x
d5301	Regulating defecation		x					

d540	Dressing	x	x	x		x	x	x
d550	Eating	x	x	x		x	x	x
d560	Drinking	x	x	x			x	x
d570	Looking after one's health	x	x	x	x	x	x	x
d5700	Ensuring one's physical comfort	x						
d5701	Managing diet and fitness	x	x					
d5702	Maintaining one's health	x	x					
Chapter 6: domestic life								
d620	Acquisition of goods and services	x	x	x	x	x	x	x
d6200	Shopping	x	x					
d630	Preparing meals	x	x	x	x	x	x	x
d6300	Preparing simple meals	x						
d640	Doing housework	x	x	x	x	x	x	x
d6401	Cleaning cooking area and utensils	x						
d6402	Cleaning living area	x	x					
d650	Caring for household objects	x	x	x	x	x		x
d660	Assisting others	x	x	x	x	x		x
Chapter 7: interpersonal interactions and relationships								
d710	Basic interpersonal interactions	x	x	x		x		x
d720	Complex interpersonal interactions	x	x	x		x		x
d730	Relating with strangers						x	x
d740	Formal relationships	x						x
d750	Informal social relationships	x	x	x	x	x		x
d760	Family relationships	x	x	x	x	x		x
d770	Intimate relationships	x	x	x	x	x	x	x
d7702	Sexual relationships	x	x					
Chapter 8: major life areas								
d825	Vocational training	x	x	x		x		x
d830	Higher education	x	x	x		x		x
d845	Acquiring, keeping, and terminating a job	x	x	x	x	x	x	
d8451	Maintaining a job		x					
d850	Remunerative employment	x	x	x	x	x	x	x
d855	Nonremunerative employment	x	x			x		
d860	Basic economic transaction	x	x	x		x		x
d865	Complex economic transactions							x
d870	Economic self-sufficiency	x	x	x	x			x
Chapter 9: community, social, and civic life								
d910	community life	x	x	x	x	x		x
d920	Recreation and leisture	x	x	x	x	x	x	x
d9201	Sports	x	x					
d9203	Crafts	x						
d9204	Hobbies	x	x					
d930	Religion and spirituality	x	x	x				x
d9300	Organized religion		x					
d940	Human rights						x	x
Environmental factors								
Chapter 1: Products and technology								

e110	Products or substances for personal consumption						x	x
e1101	Drugs	x	x	x	x			
e1108	Products or substances for personal consumption, other specified (special formulations of food to maintain safety and nutrition)	x	x	x				
e115	Products and technology for personal use in daily living	x	x	x			x	x
e1150	General products and technology for personal use in daily living	x	x					
e1151	Assistive products and technology for personal use in daily living	x	x					
e120	Products and technology for personal indoor and outdoor mobility and transportation	x	x	x	x		x	x
e1200	General products and technology for personal indoor and outdoor mobility and transportation	x						
e1201	Assistive products and technology for personal indoor and outdoor mobility and transportation	x	x					
e125	Products and technology for communication	x	x	x			x	
e135	Products and technology for employment	x	x	x				
e140	Products and technology for culture, recreation, and sport	x						
e1400	General products and technology for culture, recreation, and sport	x						
e1401	Assistive products and technology for culture, recreation, and sport	x						
e150	Design, construction, and building products and technology of buildings for public use	x	x	x	x		x	x
e1500	Design, construction, and building products and technology for entering and exiting buildings for public use	x						
e1501	Design, construction, and building products and technology for gaining access to facilities inside buildings for public use	x						
e155	Design, construction, and building products and technology of buildings for private use	x	x	x			x	x
e1550	Design, construction, and building products and technology for entering and exiting buildings for private use	x						
e160	Products and technology of land development						x	

e165	Assets	x	x	x			x	
Chapter 2: Natural environment and human-made changes to environment								
e210	Physical geography	x			x			
e215	Population	x						
e225	Climate	x			x		x	x
e2250	Temperature	x	x	x				
e2251	Humidity	x	x	x				
e2253	Precipitation	x	x	x				
e2254	Wind	x						
e240	Light							x
e250	Sound							x
e2600	Indoor air quality		x					
Chapter 3: support and relationships								
e310	Immediate family	x	x	x	x		x	x
e315	Extended family	x	x	x	x		x	x
e320	Friends	x	x	x			x	x
e325	Acquaintances, peers, colleagues, neighbors, and community members	x	x	x			x	x
e330	People in positions of authority	x	x	x				x
e340	Personal care providers and personal assistants	x	x	x			x	x
e345	Strangers						x	
e355	Health care professionals	x	x	x			x	x
e360	Other professionals	x	x	x			x	x
Chapter 4: attitudes								
e410	Individual attitudes of immediate family members	x	x	x			x	x
e415	Individual attitudes of extended family members	x	x	x			x	x
e420	Individual attitudes of friends	x	x	x			x	x
e425	Individual attitudes of acquaintances, peers, colleagues, neighbors, and community members	x	x	x				
e430	Individual attitudes of people in positions of authority	x	x	x				
e435	Individual attitudes of people in subordinate position						x	
e440	Individual attitudes of personal care providers and personal assistants	x	x	x				x
e450	Individual attitudes of health care professionals	x	x	x				x
e455	individual attitudes of health-related professionals						x	x
e460	Societal attitudes	x	x	x			x	
e465	Social norms, practices, and ideologies	x						x
Chapter 5: services, systems, and policies								
e510	Services, systems and policies for the production of consumer goods		x					

e515	Architecture and construction services, systems, and policies	x		x			x	
e520	Open-space planning services, systems, and policies	x					x	
e525	Housing services, systems, and policies	x	x	x			x	x
e530	Utilities services, systems and policies						x	
e5351	Communication systems	x						
e5352	Communication policies	x						
e540	Transportation services, systems, and policies	x	x	x	x		x	x
e5400	Transportation services		x					
e545	Civil protection services, systems, and policies	x						
e5450	Civil protection services	x						
e550	Legal services, systems, and policies	x	x	x				x
e555	Economic services, systems, and policies	x	x	x			x	
e5550	Associations and organizational services	x	x					
e565	Economic services, systems, and policies	x					x	
e5650	Economic services	x						
e570	Social security services, systems, and policies	x	x	x			x	x
e5700	Social security services		x					
e575	General social support services, systems, and policies	x	x	x				x
e5750	General social support services	x	x					
e580	Health care services, systems, and policies	x	x	x	x		x	x
e5800	Health care services	x	x					
e585	Education and training services, systems, and policies	x	x	x				x
e5850	Education and training services	x						
e590	Labor and employment services, systems, and policies	x	x	x				x
e5950	Political services		x					
		physio therapist	physicians	International experts from diff. health prof.	patient perspective ICF checklist	Patient perspective	patient perspective (qualitative analysis)	patient perspective ICF checklist

Appendix

Table 2a. Cut points for PerfOs

item#	PerfO	increment used	MDC; % change; 1/2 SD	source
264	Push up (#)	1	1/2 SD=3	Mayo et al., 2013[60]; Kuspinar et al, 2010[29]
265	Partial curl ups (#)	1	1/2 SD=5	Mayo et al., 2013[60]; Kuspinar et al, 2010[29]
265	Partial curl ups (#)	1	1/2 SD=5	Mayo et al., 2013[60]; Kuspinar et al, 2010[29]
252	Vertical jump (cm)	5	1/2 SD=5	Mayo et al., 2013[60]; Kuspinar et al, 2010[29]
259	6 minute walk test (m)	49	MDC=92.16 m	Paltamaa et al., 2008[61]
260	9 hole peg test dominant hand (s)	1.9	20% change	Schwid et al., 2002[62]
266	VO2 max (ml/kg/min)	1.9	10% change	Mayo et al., 2013[60]
267	Gait speed comfortable (m/s)	0.09	MDC=0.26 m/s	Paltamaa et al., 2008[61]
268	Gait speed fast (m/s)	0.09	MDC=0.26 m/s	Paltamaa et al., 2008[61]
271	Grip strength (kg)	10	1/2 SD=12 kg	Mayo et al., 2013[60]

MDC: minimal detectable change; SD: standard deviation

Appendix

Table 3a 136 items (PROs, PerfOs, ClinROs)

item	Scale	concept
#	**Symptom checklist**	
I0008	Loss of co-ordination or dexterity	impairment
I0009	Weakness or heaviness in your arms	impairment
I0010	Weakness or heaviness in your legs	impairment
I0011	Unsteadiness or loss of balance	impairment
I0012	Dizziness	impairment
I0013	Altered or loss of sensation	impairment
I0014	Problems with your bladder	impairment
I0015	Problems with your bowels	impairment
I0016	Fatigue or lack of energy	impairment
I0018	Choking or coughing when eating or drinking	impairment
I0019	Muscle stiffness or spasms	impairment
I0020	Blurred, double or shaky vision	impairment
I0021	Pain	impairment
I0024	Problems with sleep	impairment
I0025	Difficulties with your sexual function or performance	impairment
	RAND36: Physical function index	
I0026	Vigourous activities, such as running, lifting heavy objects, participating in strenuous sports	activity
I0027	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	activity
I0028	Lifting or carrying groceries	activity
I0029	Climbing several flights of stairs	activity
I0030	Climbing one flight of stairs	activity
I0031	Bending, kneeling, or stooping	activity
I0032	Walking more than a kilometre	activity
I0033	Walking several blocks	activity
I0034	Walking one block	activity
I0035	Bathing or dressing yourself	activity
	ED-5D	
I0036	Mobility	activity
I0037	Self-Care	activity
I0038	Usual Activities	impairment
I0039	Pain/Discomfort	impairment
	Perference based MS index V1	
I0041	How would you best describe your ability to walk with or without a walking aid.	impairment
I0042	How would you best describe your ability to go up and down stairs.	impairment

I0043	How would you best describe your ability to perform physically demanding activities.	impairment
I0044	How would you best describe your participation in recreational activites (like painting, knitting, playing careds, etc?).	activity
I0045	How would you best describe your ability to accomplish work or other activities.	activity
I0047	How would you best describe your ability to speak.	impairment
I0048	How would you best describe your ability to deal with life problems.	impairment
I0049	How would you best describe your appreciation of yourself?	impairment
I0082	How would you best describe your ability to drive a car.	activity
	Equi-balance (ClinRO)	
I0072	Sitting balance	activity
I0073	Standing balance	activity
I0074	Sit-to-stand	activity
I0075	Nudge	activity
I0076	Lean forward	activity
I0077	Pick-up	activity
I0078	Stand with eyes closed	activity
I0079	Rotate	activity
I0080	Stand with eyes closed, head extended	activity
I0081	Tandem stance	activity
	Fatigue questionnaire	
I0101	I felt that everything I did was an effort.	impairment
I0102	I could not get "going".	impairment
I0103	I felt exhausted.	impairment
I0104	I had to go to bed earlier than I would have liked to.	impairment
I0172	I had trouble starting things because I was tired.	impairment
I0173	I had trouble finishing things because I was tired.	impairment
I0174	Because of my fatigue, I needed to rest during the day.	impairment
I0175	Resting helps my fatigue.	impairment
I0176	Because of my fatigue, I have had to pace myself in my physical activities.	impairment
I0177	Because of my fatigue, I was less motivated to do anything that required physical effort.	impairment
I0178	Because of my fatigue, I took longer to do things.	impairment
I0179	I feel fit.	impairment
I0180	I feel very active.	impairment
I0181	I am rested.	impairment
I0182	Physically, I feel only able to do a little.	impairment
I0183	Physically, I can take on a lot.	impairment
I0184	Physically, I feel I am in bad condition.	impairment
I0185	I tire easily.	impairment
I0186	Physically, I feel I am in excellent condition.	impairment

	RAND36: Pain	
I0188	How much bodily pain have you had during the past 4 weeks?	impairment
I0111	How much did pain interfere with your normal work (including both work outside the home and housework)?	impairment
	Ashworth scale (ClinRO)	
I0105	clonus right	impairment
I0106	clonus left	impairment
I0198	Elbow Flexors right	impairment
I0199	Elbow Flexors left	impairment
I0200	Wrist Flexors right	impairment
I0201	Wrist Flexors left	impairment
I0202	Wrist Extensors right	impairment
I0203	Wrist Extensors left	impairment
I0204	Knee Extensors (quadriceps) right	impairment
I0205	Knee Extensors (quadriceps) left	impairment
I0206	Knee Flexors (hamstring) right	impairment
I0207	Knee Flexors (hamstring) left	impairment
I0208	Ankle Plantarflexors right	impairment
I0209	Ankle Plantarflexors left	impairment
	Disability of the Arm, Shoulder, and Hand (DASH)	
I0151	Open a tight or new jar.	activity
I0152	Write.	activity
I0153	Turn a key.	activity
I0154	Prepare a meal.	activity
I0155	Push open a heavy door.	activity
I0156	Place an object on a shelf above your head.	activity
I0157	Do heavy household chores (e.g., wah walls, wash floors).	activity
I0158	Garden or do yard work.	activity
I0159	Make a bed.	activity
I0160	Carry a shopping bag or briefcase.	activity
I0161	Carry a heavy object (over 10 lbs).	activity
I0162	Change a lightbulb overhead.	activity
I0163	Wash or blow dry your hair.	activity
I0164	Wash your back.	activity
I0165	Put on a pullover sweater.	activity
I0166	Use a knife to cut food.	activity
I0167	Recreational activities which require little effort (e.g., card playing, knitting, etc.).	activity
I0168	Recreational activities in which you take some force or impact through your arm, shoulder or hand (e.g., golf, hammering, tennis, etc.).	activity
I0169	Recreational activities in which you move your arm freely (e.g., playing Frisbee, badminton, etc.).	activity

I0170	Manage transportation needs (getting from one place to another)	activity
I0171	Sexual activities.	activity
	RAND36: Vitality	
I0189	Did you feel full of pep?	impairment
I0193	Did you have a lot of energy?	impairment
I0195	Did you feel worn out?	impairment
I0197	Did you feel tired?	impairment
	ABC Confidence scale	
I0228	Confidence-Walk outside on icy sidewalks?	activity
I0229	Confidence-Stand on a chair and reach for something?	activity
I0230	Confidence-step onto or off of an escalator while holding onto parcels such that you cannot hold onto the railing?	activity
I0231	Confidence-Stand on your tip toes and reach for something above your head?	activity
I0232	Confidence-Are bumped into by people as you walk through the mall?	activity
I0233	Confidence-Sweep the floor?	activity
I0234	Confidence-Walk in a crowded mall where people rapidly walk past you?	activity
I0235	Confidence-Step onto or off of an escalator while holding onto a railing?	activity
I0236	Confidence-Bend over and pick up a slipper from the front of a closet floor?	activity
I0237	Confidence-Walk across a parking lot to the mall?	activity
I0238	Confidence-Walk up or down a ramp?	activity
I0239	Confidence-Walk up or down stairs?	activity
I0240	Confidence-Reach for a small can off a shelf at eye level?	activity
I0241	Confidence-Walk outside the house to a car parked in the driveway?	activity
I0242	Confidence-Get into or out of a car?	activity
I0243	Confidence-Walk around the house?	activity
	Performanced based outcome (PerfO)	
I0252	Vertical jump	activity
I0258	2 minute walk test	activity
I0259	6 minute walk test	activity
I0260	9hole peg test dominant hand	activity
I0261	9hole peg test non-dominant had	activity
I0264	Push ups	impairment
I0265	Partial curl ups	impairment
I0266	VO2 max	impairment
I0267	Gait speed-comfortable	impairment
I0268	Gait speed-fast	impairment
I0270	EDSS **(ClinRO)**	impairment
I0271	hand grip	impairment

Chapter 12

Conclusion

The global aim of this thesis is to contribute evidence towards an optimal measurement approach to quantify MS disability over time.

Manuscript 1 quantified the variation in MS disease course over time. We have provided estimates of disease course and ARR contribution to disability progression. GBTM can be used to identify and cluster MS patients with similar rates of disability accumulation into manageable distinct groups. These results and the method employed can provide additional knowledge to further delineate levels of disease course activity within MS subtypes. This analysis provides supportive evidence of the variability in disease severity and the importance of having additional descriptors for more and less active disease course.

From the biomedical model perspective, we provide encouraging results showing that a large proportion of MS patients remained stable at their initial level of disability in the post-1995 cohorts for at least 15 years. Trajectories with high disability were associated with a higher ARR with reference to the lowest disability trajectory and had odds ratios and confidence intervals > 1 for both men and women. The results support the importance of the impact of relapses in MS disability trajectories.

The pre-1995 cohort had lower proportions of patients with stable trajectories. This cohort represents a period before the development of DMTs when the definite diagnosis of MS was based on clinical evidence of experiencing two clinical lesions in different locations at least one month apart.[17] However, secular changes in clinic practice have occurred that might impact on the interpretation of the pre-1995 cohort data in reference to the post-1995 cohorts. Among these changes were the introduction of EDSS training in 1997[152], the introduction of the McDonald diagnostic criteria that might have changed ascertainment

time[18], and the recording practice of relapses that might have changed as it is a requirement to be eligible for approved DMT treatments.

Due to many differences, both known and unknown, that exists in our historical cohort, we prefer to treat the pre-1995 onset cohort as a historical reference and view the results of the post-1995 onset cohorts as a reflection of the current patient management strategy at a tertiary clinic. GBTM's ability to summarize the heterogeneity in longitudinal change of MS disease course and results supporting previous reports of predictors of favorable prognosis lends credence in using this methodology to model disease progression. The availability of observational data analyzed using this statistical approach can be useful to inform clinicians of expected disease course of individual patients.

Manuscript 2 showed that the variability in scoring the FSS/EDSS by neurologist impact pooling of data from multiple sources. The majority if not all MS database registries record the EDSS along with the FSS as the standard outcome for MS disability with the goal of facilitating systematic analysis and comparison of longitudinal data across multiple sources.[271,272] The assumption that the EDSS is consistently interpreted and scored by all raters such that data from multiple sources can be pooled was tested.

Rasch analysis provides a strong statistical methodology to assess the psychometric properties of the EDSS/FSS, a technique to detect data heterogeneity, and a method to harmonize data from different sources when heterogeneity is found. Using Rasch analysis we were able to show that neurologists were not able to consistently apply the item response options in a monotonic fashion resulting in disordered thresholds. Rasch analysis was able to identify and eliminate sources that impact on threshold order, improving item reliability and decreasing data noise. The DIF analysis offers a sophisticated method to identify inter-rater item bias among the neurologists and was able to adjust for it. We showed that different raters at a single clinic all certified on the Neurostatus© (training for the FSS/EDSS) were not scoring or interpreting the FSS items in a similar manner (response bias as shown by DIF). Reasons for this may be rater training on using the EDSS, experience as a neurologist, and/or the complexity of the scoring rules of the tool. We have provided a solution to control data heterogeneity to ensure data harmonization when

pooling FSS/EDSS data from multiple scores, which may be important considering its continual use in MS registries.

Manuscript 3 presented the development of a prototype physical disability measure for MS (MS-PDM). Although we were able to find a solution to adjust for the psychometric limitations of the FSS/EDSS, it still only measures a narrow range of impairments and is weighted heavily towards ambulation[163]. A more comprehensive measure of MS disability is needed. Currently there is no "core set" of outcome measures used to assess MS disability.[273] The APTA identified 120 outcome measures that have been used to assess MS disability. Recommendations were made on 63 outcome measures on the appropriate use of each tool.[198] This makes selection of appropriate outcomes difficult. To comprehensively assess patients on a domain of interest (e.g. physical function) requires multiple health indices and can represent a significant response burden for the patient.[70] Having multiple outcomes also makes it difficult to interpret the multiple results.

To resolve these issues, researchers have recently combined multiple outcomes including both objective measures from PerfOs and PROs to represent the patients' perspective to form a single measure.[70,262] Manuscript 3 presented the results of developing a physical disability measure for MS (MS-PDM) from commonly used health indices. The ICF conceptual framework was used to select the appropriate domains and categories to include in the measure whereas Rasch analysis provided the methods to create a unidimensional hierarchical continuous linear measure. Outcomes related to body function impairment and activity limitations were included in the MS-PDM. We were able to provide evidence that self-reported outcome and performance-based tasks can both be used to form a single measure of physical disability. PerfOs were located solely at the low disability end of the continuum whereas only PROs were located at the high disability end supporting previous reports that PerfOs can detect milder changes in ADL.[274] We provided a solution on how to reduce the assessment burden to both patients and clinicians by removing redundant and non-relevant items. We also provided a method to interpret multiple indices on a single measurement scale (ruler).

The extensive literature review provided the foundation to understand MS care and research from the perspectives of the biomedical model and the biopsychosocial model. Medicine is largely based on the biomedical model.[73] The underlying belief is that a disease is due to an anomaly within the body, and identifying and eliminating the "cause" would result in curing or improvement of the patient.[73,74] The model views health as the absence of disease and the patient as a passive recipient of treatment.[75] This model for medicine has been very successful when applied to acute conditions and/or in medical and surgical patient care where a "cause" is identified and a "cure" is effective.[76-78] The majority of medical research in MS also relates to the biomedical model as was revealed in the scoping review of MS relapses in chapter 3.

The scoping review provided a descriptive summary of the key concepts of MS relapse research and helped identify the measurement challenges of relapses. A majority of research related to MS relapse followed the biomedical perspective. One third of all research focused on drug development with the goal to modify the disease course. The primary endpoint indicating drug efficacy is often relapse frequency. Other endpoints are delaying time to sustained progression (as measured by the EDSS) and decreased MRI activity. Other research directly related to relapses focused on measuring and understanding relapse (disease activity) using MRI and immunology studies.

The diagnosis of MS or a relapse is largely based on objective neurological signs with limited input from the patient in reporting symptoms that are consistent with MS. Previously, it took an average of seven years[18] to obtain a definite diagnosis of MS after a patient experienced two relapses with lesions in different locations at least 30 days apart.[17,22] The criterion to diagnose MS has been updated to include MRI findings.[19-21] This has changed the time it takes to make a definite diagnosis of MS to an average of seven months.[18] Diagnosing MS earlier will lead to earlier treatment with the rationale that "time is brain".[275]

From the perspective of the biomedical model, medical research in MS has been successful. With the advancements in DMTs and earlier detection of MS, there have been discussions of MS patients entering "remission" from disease activity. This has been termed "no evidence

of disease activity" with the acronym NEDA. This status has been observed in rheumatoid arthritis, Crohn's disease, and psoriasis.[276] Generally, NEDA means there is no observed change in EDSS, relapse activity, or MRI activity.[277] From the perspective of the treating neurologist, until a cure is found, reaching a goal of "NEDA" is the next best outcome.[278] Some researchers have commented that perhaps a patient with 15 years of NEDA might be considered as a working definition of a "cure".[58]

The EDSS and MRI are very important tools to monitor disease activity for the treating neurologist. The latest recommendations to classify MS disease phenotype[279] and optimal treatment[280] are based on relapses, clinical outcome using the EDSS, and MRI findings. A Canadian expert panel recommends the incorporation of MRI to monitor patient status and aid in treatment decision making. [174,281]

However, there are limitations in using these tools to assess MS disability. The EDSS has been criticized as being too heavily weighted towards ambulation[163] and does not represent the patients' interests.[178] There is weak correlation between MRI metrics and clinical status as measured by the EDSS.[175,176] Gd-enhancing Ti-W1 appears to detect 5-10 times more activity than clinical observation (relapses).[33] Ninety percent of Gd-enhancing lesions are not associated with identifiable signs or symptoms.[32] Zivadinov et al., concluded that conventional MRI such as Gd-enhancing T1 and T2 lesion have only limited value for predicating clinical status in MS due to their poor sensitivity and specificity for the underlying pathophysiologic process and feel that newer techniques may be better.[177] These newer techniques have only been used in a research setting.[177] MRIs (in a clinical setting) are logistically difficult to obtain in Canada and costly with little return.[282] The challenge remains to measure a relapse (disease activity) clinically without a good biomarker.

Currently, under the biomedical model, impairments that are assessed by PROs are less well studied. Fatigue and cognitive function are recognized as important aspects of MS disabilities but are still not included in some MS management models.[24] Fatigue, depression, or cognitive dysfunction have been referred to as "soft" and "invisible" symptoms since they are unseen by others, and are difficult to measure or attribute solely to MS or relapses[41,140,283] Many

sensory symptoms are not measured and have been referred to as a "hidden reservoir of morbidity".[284]

The perspective of MS care and research from the biopsychosocial model.

The goal of rehabilitation is to maintain the person's autonomy, minimize disability, and maximize function.[186] The ICF is the accepted model for rehabilitation[185] and the conceptual framework for disability for this thesis. The ICF MS core sets[96-98,187,189-191] identified all the relevant body structure and function impairments, activity limitations, and participation restrictions associated with MS disability and can be used to help guide how it is measured.

The trajectory results from manuscript 1 were based on an outcome measure that represents a narrow range of impairments. From a biopsychosocial perspective, whether the trajectory patterns would be similar using an outcome measure of disability that is more relevant from the patients' perspective is unknown.

As discussed in manuscript 3 there is a need for a prototype measure such as the MS-PDM. The principal role of rehabilitation is to maximize remaining function by reducing residual symptoms and activity limitations or participation restrictions to achieve maximal autonomy through rehabilitation interventions.[59] A more relevant outcome including the patients' perspective such as the MS-PDM should be included to assess the impact of DMTs and rehabilitation interventions. Applying the same rationale as including the EDSS in MS database registries, having a common measure that includes important additional domains to measure the health status of MS patients is appealing. This would facilitate the inclusion of the outcome measure in database registries allowing for the collection of longitudinal data to answer the question of the long-term benefit of DMTs based on an outcome more relevant from the patients' perspective.

The common goal whether from the perspective of the biomedical or biopsychosocial model is the well-being of people with MS. However from the viewpoint of the treating neurologists, patients that fulfill the requirements of NEDA are considered stable and the "level of concern would be low".[280] At this point there is very little for the neurologist to do

other than to monitor for change in disease status. From a biopsychosocial perspective, patients in "remission" from disease activity might have a greater benefit from a rehabilitation intervention. Where NEDA is possibly the end goal under the biomedical model, it might be a good starting point for additional interventions in the biopsychosocial model. Having stable disease would give the patient an opportunity to maximize remaining function by reducing residual symptoms and activity limitations or participation restrictions to achieve maximal autonomy through rehabilitation interventions.[59]

Rasch analysis

Rasch analysis provided a useful methodology to develop new measurement tools. Rasch analysis was applied for a different purpose in manuscript 2, to harmonize data generated from multiple sources. the psychometric limitations of the FSS/EDSS were identified and the scores were adjusted in order to pool data from multiple sources. For manuscript 3 we were able to repurpose existing multiple measures of body function impairment and activity limitations to form a single measure. Also, we were able to combine PerfOs, PROs, and ClinROs. In doing so, we are able to show how existing items or tasks relate to each other on the same scale.

Future work

The results of the Rasch analysis provides evidence for a core set of physical ability items that work together. Similar work can be undertaken in cognition. The methodology applied in this thesis is a starting point in learning how to harmonize existing data from MS registries to answer questions of MS disease course. Using the modern psychometric method of Rasch analysis has provided a useful tool to continually assess or reassess the usefulness of existing measures.

The measure of disability needs further testing. It could serve as a way of patients to assess themselves and act to instigate strategies before declining to the point of no return. Self-assessment is a key component of self-management.[285] Future work would be fruitful in this area.

REFERENCES

1. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapse and progression of disability in multiple sclerosis. The New England Journal of Medicine 2000;343:1430-8.
2. Murray TJ. Diagnosis and treatment of multiple sclerosis. BMJ 2006;332:525-7.
3. Statistics Canada. Canadian Communifty Health Survey: Neurological conditions prevalence files, 2010/2011: Statistics Canada; 2012.
4. Beck CA, Metz LM, Svenson LW, Patten SB. Regional variation of multiple sclerosis prevalence in Canada. Multiple Sclerosis 2005;11:516-9.
5. Orton S-M, Herrera BM, Yee IM, et al. Sex ratio of multiple sclerosis in Canada: a longitudinal study. Lancet Neurology 2006;5:932-6.
6. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG. Multiple Sclerosis. The New England Journal of Medicine 2000;343:938-52.
7. Kingwell E, van der Kop M, Zhao Y, et al. Relative mortality and survival in multiple sclerosis: findings from British Columbia, Canada. Journal of Neurology, Neurosurgery & Psychiatry 2011.
8. Ebers GC, Heigenhauser L, Daumer M, Lederer C, Noseworthy JH. Disability as an outcome in MS clinical trials. Neurology 2008;71:624-31.
9. Trapp BD, Nave K-A. Multiple Sclerosis: An Immune or Neurodegenerative Disorder? Annual Review of Neuroscience 2008;31:247-69.
10. Brück W, Bitsch A, Kolenda H, Brück Y, Stiefel M, Lassmann H. Inflammatory central nervous system demyelination: Correlation of magnetic resonance imaging findings with lesion pathology. Annals of Neurology 1997;42:783-93.
11. Weiner HL. The challenge of multiple sclerosis: How do we cure a chronic heterogeneous disease? Annals of Neurology 2009;65:239-48.
12. Martin R, Bielekova B, Hohlfield R, Utz U. Biomarkers in multiple sclerosis. Disease Markers 2006;22:183-5.
13. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: Results of an international survey. Neurology 1996;46:907-11.
14. Weinshenker BG, Bass B, Rice GPA, et al. The Natural History of Multiple Sclerosis: A Geographically Based Study 1. Clinical Course and Disability. Brain 1989;112:133-46.
15. Lublin FD, Baier M, Cutter G. Effect of relapses on development of residual deficit in multiple sclerosis. Neurology 2003;61:1528-32.
16. Rovaris M, Confavreux C, Furlan R, Kappos L, Comi G, Filippi M. Secondary progressive multiple sclerosis: current knowledge and future challenges. Lancet Neurology 2006;5:343-54.
17. Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: Guidelines for research protocols. Annals of Neurology 1983;13:227-31.
18. Marrie RA, Cutter G, Tyry T, Hadjimichael O, Campagnolo D, Vollmer T. Changes in the ascertainment of multiple sclerosis. Neurology 2005;65:1066-70.
19. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. Annals of Neurology 2001;50:121-7.
20. Polman CH, Reingold SC, Edan G, et al. Diagnostic Criteria for Multiple Sclerosis: 2005 Revisions to the "McDonald Criteria". Annals of Neurology 2005;58:840-6.
21. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. Annals of Neurology 2011;69:292-302.

22. Schumacher GA, Beebe G, Kibler RF, et al. Problems of experimental trials of therapy in multiple sclerosis: report by the panel on the evaluation of experimental trials of therapy in multiple sclerosis. Annals of the New York Academy of Sciences 1965;122:552-68.
23. Goldman MD, Motl RW, Rudick RA. Possible clinical outcome measures for clinical trials in patients with multiple sclerosis. Therapeutic Advances in Neurological Disorders 2010;3:229-39.
24. Sharief MK. MS patient management: the Analog Model. Journal of Neurology 2004;251:v74-v8.
25. Freedman MS, Patry DG, Grand'Maison F, et al. Treatment optimization in multiple sclerosis. Can J Neurol Sci 2004;31:157-68.
26. Lhermitte F, Marteau R, Gazengel J, Dordain G, Deloche G. The frequency of relapse in multiple sclerosis. A study based on 245 cases. Z Neurol 1973;205:47-59.
27. The IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis I.Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. Neurology 1993;43:655-61.
28. Jacobs L, Cookfair D, Rudick R, et al. A phase III trial of intramuscular recombinant interferon beta as treatment for exacerbating-remitting multiple sclerosis: design and conduct of study and baseline characteristics of patients. Multiple Sclerosis 1995;1:118-35.
29. PRISMS Study Group. Randomised double-blind placebo-controlled study of interferon ß-1a in relapsing/remitting multiple sclerosis. The Lancet 1998;352:1498-504.
30. Johnson KP, Brooks BR, Cohen JA, et al. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis. Neurology 1995;45:1268-76.
31. Weinshenker BG, Ebers GC. The Natural History of Multiple Sclerosis. The Canadian Journal of Neurological Sciences 1987;14:255-61.
32. Vollmer T. The natural history of relapses in multiple sclerosis. Journal of the Neurological Sciences 2007;256:S5-S13.
33. Zivadinov R. Can imaging techniques measure neuroprotection and remyelination in multiple sclerosis? Neurology 2007;68:S72-S82.
34. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 1983;33:1444-52.
35. Ebers GC. The natural history of multiple sclerosis. Neurol Sci 2000;21:S815-7.
36. Weinshenker BG, Bass B, Rice GPA, et al. The Natural History of Multiple Sclerosis: A Geographically Based Study 2. Predictive Value of the Early Clinical Course. Brain 1989;112:1419-28.
37. Kantarci O, Siva A, Eraksoy M, et al. Survival and predictors of disability in Turkish MS patients. Neurology 1998;51:765-72.
38. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003;126:770-82.
39. Tremlett H, Yousefi M, Devonshire V, Rieckmann P, Zhao Y, Neurologists UBC. Impact of multiple sclerosis relapses on progression diminishes with time. Neurology 2009;73:1616-23.
40. Kurtzke JF, Beebe GW, Nagler B, Kurland LT, Auth TL. Studies on the natural history of multiple sclerosis 8: Early prognostic features of the later course of the illness. Journal of Chronic Diseases 1977;30:819-30.

41. Williams G. Chapter 16: Management of patients who have relapses in multiple sclerosis. In: Woodward S, ed. Neuroscience Nursing: Assessment & Patient Management. London, UK: Quay Books Mark Allen Group; 2006:217-26.
42. Halper J. The psychosocial effect of multiple sclerosis: The impact of relapses. Journal of the Neurological Sciences 2007;256, Supplement 1:S34-S8.
43. Sorensen PS. New management algorithms in multiple sclerosis. Current Opinion in Neurology 2014;27:246-59.
44. Willis MD, Robertson NP. Alemtuzumab for the treatment of multiple sclerosis. Therapeutics and Clinical Risk Management 2015;11:525-34.
45. Hirst C, Ingram G, Pearson O, Pickersgill T, Scolding N, Robertson N. Contribution of relapses to disability in multiple sclerosis. J Neurol 2008;255:280-7.
46. Coyle PK. Early treatment of multiple sclerosis to prevent neurologic damage. Neurology 2008;71:S3-7.
47. Polman CH, O'Connor PW, Havrdova E, et al. A Randomized, Placebo-Controlled Trial of Natalizumab for Relapsing Multiple Sclerosis. The New England Journal of Medicine 2006;354:899-910.
48. Major EO, Douek DC. Risk factors for rare diseases can be risky to define: PML and natalizumab. Neurology 2013;81:858-9.
49. Kappos L, Radue E-W, O'Connor P, et al. A Placebo-Controlled Trial of Oral Fingolimod in Relapsing Multiple Sclerosis. New England Journal of Medicine 2010;362:387-401.
50. Kappos L, Antel J, Comi G, et al. Oral Fingolimod (FTY720) for Relapsing Multiple Sclerosis. New England Journal of Medicine 2006;355:1124-40.
51. O'Connor P, Wolinsky JS, Confavreux C, et al. Randomized Trial of Oral Teriflunomide for Relapsing Multiple Sclerosis. New England Journal of Medicine 2011;365:1293-303.
52. Confavreux C, O'Connor P, Comi G, et al. Oral teriflunomide for patients with relapsing multiple sclerosis (TOWER): a randomised, double-blind, placebo-controlled, phase 3 trial. The Lancet Neurology 2014;13:247-56.
53. Fox RJ, Miller DH, Phillips JT, et al. Placebo-Controlled Phase 3 Study of Oral BG-12 or Glatiramer in Multiple Sclerosis. New England Journal of Medicine 2012;367:1087-97.
54. Gold R, Kappos L, Arnold DL, et al. Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis. New England Journal of Medicine 2012;367:1098-107.
55. Cohen JA, Coles AJ, Arnold DL, et al. Alemtuzumab versus interferon beta 1a as first-line treatment for patients with relapsing-remitting multiple sclerosis: a randomised controlled phase 3 trial. The Lancet 2012;380:1819-28.
56. Coles AJ, Twyman CL, Arnold DL, et al. Alemtuzumab for patients with relapsing multiple sclerosis after disease-modifying therapy: a randomised controlled phase 3 trial. The Lancet 2012;380:1829-39.
57. Hartung DM, Bourdette DN, Ahmed SM, Whitham RH. The cost of multiple sclerosis drugs in the US and the pharmaceutical industry: Too big to fail? Neurology 2015;84:2185-92.
58. Banwell B, Giovannoni G, Hawkes C, Lublin F. Editors' welcome and a working definition for a multiple sclerosis cure. Multiple Sclerosis and Related Disorders 2013;2:65-7.

59. Beer S, Khan F, Kesselring J. Rehabilitation interventions in multiple sclerosis: an overview. Journal of Neurology 2012;259:1994-2008.
60. Motl RW, McAuley E, Snook EM. Physical activity and multiple sclerosis: a meta-analysis. Multiple Sclerosis 2005;11:459-63.
61. Asano M, Dawes D, Arafah A, Moriello C, Mayo N. What does a structured review of the effectiveness of exercise interventions for persons with multiple sclerosis tell us about the challenges of designing trials? Multiple Sclerosis 2009;15:412-21.
62. Mayo N, Bayley M, Duquette P, Lapierre Y, Anderson R, Bartlett S. The role of exercise in modifying outcomes for people with multiple sclerosis: a randomized trial. BMC Neurology 2013;13:69.
63. Dalgas U, Stenager E. Exercise and disease progression in multiple sclerosis: can exercise slow down the progression of multiple sclerosis? Therapeutic Advances in Neurological Disorders 2012;5:81-95.
64. Rietberg MB, Brooks D, Uitdehaag BMJ, Kwakkel G. Exercise therapy for multiple sclerosis. The Cochrane Database of Systematic Reviews 2004:CD003980.
65. World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva: World Health Organization; 2001.
66. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institution for Educational Research; 1960.
67. Andrich D. Rasch models for measurement: SAGE; 1988.
68. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care & Research 2007;57:1358-62.
69. Hsueh I-P, Wang W-C, Sheu C-F, Hsieh C-L. Rasch Analysis of Combining Two Indices to Assess Comprehensive ADL Function in Stroke Patients. Stroke 2004;35:721-6.
70. Finch LE, Higgins J, Wood-Dauphinee SL, Mayo NE. A Measure of Physical Functioning to Define Stroke Recovery at 3 Months: Preliminary Results. Archives of physical medicine and rehabilitation 2009;90:1584-95.
71. Meyer-Moock S, Feng Y-S, Maeurer M, Dippel F-W, Kohlmann T. Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis. BMC Neurology 2014;14:58.
72. Lhermitte J. Training of the neurologist. Arch Neurol Psychiatry 1933;30:405-12.
73. Engel G. The need for a new medical model: a challenge for biomedicine. Science 1977;196:129-36.
74. Ludwig AM. The Psychiatrist as Physician. JAMA: The Journal of the American Medical Association 1975;234:603-4.
75. Wade DT, Halligan PW. Do Biomedical Models Of Illness Make For Good Healthcare Systems? BMJ: British Medical Journal 2004;329:1398-401.
76. McCollum L, Pincus T. A Biopsychosocial Model to Complement a Biomedical Model: Patient Questionnaire Data and Socioeconomic Status Usually Are More Significant than Laboratory Tests and Imaging Studies in Prognosis of Rheumatoid Arthritis. Rheumatic Disease Clinics of North America 2009;35:699-712.
77. Sokka T, Häkkinen A. Poor physical fitness and perfermance as predictors of mortality in normal populations and patients with rheumatic and other disease. Clinical and Experimental Rheumatology 2008;26:S14-S20.

78. Weiner BK. Difficult medical problems: On explanatory models and a pragmatic alternative. Medical Hypotheses 2007;68:474-9.
79. Wilson IB, Cleary PD. Linking Clinical Variables With Health-Related Quality of Life. JAMA: The Journal of the American Medical Association 1995;273:59-65.
80. World Health Organization. Towards a Common Language for Functioning, Disability and Health. wwwwhoint/classifications/icf/training/icfbeginnersguidepdf 2002.
81. Bury M. What is Health? In: Health and Illness. Malden: Polity Press; 2005.
82. Grad FP. The Preamble of the Constitution of the World Health Organization. Bulletin of the World Health Organization 2002;80:981-4.
83. Kesselring J, Coenen M, Cieza A, Thompson A, Kostanjsek N, Stucki G. Developing the ICF Core Sets for multiple sclerosis to specify functioning. Multiple Sclerosis 2008;14:252-4.
84. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. Journal of Nursing Scholarship 2005;37:336-42.
85. Sousa K, Kwok O-M. Putting Wilson and Cleary to the Test: Analysis of a HRQOL Conceptual Model using Structural Equation Modeling. Quality of Life Research 2006;15:725-37.
86. Wyrwich KW, Harnam N, Locklear JC, Svedsater H, Revicki DA. Understanding the relationships between health outcomes in generalized anxiety disorder clinical trials. Quality of Life Research 2011;20:255-62.
87. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 2007;60:34-42.
88. Heo S, Moser DK, Riegel B, Hall LA, Christman N. Testing a published model of health-related quality of life in heart failure. Journal of Cardiac Failure 2005;11:372-9.
89. Miller D, Rudick RA, Hutchinson M. Patient-centered outcomes: Translating clinical efficacy into benefits on health-related quality of life. Neurology 2010;74:S24-S35.
90. Patrick DL, Burke LB, Powers JH, et al. Patient-Reported Outcomes to Support Medical Product Labeling Claims: FDA Perspective. Value in Health 2007;10, Supplement 2:S125-S37.
91. Sprangers MAG, Sloan JA, Barsevick A, et al. Scientific imperatives, clinical implications, and theoretical underpinnings for the investigation of the relationship between genetic variables and patient-reported quality-of-life outcomes. Quality of Life Research 2010;19:1395-403.
92. U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims; 2009.
93. Hahn EA, Cella D, Chassany O, Fairclough DL, Wong GY, Hays RD. Precision of Health-Related Quality-of-Life Data Compared With Other Clinical Measures. Mayo Clinic Proceedings 2007;82:1244-54.
94. Ferrans CE. Differences in What Quality-of-Life Instruments Measure. JNCI Monographs 2007;2007:22-6.
95. Forbes A, While A, Taylor M. What people with multiple sclerosis perceive to be important to meeting their needs. Journal of Advanced Nursing 2007;58:11-22.

96. Coenen M, Cieza A, Freeman J, et al. The development of ICF Core Sets for multiple sclerosis: results of the International Consensus Conference. Journal of Neurology 2011;258:1477-88.
97. Berno S, Coenen M, Leib A, Cieza A, Kesselring J. Validation of the Comprehensive International Classification of Functioning, Disability, and Health Core Set for multiple sclerosis from the perspective of physicians. Journal of Neurology 2012:1-14.
98. Coenen M, Basedow-Rajwich B, König N, Kesselring J, Cieza A. Functioning and disability in multiple sclerosis from the patient perspective. Chronic Illness 2011;7:291-310.
99. Stucki G, Cieza A, Ewert T, Kostanjsek N, Chatterji S, Bedirhan ÜT. Application of the International Classification of Functioning, Disability and Health (ICF) in clinical practice. Disability and Rehabilitation 2002;24:281-2.
100. Cerniauskaite M, Quintas R, Boldt C, et al. Systematic literature review on ICF from 2001 to 2009: its use, implementation and operationalisation. Disability and Rehabilitation 2011;33:281-309.
101. Üstün B, Chatterji S, Kostanjsek N. Comments from WHO for the Journal of Rehabilitation Medicine Special Supplement on ICF Core Sets. Journal of Rehabilitation Medicine 2004;36:7-8.
102. Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. Disability and Rehabilitation 2002;24:932-8.
103. National Institute for Clinical Excellence. Management of multiple sclerosis in primary and secondary care. London: NICE; 2003.
104. Solari A. Role of health-related quality of life measures in the routine care of people with multiple sclerosis. Health and Quality of Life Outcomes 2005;3:16.
105. Guyatt GH, Feeny DH, Patrick DL. Measuring Health-Related Quality of Life. Annals of Internal Medicine 1993;118:622-9.
106. Poolman RW, Swiontkowski MF, Fairbank JCT, Schemitsch EH, Sprague S, de Vet HCW. Outcome Instruments: Rationale for Their Use. The Journal of Bone and Joint Surgery 2009;91:pg41-9.
107. Acquadro C, Berzon R, Dubois D, et al. Incorporating the Patient's Perspective into Drug Development and Communication: An Ad Hoc Task Force Report of the Patient-Reported Outcomes (PRO) Harmonization Group Meeting at the Food and Drug Administration, February 16, 2001. Value in Health 2003;6:522-31.
108. Gruenewald DA, Higginson IJ, Vivat B, Edmonds P, Burman RE. Quality of life measures for the palliative care of people severely affected by multiple sclerosis: a systematic review. Multiple Sclerosis 2004;10:690-725.
109. Osoba D. Translating the science of patient-reported outcomes assessment into clinical practice. Journal of the National Cancer Institute 2007;Monographs.:5-11.
110. Valderas J, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. Quality of Life Research 2008;17:1125-35.
111. World Health Organization. Training Manual on Disability Statistics. In: Pacific WHOUNEaSCfAat, ed. Bangkok: United Nations; 2008.
112. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005;8:19-32.

113. Miller CE, Jezewski MA. Relapsing MS patients' experiences with glatiramer acetate treatment: a phenomenological study. Journal of Neuroscience Nursing 2006;38:37-41.
114. Miller CM. The lived experience of relapsing multiple sclerosis: a phenomenological study. Journal of Neuroscience Nursing 1997;29:294-304.
115. Vickers MH. Life and work with multiple sclerosis (MS): the role of unseen experiential phenomena on unreliable bodies and uncertain lives. Illness, Crisis & Loss 2009;17:7-21.
116. Vogt MHJ, Floris S, Killestein J, et al. Osteopontin levels and increased disease activity in relapsing-remitting multiple sclerosis patients. J Neuroimmunol 2004;155:155-60.
117. Okuda DT, Kozma C, Dickson M, Meletiche D. Treatment gaps and incidence of severe multiple sclerosis relapses. International Journal of MS Care 2008;10:27-.
118. Meletiche D, Kozma C, Bennett R, Al-Sabbagh A. Treatment adherence and severe relapses in multiple sclerosis patients. International Journal of MS Care 2008;10:24-5.
119. Meletiche D, Kozma C. Multiple sclerosis relapse prevalence: effect of varying definitions from claims database. International Journal of MS Care 2008;10:24-.
120. Vercellino M, Romagnolo A, Mattioda A, et al. Multiple sclerosis relapses: a multivariable analysis of residual disability determinants. Acta Neurologica Scandinavica 2009;119:126-30.
121. Hirst CL, Ingram G, Pickersgill TP, Robertson NP. Temporal evolution of remission following multiple sclerosis relapse and predictors of outcome. Mult Scler 2012;18:1152-8.
122. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. New England Journal of Medicine 2000;342:1887-92.
123. Hurwitz BJ. Analysis of current multiple sclerosis registries Neurology 2011;76:S7-S13.
124. Trojano M, Russo P, Fuiani A, et al. The Italian Multiple Sclerosis Database Network (MSDN): The risk of worsening according to IFNbeta exposure in multiple sclerosis. Mult Scler 2006;12:578-85.
125. Tremlett H, Zhao Y, Joseph J, Devonshire V, the UCN. Relapses in multiple sclerosis are age- and time-dependent. J Neurol Neurosurg Psychiatry 2008;79:1368-74.
126. Weinshenker BG. Databases in MS research: pitfalls and promises. Multiple Sclerosis 1999;5:206-11.
127. Inusah S, Sormani MP, Cofield SS, et al. Assessing changes in relapse rates in multiple sclerosis. Multiple Sclerosis 2010;16:1414-21.
128. Flachenecker P, Stuke K. National MS registries. Journal of Neurology 2008;255:102-8.
129. Hurwitz BJ. Registry studies of long-term multiple sclerosis outcomes. Neurology 2011;76:S3-S6.
130. The Once Weekly Interferon for MS Study Group. Evidence of interferon β-1a dose response in relapsing-remitting MS. Neurology 1999;53:679-86.
131. Comi G, Filippi M, Wolinsky JS. European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging–measured disease activity and burden in patients with relapsing multiple sclerosis. Annals of Neurology 2001;49:290-7.

132. Comi G, Pulizzi A, Rovaris M, et al. Effect of laquinimod on MRI-monitored disease activity in patients with relapsing-remitting multiple sclerosis: a multicentre, randomised, double-blind, placebo-controlled phase IIb study. The Lancet 2008;371:2085-92.
133. Comi G, Jeffery D, Kappos L, et al. Placebo-Controlled Trial of Oral Laquinimod for Multiple Sclerosis. New England Journal of Medicine 2012;366:1000-9.
134. Vollmer TL, Sorensen PS, Selmaj K, et al. A randomized placebo-controlled phase III trial of oral laquinimod for multiple sclerosis. Journal of Neurology 2014;261:773-83.
135. Giovannoni G, Comi G, Cook S, et al. A Placebo-Controlled Trial of Oral Cladribine for Relapsing Multiple Sclerosis. New England Journal of Medicine 2010;362:416-26.
136. Gold R, Giovannoni G, Selmaj K, et al. Daclizumab high-yield process in relapsing-remitting multiple sclerosis (SELECT): a randomised, double-blind, placebo-controlled trial. The Lancet 2013;381:2167-75.
137. Calabresi PA, Kieseier BC, Arnold DL, et al. Pegylated interferon beta-1a for relapsing-remitting multiple sclerosis (ADVANCE): a randomised, phase 3, double-blind study. The Lancet Neurology 2014;13:657-65.
138. Filippi M, Wolinsky JS, Comi G, Group TCS. Effects of oral glatiramer acetate on clinical and MRI-monitored disease activity in patients with relapsing multiple sclerosis: a multicentre, double-blind, randomised, placebo-controlled study. Lancet Neurology 2006;5:213-20.
139. Sorensen PS, Koch-Henriksen N, Ravnborg M, et al. Immunomodulatory treatment of multiple sclerosis in Denmark: a prospective nationwide survey. Multiple Sclerosis 2006;12:253-64.
140. Stuke K, Flachenecker P, Zettl U, et al. Symptomatology of MS: results from the German MS Registry. Journal of Neurology 2009;256:1932-5.
141. Confavreux C, Aimard G, Devic M. Course and prognosis of multiple sclerosis assessed by the computerized data processing of 349 patients. Brain 1980;103:281-300.
142. Waubant E, Vukusic S, Gignoux L, et al. Clinical characteristics of responders to interferon therapy for relapsing MS. Neurology 2003;61:184-9.
143. Debouverie M, Laforest L, Van Ganse E, Guillemin F, Group ftL. Earlier disability of the patients followed in Multiple Sclerosis centers compared to outpatients. Multiple Sclerosis 2009;15:251-7.
144. Binquet C, Quantin C, Le Teuff G, Pagliano JF, Abrahamowicz M, Moreau T. The Prognostic Value of Initial Relapses on the Evolution of Disability in Patients with Relapsing-Remitting Multiple Sclerosis. Neuroepidemiology 2006;27:45-54.
145. Trojano M, Russo P, Fuiani A, et al. The Italian Multiple Sclerosis Database Network (MSDN): the risk of worsening according to IFNβ exposure in multiple sclerosis. Multiple Sclerosis 2006;12:578-85.
146. Trojano M, Pellegrini F, Fuiani A, et al. New natural history of interferon-β–treated relapsing multiple sclerosis. Annals of Neurology 2007;61:300-6.
147. Jacobs LD, Wende KE, Brownscheidle CM, et al. A profile of multiple sclerosis: The New York State Multiple Sclerosis Consortium. Multiple Sclerosis 1999;5:369-76.
148. Fromont A, Debouverie M, Le Teuff G, Quantin C, Binquet C, Moreau T. Clinical parameters to predict response to interferon in relapsing multiple sclerosis. Neuroepidemiology 2008;31:150-6.

149. Goodkin DE, Hertsgaard D, Rudick RA. Exacerbation Rates and Adherence to Disease Type in a Prospectively Followed-up Population With Multiple Sclerosis: Implications for Clinical Trials. Arch Neurol 1989;46:1107-12.
150. Myhr K, Tiise T, Vedeler C, et al. Disability and prognosis in multiple sclerosis: demographic and clinical variables important for the ability to walk and awarding of disability pension. Multiple Sclerosis 2001;7:59-65.
151. Kurtzke JF, Beebe GW, Nagler B, Auth TL, Kurland LT, Nefzger MD. Studies on natural history of multiple sclerosis. 4. Clinical Features of the Onset Bout. Acta Neurologica Scandinavica 1968;44:467-94.
152. Kappos L, Lechner-Scott J, Lienert C. Neurostatus.net. 2007.
153. Hawkes CH, Giovannoni G. The McDonald Criteria for Multiple Sclerosis: time for clarification. Multiple Sclerosis 2010;16:566-75.
154. Morrow SA, Kremenchutzky M. The role of the primary-care physician in treating relapses in multiple sclerosis patients. International Journal of MS Care 2009;11:122-6.
155. McHugh JC, Galvin PL, Murphy RP. Retrospective comparison of the original and revised McDonald criteria in a general neurology practice in Ireland. Multiple Sclerosis 2008;14:81-5.
156. Nicholas R, Straube S, Schmidli H, Schneider S, Friede T. Trends in annualized relapse rates in relapsing-remitting multiple sclerosis and consequences for clinical trial design. Multiple Sclerosis Journal 2011;17:1211-7.
157. Held U, Heigenhauser L, Shang C, Kappos L, Polman C, Sylvia Lawry Centre for MSR. Predictors of relapse rate in MS clinical trials. Neurology 2005;65:1769-73.
158. Klawiter EC, Cross AH, Naismith RT. The present efficacy of multiple sclerosis therapeutics: Is the new 66% just the old 33%? Neurology 2009;73:984-90.
159. Barkhof F, Hulst HE, Drulović J, et al. Ibudilast in relapsing-remitting multiple sclerosis: A neuroprotectant? Neurology 2010;74:1033-40.
160. Kurtzke JF. A New Scale for Evaluating Disability in Multiple Sclerosis. Neurology 1955;5:580-3.
161. Kurtzke JF. On the evaluation of disability in multiple sclerosis. Neurology 1961;11:686-94.
162. Kurtzke JF. Disability Rating Scales in Multiple Sclerosis. Annals of the New York Academy of Sciences 1984;436:347-60.
163. Sharrack B, Hughes RAC. Clinical scales for multiple sclerosis. Journal of the Neurological Sciences 1996;135:1-9.
164. Sharrack B, Hughes RAC, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999;122:141-59.
165. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000;123:1027-40.
166. Bowen J, Gibbons L, Gianas A, Kraft GH. Self-administered Expanded Disability Status Scale with functional system scores correlates well with a physician-administered test. Multiple Sclerosis 2001;7:201-6.
167. Gaspari M, Roveda G, Scandellari C, Stecchi S. An expert system for the evaluation of EDSS in multiple sclerosis. Artificial Intelligence in Medicine 2002;25:187-210.
168. Cutter GR, Baier ML, Rudick RA, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999;122:871-82.
169. Goodkin DE. EDSS reliability. Neurology 1991;41:332.

170. Kragt JJ, Nielsen JM, van der Linden FA, Uitdehaag BM, Polman CH. How similar are commonly combined criteria for EDSS progression in multiple sclerosis? Multiple Sclerosis 2006;12:782-6.
171. Fischer JS, Rudick RA, Cutter GR, Reingold SC, Force NMSCOAT. The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. Multiple Sclerosis 1999;5:244-50.
172. World Health Organization. International Classification of Impairments, Disabilities and Handicaps (ICIDH). Geneva: World Health Organization; 1980.
173. Dickson HG. Problems with the ICIDH definition of impairment. Disability and Rehabilitation 1996;18:52-4.
174. Traboulsee A, Létourneau-Guillon L, Freedman MS, et al. Canadian Expert Panel Recommendations for MRI Use in MS Diagnosis and Monitoring. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 2015;42:159-67.
175. Goodin DS. Magnetic resonance imaging as a surrogate outcome measure of disability in multiple sclerosis: Have we been overly harsh in our assessment? Annals of Neurology 2006;59:597-605.
176. Neema M, Stankiewicz J, Arora A, Guss Z, Bakshi R. MRI in Multiple Sclerosis: What's Inside the Toolbox? Neurotherapeutics 2007;4:602-17.
177. Zivadinov R, Stosic M, Cox JL, Ramasamy DP, Dwyer MG. The place of conventional MRI and newly emerging MRI techniques in monitoring different aspects of treatment outcome. Journal of Neurology 2008;255:61-74.
178. Foley JF, Brandes DW. Redefining functionality and treatment efficacy in multiple sclerosis. Neurology 2009;72:S1-S11.
179. Richards RG, Sampson FC, Beard SM, Tappenden P. A review of the natural history and epidemiology of multiple sclerosis: implications for resource allocation and health economic models. Health Technology Assessment 2002;6:1-73.
180. Lapierre Y, Hum S. Treating Fatigue. The International MS Journal 2007;14:64-71.
181. Krupp LB, Alvarez LA, LaRocca NG, Scheinberg LC. Fatigue in Multiple Sclerosis. Archives of Neurology 1988;45:435-7.
182. Paltamaa J, Sarasoja T, Leskinen E, Wikstrom J, Malkia E. Measures of physical functioning predict self-reported performance in self-care, mobility, and domestic life in ambulatory persons with multiple sclerosis. Archives of Physical Medicine & Rehabilitation 2007;88:1649-57.
183. Rao SM, Leo GJ, Ellington L, Nauertz T, Bernadin L, Unverzagt F. Cognitive dysfunction in multiple sclerosis. II. Impact on employment and social functioning. Neurology 1991;41:692-6.
184. Hakim EA, Bakheit AMO, Bryant TN, et al. The social impact of multiple sclerosis - a study of 305 patients and their relatives. Disability & Rehabilitation 2000;22:288-93.
185. Stucki G. International Classification of Functioning, Disability, and Health (ICF): A Promising Framework and Classification for Rehabilitation Medicine. American Journal of Physical Medicine & Rehabilitation 2005;84:733-40.
186. Stevenson VL, Playford ED. Rehabilitation and MS. International MS Journal 2007;14:85-92.
187. Khan F, Pallant JF. Use of International Classification of Functioning, Disability and Health (ICF) to describe patient-reported disability in multiple sclerosis and identification of relevant environmental factors. Journal of Rehabilitation Medicine 2007;39:63-70.

188. Svestkova O, Angerova Y, Sladkova P, Keclikova B, Bickenbach JE, Raggi A. Functioning and disability in multiple sclerosis. Disability & Rehabilitation 2010;32 Suppl 1:S59-67.
189. Conrad A, Coenen M, Schmalz H, Kesselring J, Cieza A. Validation of the Comprehensive ICF Core Set for Multiple Sclerosis From the Perspective of Physical Therapists. Physical Therapy 2012;92:799-820.
190. Karhula ME, Kanelisto KJ, Ruutiainen J, Hämäläinen PI, Salminen A-L. The activities and participation categories of the ICF Core Sets for multiple sclerosis from the patient perspective. Disability and Rehabilitation 2013;35:492-7.
191. Holper L, Coenen M, Weise A, Stucki G, Cieza A, Kesselring J. Characterization of functioning in multiple sclerosis using the ICF. Journal of Neurology 2010;257:103-13.
192. Khan F, Amatya B, Turner-Stokes L. Symptomatic Therapy and Rehabilitation in Primary Progressive Multiple Sclerosis. Neurology Research International 2011;2011.
193. Rasova K, Feys P, Henze T, et al. Emerging evidence-based physical rehabilitation for Multiple Sclerosis - Towards an inventory of current content across Europe. Health and Quality of Life Outcomes 2010;8:76.
194. Pollard B, Dixon D, Dieppe P, Johnston M. Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. Health and Quality of Life Outcomes 2009;7:41.
195. Grill E, Stucki G. Scales could be developed based on simple clinical ratings of International Classification of Functioning, Disability and Health Core Set categories. Journal of Clinical Epidemiology 2009;62:891-8.
196. Johansson S, Ytterberg C, Back B, Holmqvist LW, von Koch L. The Swedish Occupational Fatigue Inventory in people with multiple sclerosis. Journal of Rehabilitation Medicine 2008;40:737-43.
197. Cano SJ, Hobart JC. The problem with health measurement. Patient Preference and Adherence 2011;5:279-90.
198. Potter K, Cohen ET, Allen DD, et al. Outcome Measures for Individuals With Multiple Sclerosis: Recommendations From the American Physical Therapy Association Neurology Section Task Force. Physical Therapy 2013.
199. Clinical Outcome Assessment Qualification Program. 2014. (Accessed at http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm.)
200. Kragt J, Nielsen J, van der Linden F, Polman C, Uitdehaag B. Disease progression in multiple sclerosis: combining physicians' and patients' perspectives? Multiple Sclerosis Journal 2011;17:234-40.
201. Bamer AM, Cetin K, Amtmann D, Bowen JD, Johnson KL. Comparing a self report questionnaire with physician assessment for determining multiple sclerosis clinical disease course: a validation study. Multiple Sclerosis 2007;13:1033-7.
202. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;314:1580.
203. Guralnik JM, Branch LG, Cummings SR, Curb JD. Physical Performance Measures in Aging Research. Journal of Gerontology 1989;44:M141-M6.

204. Stratford PW, Kennedy D, Pagura SMC, Gollish JD. The relationship between self-report and performance-related measures: Questioning the content validity of timed tests. Arthritis Care & Research 2003;49:535-40.
205. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. second edition ed. Oxford: Oxford University Press; 1995.
206. Goverover Y, Kalmar J, Gaudino-Goering E, et al. The Relation Between Subjective and Objective Measures of Everyday Life Activities in Persons With Multiple Sclerosis. Archives of Physical Medicine and Rehabilitation 2005;86:2303-8.
207. Nagin D. Group-Based Modeling of Development. Cambridge: Harvard University Press; 2005.
208. Horton NJ, Switzer SS. Statistical Methods in the Journal. New England Journal of Medicine 2005;353:1977-9.
209. Cnaan A, Ryan L. Survival analysis in natural history studies of disease. Statistics in Medicine 1989;8:1255-68.
210. Kappos L, Freedman MS, Polman CH, et al. Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. The Lancet 2007;370:389-97.
211. Tintoré M, Rovira A, RÍo J, et al. Baseline MRI predicts future attacks and disability in clinically isolated syndromes. Neurology 2006;67:968-72.
212. Comi G, Filippi M, Barkhof F, et al. Effect of early interferon treatment on conversion to definite multiple sclerosis: a randomized study. The Lancet 2001;357:1576-82.
213. Jacobs LD, Beck RW, Simon JH, et al. Intramuscular interferon beta-1a therapy initiated during a first demyelinating event in multiple sclerosis. The New England Journal of Medicine 2000;343:898-904.
214. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurological Sciences 2000;21:S831-8.
215. Cottrell DA, Kremenchutzky M, Rice GPA, Hader W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: a geographically based study: 6. Applications to planning and interpretation of clinical therapeutic trials in primary progressive multiple sclerosis. Brain 1999;122:641-7.
216. Ebers GC, Koopman WJ, Hader W, et al. The natural history of multiple sclerosis: a geographically based study: 8: Familial multiple sclerosis. Brain 2000;123:641-9.
217. Pittock SJ, Mayr WT, McClelland RL, et al. Disability profile of MS did not change over 10 years in a population-based prevalence cohort. Neurology 2004;62:601-6.
218. Scalfari A, Neuhaus A, Degenhardt A, et al. The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. Brain 2010;133:1914-29.
219. Liu C, Blumhardt LD. Disability outcome measures in therapeutic trials of relapsing-remitting multiple sclerosis: effects of heterogeneity of disease course in placebo cohorts. Journal of Neurology, Neurosurgery & Psychiatry 2000;68:450-7.
220. Gray O, Butzkueven H. Measurement of disability in multiple sclerosis. Neurology Asia 2008;13:153-6.
221. Mandel M, Gauthier SA, Guttmann CRG, Weiner HL, Betensky RA. Estimating Time to Event From Longitudinal Categorical Data: An Analysis of Multiple Sclerosis Progression. Journal of the American Statistical Association 2007;102:1254-66.
222. Mathew A, Pandey M, Murthy NS. Survival analysis: caveats and pitfalls. European Journal of Surgical Oncology (EJSO) 1999;25:321-9.

223. Twisk JWR. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge: Cambridge University Press; 2003.
224. Tu XM, Kowalski J, Zhang J, Lynch KG, Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. Statistics in Medicine 2004;23:2799-815.
225. Twisk JWR. Longitudinal Data Analysis. A Comparison Between Generalized Estimating Equations and Random Coefficient Analysis. European Journal of Epidemiology 2004;19:769-76.
226. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.
227. Zeger SL, Liang K-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics 1986;42:121-30.
228. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. Biometrics 1982;38:963-74.
229. Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press; 2003.
230. Maxwell SE, Delaney HD. Designing Experiments and Analyzing Data: A Model Comparison Perspective. Mahwah: Lawrence Erlbaum Associates, Publishers; 2004.
231. Atkins DC. Using Multilevel Models to Analyze Couple and Family Treatment Data: Basic and Advanced Issues. Journal of Family Psychology;Journal of Family Psychology 2005;19:98-110.
232. Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. Statistics in Medicine 1993;12:1723-32.
233. Zorn CJW. Generalized Estimating Equation Models for Correlated Data: A Review with Applications. American Journal of Political Science 2001;45:470-90.
234. Stoel RD, Hox JJ, Wittenboer Gvd. Analyzing longitudinal data using multilevel regression and latent growth curve analysis. Metodologia de las Ciencas del Comportamiento 2003;5:1-21.
235. DeLucia C, Pitts SC. Applications of Individual Growth Curve Modeling for Pediatric Psychology Research. Journal of Pediatric Psychology 2006;31:1002-23.
236. Nagin DS, Odgers CL. Group-Based Trajectory Modeling in Clinical Research. Annual Review of Clinical Psychology 2010;6:109-38.
237. McLachlan G, Peel D. Finite Mixture Models. New York: John Wiley & Sons, Inc; 2000.
238. Nagin DS, Tremblay RE. DEVELOPMENTAL TRAJECTORY GROUPS: FACT OR A USEFUL STATISTICAL FICTION?*. Criminology 2005;43:873-904.
239. Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. Psychological Methods 1999;4:139-57.
240. Andrich D. Rating scales and Rasch measurement. Expert Review of Pharmacoeconomics & Outcomes Research 2011;11:571-85.
241. Cano S, Klassen AF, Scott A, Feeny D. Health outcome and economic measurement in breast cancer surgery: challenges and opportunities. Expert Review of Pharmacoeconomics & Outcomes Research 2010;10:583-94.
242. Mayo NE. Dictionary of Quality of Life and Health Outcomes Measurement, Version 1. Version 1 ed. ISQOL; 2015.
243. Tennant A, McKenna SP, Hagell P. Application of Rasch Analysis in the Development and Application of Quality of Life Instruments. Value in Health 2004;7:S22-S6.

244. Baghaei P. Local Dependency and Rasch Measures. Rasch Measurement Transactions 2008;21:1105-6.
245. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technology Assessment 2009;13:1-177.
246. McHorney CA. Ten Recommendations for Advancing Patient-Centered Outcomes Measurement for Older Persons. Annals of Internal Medicine 2003;139:403-9.
247. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology 2007;6:1094-105.
248. Andrich D, Luo G, Sheridan B. Interpreting RUMM2020 Part I Dichotomous Data. In. ninth ed. Perth, Western Australia: RUMM Laboratory Pty Ltd.; 2004.
249. University of Leeds. Introductory Rasch Analysis. In: Psylab Group University of Leeds; 2013.
250. Pallant JF, Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology 2007;46:1-18.
251. Ramp M, Khan F, Misajon R, Pallant J. Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). Health and Quality of Life Outcomes 2009;7:58.
252. Vanhoutte EK, Faber CG, van Nes SI, et al. Modifying the Medical Research Council grading system through Rasch analyses. Brain 2012;135:1639-49.
253. Muller S, Roddy E. A Rasch Analysis of the Manchester Foot Pain and Disability Index. Journal of Foot and Ankle Research 2009;2:29.
254. Bartram D, Sinclair J, Baldwin D. Further validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in the UK veterinary profession: Rasch analysis. Quality of Life Research 2013;22:379-91.
255. Mills RJ, Young CA, Nicholas RS, Pallant JF, Tennant A. Rasch analysis of the Fatigue Severity Scale in multiple sclerosis. Multiple Sclerosis 2009;15:81-7.
256. Smith EV. Detecting and Evaluating the Impact of Multidimensionallity using Item Fit Statistics and Principal Component Analysis of Residuals. Journal of Applied Measurement 2002;3:205-31.
257. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. J Rehabil Med 2011;43:884-91.
258. Finch L, Higgins J, Wood-Dauphinee S, Mayo N. Development of a measure of functioning for stroke recovery: The functional recovery measure. Disability & Rehabilitation 2008;30:577-92.
259. Tennant A, Penta M, Tesio L, et al. Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales through Differential Item Functioning within the Framework of the Rasch Model: The PRO-ESOR Project. Medical Care 2004;42:I37-I48.
260. Lundgren-Nilsson A, Tennant A, Grimby G, Sunnerhagen K. Cross-diagnostic validity in a generic instrument: an example from the Functional Independence Measure in Scandinavia. Health and Quality of Life Outcomes 2006;4:55.
261. Wright BD. Rack and Stack: Time 1 vs. Time 2. Rasch Measurement Transactions 2003;17:905-6.

262. Johnston MV, Shawaryn MA, Malec J, Kreutzer J, Hammond FM. The structure of functional and community outcomes following traumatic brain injury. Brain Injury 2006;20:391-407.
263. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561-73.
264. Masters GN. A rasch model for partial credit scoring. Psychometrika 1982;47:149-74.
265. Andrich D, Luo G, Sheridan B. Interpreting RUMM2020 Part II Polytomous Data. In. ninth ed. Perth, Western Australia: RUMM Laboratory Pty Ltd.; 2005.
266. Andrich D, Lyne A, Sheridan B, Luo G. Rasch Unidimensional Measurement Models (RUMM2020 Version 4.1). In. Duncraig, Western Australia: Rumm Laboratory Pty Ltd; 2003.
267. Scott N, Fayers P, Aaronson N, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health and Quality of Life Outcomes 2010;8:81.
268. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Medical Care 2006;44:S115-S23 10.1097/01.mlr.0000245183.28384.ed.
269. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310.
270. Marais I. Local Dependence. In: Christensen KB, Kreiner S, Mesbah M, eds. Rasch Models in Health. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2013:111-30.
271. Butzkueven H, Chapman J, Cristiano E, et al. MSBase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. Multiple Sclerosis 2006;12:769-74.
272. Lublin FD, Cofield SS, Cutter GR, et al. Randomized study combining interferon and glatiramer acetate in multiple sclerosis. Annals of Neurology 2013;73:327-40.
273. Lamers I, Kelchtermans S, Baert I, Feys P. Upper Limb Assessment in Multiple Sclerosis: A Systematic Review of Outcome Measures and their Psychometric Properties. Archives of Physical Medicine and Rehabilitation 2014;95:1184-200.
274. Rozzini r, Frisoni GB, Ferrucci L, Barbisoni P, Bertozzi B, Trabucchi M. The effect of chronic diseases on physical function. Comparison between activities of daily living scales and the Physical Performance Test. Age and Ageing 1997;26:281-7.
275. Freedman MS. "Time is brain" also in multiple sclerosis. Multiple Sclerosis 2009.
276. Bevan CJ, Cree BC. Disease activity free status: A new end point for a new era in multiple sclerosis clinical research? JAMA Neurology 2014;71:269-70.
277. Stangel M, Penner IK, Kallmann BA, Lukas C, Kieseier BC. Towards the implementation of 'no evidence of disease activity' in multiple sclerosis treatment: the multiple sclerosis decision model. Therapeutic Advances in Neurological Disorders 2015;8:3-13.
278. Giovannoni G, Turner B, Gnanapavan S, Offiah C, Schmierer K, Marta M. Is it time to target no evident disease activity (NEDA) in multiple sclerosis? Multiple Sclerosis and Related Disorders 2015;4:329-33.
279. Lublin FD, Reingold SC, Cohen JA, et al. Defining the clinical course of multiple sclerosis: The 2013 revisions. Neurology 2014;83:278-86.

280. Freedman MS, Selchen D, Arnold DL, et al. Treatment Optimization in MS: Canadian MS Working Group Updated Recommendations. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 2013;40:307-23.
281. Arnold DL, Li D, Hohol M, et al. Evolving role of MRI in optimizing the treatment of multiple sclerosis: Canadian Consensus recommendations. Multiple Sclerosis Journal – Experimental, Translational and Clinical 2015;1.
282. Emery DJ, Forster AJ, Shojania KG, Magnan S. Management of MRI Wait Lists in Canada. Healthcare Policy 2009;4:76-86.
283. Kieseier BC, Wiendl H, Hartung H-P, Leussink V-I, Stüve O. Risks and benefits of multiple sclerosis therapies: need for continual assessment? Current Opinion in Neurology 2011;24:238-43 10.1097/WCO.0b013e32834696dd.
284. Rae-Grant AD, Eckert NJ, Bartz S, Reed JF. Sensory symptoms of multiple sclerosis: a hidden reservoir of morbidity. Multiple Sclerosis 1999;5:179-83.
285. Lorig KR, Sobel DS, Stewart AL, et al. Evidence Suggesting That a Chronic Disease Self-Management Program Can Improve Health Status While Reducing Hospitalization: A Randomized Trial. Medical Care 1999;37:5-14.
286. Hobart J, Freeman J, Lamping D, Fitzpatrick R, Thompson A. The SF-36 in multiple sclerosis: why basic assumptions must be tested. Journal of Neurology, Neurosurgery & Psychiatry 2001;71:363-70.
287. Fisk JD, Brown MG, Sketris IS, Metz LM, Murray TJ, Stadnyk KJ. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. Journal of Neurology, Neurosurgery & Psychiatry 2005;76:58-63.
288. Cella DF, Dineen K, Arnason B, et al. Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. Neurology 1996;47:129-39.
289. Fisk JD, Pontefract A, Ritvo PG, Archibald CJ, Murray JT. The Impact of Fatigue on Patients with Multiple Sclerosis. Canadian Journal of Neurological Sciences 1994;21:9-14.
290. Smets EMA, Garssen B, Bonke B, De Haes JCJM. The multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. Journal of Psychosomatic Research 1995;39:315-25.
291. Gold SM, Schulz H, Mönch A, Schulz K-H, Heesen C. Cognitive impairment in multiple sclerosis does not affect reliability and validity of self-report health measures. Multiple Sclerosis 2003;9:404-10.
292. Shawaryn MA, Schiaffino KM, LaRocca NG, Johnston MV. Determinants of health-related quality of life in multiple sclerosis: the role of illness intrusiveness. Multiple Sclerosis 2002;8:310-8.
293. Cano S, Barrett L, Zajicek J, Hobart J. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. Multiple Sclerosis Journal 2011;17:214-22.
294. Ritvo PG, Fischer JS, Miller DM, Andrews H, Paty DW, LaRocca NG. Multiple Sclerosis Quality of Life Inventory: A User's Manual. New York: National Multiple Sclerosis Society; 1997.
295. Lovera J, Bagert B, Smoot KH, et al. Correlations of Perceived Deficits Questionnaire of Multiple Sclerosis Quality of Life Inventory and Beck Depression Inventroy and neuropsychological tests. Journal of Rehabilitation Research & Development 2006;43:73-82.

296. Poissant L, Mayo NE, Wood-Dauphinee S, Clarke AE. The development and preliminary validation of a Preference-Based Stroke Index (PBSI). Health and Quality of Life Outcomes 2003;1.
297. Solari A, Radice D, Manneschi L, Motti L, Montanari E. The multiple sclerosis functional composite: different practice effects in the three test components. Journal of the Neurological Sciences 2005;228:71-4.
298. Feys P, Duportail M, Kos D, Van Aschand P, Ketelaer P. Validity of the TEMPA for the measurement of upper limb function in multiple sclerosis. Clinical Rehabilitation 2002;16:166-73.
299. Benedict RHB, Cookfair D, Gavett R, et al. Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). Journal of the International Neuropsychological Society 2006;12:549-58.
300. Goldman MD, Marrie RA, Cohen JA. Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls. Multiple Sclerosis 2008;14:383-90.
301. Schwid SR, Thornton CA, Pandya S, et al. Quantitative assessment of motor fatigue and strength in MS. Neurology 1999;53:743.
302. Canadian Society for Exercise Physiology. The Canadian Physical Activity Fitness & Lifestyle Approach (CPAFLA) :CSEP-Health & Fitness Program's Health-Related Appraisal and Counselling Strategy. 3rd ed; 2003.
303. Tesio L, Perucca L, Franchignoni FP, Battaglia MA. A short measure of balance in multiple sclerosis: Validation through Rasch analysis. Functional Neurology 1997;12:255-65.
304. Graham JE, Ostir GV, Fisher SR, Ottenbacher KJ. Assessing walking speed in clinical research: a systematic review. Journal of Evaluation in Clinical Practice 2008;14:552-62.
305. Vaney C, Blaurock H, Gattlen B, Meisels C. Assessing mobility in multiple sclerosis using the Rivermead Mobility Index and gait speed. Clinical Rehabilitation 1996;10:216-26.
306. Nuyens G, De Weerdt W, Ketelaer P, et al. Inter-rater reliability of the Ashworth scale in multiple sclerosis. Clinical Rehabilitation 1994;8:286-92.
307. Pandyan AD, Price CI, Barnes MP, Johnson GR. A biomechanical investigation into the validity of the modified Ashworth Scale as a measure of elbow spasticity. Clinical Rehabilitation 2003;17:290-4.
308. Kuspinar A, Andersen RE, Teng SY, Asano M, Mayo NE. Predicting Exercise Capacity Through Submaximal Fitness Tests in Persons With Multiple Sclerosis. Archives of physical medicine and rehabilitation 2010;91:1410-7.
309. Langeskov-Christensen M, Langeskov-Christensen D, Overgaard K, Møller AB, Dalgas U. Validity and reliability of VO2-max measurements in persons with multiple sclerosis. Journal of the Neurological Sciences 2014;342:79-87.

Appendix

Table 3.1a: Randomized Control Trials in MS

ref id	Author	year pub	enrol	# centres and countries	n	type	drug	age	sex ratio	EDSS
.[27]	IFNB GRP	1993	Jun 1988- May 1990	11 centres in USA and Canada	372	RR	placebo; betaseron 1.6MIU/8 MIU	18-50	2:1	0.0-5.5
.[28]	Jacobs	1995	Nov 1990	4 centres in USA	301	RR	avonex vs placebo	18-55	3:1	1.0-3.5
.[30]	Johnson	1995	Oct 1991	11 centres in USA	251	RR	copaxone vs placebo	18-45	2.7:1	0.0-5.0
.[29]	Ebers	1998	May 1994 - Feb 1995	22 centres in 9 countries	560; (533)	RR	placebo; rebif 22 tiw/rebif 44 tiw	not stated	2.2:1	0.0-5.0
.[130]	OWIMS	1999	Mar 1995 - Nov 1995	11 centres in 5 countries	293	RR	OWIMS (1 yr) placebo; rebif 22 qw/ rebif 44 qw	18-50	2.44:1	0.0-5.0
.[131]	Comi	2001	Feb 1997 - Nov 1997	29 centres in 6 Europe and Canada	239	RR	copaxone vs placebo	18-50	unk	0.0-5.0
.[138]	Filippi	2006	Mar 2000 - Sept 2000	158 centres worldwide	1651; (1644)	RR	oral copaxone 5mg or 50mg vs placebo	18-50	2.87:1	0.0-5.0
.[47]	Polman	2006	Nov 2001	99 centres in Europe, North America, Autralia, and New Zealand	942	RR	AFFIRM: Natalizumab vs placebo	18-50	2.33:1	0.0-5.0
.[50]	Kappos	2006	May 2003 - April 2004	32 centres in 10 European countries and Canada	281; (277)	RR; SP	FTY720 1.25mg or 5mg vs placebo	18-60	3.4:1	0.0-6.0
.[132]	Comi	2008	Mar 2005 - Oct 2005	51 centres in 9 countreis	306	RR	Laquinimod 0.3mg or 0.6mg vs placebo	18-50	unk	1.0-5.0

.[135]	Giovannoni	2010	Apr 2005 - Jan 2007	155 centres in 32 countries	1326	RR	Oral Cladribine 3.5mg/kg or 5.25mg/kg vs placebo	18-65	1.93:1	0.0-5.5
.[49]	Kappos	2010	Jan 2006 - Aug 2007	138 centres in 22 countries	1272	RR	FTY720 1.25mg or 5mg vs placebo	18-55	2.32:1	0.0-5.5
.[51]	O'Connor	2011	Sept 2004 -Mar 2008	127 centres in 21 countries	1086	RR; SP; PR	Teriflunomide 14mg, 7mg, vs placebo	18-55	2.59:1	0.0-5.5
.[159]	Barkhof	2010	unk	19 centres in Europe	297	RR; SP	Ibudilast	18-55	1.9:1	0.0-5.0
.[133]	Comi	2012	Nov 2007 - Nov 2008	139 centres in 24 countries	1106	RR	Laquinimod 0.6mg vs placebo	18-55	2.2:1	0.0-5.5
.[53]	Fox	2012	June 2007 from clinicaltrials.gov	200 centres in 28 countries	1417	RR	BG-12 240mgBID vs BG-12 240mgTID vs GA vs placebo	18-55	2.3:1	0.0-5.0
.[54]	Gold	2012	Jan 2007 from clinicaltrials.gov	198 centres in 28 countries	1234	RR	BG-12 240mgBID vs BG-12 240mgTID vs placebo	18-55	2.8:1	0.0-5.0
.[136]	Gold	2013	Feb 2008 - May 2010	76 centres in 9 countries	621	RR	Daclizumab 150mg vs 300mg vs placebo	18-55	1.8:1	0.0-5.0
.[137]	Calabresi	2014	June 2009 from clinicaltrials.gov	183 centres in 26 countries	1512	RR	peginterferon beta-1a 125ug 2wks vs 125ug 4wks vs placebo	18-65	2.4:1	0.0-5.0
.[52]	Confavreux	2014	Sept 2008 - Feb 2011	189 centres in 26 countries	1169	RR	teriflunomide 7mg vs 14mg vs placebo	18-55	2.5:1	0.0-5.5
.[134]	Vollmer	2014	Spril 2008 - June 2011	155 centres in 18 countries	1331	RR	Laquinimod 0.6mg vs IFN beta1a IM 30 ug vs placebo	18-55	2.2:1	0.0-5.5

Table 3.1b: Randomized Control Trials in MS

ref id	pre-relapse rate	pre ARR	primary outcome (ARR=annual relapse rate)	EDSS endpoint	min relapse time	relapse exam time	tools
.[27]	2 in 2yrs	unk	ARR; prop. relapse free	secondary endpoint change in EDSS or NRS	24	call clinic	EDSS; Scripps
.[28]	>=0.67/yr for previous 3yrs OR 1relapse/yr if less than 3 yrs of disease	1.2	=>1 EDSS for 6 months (ARR is secondary outcome)	primary outcome sustained EDSS ≥ 1 for 6 months	48	24 hrs	EDSS
.[30]	2 in 2yrs	1.45	mean # of relapse in 2 yrs	secondary endpoint sustained EDSS ≥ 1 for 3 months	48	7 days	EDSS
.[29]	2 in 2yrs	1.5	relapse count over study (ARR)	secondary endpoint sustained EDSS ≥ 1 for 3 months	24	7 days	EDSS; Scripps; ADL
.[130]	1 relapse in last 24 months	1.2	# of active PD/T2 OR T1-Gd at 24wks (ARR is secondary endpoint)	EDSS and SNRS disability measure	24	7 days	EDSS; Scripps
.[131]	1 relapse in last 24 months and at least 1 Gad-enh lesion on MRI	prior 2 yrs:2.65	# of Gad-Enh lesions (ARR is tertiary endpoint)	EDSS tertiary endpoint	48	7 days	EDSS
.[138]	1 relapse in last 12 months	1yr:1.5; 2yr:2.17	ARR	secondary endpoint is # of relapses treated with steroids; EDSS tertiary endpoint	48	7 days	EDSS
.[47]	1 relapse in last 12 months	1.52	at 1 yr:ARR; at 2 yrs primary endpoint was sustained prog. for 3 months of EDSS ≥ 1.0 with baseline EDSS=1 or more or EDSS ≥ 1.5 with baseline EDSS= 0	at 2 years,second. endpoint were ARR; vol of T2 lesions; T1 lesions; MSFC	24	72 hrs	EDSS; MSFC

.[50]	2 in 2 yrs OR 1 in 1 yr and 1 or more gado-enh. lesions on MRI screening	1.26	# of gado-enh lesion per patient recorded on T1-wt MRI monthly for 6 months (ARR is tertiary endpoint)	EDSS and FSS change to confirm relapse	24	not stated	EDSS; (MSFC: not for relapse)
.[132]	1 relapse in last 12 months and 1 gad-enh lesion	1.45	# of Gd-enh lesion (ARR is secondary endpoint)	additional clinical outcome was change in EDSS	48	7 days	EDSS
.[135]	1 relapse in last 12 months	1.35	ARR	second. endpoint was sustained prog. for 3 months of EDSS ≥ 1.0 with baseline EDSS=1 or more or EDSS ≥ 1.5 with baseline EDSS= 0	24	7 days	EDSS
.[49]	1 or more relapse in last 12 months or 2 or more relapse in last 24 months	prior 2 yrs:1.47	ARR	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS > 5.5	not stated	7 days	EDSS; (MSFC: not for relapse)
.[51]	2 in yrs or 1 in last year	1 yr=1.4; 2 yr=2.2	ARR	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS >5.5	24	7 days	EDSS
.[159]	1 relapse in last 12 months; 1 or more Gd eh lesion	1 in 1 yr	new Gd enh lesions on MRI (ARR is secondary outcome)	additional outcomes: change in EDSS at 12 and 24 months and confirmed EDSS progrss ≥1 for ≥ 4 months	unk	unk	EDSS

.[133]	1≥relapse in last year, or 2 ≥ in last 2 years, or 1 relapse in 1 and 2 years and 1≥ Gd enh lesion in last year	1.3 in 1 yr;1.9 in 2 yrs	number of confirmed relapses (ARR)	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS > 5.5	48	contact within 48 hrs and evaluation by examining neurologist within 7 days	EDSS; MSFC
.[53]	1≥relapse in last 12 months or 1≥Gd enh lesion in last 6 weeks	1.4 in 1 yr	ARR at 2 years	second. endpoint was sustained prog. for 3 months of EDSS ≥ 1.0 with baseline EDSS=1 or more or EDSS ≥ 1.5 with baseline EDSS= 0	24	contact within 48 hrs; evaluation within 72 hrs by treating neurologist and 5 days by examining neurologist	EDSS
.[54]	1≥relapse in last 12 months or 1≥Gd enh lesion in last 6 weeks	1.3 in 1 yr	ARR at 2 years	second. endpoint was sustained prog. for 3 months of EDSS ≥ 1.0 with baseline EDSS=1 or more or EDSS ≥ 1.5 with baseline EDSS= 0	24	contact within 48 hrs; evaluation within 72 hrs by treating neurologist and 5 days by examining neurologist	EDSS
.[136]	1≥relapse in last 12 months or 1≥Gd enh lesion in last 6 weeks	1.3 in 1 yr	ARR at week 52	tertairy endpoint was sustained progression for 3 months of EDSS ≥ 1.0 with baseline EDSS=1 or more or EDSS ≥ 1.5 with baseline EDSS= 0	24	not stated	EDSS
.[137]	2≥ relapses in 3years and 1 in last 12 months	1.6 in 1 yr; 2.6 in 3 yrs	ARR at week 48	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS > 5.5	24	not stated	EDSS

.[52]	1≥relapse in last 12 months or 2≥relapses in last 24 months	1.4 in 1 yr; 2.1 in 2 yrs	ARR between 48 and 152 weeks	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS > 5.5	24	not stated	EDSS
.[134]	1≥relapse in last year, or 2≥in last 2 years, or 1 relapse in 1 and 2 years and 1≥ Gd enh lesion in last year	1.0 in 1 yr; 2.0 in 2 yrs	ARR over the 24 month treatment period	second. endpoint was sustained prog. For 3 months of EDSS ≥ 1.0 with baseline EDSS=0-5.0 or more or EDSS ≥ 0.5 with baseline EDSS > 5.5	48	contact within 48 hrs and evaluation by examining neurologist within 7 days	EDSS; MSFC

Table 3.1c: Randomized Control Trials in MS

ref id	duration	duration; severity	length of study	n	placebo relapse rate	MS diag crit	definition of a relapse
.[27]	yes not reported	duration recorded but not reported; Scripps; Hospital-ization	2yrs	112	1.27 (95%CI; 1.12-1.43)	Poser	An exacerbation was defined as the appearance of a new symptom or worsening of an old symptom, attributable to MS; accompanied by an appropriate new neurologic abnormality; lasting at least 24 hours in the absence of fever; and preceded by stability or improvement for at least 30 days. Documentation of an exacerbation implied that the investigator thought there was at least one new MS lesion or enlargement of an old one.
.[28]	not done	duration not done; severity not done	2yrs	143	0.82 (0.90 over 2 yrs)	Poser	...development of new neurologic symptoms, or worsening of pre-existing neurologic symptoms which lasted at least 48 h in a patient who had been neurologically stable or improving for the previous 30 days….objective changes had to be evident on neurologic examination, as defined by a deterioration of 0.5 points on the EDSS or a worsening by at least 1 point on the pyramidal, cerebellar, brainstem or visual FSS. Patients who developed trigeminal neuralgia or paroxysmal dystonia that was sustained for at least 48 h could be considered to have had on-study exacerbations even if they did not fulfill the FSS or EDSS criteria. We excluded exacerbations consisting of sensory symptoms, bladder or bowel dysfunction, alterations in cognitive function or mood, Lhermitte's phenomenon, Uhthoff s phenomenon, fatigue or depression if there was no objective change on neurologic examination.
.[30]	not done	duration not done; severity not done	2yrs	126	0.84	Poser	A relapse was defined as the appearance or reappearance of one or more neurologic abnormalities persisting for at least 48 hours and immediately preceded by a relatively stable or improving neurologic state of at least 30 days. A relapse was confirmed only when the patient's symptoms were accompanied by objective changes on the neurologic examination

							consistent with an increase of at least a half a step on the EDSS, two points on one of the seven functional systems, or one point on two or more of the functional systems. Events associated with fever were excluded. A change in bowel bladder or cognitive function could not be solely responsible for the changes in either the EDSS or the functional system scores.
.[29]	unk	duration not done; severity: Scripps; steroid use; hospitalization; act. of daily living	2yrs	187	2.56 relapses over 2 years	Poser	Relapse, defined by Schumacher and colleagues, required the appearance of a new symptom or worsensing of an old symptom over at least 24h that could be attributed to MS activity and was preceded by stability or improvement for at least 30 days
.[130]	not done	duration not done; severity: Scripps; steroid use; hospitalization	1 yr	187	1.08; SD=1.15	Poser	An exacerbation was defined as the appearance of a new symptom or worsening of an old symptom, attributable to MS, accompanied by an appropriate new neurologic abnormality or focal neurologic dysfunction lasting at least 24 hours in the absence of fever and preceded by stability or improvement for at least 30 days
.[131]	not done	duration not done; severity: steroid use; hospitalization	9 months	120	1.21	Poser	A relapse was defined as the appearance of one or more new neurological symptoms, or the reappearance of one or more previously experienced ones. Neurological deterioration had to last at least 48 hours and be preceded by a relatively stable or improving neurological state in the prior 30 days. An event was counted as a relapse only when the patient's symptoms were accompanied by objective changes in the neurological examination corresponding to an increase of at least 0.5 points on the EDSS, or one grade in the score of two or more Functional Systems (FS), or two grades in one FS. Deterioration associated with fever or infection that can cause transient, secondary impairment of neurological function in MS patients were not considered relapses. Nor was a change in bowel, bladder, or cognitive function alone accepted as a relapse.

.[138]	not done	duration not done; severity: steroid use; hospitalization	14 months; 56 WEEKS	548	0.61	Poser	…appearance of one or more new neurological symptoms or the reappearance of one or more previously experienced symptoms. Neurological deterioration had to last at least 48 h and be preceded by a relatively stable or improving neurological state in the prior 30 days. An event was counted as a relapse only when the patient's symptoms were accompanied by objective changes in the neurological examination corresponding to an increase of at least 0·5 points on the EDSS, or one grade in the score of two or more functional systems or two grades in one functional system. Deterioration associated with fever or infections that can cause transient, secondary impairment of neurological function in patients with multiple sclerosis was not regarded as a relapse. Change in bowel, bladder, or cognitive function alone was not accepted as a relapse. (makes distinction between confirmed and unconfirmed relapse)
.[47]	not done	duration not done; severity not done	at 1 yr	315	at yr1=0.78; at yr2=0.73	mcd2001	Relapses were defined as new or recurrent neurologic symptoms not associated with fever or infection that lasted for at least 24 hours and were accompanied by new neurologic signs found by the examining neurologist.
.[50]	not done	duration not done; severity not done	0.5-1yrs	92	0.77	mcd2001	Confirmed relapse was defined as the occurrence of new symptoms or worsening of previously stable or improving symptoms and signs not associated with fever, lasting more than 24 hours and accompanied by an increase of at least half a point in the EDSS score or 1 point in the score for at least one of the functional systems (excluding the bowel and bladder and mental systems). Neurologic deterioration that was classified by the treating physician as a relapse but that did not fulfill these criteria was documented as an unconfirmed relapse.

.[132]	not done	duration not done; severity not done	36 weeks	102	0.77	mcd2005	A relapse was defined as the appearance of one or more new neurological symptoms or the reappearance of one or more previous symptoms lasting at least 48 h, not accompanied by fever or infection, and preceded by a stable or improving neurological state during the previous 30 days. Patients were instructed to notify the study centre of a potential change in neurological status immediately, and an unscheduled visit was done within 7 days of notification. An event was counted as a relapse only when the patient's symptoms were accompanied by objective changes corresponding to an increase of at least 0·5 points on the EDSS, one grade in two or more functional system scores, or two grades in one functional system score. Isolated changes in bowel, bladder, and cognitive function did not qualify as relapses. The treating neurologist established whether the change in symptoms qualified as an on-study relapse
.[135]	not done	duration not done; severity not done	96 weeks	437	0.33	mcd2001	A relapse was defined as an increase of 2 points in at least one functional system of the EDSS or an increase of 1 point in at least two functional systems (excluding changes in bowel or bladder function or cognition) in the absence of fever, lasting for at least 24 hours and to have been preceded by at least 30 days of clinical stability or improvement.
.[49]	not done	duration not done; severity not done	24 months	418	0.4	mcd2005	To constitute a confirmed relapse, the symptoms must have been accompanied by an increase of at least half a point in the EDSS score, of one point in each of two EDSS functional- system scores, or of two points in one EDSS functional-system score (excluding scores for the bowel–bladder or cerebral functional systems).

.[51]	not done	duration not done; severity not done	108 weeks	363	0.54	mcd2001	A relapse was defined as the appearance of a new clinical sign or symptom, or clinical worsening of a previous sign or symptom that had been stable for at least 30 days and that persisted for a minimum of 24 hours in the absence of fever. Confirmed relapses required an increase of 1 point in each of two EDSS functional-system scores or of 2 points in one EDSS functional-system score (excluding bowel and bladder function and cerebral function) or an increase of 0.5 points in the EDSS score from the previous clinically stable assessment.
.[159]	not done	duration not done; severity not done	12 months with 12 months extension	100	0.9; SD=1.0	mcd2001	not provided
.[133]	not done	duration not done; severity not done	24 months	556	0.39; SE=0.03	mcd2005	A relapse was defined as the appearance of one or more new neurologic abnormalities or the reappearance of one or more previously observed neurologic abnormalities lasting for at least 48 hours and occurring after an improved neurologic state for at least 30 days. An event was counted as a relapse if the patient's symptoms were accompanied by objective neurologic changes as indicated by at least one of the following: an increase of at least 0.5 points in the EDSS score, an increase of one grade in two or more of the seven functional systems that are graded in the EDSS (pyramidal, cerebellar, brain stem, sensory, bowel and bladder, visual, and cerebral), or an increase of two grades in one functional system.
.[53]	not done	duration not done; severity not done	96 weeks	363	0.40 (95%CI; 0.33-0.49)	mcd2005	protocol- defined relapses (new or recurrent neurologic symptoms not associated with fever or infection, lasting at least 24 hours, accompanied by new objective neurologic findings, and separated from the onset of other confirmed relapses by at least 30 days) that were confirmed by the independent neurologic evaluation committee.

.[54]	not done	duration not done; severity not done	96 weeks	408	0.36 (95%CI; 0.30-0.44)	mcd2005/ Lublin1996	Protocol- defined relapses were new or recurrent neurologic symptoms, not associated with fever or infection, that lasted for at least 24 hours and that were accompanied by new objective neurologic findings according to the examining neurologist's evalua- tion. All protocol-defined relapses were evaluated by an independent neurologic evaluation committee
.[136]	not done	duration not done; severity not done	52 weeks	196	0.46 (95%CL; 0.37-0.57)	mcd2005	We defined relapses as new or recurrent neurological symptoms (not associated with fever or infection) lasting 24 h or more, accompanied by new neurological findings at assessment by the examining neurologist. Three members of an independent neurology assessment committee, consisting of multiple sclerosis neurologists who were masked to group assignment, adjudicated whether the protocol definition of relapse was satisfied.
.[137]	not done	duration not done; severity not done	48 weeks	500	0.40 (95%CI; 0.33-0.48)	mcd2005	Relapse was defined as new or recurrent neurological symptoms not associated with fever or infection, lasting for at least 24 h, accompanied by new objective neurological findings confirmed by the independent neurological evaluation committee, and separated from the onset of other confirmed relapses by at least 30 days.
.[52]	not done	duration not done; severity not done	48 to 152 weeks	389	0.50 (95%CI; 0.43 - 0.58)	mcd2005	Relapse was defined as new or worsening clinical signs or symptoms lasting at least 24 h without fever. Protocol-defined relapses constituted an increase of either 1 point in at least two EDSS functional system scores, or 2 points in one EDSS functional system score (excluding bowel and bladder function, and cerebral function), or 0·5 points in total EDSS score from a previous clinically stable assessment.

.[134]	not done	duration not done; severity not done	24 months	450	0.34; SE=0.03	mcd2005	A confirmed relapse was defined as the appearance of one or more new neurological abnormalities, or reappearance of one or more previously observed neurological abnormalities, in the absence of fever, persisting for ≥ 48 h, preceded by > 30 days of a stable or improving condition, and accom- panied by at least one of the following: an increase of at least 0.5 point in EDSS score, an increase of one grade in the score of two of the seven functional systems (FS) on the EDSS, or an increase of two grades in one FS.

Table 3.2a: Longitudinal Cohort/Database Registry Studies in MS

ref id	Author	pub year	enrol.	database; cohort	database	type clinic location	sex ratio (total sample)	min. relapse length
.[26]	Lhermitte	1973	Jan 1965-Oct 1970	Hopital de la Salpetriere	retro./prosp.	Service de Neurologie et Neuropsychologie (Höpital de la Salpêtrière, Paris)	2.2:1	no time
.[40]	Kurtzke	1977	1942-1951	Army and Veterans Admin.	retrospective	Army and Veterans Administration	all men	24 hrs
.[141]	Confavreux	1980	1957-1976	EDMUS-Lyon	retro./prosp.	Lyons Multiple Sclerosis Cohort	1.49:1	24 hrs
.[36]	Weinshenker	1989	1972-1984	London, ON	prosp. (inception); retro./prosp.; retro./prosp.	estimated 90% of MS population in catchment area London, Ontario (teriary care)	1.92:1	24 hrs
.[149]	Goodkin	1989	May 1983-July 1988	Cleveland	database source mostly prospective	MS multidisciplinary clinic (Not stated but most likely Cleveland Ohio)	2.13:1	5 days
.[147]	Jacobs	1999	1996-July 1998	NYSMSC	retro./prosp.	12 MS Clinics in New York State	2.79:1	24 hrs
.[150]	Myhr	2001	1976-1987	Hordaland County-Western Norway	retrospective	Haukeland University Hospital (Western Norway)	1.65:1	24 hrs
.[142]	Waubant	2003	1990-March 2001	EDMUS-Lyon	prospective	Lyons Multiple Sclerosis Cohort	3.17:1 (n=200)	24 hrs
.[144]	Binquet	2006	1990-2003	EDMUS-Burgundy	retro./prosp.	private/public (Burgundy region)	2.74:1	24 hrs
.[139]	Sorensen	2006	1948 (1996-	Danish Registry	prospective	14 countys national registry	2.31:1	24 hrs

			2003)					
.[145]	Trojano	2006	2006	Italian MS Database Network	prospective	25 MS Clinics national registry	2.13:1	24 hrs
.[146]	Trojano	2007	onset to 7 yrs	MSDN-iMED and EDMUS	prospective	MS Centres in Bari and Florence	2.21:1	24 hrs
.[148]	Fromont	2008	database created in 1996 and 2000	EDMUS-Burgundy-Lorraine	database source	Neurology department of Nancy (Burgundy Lorraine region)	3.54:1	24 hrs
.[8]	Ebers	2008	31 RCT unk	SLCMSR database	from RCT placebo arm	Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR) (31 RCT database)	2.85:1 (n=516)	N/A
.[125]	Tremlett	2008	July 1988-July 2003	UBC-database	database source mostly prospective	University of British Columbia (estimated 80% of MS population in BC)	2.68:1	24 hrs
.[143]	Debouverie	2009	unk to Feb 2008	EDMUS-LORSEP	retro./prosp.	MS Centres(MSC) vs NonMS Centres (NMSC); (LORSEP=Lorraine Multiple Sclerosis Regional Network of neurologist)	2.62:1	24 hrs
.[140]	Stuke	2009	2009	German MS Registry	prospective	100 German MS Clinic national registry	2.45	unk

Table 3.2b: Longitudinal Cohort/Database Registry Studies in MS

ref id	tool	duration	severity	data size	ARR period	ARR	diag crit./ relapse crit.	definition
.[26]	other	not reported	not reported	245 (RR:212)	whole; 0-5yr; 0.1-5yr; 6-15yr	0.75; 0.58; 0.30; 0.33	MacAlpine/ MacAlpine	The criteria defining the attacks are those of McAlpine [11]; they exclude the transient exacerbations or temporary worsening of the functional disability when an intercurrent disorder occurs. [delete onset relapse in ARR estimate]
.[40]	DSS	duration measured (in mths) but did not indicate how it was done	severity was measured but did not indicate how it was done (no data)	527 (RR:unk)	5 yrs after onset	reported but not as ARR	Schumacher/ Schumacher	A bout,or period of worsening, or exacerbation, was defined as the period in which neurologic symptoms progressed. Thus the bout ended with the attainment of maximum severity of symptoms, or with the date of neurologic observation if the patient was hospitalized. The disappearance of some symptoms while the remainder progressed did not constitute a separate bout. Fluctuations were ignored insofar as possible, and, in a slowly progressive course, only the onset of new symptoms, or a sudden and definite worsening of old symptoms, was considered adequate evi- dence of a separate bout. No symptom lasting less than 24 hours was counted as a bout.
.[141]	other	stated duration was measured but did not indicate how it was done	stated that severity was measured but did not indicate how it was done	349 RR:147	9 yrs	0.31 0.95 (1.00)	MacAlpine/ MacAlpine	Relapses were characterized by either the rapid appearance of new symptoms or the sudden worsening of old symptoms, lasting longer than twenty-four hours and occurring at least one month after the preceding relapse. Relapses could be further classified into 'pure relapses' with a complete regression of symptoms and 'relapses with sequelae' with lasting symptoms of a chronic disability grading of 2 or more (MacAlpine, 1961). The transient neurological symptoms at the time of hyperthermia were not classified as relapses. Following a relapse there was a complete or partial regression of the symptoms to a stable state which did not alter until the next relapse. [delete onset relapse in ARR estimate]
.[36]	DSS	not reported	not reported	1099 (RR:722) 119; 119; 358; 358	1 yr; 2 yrs; 1 yr; 2 yrs	1.8 (0.10); 0.55 (0.08); 1.57 (0.05); 0.35 (0.04)	Poser (only for prosp. Data)/ Poser	The onset of MS was taken to be the date of occurrence of the first unequivocal symptoms of MS. An attack was defined as acute development of new symptoms or worsening of existing symptoms, the duration of which was greater than 24 h (Poser et al., 1983). The attack frequency could only be determined retrospectively for most patients in the total population so mat this definition could be used only as a working guideline. Generally, attack frequency was recorded prospectively for the 'seen from onset' subgroup.

.[149]	EDSS; Ambulation Index	not reported	not reported	425 (RR:203)	1 yr; 2yr; 3yr	0.65 (0.91); 0.61(0.82); 0.65(0.89)	Poser/ defined	Exacerbation was defined as a perceived worsening of old symptoms or appearance of new symptoms accompanied by a worsening of more than 0.5 points on the EDSS or more than 1.0 points on the AI lasting more than 5 and less than 60 days.
.[147]	EDSS	not reported	not reported	3019 (RR:1657)	not reported	not reported	Poser;Schumacher/ Schumacher	A relapse or exacerbation is defined as the development of neurological symptoms or worsening of preexisting neurological symptoms lasting for at least 24 hours, accompanied by objective changes on neurological examination.
.[150]	EDSS	not reported	not reported	220 (RR:179)	-	0.32 (0.02)	Poser/ Poser	defined as the appearance of new symptoms from the CNS or a worsening of preexisting symptoms lasting for at least 24 h in the absence of fever in a patient who had been neurological stable or improving for the previous 30 days.
.[142]	DSS	not reported	not reported	3177 (RR:200)	1 yr; 2 yrs;	1.7 (1.20); 1.40 (0.90)	Poser/ not stated but cited 834 & 19	no definition of a relapse was stated [But cited the original source of the data ...EDMUS-Lyon Confavreux et al., 1992 which stated Schumacher definition]
.[144]	EDMUS-GS	not reported	not reported	527 (RR:487) 288	2 yrs	median 2 [IQR(1-3)]	Poser/ Schumacher	An MS attack was defined as the occurrence, the recurrence or the worsening of symptoms of neurological dysfunction that lasted more than 24 h and that stabilized or resolved either par tially or completely [7]. Fatigue alone or a transient fever-related worsening of symptoms were not considered as a specific MS attack. Symptoms occurring within a month after the initial symptoms of an MS attack were considered to be a part of the same episode. Indeed, to be considered as distinct, 2 MS attacks had to be separated by at least 1 month.
.[139]	EDSS	not reported	not reported	14441 RR:2393	2 yrs pre tx.	1.3	Poser/ not stated	no definition of a relapse was stated
.[145]	EDSS	not reported	not reported	10078 2090; RR:1888; SP:202	7.4yrs	1.3	not stated but defined in Imed/ defined	A relapse was defined as a new symptom or worsening of an old symptom of MS lasting]/24 h without fever, accompanied by a new neurologic abnormality, and preceded by stability or improvement for]/30 days.
.[146]	EDSS	not reported	not reported	1504 (RR:401)	1 yr	0.6 (0.70)	Poser or McDonald criteria/ not stated but defined in Imed	no definition of a relapse was stated (most likely the same as MS diagnostic criteria of Poser or Mcdonald)

.[148]	DSS	not reported	not reported	2645 (RR:751) 555; 555; 196; 196	1 yr; 2 yrs; 1 yr; 2 yrs	1.6 (0.90); 1.10 (0.50); 1.3 (1.00); 1.00 (0.60)	Poser/ Schumacher	A relapse is defined as the occurrence of new neurological symptoms or the worsening of pre-existing signs, apart from fever, for more than 24 h [17].
.[8]	EDSS	not applicable	not applicable	1344 in the open dataset (RR:516)	2 years	1.50	not applicable/ not applicable	on specific definition-SLCMSR is a repository of data from RCTs
.[125]	EDSS	not reported	not reported	5727 (RR:2477)	5 to 30 yrs	0.10 to 0.33	Poser/ not stated but definition provided	All relapses were confirmed by an MS specialist neurologist, and were defined by new or worsening symptoms lasting more than 24 hours in the absence of fever or infection. Episodes occurring within 30 days of each other were considered to be part of the same relapse. [delete onset relapse in ARR estimate]
.[143]	EDSS	not reported	not reported	3602 (RR:2132) 519 MSC; 1613NMSC	first 5 yrs of MS onset	0.58; 0.56	pre-2002 was Poser; post-2002 was McDonald/ McDonald	A relapse of MS was defined as the occurrence, recurrence, or worsening of symptoms of neurological dysfunction lasting more than 24 h and usually ending with a partial or complete remission. Fatigue alone and transient fever-related worsening of symptoms were not considered as a relapse [18]. Symptoms occurring within 1 month were considered as part of the same relapse.
.[140]	EDSS	not reported	not reported	16554; type unk	unk	unk	McDonald/ not stated	no definition of a relapse was stated

Appendix

Measurements selected for this study: Below are all the health indices available to this study to comprehensively assess MS disability. The procedures in selecting these health indices are described above using the ICF model. A description of each measure is provided below:

Expanded Disability Status Scale (EDSS):

The most common and widely used tool to assess MS patient neurological impairment is the EDSS. It is based on a standard neurological exam of 8 functional systems (FS) [pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual, cerebral plus "other neurological findings attributable to MS"]. The EDSS is scored using a rubric of the FS ordinal scale. Each FS score ranging from 0 to 5 or 6. The EDSS is an ordinal clinical rating scale ranging from 0 (normal neurologic examination) to 10 (death due to MS) in half-point increments. EDSS scores above 4.0 also require an assessment of ambulation. Both intra-rater and inter-rater reliability was high with Kappa=0.65;ICC=0.99 and Kappa=0.70;ICC=0.99 respectively for the EDSS. Intra-rater and inter-rater reliability for the FS were Kappa range from 0.41 to 67; ICC range from 81 to 95 and Kappa range from 0.42 to 0.64; ICC range from 0.67 to 0.92 respectively.[164] Others have found that intra-rater, but not inter-rater reproducibility was adequate. Convergent and discriminant validity was supports.[165] Responsiveness was poor.[164,165]

The RAND 36-item health survey 1.0: The RAND-36 is a widely used generic health related quality of life measure that contains 8 domains (physical function, role physical, body pain, general health, vitality, social functioning, role emotional, and mental health). This self-reported questionnaire consists of 36 items with 2, 3, 5, and 6 point response scales. The domains can be used to generate 2 summary scales, comprising of the Physical Component Summary (PCS) and the Mental Component Summary (MCS). The total scores on the RAND-36 for the subscales and the summary scores have a range from 0 to 100 (with 100 representing the best health). The RAND-36 has excellent reliability in a MS population with internal consistency Cronbach's α range between 0.81 to 0.94. It also demonstrated convergent validity.[286]

EQ-5D: The EQ-5D is a generic health relate quality of life self-reported measure. It is used to assess mobility, self-care, usual activities (role limitations), pain/discomfort, and anxiety / depression. Each dimension has three levels, generating a total of 245 theoretical possible health states. Test-retest reliability was good with an ICC=0.81. Construct validity has been assessed between EQ-5D and SF-6D and HUI Mark III in MS.[287]

Fatigue questionnaire: The fatigue questionnaire contained 21 items using 4 and 5 point response options. Two of the items used time (mins) as the response. The items were selected from fatigue questionnaires commonly used in MS (Functional Assessment of Multiple Sclerosis (FAMS)[288] Cronbach's α=0.90, Modified Fatigue Impact Scale (MFIS)[289] Cronbach's α=0.93 and Multidimensional Fatigue Inventory (MFI)[290] Cronbach's α=0.84) by conducting a focus group of the neurologist involved in this study. This composite fatigue questionnaire as a whole has not been assessed for reliability or validity at this time.

HADS: The Hospital Anxiety and Depression Scale (HADS) is a self-reported questionnaire used to assess the presence of anxiety and depression. It contained 14 items, 7 for anxiety and 7 for depression both with score ranging from 0 to 21. Test–retest reliability in two MS

population were α=0.81 or 0.84 for anxiety and α=0.74 or 0.83 for depression. Convergent and discriminant construct validity was tested in a MS population with expected patterns.[291]

Illness intrusiveness: The degree to which the patient perceived certain disease-related factors as disruptive to their lifestyle, activities, and interests was assessed with the Illness Intrusiveness Ratings Scale. This self-reported questionnaire assessed disease-related factors that interferes with their lifestyle, activities, and participation. The 13 items assess the domains related to quality of life, including health, diet, activities, and relationships. It uses a 7-point likert scale responses options ranging from 1 (not very much) to 7 (very much). Total score ranges from 13 to 91. It had excellent internal consistency with a Cronbach's α=0.90 and a test-retest score ranging from 0.80 to 0.85. Good construct validity was obtained across chronic illness population.[292] Five additional items were added to the existing questionnaire.

Disability of the Arm Shoulder and Hand (DASH) questionnaire: This is a 30 item, self-report rating scale used to assess upper limb physical function and symptoms in orthopaedic and neurologic disorders. It uses a 5-point likert type response options providing a transformed total raw scores out of 0 (no disability) to 100 (severe disability). Its psychometric properties has recently been tested in MS patients with high Cronbach's α=0.98. Convergent validity was tested against ABLIHAND and MSIS-29 $r>0.7$ and Divergent validity was tested against MSWS-12 $r<0.7$.[293]

Perceived Deficit Questionnaire (PDQ): The PDQ is used specifically in MS and is part of the MSQLI (Multiple Sclerosis Quality of Life Inventory) and used to assess cognitive function (on 4 domains: attention, retrospective memory, prospective memory, and planning/organization). It contains a total of 20 items with 5 items per domain. Every item uses a 5-point likert scale. Internal consistency Cronbach's α is 0.82, 0.86, 0.74, and 0.85 for the respective domains.[294] Construct validity should that the PDQ correlated moderately strongly with the BDI.[295]

Symptom checklist: This straightforward 18 item list of common symptoms in MS with a binary (yes/no) response option. This checklist was not used for evaluation.

Preference Based Multiple Sclerosis Index-V1 (PBMSI-V1): The PBMSI-V1 was developed for use in stroke and called the "Preference Based Stroke Index". This questionnaire was originally developed as a stroke-specific health index that would take into account the person's preferences for stroke relevant health states. It contained 11 items, 10 items used a 3-point response options with the remaining item having a 4-point option. Internal consistency in a stroke population had a Cronbach's α=0.84.[296]

9-Hole Peg Test (9-HPT): The 9-hole peg test (9-HPT) is used to assess upper limb function and motor speed. It is a timed task that requires the person to put nine pegs one at a time into a pegboard in any order and then removed them one at a time as quickly as possible. Scores used for the analyses are averaged over 2 trials for each dominant and non-dominant hand. Inter-rater reliability in a MS population was ICC=0.93 and intra-rater reliability was ICC=0.96.[297] Concurrent validity of the 9-HPT was tested with TEMPA in MS patients.[298]

Pace Auditory Serial Addition Test (PASAT-3"): Levels of cognitive functioning was assessed with the 3 sec inter-stimulus interval version of the Paced Auditory Serial Addition Test (PASAT). It is a serial addition task that assesses the rate of information

processing and sustained attention and working memory. The PASAT included 60 trials the raw score of 60 trials is recorded and transformed to percentages.[292,299] There was good inter-rater reliability (ICC=0.96) and intra-rater reliability (ICC=0.93). Construct validity was demonstrated in RRMS and SPMS patients.[299]

6 Minute walk test (6MWT): The 6MWT was been used to assess ambulation in many health conditions and has recently been validated as a measure of ambulatory capacity in MS.[300] The distance walked during each minute and total distance in 6 minutes were recorded to the closest meter. Reliability was excellent with inter-rater ICC=0.95 and intra-rater ICC=0.91. Construct validity was assessed. The 6MWT correlated well with subjective measures of general and physical fatigue, physical health status, perceived walking ability, the EDSS and MSFC.[300]

Strength: Strength was assessed with 4 standard tests (grip strength, partial curl-ups, push-ups, and vertical jump) from The Canadian Physical Activity Fitness and Lifestyle Appraisal (CPAFLA). Grip strength for each hand was measured 3 times with results expressed in kg. Test-retest reliability was excellent with ICC=0.90 for the left hand and ICC=0.94 for the right hand. Grip strength has been tested for predictive and construct validity has been tested.[301] Push-ups are used to measure muscle endurance. The patient is asked to perform as many push-ups until fatigued with no time limit. The number of push-up was recorded. Partial curl-ups also test muscle endurance. Patients are asked to perform as many partial curl-ups at a rate of 25/min (set by a metronome) for a maximum of 1 minute. Vertical jump was used to measured lower extremity power. The vertical jump was recorded in cm and repeated in 3 trials.[302]

Equi-scale (Balance): The Equi-scale is used to assess balance and is derived from the Performance Oriented Balance Scale and the Berg Balance Scale. It was Rasch modeled using a MS population. It contains 10 items listing in order by difficulty with question 1 as the easiest. Each item has a 3-point response option. The total score is 20. The items met the model requirements of unidimensionality and reliability.[303]

Gait speed: Gait speed was used to assess walking ability. Gait speed was timed for 5 metres, and acceleration and deceleration distances were each assessed at 2 metres. There is considerable variation in testing procedures but all have demonstrated high (>0.90) test-retest and inter-rater reliabilities.[304] Gait speed construct validity has been confirmed with the Rivermead Mobility Index in MS.[305]

Modified Ashworth Scale (Spasticity): The Modified Ashworth Scale is often used to score muscle spasticity by a trained physiotherapist. Upper and lower limb muscle groups were assessed on this 6-point scale. Clonis was assessed using a 4-point scale. In a MS population, inter-rater reliability range from a Kendall's tau ranged from 0.239 to 0.857 when assessing different muscle group. The Kendall's tau for the total score was 0.71.[306] Unfortunately, when tested on a MS population, there was some evidence of insufficient validity (criterion) to be used as a six-point ordinal level measure of spasticity.[307]

VO_2 max: Exercise capacity was tested by VO_2max using a graded cycle ergometer test.[308] Heart rate was recorded every minute. Each person started with a minimal workload of 10W with incremental increase of 10W per minute. The instruction for pedal rate was to maintain 60rpm. The test was terminated when pedal rate was below 45rpm. VO_2max was recorded as ml/kg/min. Reliability was has been estimated to be r=0.98 in a sample of 20 RR/SP MS patients.[309]