

Imperial College London
Department of Computing

IMPERIAL

**Exploring the Feasibility and Validity
of Online Assessment Tools for
Motor and Cognitive Impairment
in Multiple Sclerosis**

Annalaura Lerede

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London
February 2024, London, United Kingdom

Statement of Originality

I declare that the work presented in this thesis is my own, except where otherwise acknowledged. Parts of this research have been conducted in collaboration with other researchers. This is clearly indicated where relevant and here summarised.

Chapter 3 introduces an analysis pipeline for online-collected cognitive data which was developed in collaboration with Valentina Giunchiglia:

- **Pipeline design:** Annalaura Lerede and Valentina Giunchiglia
- **Pipeline implementation:** Annalaura Lerede and Valentina Giunchiglia
- **Pipeline application to control/patient data:** Annalaura Lerede

Chapter 4 is based on 11 years of data collected by the UK MS Register. These are the respective contributions for this chapter:

- **Data collection:** UK MS Register group
- **Data analysis:** Annalaura Lerede
- **Data visualisation:** Annalaura Lerede
- **Results interpretation:** Annalaura Lerede, Adam Hampshire and Richard Nicholas

Chapter 5 and Chapter 6 are based on data collected in the context of this thesis. These are the respective contributions for these chapters:

- **Software:** Adam Hampshire, Pete Hellyer, William Trender and H2 Cognitive Designs
- **Study design:** Annalaura Lerede, Adam Hampshire and Richard Nicholas
- **Implementation on the UK MS Register website:** Annalaura Lerede, UK MS Register group and H2 Cognitive Designs
- **Patient involvement, recruiting and support:** Annalaura Lerede, Rod Middleton, Katherine Tuite-Dalton and UK MS Register group
- **Data analysis:** Annalaura Lerede
- **Data visualisation:** Annalaura Lerede
- **Results interpretation:** Annalaura Lerede, Adam Hampshire and Richard Nicholas

Finally, Chapter 7 is based on data already introduced in the previous experimental chapters. The same contributions and collaborations apply for the respective data collection, analysis, visualisation and interpretation.

The work in Chapter 4 has been presented at Brain Conference 2022 and published in Brain Communications:

“Annalaura Lerede, Jeff Rodgers, Rod M Middleton, Adam Hampshire, Richard Nicholas, and The UK MS Register Research Group. (Aug. 2023). Patient-reported outcomes in multiple sclerosis: a prospective registry cohort study. *Brain Communications* 5(4):fcad199”

The work in Chapter 5 has been presented at Cognitive Neuroscience Society (CNS) Annual Meeting 2022, Association of British Neurologists (ABN) Annual Meeting 2022 and European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) Congress 2022.

The work in Chapter 6 has been presented at ECTRIMS Congress 2023.

The work in Chapter 5, 6 and 7 is being prepared for submission to a journal.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Multiple Sclerosis (MS) can result in both motor and cognitive impairment. This thesis is dedicated to the development and validation of online assessment tools to measure these impairments.

The first part introduces a novel approach for analysing 11 years of patient-reported outcome (PRO) data from 15,976 users of the UK MS Register. Using this approach, robust associations were shown between physical disability and disease subtype and duration. Results indicated progressive deterioration over time for most subtypes and significant differences among subtypes, underscoring the potential of PROs for capturing physical disability progression.

The second part focuses on developing and validating an online cognitive assessment tool for people with MS (pwMS) and comprises two studies. The first study evaluated 22 online cognitive tasks on a cohort of 3,066 pwMS to demonstrate the feasibility of online cognitive assessment in pwMS and form a concise battery for measuring cognitive deficits in this population. The second study validated this battery on an overlapping cohort of 2,696 pwMS. The high participation and completion rates demonstrated the feasibility of the tool in pwMS. The tasks selected for the optimal battery exhibited consistent sensitivity to MS-related deficits across timepoints and aligned with cognitive domains reported to be affected in the literature. Additionally, data collected with the battery enabled confirming known associations with disease subtype and duration with greater statistical power.

In the final part, PROs and cognitive performance measures were combined to derive symptoms-driven subtypes. Clustering analysis identified four clusters with diverse combinations and severities of motor and cognitive symptoms, distinct from traditional clinical phenotypes.

Altogether these findings highlight the value of online assessment tools in MS and provide a battery of online cognitive tasks for longitudinal monitoring of MS-related deficits. The integration of these tools within clinical registries holds tremendous potential to advance our understanding of MS.

List of Acronyms

BN Benign

DfE Deviation from Expected

EDSS Expanded Disability Status Scale

EQ5D EuroQoL - 5 Dimensions

FSS Fatigue Severity Scale

GBIT Great British Intelligence Test

HADS Hospital Anxiety and Depression Scale

MS Multiple Sclerosis

MSIS-29 Multiple Sclerosis Impact Scale

MSWS-12 Multiple Sclerosis Walking Scale

PP Primary Progressive

PRO Patient-reported outcome

pwMS people with Multiple Sclerosis

RR Relapsing-Remitting

SP Secondary Progressive

UKMSR United Kingdom Multiple Sclerosis Register

Contents

Statement of Originality	2
Copyright Declaration	4
Abstract	5
List of Acronyms	6
List of Figures	15
List of Tables	20
1 Introduction	23
1.1 Aims and objectives	24
1.2 Hypotheses	24
1.3 Chapters summary	25
2 Literature review	26
2.1 Definition of MS	27
2.2 Epidemiology of MS	27
2.3 Economic, health and social burden of MS	28
2.4 Clinical disease course in MS	29
2.5 Personalised medicine in MS	32
2.6 MS Registries	32
2.7 Patient-reported outcomes	34
2.8 Cognition in MS	36
2.8.1 Impact of cognitive impairment in people with MS	36
2.8.2 Prevalence of cognitive impairment in MS	36

2.8.3	Cognitive domains affected in MS	38
2.9	Traditional in-person cognitive assessments in MS	39
2.10	Online cognitive assessments	40
2.10.1	Benefits	40
2.10.2	Challenges	41
2.10.3	Cognitron	43
2.11	Online cognitive assessments in MS	44
3	General Methods	46
3.1	Patient-reported outcomes	47
3.1.1	Multiple Sclerosis Impact Scale - MSIS-29	47
3.1.2	Multiple Sclerosis Walking Scale - MSWS-12	47
3.1.3	Hospital Anxiety and Depression Scale - HADS	47
3.1.4	Fatigue Severity Scale - FSS	47
3.1.5	EuroQoL-5 Dimensions - EQ5D	48
3.1.6	Web Expanded Disability Status Scale - webEDSS	48
3.2	Online cognitive assessment using Cognitron	48
3.2.1	The online cognitive tasks	48
3.2.2	Information collected for each task	53
3.2.3	Normative data	54
3.3	Pre-processing pipeline for online-collected large-scale cognitive data	54
3.3.1	Pre-processing pipeline overview	55
3.3.2	Extraction of summary performance metrics	55
3.3.2.1	Summary accuracy metrics	56
3.3.2.2	Summary latency metrics	56
3.3.2.3	Additional summary performance metrics	56
3.3.3	Data quality checks	57
3.3.4	Distributions' tails cleaning algorithms	58
3.3.4.1	Right-tail cleaning of reaction time distribution	58
3.3.4.2	Left-tail cleaning of reaction time distribution	59
3.3.4.3	Left-tail cleaning of accuracy distribution	60

3.3.4.4	Adaptation to patient data: where to evaluate the thresholds	61
3.3.5	Cleaning based on sociodemographics	62
3.3.6	Conclusions	62
4	Validity and utility of patient-reported outcomes to assess physical disability in Multiple Sclerosis	64
4.1	Introduction	65
4.2	Methods	66
4.2.1	Data collection	66
4.2.2	Patient-reported outcomes	66
4.2.3	Data curation and handling sparsity	67
4.2.4	Assignment of MS subtype labels at patient-reported outcome timepoints . . .	67
4.2.5	Patient-reported outcome trajectories with disease duration	70
4.2.6	Statistical analysis	70
4.3	Results	72
4.3.1	Descriptive analysis of user engagement with the UKMSR	72
4.3.2	Validity of the patient-reported outcomes	72
4.3.2.1	Patient-reported outcomes evolve differently across subtypes	72
4.3.2.2	Disease duration influences patient-reported outcomes for all subtypes but Benign	73
4.3.2.3	Patient-reported outcomes are sensitive to subtype within all disease time bins	74
4.3.2.4	Pairwise comparisons across subtypes	77
4.3.3	Deriving new insights using the patient-reported outcomes	79
4.4	Discussion	81
4.4.1	Conclusions	83
5	Feasibility and sensitivity of online cognitive assessment in people with Multiple Sclerosis	84
5.1	Introduction	85
5.2	Methods	86
5.2.1	Study design	86
5.2.2	Procedures	87

5.2.2.1	Online cognitive assessment: the software and the tasks	87
5.2.2.2	Sampling of online cognitive tasks	87
5.2.2.3	Questionnaires	87
5.2.2.4	Patient-reported outcomes	88
5.2.2.5	Normative data	88
5.2.3	Statistical analysis	90
5.2.3.1	Summary performance measures and pre-processing	91
5.2.3.2	Task sensitivity to MS	91
5.2.3.3	Correlations with patient-reported outcomes	91
5.2.3.4	Effect of device variability	92
5.2.3.5	Factor analysis	92
5.2.3.6	Task selection for the concise optimal battery	92
5.3	Results	93
5.3.1	Feasibility of online cognitive assessment in people with MS	93
5.3.1.1	Participation rates	93
5.3.1.2	Completion rates	94
5.3.1.3	Dropout rates	94
5.3.1.4	Engagement of active users	94
5.3.2	User feedback on testing time	97
5.3.3	Participants	98
5.3.4	Sensitivity to MS of the evaluated online cognitive tasks	100
5.3.4.1	Performance accuracy scores	100
5.3.4.2	Performance latency scores	100
5.3.4.3	Performance costs in accuracy and latency	100
5.3.5	Correlations with patient-reported outcomes	102
5.3.5.1	Correlations between accuracy scores and patient-reported outcomes .	102
5.3.5.2	Correlations between latency scores and patient-reported outcomes .	102
5.3.5.3	Correlations between costs and patient-reported outcomes	103
5.3.6	Selection of online cognitive tasks for the optimal battery	104
5.3.6.1	Task sensitivity to MS	104
5.3.6.2	Task sensitivity to device	105

5.3.6.3	Factor analysis	105
5.3.6.4	Task selection	105
5.4	Discussion	108
5.4.1	Summary of findings	108
5.4.2	Evidence in favour of the feasibility of online cognitive assessment in pwMS . .	108
5.4.3	Benefits of online cognitive assessment	109
5.4.4	Validity of the evaluated assessment and cognitive domains affected in MS . .	110
5.4.5	Comparison with traditional neuropsychological batteries	111
5.4.6	Identification of the optimal battery of online cognitive tasks	111
5.4.7	The confounding effect of visuo-motor impairment on cognitive performance measures	112
5.4.8	Conclusions	113

6 Optimal battery for online cognitive assessment in MS: validation, application and insights on disease-related associations 114

6.1	Introduction	115
6.2	Methods	116
6.2.1	Study design	116
6.2.2	Procedures	117
6.2.2.1	Online cognitive assessment: the software and the tasks	117
6.2.2.2	Questionnaires	119
6.2.2.3	Patient-reported outcomes	119
6.2.2.4	Normative data	120
6.2.2.5	Timepoint 1 data	121
6.2.3	Statistical analysis	122
6.2.3.1	Summary performance measures and pre-processing	123
6.2.3.2	Impact of task modifications on performance	123
6.2.3.3	Replication of timepoint 1 results	123
6.2.3.4	Definition of global cognitive impairment	124
6.2.3.5	Definition of task-level cognitive impairment	125
6.2.3.6	Sensitivity to MS and to early-stage MS	125
6.2.3.7	Sensitivity to disease subtype and duration	125

6.2.3.8	Factor analysis	126
6.2.3.9	Longitudinal analysis	126
6.3	Results	127
6.3.1	User engagement	127
6.3.1.1	Participation and return rates	127
6.3.1.2	Completion rates	127
6.3.1.3	Engagement of active users	128
6.3.2	Participants	131
6.3.3	Analyses on independent sample	133
6.3.3.1	Task orientation does not impact performance	133
6.3.3.2	Task sensitivity to MS replicates across timepoints	134
6.3.4	Analyses on complete sample	135
6.3.4.1	Sensitivity to MS of tasks in the optimal battery	135
6.3.4.2	Prevalence rates of cognitive impairment in pwMS: global-level	136
6.3.4.3	Prevalence rates of cognitive impairment in pwMS: task-level	138
6.3.4.4	Sensitivity to disease subtype and duration	140
6.3.4.5	Sensitivity to early-stage MS	142
6.3.4.6	Associations of underlying cognitive domains with disease subtype and duration	143
6.3.5	Analyses on overlapping sample	145
6.3.5.1	Longitudinal analysis	145
6.4	Discussion	146
6.4.1	Summary of findings	146
6.4.2	High participation and completion rates	146
6.4.3	Tested sample was representative of broader MS population	147
6.4.4	People with MS who participated in TP1 returned for TP2	147
6.4.5	Validation of the content of the optimal battery	148
6.4.6	Additions to the optimal battery	148
6.4.7	Design of the optimal battery	149
6.4.8	Observed prevalence of cognitive impairment further validates the optimal battery	149

6.4.9	Data from the optimal battery can be used to evaluate the impact of disease subtype and duration on cognitive function	149
6.4.10	Limitations and future directions	150
6.4.11	Conclusions	151

7 Unsupervised clustering of people with MS based on digital markers of motor and cognitive impairment 152

7.1	Introduction	153
7.2	Methods	154
7.2.1	Study design	154
7.2.2	Participants	154
7.2.3	Procedures	154
7.2.3.1	Online cognitive data	154
7.2.3.2	Sociodemographic data and clinical factors	155
7.2.3.3	Patient-reported outcome data	155
7.2.4	Statistical analysis	156
7.2.4.1	Pre-processing	156
7.2.4.2	Agglomerative hierarchical clustering	156
7.2.4.3	Associations of cluster label with clustering and clinical variables . . .	157
7.2.4.4	Prevalence of cognitive impairment across clusters	157
7.2.4.5	Associations of cluster label with sociodemographic variables and clinical factors	157
7.3	Results	158
7.3.1	Participants	158
7.3.2	Four clusters were selected as the optimal number of clusters	159
7.3.3	Labelling of clusters examining patterns in clustering variables	159
7.3.3.1	Patterns in motor function across clusters	160
7.3.3.2	Patterns in cognitive function across clusters	160
7.3.4	Distribution of sociodemographics and clinical factors across clusters	164
7.3.5	Characterisation of clusters based on unseen clinical variables	165
7.3.5.1	Patterns in mental health across clusters	165
7.3.5.2	Patterns in self-reported fatigue across clusters	166

7.3.5.3	Patterns in quality of life and disability severity across clusters	166
7.4	Discussion	168
7.4.1	Summary of findings	168
7.4.2	Independence of motor and cognitive impairment	168
7.4.3	Insights in the No Motor + Mild Cognitive cluster	169
7.4.4	Insights in the other clusters	169
7.4.5	Do clusters represent different stages of MS?	170
7.4.6	Clinical variables are associated with motor function	170
7.4.7	Limitations and future directions	171
7.4.8	Conclusions	171
8	Discussion	172
8.1	Summary of findings	173
8.1.1	Chapter 4	173
8.1.2	Chapter 5	173
8.1.3	Chapter 6	174
8.1.4	Chapter 7	174
8.2	Interpretation	175
8.3	Limitations	177
8.4	Future directions	178
8.5	Concluding remarks	180

List of Figures

2.1	Atlas of MS. Atlas showing geographic variations in MS estimated prevalence rates (prevalence per 100,000 people) across countries as of 2020. Prevalence rates are higher in countries with higher degrees of investigation and located at higher latitudes (Multiple Sclerosis International Federation 2020).	28
2.2	Clinical disease course in MS. Visual representation of a possible RRMS (A), SPMS (B), and PPMS (C) disease course adapted from Klineova et al. 2018.	31
3.1	Latency distribution including non-compliant users. Illustration of typical latency distribution shape in case of non-compliant behaviours.	59
4.1	Schematic of disease subtype labels assignment. Illustration of how MS subtype labels at questionnaire completion dates were derived from the time course of recorded labels for each user and of how the records were subsequently grouped based on these labels.	68
4.2	Flowchart illustrating how the studied subgroups were derived from the overall UKMSR population. For each subgroup the number of unique users (n) and the total number of MSIS-29 motor (m) and MSWS-12 (w) questionnaires completed by the users are reported. The flowchart also reports how many of the MSIS-29 motor questionnaires completed were, respectively, version 1 (mv1) and version 2 (mv2) of the questionnaire. The minimal requirements to be included in the study consist of having completed at least one PRO, having reported a legitimate year of birth and of symptoms onset and having disease subtype label at completion date known or derivable.	69
4.3	Schematic explaining how permutation testing was carried out.	71
4.4	Analysis of user engagement with the UKMSR. (A) Frequency of users who completed a specific number of times MSIS-29 motor and MSWS-12. 62% of the users for MSIS-29 motor and 64% for MSWS-12 completed between 1 and 4 PROs. (B) Distribution of the time intervals elapsed between consecutive completions of MSIS-29 motor ($time_{mean} = 7months$, $time_{median} = 5months$, $time_{max} = 131months$) and MSWS-12 ($time_{mean} = 7months$, $time_{median} = 4months$, $time_{max} = 113months$) across all users. (C) Number of times MSIS-29 motor and MSWS-12 were completed at the different disease time bins.	73

4.5	Mean score trajectories with 95% CI over disease duration for the different disease subtypes for MSIS-29 motor and MSWS-12. Performing permutational multivariate analysis of variance ($n = 10000$ permutations), a robust main effect of disease subtype was found on the scores (MSIS-29 motor: $F_{real} = 2698$, $p < 0.0001$; MSWS-12: $F_{real} = 3007$, $p < 0.0001$) as well as an interaction of disease subtype and duration (MSIS-29 motor: $F_{real} = 3.9$, $p < 0.0001$; MSWS-12: $F_{real} = 3.5$, $p < 0.0001$).	74
4.6	MSIS-29 motor (top panel) and MSWS-12 (bottom panel) probability density functions stratified by subtype for the different disease time bins considered. The probability density curve for the BN subgroup between 0 and 5 years from onset is dotted as this label was given retrospectively and was based on a maximum of 11-year follow-up. Performing permutational multivariate analysis of variance ($n=10000$ permutations), a robust main effect of disease subtype was found on the scores at disease onset and throughout all time bins considered (MSIS-29 motor: all $p < 0.0001$; MSWS-12: all $p < 0.0001$).	76
4.7	Mean score trajectories with 95% CI over disease duration for transitioning and non-transitioning individuals in the UKMSR cohort taken in their relapsing phase for MSIS-29 motor (left panel) and MSWS-12 (right panel). Being RRTrans was associated with higher scores on average in both MSIS-29 motor ($t_{real} = -23.5$, $p < 0.0001$) and MSWS-12 ($t_{real} = -26.2$, $p < 0.0001$) when performing permutation testing with t statistic and $n=10000$ permutations. Mean score trajectories with 95% CI for BN, PPMS and SPMS individuals in their secondary progressive phase are also displayed here in transparency for comparison and were not included in the statistical tests.	80
5.1	Diagram illustrating the study design. Participants were recruited via email from the UK MS Register. They could perform the assessment on any device of their choice. The online cognitive assessment consisted of 12 tasks: 3 fixed tasks and 9 tasks random sampled from a superset of 19 tasks. Overall, 22 tasks were evaluated. Participants were also asked to complete a sociodemographic questionnaire at the beginning, and a feedback questionnaire at the end of the assessment. The resulting cognitive performance scores and sociodemographic variables were linked to data already available on the Register (e.g. patient-reported outcomes) for further analysis.	88
5.2	Engagement of people with MS with online cognitive assessment. (A) Percent of participants who completed, overall, in one session, in one day, or in more days, and who did not complete. (B) Percent of participants who also completed the different PROs on the UK MS Register website at the data collection windows either prior to or subsequent to the cognitive data collection. (C) Percent of participants who dropped out after completing a certain number of tasks (above) or a particular task (below).	95
5.3	Feedback on online cognitive assessment duration. Feedback on testing time and how it varied across three groups: everyone who completed, participants who completed in one session and participants who completed in more sessions.	98

5.4	Sensitivity to MS across tasks. Median DfE score for accuracy, latency (A), and extra performance metrics (B) across tasks with 95% confidence intervals. P values are obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.	101
5.5	Correlations between summary performance measures and patient-reported outcomes. Kendall's rank correlations of accuracy (A) and latency (B) DfE scores across tasks with patient-reported outcomes. Only significant correlations ($p \leq 0.05$) are reported to increase readability.	103
5.6	Task selection for the optimal battery. (A) Tasks selected for the optimal MS reduced battery with underlying cognitive domain they correspond to. (B) Scree plot of eigenvalues derived from factor analysis of most representative performance metrics across tasks. (C) Loadings of the tasks onto the different rotated factors, which represent underlying cognitive domains. (D) The device effect size for accuracy and latency measures across tasks, measured in terms of eta squared. Values ≤ 0.06 are small and values ≤ 0.01 are negligible.	107
6.1	Diagram illustrating the total number of people with MS having participated in timepoint 1 and 2, how they were grouped, how many users were in each group, and which combination of groups was used for each analysis presented in this chapter. Numbers refer to pwMS retained for analysis after pre-processing procedures were applied and represent the total number of users in each group. The number of users available for each task varies across tasks. Overall users from timepoint 2 were initially grouped based on whether they had participated or not in timepoint 1 (and vice versa), resulting in one independent and one overlapping sample. For the overlapping samples, the number of users who had valid scores at each timepoint, as well as the number of users who had valid scores at both timepoints, were reported. The overlapping sample from timepoint 2 was further divided for each task based on whether users had received or not that specific task in their random sampled battery during timepoint 1. Moreover, two further groups were derived from the overlapping samples from timepoint 1 and 2, including users who had filled in the PROs during the collection windows prior to both cognitive assessments (in Autumn 2021 and 2022). The number of users in these groups varied based on the specific task-PRO combination considered.	118
6.2	Engagement of people with MS with online cognitive assessment. (A) Percent of participants from TP2 who completed, overall, in one session, in one day, or in more days, and who did not complete. (B) Percent of participants from TP1 who did vs did not return and percent of participants from TP2 who did vs did not do TP1. (C) Percent of participants from TP2 who completed the different PROs on the UK MS Register website at the data collection window prior to the cognitive data collection. (D) Percent of participants who dropped out after completing a particular task. . . .	129

6.3	Impact of task modifications on performance. Standardised regression coefficients for task modifications resulting from multiple linear regression models predicting performance scores across tasks from sociodemographic variables, device label and task modifications.	133
6.4	Replication of task sensitivity to MS across timepoints. Scatter plots of sensitivity to MS in terms of accuracy (left) and latency (right) across tasks for timepoint 1 (x-axis) vs timepoint 2 (y-axis). Sensitivity to MS is defined as the median DfE score for each task. Overall correlations between sensitivities at timepoint 1 vs at timepoint 2 for accuracy and latency are reported on the plots.	134
6.5	Sensitivity to MS across tasks. Median DfE score for accuracy, latency, and extra performance metrics across tasks with 95% confidence intervals. P values were obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.	136
6.6	Prevalence of cognitive impairment in people with MS and controls across tasks. (A) Violin plots of DfE scores across tasks for pwMS (blue) and controls (grey) with liberal, fair, and conservative thresholds for impairment indicated on the plot. The thresholds are positive for tasks for which latency was chosen as primary outcome measure, and negative for tasks for which accuracy was chosen. (B) Barplots of the percent of users impaired for pwMS and controls, as well as (C) of the difference in percent of users impaired between the two groups, across tasks and using the different thresholds.	139
6.7	Associations of performance across tasks with disease subtype and duration. (A) Mean DfE scores with 95% confidence intervals for the primary outcome metrics for all tasks plotted across subtypes (above) and disease durations (below). 2-way permutational ANOVA was applied separately for each task to test whether the DfE scores varied significantly with disease subtype, duration, and their interaction. Resulting p values are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. (B) P-values resulting from post-hoc pairwise comparisons of DfE scores across subtypes for tasks that showed a significant association with subtype. . .	141
6.8	Sensitivity to early-stage MS across tasks. Median DfE scores with 95% confidence intervals for primary outcome metrics across tasks for people with early-stage MS (between 0 and 5 years from disease onset). P values were obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.	142

6.9	Underlying cognitive domains and their associations with disease subtype and duration. (A) Loadings of the tasks onto the different rotated factors, which represent underlying cognitive domains. (B) Mean scores for factors across subtypes and (C) disease durations. 2-way ANOVA was applied separately for each factor to test whether the scores on that factor varied significantly with disease subtype and duration. Resulting p values are reported in the figures as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. (D) Scree plot of eigenvalues derived from factor analysis of most representative outcome metrics across tasks. (E) Standardised regression coefficients across disease subtypes resulting from multiple linear regression models predicting factors' scores from them.	144
6.10	Change in cognitive function over 1-year in people with MS. Mean standardised performance deltas in terms of accuracy (left) and latency (right) scores between timepoint 1 and timepoint 2 with 95% confidence intervals. P values were obtained performing t-tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.	146
7.1	Optimal number of clusters. Agglomerative hierarchical clustering dendrogram generated using the Ward linkage method. The dendrogram provides a visual representation of how pwMS were grouped into clusters based on their similarities in the selected motor and cognitive features. Only the nodes with up to 4 merges from the final merge are shown for clarity. The x-axis represents the number of pwMS in each leaf node. The y-axis represents the distance between clusters. The vertical lines represent the fusion of clusters, with the height of each line indicating the dissimilarity at which the fusion occurred.	159
7.2	Patterns in clustering variables across clusters. Median scores with 95% confidence intervals across clusters for clustering variables (motor and cognitive features) as well as factors derived factor-analysing the cognitive features.	162
7.3	Pairwise comparisons heatmaps. P-values resulting from post-hoc pairwise comparisons of scores across clusters for motor features, cognitive features, and factors. A series of Dunn's tests was performed to obtain the p-values and the Bonferroni method was applied to correct for multiple comparisons.	163
7.4	Patterns in sociodemographics and disease factors across clusters. Gender, education level, and disease subtype composition of each cluster. Bars represent the percentage of pwMS in the different categories.	165
7.5	Patterns in clinical variables across clusters. (A) Median scores with 95% confidence intervals across clusters for clinical variables. (B) Prevalence of mental health problems in each cluster. Bars represent the percentage of pwMS in the different categories. (C) P-values resulting from post-hoc pairwise comparisons of scores across clusters for clinical variables. A series of Dunn's tests was performed to obtain the p-values and the Bonferroni method was applied to correct for multiple comparisons. .	167

7.6	Pairwise Pearson's correlations among the G factor, the MSIS-29 motor, the MSWS-12, the HADS-depression, the HADS-anxiety, the FSS, the EQ5D, and the webEDSS.	168
-----	--	-----

List of Tables

3.1	Measures collected at the time of task performance, including a description of what they are and of the way in which they are used.	53
4.1	Characterisation of PRO responses of the UKMSR population with the minimal requirements at all timepoints. MSIS-29 / MSWS-12 record counts are reported.	71
4.2	Mean, standard deviation, and skew of the distributions of MSIS-29 motor and MSWS-12 across subtypes and disease time bins.	75
4.3	Real F statistics and p values obtained randomly reassigning the disease subtype labels of the MSIS-29 motor and MSWS-12 separately at each time bin.	77
4.4	Real t statistics and p values obtained from pairwise comparisons of the MSIS-29 motor and MSWS-12 of subtypes overall and at each disease time bin.	78
4.5	Real t statistics and p values obtained from pairwise comparisons of the MSIS-29 motor and MSWS-12 of relapsing individuals who did vs did not transition during the follow-up, overall and at each time bin.	80
5.1	Number of observations available for each task and number of controls who completed each task, before and after selection criteria and pre-processing procedures were applied.	89
5.2	Sociodemographic characteristics of participants from the GBIT study retained as controls for this study.	90
5.3	Demographics, clinical phenotype, disease duration, and disease severity across participants who completed, in one or more sessions, participants who did not complete, participants overall and overall UK MS Register population. P-values were obtained using t-tests, ANOVA or chi-squared tests where appropriate.	96
5.4	Number of participants who completed each task, before and after pre-processing procedures were applied.	97
5.5	Sociodemographic characteristics of pwMS from the UK MS Register who participated in the online cognitive assessment and were retained for analysis after pre-processing procedures were applied.	99
5.6	Number of observations available for each task-questionnaire combination, on which the pairwise correlation calculation is based.	104

5.7	List of the evaluated online cognitive tasks with the metric selected as most representative for each. Additionally, the median DfE score for this metric, along with its comparative ranking based on its value are reported. Moreover, for each task it is indicated whether its most representative metric is sensitive to device and the underlying cognitive domain onto which it primarily loads. Selected tasks are highlighted in grey.	106
6.1	Version of each task from which the selected normative data were derived. Number of observations available and number of controls for the selected version of each task, before and after selection criteria and pre-processing procedures were applied.	121
6.2	Sociodemographic characteristics of participants from the GBIT study retained as controls for this study.	122
6.3	Demographics, clinical phenotype, disease duration, and disease severity across: i) participants from TP2 who completed, in one or more sessions, vs who did not complete; ii) participants from TP1 who did vs did not return to take part in TP2; iii) overall participants from TP1 vs TP2. P-values were obtained using t-tests, ANOVA or chi-squared tests where appropriate.	130
6.4	Number of participants who completed each task, both overall and separately for participants who were doing each task for the first time, before and after pre-processing procedures were applied.	131
6.5	Sociodemographic characteristics of pwMS from the UK MS Register who participated in the online cognitive assessment and were retained for analysis after pre-processing procedures were applied.	132
6.6	Median and mean DfE scores for accuracy and latency performance measures across tasks at timepoint 1 (TP1) and at timepoint 2 (TP2).	135
6.7	Prevalence, demographics, clinical phenotype, and disease duration of people with MS classified as having or not cognitive impairment using three criteria from the literature with different stringency. The criteria were based on a combination of different abnormality cut-offs and critical numbers of abnormal parameters. P-values were obtained using t-tests or chi-squared tests where appropriate.	137
7.1	Sociodemographic characteristics of pwMS from the UKMSR who participated in the online cognitive assessment at timepoint 2 and were retained for this study.	158
7.2	Sociodemographics, disease subtype, disease duration, and age at onset for pwMS across clusters. P-values were obtained using t-tests or chi-squared tests where appropriate. .	164

Chapter 1

Introduction

In this chapter, I state the aims, objectives and hypotheses underlying this thesis and provide an outline of the remaining chapters.

1.1 Aims and objectives

This thesis aims to develop and investigate the validity and utility of digital markers of disease-related disability in MS focusing on motor and cognitive impairment. Two patient-reported outcomes already collected as part of the UK MS Register will be investigated as digital markers of motor impairment in MS. For cognitive impairment, an optimal battery of online cognitive tasks specific to MS will be first developed and then validated on a sample of people from the UK MS Register. Performance scores on the selected online cognitive tasks will be investigated as digital markers of cognitive impairment in MS. Finally, this thesis aims to derive symptoms-driven subtypes using only these digital markers of disability.

The main objectives of this thesis are:

1. To provide further evidence of the validity of two patient-reported outcomes, the Multiple Sclerosis Impact Scale – motor component and the Multiple Sclerosis Walking Scale, in measuring motor impairment in people with MS as well as investigate their utility in offering novel insights in the developmental time course of the disease.
2. To assess the feasibility of online cognitive assessment in people with MS.
3. To identify an optimal battery of online cognitive tasks to monitor cognitive function in people with MS.
4. To validate the optimal battery of online cognitive tasks.
5. To showcase the utility of the cognitive data collectable using the optimal battery.
6. To derive symptoms-driven MS subtypes solely based on digital markers of motor and cognitive impairment (i.e., on patient-reported outcomes and cognitive performance measures).

1.2 Hypotheses

The main hypotheses of this thesis are:

1. The Multiple Sclerosis Impact Scale – motor component and the Multiple Sclerosis Walking Scale will be confirmed as valid measures of physical disability in MS and will provide novel insights into the disease when applied longitudinally at a large scale.
2. Online cognitive assessment will be feasible and accessible for people with MS.
3. People with MS will perform significantly worse than matched controls on at least a subset of the evaluated online cognitive tasks.
4. The online cognitive tasks that will show the highest sensitivity to MS will be those that measure cognitive domains known to be affected in the MS literature.
5. Symptoms-driven subtypes will not perfectly align with clinical subtypes given the heterogeneity of symptoms experienced within and across clinical subtypes.

1.3 Chapters summary

In Chapter 2, I cover the background literature on Multiple Sclerosis as well as existing methods to assess disease-related disability remotely.

In Chapter 3, I describe the overarching methods employed throughout my experimental chapters.

In Chapter 4, I analyse 11 years of PRO data from the UKMSR. I develop a novel approach to handle complexities inherent in registry-like data whilst maximising the number of available records retained. I use this approach to investigate differences in the two of the motor-related PROs, the MSIS-29 motor and the MSWS-12, as a function of subtype, disease duration, and future transition to a secondary progressive course.

In Chapter 5, I present an exploratory study done in collaboration with the UKMSR to evaluate the feasibility and sensitivity of online cognitive assessment in pwMS. I use a random-sampling strategy to evaluate the sensitivity to MS of 22 candidate online cognitive tasks from Cognitron on a large cohort of pwMS. I select a subset of these tasks to form an MS-optimised battery deployable on any device.

In Chapter 6, I validate the optimal battery of online cognitive tasks derived in Chapter 4 on a partially overlapping sample of pwMS using a fixed study design. I replicate the results from Chapter 4 on the independent component of the sample. With the available complete data, I evaluate the prevalence of cognitive impairment in the sample of pwMS and investigate its associations with disease-related factors.

In Chapter 7, I present an exploratory study investigating an alternative, data-driven way to cluster pwMS solely based on disability markers obtained using online tools. I combine data collected in collaboration with the UKMSR using PROs and online cognitive assessment. I characterise the resulting clusters investigating patterns in sociodemographic, clustering and unseen clinical variables.

In Chapter 8, I summarise and interpret the main findings of my thesis. I discuss their limitations and potential impact on the field. I conclude by suggesting possible directions for future research on the basis of my work.

Chapter 2

Literature review

In this chapter, I cover the background literature on Multiple Sclerosis focusing on epidemiology, disease burden on society, and possible disease courses. Moreover, I review methods to assess disease-related disability remotely focusing on patient-reported outcomes for physical disability and online cognitive assessments for cognitive impairment.

2.1 Definition of MS

Multiple Sclerosis (MS) is a chronic autoimmune disease of the central nervous system in which the body's immune system attacks the myelin stratus coating the nerves disrupting their ability to conduct electrical impulses within the brain. People with MS (pwMS) experience a broad and diverse range of motor, cognitive, and perceptual symptoms depending on the spectrum of brain damage (sites where the myelin is lost) and of consequent functional losses (Compston et al. 2008; Dobson et al. 2019). Common symptoms include upper and lower extremity weakness, visual disturbances, balance and coordination problems, spasticity, altered sensation, abnormal speech, swallowing disorders, fatigue, bladder and bowel problems, sexual dysfunction, and cognitive and emotional disturbances (Trisolini et al. 2010). The causes that trigger MS and the mechanisms underlying demyelination remain the topic of intensive research efforts. This makes the disease hard to diagnose and unpredictable in its course.

2.2 Epidemiology of MS

Over 2.8 million people worldwide are affected by MS, as per the Multiple Sclerosis International Federation estimate of 2020 (Multiple Sclerosis International Federation 2020). This number has been steadily growing since 2008 when the estimate was 2.1 million people, and 2013 when the estimate was 2.3 million (Multiple Sclerosis International Federation 2013). The global prevalence rate has increased from 30 (in 2008) to 33 (in 2013), up to 35.9 (in 2020) per 100,000 people. This increase can be attributed in part, but not solely, to population growth, longer survival, and improvements in the diagnosis and reporting of MS, also thanks to the increase in the number of neurologists and MRI machines (Multiple Sclerosis International Federation 2013, 2020; Wallin et al. 2019). The average incidence rate is 2.1 per 100,000 people per year (107,000 people), resulting in someone being diagnosed with MS somewhere every 5 minutes (Multiple Sclerosis International Federation 2020).

The estimated prevalence rate varies greatly across countries and is higher for countries located at higher latitudes (Simpson et al. 2019). These differences across countries can be partially attributed to other factors, mainly the differences in the degree of disease investigation, which positively correlates with prevalence rates. Nonetheless, prevalence rates have increased in every region of the world since 2013 (Multiple Sclerosis International Federation 2013, 2020), and the latitude gradient has been seen both across world regions (Fig. 2.1) as well as within countries (Van Der Mei et al. 2001). Differences within countries cannot be solely explained by reporting differences. The positive latitude gradient in MS prevalence is strongly correlated with lower levels of sun exposure and consequently lower vitamin D levels (Sintzel et al. 2018; Tao et al. 2016), which, together with smoking and Epstein-Barr virus, have been identified as environmental factors that play a role in both the risk of MS onset and development (Ward et al. 2022).

Overall, the estimated prevalence rate of MS in Europe was 83 per 100,000 in 2006 (already higher than the global estimate for 2020), with northern countries presenting higher rates (Pugliatti et al. 2006). As of 2022, it has been estimated that over 130,000 people live with MS in the UK (199 per 100,000 people), with nearly 7,000 people being diagnosed every year. 73% of these people are women, and 33 years is the average age at MS diagnosis (MS Society 2022).

MS is the most common cause of neurological disability in young adults, with diagnosis generally occurring in early adulthood, between 20 and 40 years of age (the global average age of MS diagnosis is 32 years). Paediatric onset (under the age of 18) and late-onset (over the age of 50) are also possible, albeit rarer. There are over 30,000 people living with MS who are under 18 (as of 2020), with between 2-5% of pwMS being diagnosed under the age of 18 (Chitnis et al. 2009; Multiple Sclerosis International Federation 2020). Finally, MS has at least double the prevalence in females (69%) compared to males (31%), with the ratio reaching 4:1 in some countries. The estimates for gender prevalence and age at onset have remained consistent since 2008 (Multiple Sclerosis International Federation 2013, 2020).

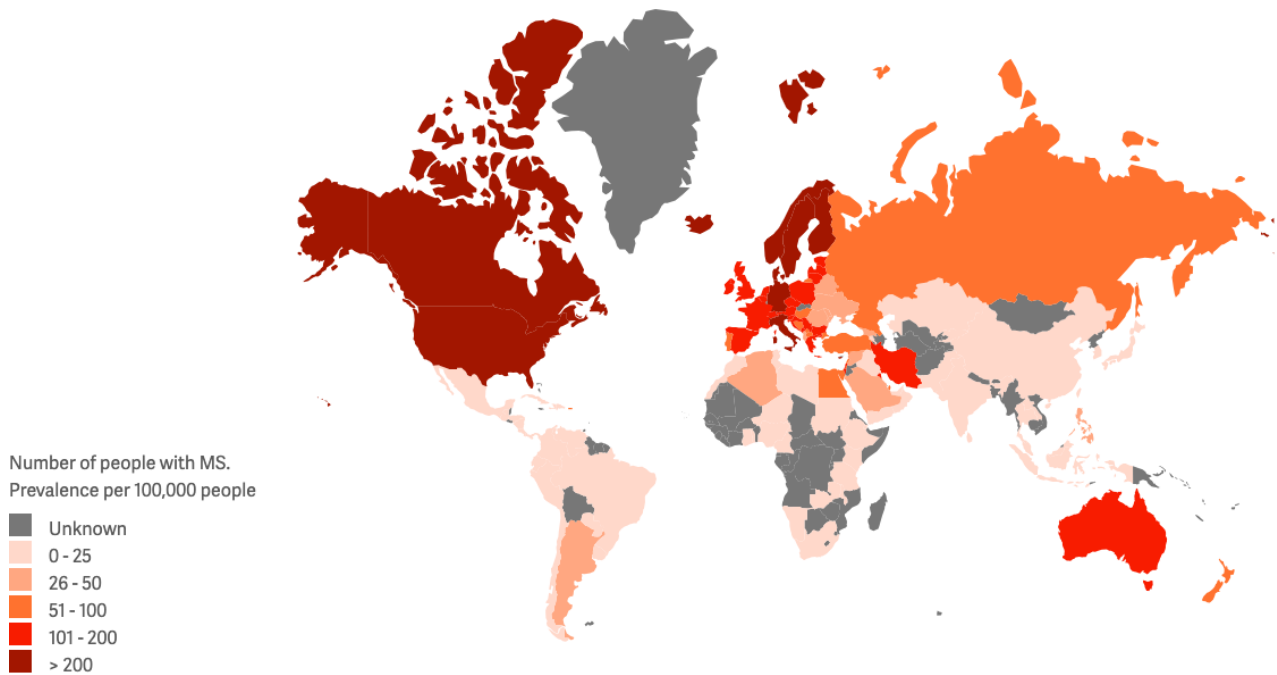


Figure 2.1: Atlas of MS. Atlas showing geographic variations in MS estimated prevalence rates (prevalence per 100,000 people) across countries as of 2020. Prevalence rates are higher in countries with higher degrees of investigation and located at higher latitudes (Multiple Sclerosis International Federation 2020).

2.3 Economic, health and social burden of MS

Epidemiological estimates are needed for policy-makers and advocacy organisations to determine disease burden and cost to society and consequently allocate resources and funding for MS research and services (Trisolini et al. 2010). This is especially crucial for pwMS because the disease usually manifests early in life, severely interfering with a person's most active and productive years: these are the years when people establish their careers, find committed relationships, have children, purchase homes, and generally organise their personal, professional, and social lives. Consequently, appropriate management strategies need to be put in place to support pwMS through their disease as well as this delicate phase of their lives.

Although MS only shortens life expectancy by 5-10 years (Brønnum-Hansen et al. 2004), it causes progressive debilitation, which disrupts patients' physical autonomy, mobility, social interaction, and

participation in leisure activities. In fact, MS is associated with a worse quality of life with respect to friendships, relationships and occupational status (Morales-González et al. 2004). Unemployment is common in pwMS, with estimates ranging from 22% to 80%. The estimates increase with disease duration, but high rates of unemployment are present also early on (Cadden et al. 2015; Strober 2020). Moreover, pwMS report trouble performing daily activities, worse emotional well-being (Bass et al. 2020; Dehghani et al. 2019), and experience a higher rate of comorbidities compared to the normal population, including worse mental health (anxiety and depression), cardiovascular disorders and autoimmune disorders (e.g. diabetes and thyroid) (Magyari et al. 2020).

This results in considerable overall costs to society. MS poses both direct and indirect costs to society. Medical and nonmedical direct costs arise from diagnosis, treatment, and accommodation of pwMS. Prescription medications, physician services, nursing care, social services, and inpatient and outpatient care are examples of direct medical costs. Modifications made to homes and cars, informal care given by friends and family, and the majority of in-home and community-based services are examples of direct nonmedical costs. The losses in labour productivity and wages resulting from patients' short- or long-term sick leave, disability pensions, early retirement due to health issues, and permanent losses from premature death are all indirect costs. Caregivers' time spent caring for patients can also result in indirect costs if, for example, they are forced to leave their jobs. Besides these, there are intangible costs resulting from the humanitarian losses due to things such as pain, anxiety, and suffering (Paz-Zulueta et al. 2020; Trisolini et al. 2010).

Many research efforts are directed toward estimating the economic burden of MS in different countries to inform decision-making on resource allocation and healthcare policies. Paz-Zulueta et al. 2020 performed a systematic review of studies investigating the economic burden of MS in different countries. Pooling the results from 20 studies, they estimated that the total annual costs per patient in Europe were, on average, 40,300€. This estimate varied across macro-areas in Europe, with Northern and Western countries facing higher costs compared to Southern and Eastern European countries. The latter faced the lowest costs: less than half the costs compared to all the other countries. Consistent across macro-areas and countries was instead the notation that the total costs increased with disease severity. Kobelt et al. 2017 estimated the total annual costs per patient across different disease severities based on a sample of 16,808 pwMS from 16 different European countries. These were 22,800€, 37,000€, and 57,500€ for patients grouped as mild, moderate, and severe disease based on their Expanded Disability Status Scale (EDSS) (Kurtzke 1983) scores. Moreover, they found that costs were dominated by production losses, non-healthcare costs and Disease Modifying Therapies (DMTs). Ernstsson et al. 2016 in another systematic review showed that the factors driving the costs changed with disease severity, with DMTs being the main driver for mild patients and production losses and informal care for more severe patients.

2.4 Clinical disease course in MS

Different disease courses are possible in MS, and even with similar disease courses, the rate at which disability accumulates, the symptoms, and the outcome vary greatly amongst patients. The different clinical phenotypes arise from different combinations of relapses and progression, which are the two main features of MS, that can present alone or in combination.

Relapses are driven by inflammatory activity in the central nervous system and manifest clinically with episodic, acute neurological symptoms of various kinds depending on the location of the inflammation. They can be acute or subacute and are followed by remission periods with partial or complete recovery of the functional impairment. A gradual, nearly continuous accumulation of irreparable disability over a minimum of one year is instead indicative of progression (Compston et al. 2008; Lublin et al. 1996; Thompson et al. 2018).

A “relapsing-remitting” (RRMS) disease course is experienced by about 85% of pwMS (Multiple Sclerosis International Federation 2020). This course is characterised by periods of relapse during which symptoms flare up either acutely or subacutely and periods of remission during which partial or complete recovery happens (Klineova et al. 2018). Most people with RRMS (about 80% (Jellinger 2003)) eventually transition to a “secondary progressive” (SPMS) disease course, which is characterised by permanent disability that builds up over time with or without acute exacerbations (Klineova et al. 2018). A reasonable estimate of the median time of transition to the progressive phase is 19 years after onset (Rovaris Marco et al. 2006). However, some patients remain free from progression even after 50 years of disease course (Skoog et al. 2014). Another study estimates that about 50% of individuals with untreated RRMS will develop SPMS within 10–15 years from onset (Weinshenker et al. 1989b).

10-15% of pwMS have a “primary progressive” (PPMS) disease course, which is marked by progression since onset independent of clinical relapses, with occasional plateau and temporary minor improvement being possible (Klineova et al. 2018). People with PPMS can experience superimposed acute relapses, with full or partial recovery and with continuous progression in between relapses. They were initially regarded as a separate clinical phenotype (“progressive-relapsing” forms – PRMS), but later considered one with PPMS as the assessment of disease activity and progression was added to the clinical phenotype description.

To further characterise the clinical phenotypes and better represent the dynamic nature of MS, all disease courses over a certain period of time can be defined as “active” or “not active” based on the presence of clinical relapses or new MRI activity, while the PPMS and SPMS disease courses can be further defined “with progression” or “without progression” based on clinical evidence of disease worsening (e.g. change in EDSS) over a specific period of time (Klineova et al. 2018; Lublin et al. 2014). A visual representation of a possible disease course for the different clinical phenotypes is reported in Fig. 2.2.

Individuals who have experienced their first clinical relapse but do not yet match the diagnostic criteria for RRMS are referred to as having “clinically isolated syndrome” (CIS). The reported risk of progression to RRMS varies in different ranges with the type of clinical presentation; the most common ones are related to lesions in the optic nerve, spinal cord, brainstem or cerebellum (Miller et al. 2012). A recent study has shown that within 20 years, 63% of individuals with CIS experience additional demyelinating episodes and are diagnosed with MS (Fisniku et al. 2008). When evidence of MRI abnormalities suggestive of inflammatory demyelination is incidentally found, but there are no clinical signs or symptoms, people are referred to as having “radiologically isolated syndrome” (RIS) and should be followed prospectively as they are at great risk of showing MS clinical symptoms in the future (Lublin et al. 2014; Okuda et al. 2009; Siva et al. 2009).

Finally, individuals who have had RRMS for a minimum of 10 years still showing minimal disability and little or no disease progression are referred to as having a “Benign” (BNMS) form of MS, characterised

by a milder disease course (Glad et al. 2006). The existence of this disease course is controversial as its definition is primarily based on physical disability, in particular ambulation, neglecting other aspects that might affect disease severity and disable people such as fatigue, pain, depression, and cognitive impairment (Correale et al. 2012; Ton et al. 2017).

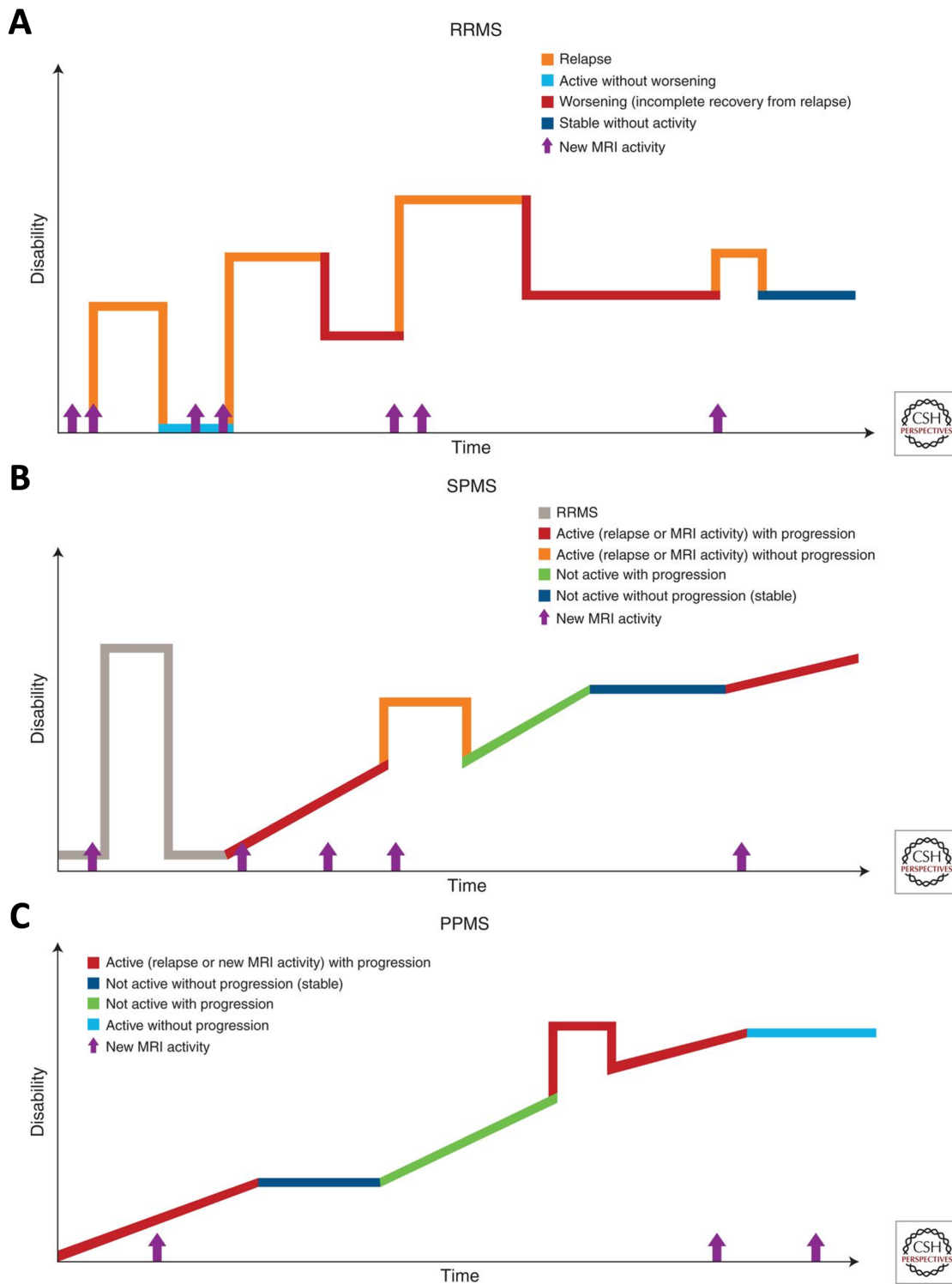


Figure 2.2: Clinical disease course in MS. Visual representation of a possible RRMS (A), SPMS (B), and PPMS (C) disease course adapted from Klineova et al. 2018.

2.5 Personalised medicine in MS

The heterogeneity of MS symptoms, progression, and outcomes across but also within clinical phenotypes, together with the increasing availability of a range of DMTs and the frequent presence of a diverse set of comorbidities highlight the importance of personalised medicine in MS. It is important to tailor the right treatment and intervention at the right time to each patient based on their individual characteristics and pathology to optimise the outcome and improve the patient’s quality of life (Gafson et al. 2017; Giovannoni 2017; Pathak 2023).

DMTs are designed to reduce the frequency and severity of relapses but often come with adverse effects and risks for the patients, risks that are higher when using more aggressive DMTs (Gholamzad et al. 2019; Vollmer et al. 2022). These risks are not worth taking if, for example, the patient will develop a Benign disease course. Moreover, early aggressive treatment is emerging as the optimal strategy in the management of MS: it allows to reduce relapses and slow down progression before the accumulation of substantial disability, improving patients’ long-term outcomes and consequently quality of life while diminishing their economic and social burden on society (Cerqueira et al. 2018).

Therefore, it is key to be able to predict early in the disease who is at greater risk of developing more aggressive forms of MS to target more aggressive treatments early on, and vice versa. Nonetheless, the disease course unfolds over the long-term, so it is not currently possible to say who, for example, is going to have a Benign course before at least 10-15 years since onset have passed (Glad et al. 2006).

Research efforts are being directed towards the identification of additional biomarkers of prognosis and treatment response, to allow for the estimation of patients’ long-term risk and the targeting of optimal treatment early on (Derfuss 2012; Gafson et al. 2017). Data collected in the context of randomised controlled trials are not suitable for this task as they are characterised by short follow-up, strict entry criteria, and limited ability to determine comparative effectiveness (Wingerchuk et al. 2002). Instead, the advent of regional and national MS patient registries provides larger-scale, longer-term comprehensive retrospective real-world data, which could play a key role in overcoming these challenges by providing a picture of longer-term disability and outcome.

2.6 MS Registries

The MS International Federation gives the following definition of registry in the 3rd edition of the MS Atlas from 2020: “A registry is a systematic, ongoing collection of data to evaluate specified outcomes for a defined population in order to serve one or more predetermined scientific, clinical or policy purposes. Registries differ from studies in that the collection of data is ongoing rather than time-limited. However, the data collected as part of registries may be utilised for individual studies.” They investigated the availability of MS registries in the world and reported that a national registry covering the whole country was found in 37% of the 102 countries who responded to the survey, with the prevalence being higher in higher-income countries (Multiple Sclerosis International Federation 2020).

Data collected by registries include but are not restricted to sociodemographic information, clinical features, symptoms, disease course, treatment and medication history, comorbidities, lifestyle factors,

environmental exposures, and patient-reported outcomes (PROs) (Ezabadi et al. 2022; Flachenecker et al. 2008). The data are collected without a specific research question but can be retrospectively analysed, for example, to compare outcomes based on a particular exposure or to evaluate the prevalence of different clinical phenotypes in the population. The most common research questions answered with registry data regard epidemiology, long-term therapy outcomes, healthcare research, and support for clinical trials (Flachenecker et al. 2014). Moreover, registers provide insights into the burden of disease, for example, collecting information on the patient's quality of life, which can help inform healthcare policies and resource allocation (Flachenecker et al. 2008).

Whenever national registers are present, it is possible to run studies in parallel and compare the results to whatever research question across different countries. This allows to evaluate disease traits and management practices, which are distinct in different countries, and to make inferences based on the observed differences (Flachenecker 2014). In fact, a lot of national MS registers collaborate internationally, allowing for data sharing and comparison across different populations and healthcare systems (Fink 2016). This is particularly important when looking at prevalence/epidemiological trends, effectiveness of long-term treatments, outcomes, and healthcare quality. The EUREMS, for example, is an ongoing initiative to establish a Europe-wide platform for systematic analysis of longitudinally collected MS data in Europe, building on existing national and regional MS registries or databases (Flachenecker et al. 2014; The Lancet Neurology 2011). Nonetheless, limitations in comparing across countries arise from differences in scope, structure, type of data collected, frequency of data collection, and duration of follow-up across registers, which in turn highlight the need to improve inter-registry consistency and determine the best techniques to interpret and utilise registry data (Flachenecker et al. 2014; Hurwitz 2011).

Registers can either be physician- or patient-driven. Traditional registries are usually physician-driven, with data being collected during routine clinical practice, most commonly face-to-face, and submitted periodically by the physician on an obligatory or voluntary basis.

Traditional registries have been extended by online registries that have widened the range and increased the scale and frequency of data that can be collected. Online registries are usually patient-driven or complemented by PROs and technologically driven outcomes that can be periodically filled in by the patients on a voluntary basis without having to rely on healthcare professionals. A good example of this is the UK MS Register (UKMSR), which combines clinical data collected by physicians with health-related quality of life data collected directly from patients through an online portal (Flachenecker et al. 2014). The downside of patient-driven registers is the possible presence of errors due to the lack of physician screening and confirmation (Hurwitz 2011). Nonetheless, Marrie et al. 2007 have shown that patient-driven registers can reach 90% or more accuracy in MS diagnosis. The collection of PROs together with the involvement of patients in the registers design and operation ensure that the scope of the registers is relevant and beneficial for pwMS. Moreover, online registries offer the opportunity for cost-effective home-based longer-term follow-up. In this sense, their usefulness has been highlighted by the recent COVID-19 pandemic where the risks to those with disabilities and on immunosuppressive disease modifying therapies attending hospital visits favoured online assessment (Evangelou et al. 2021; Landtblom et al. 2021).

Examples of successful national MS registry initiatives in Europe are the UK MS Register (Heather et al. 2023; Middleton et al. 2018; Moore et al. 2022), the Swedish MS Registry (Andersen 2012;

Hillert et al. 2015; Landtblom et al. 2021), the Danish MS Registry (Brønnum-Hansen et al. 2004; Koch-Henriksen et al. 2011; Mason et al. 2012), the Norwegian MS Registry (Myhr et al. 2015), the German MS Registry (Blaschke et al. 2022; Stuke et al. 2009), and the Italian MS Register (Trojano et al. 2019). The NARCOMS Registry is an analogous international initiative in the USA (Marrie et al. 2007).

The UK MS Register was launched in 2011 and is funded by the MS Society. It involves almost 20,000 people with MS in the UK. It includes both a cohort of unverified pwMS who respond to PROs through the internet portal and a cohort of confirmed pwMS for whom clinical data are available and reported by MS Specialist Treatment centres across the UK. The online cohort has proven to be a representative sample of pwMS in the UK and can be treated as a valid cohort of pwMS also in the absence of supporting clinical validation (Middleton et al. 2018).

Overall, registries bridge the gap between clinical research, real-world clinical practice, and patient perspectives, providing a comprehensive picture of MS and its impacts.

2.7 Patient-reported outcomes

The use of PROs in MS research and clinical trials has become increasingly popular. This happened in parallel with the advent of personalised medicine and e-health, and with the formation of patient registries, in particular online registries, which provided the perfect avenue where PROs could be administered at large-scale and longitudinally in a time- and cost-effective way. Altogether this favoured a paradigm shift in MS healthcare towards a more patient-centric approach, although a gold standard of PROs to use for different purposes is still lacking (D’Amico et al. 2019).

PROs are self-reported assessments, usually in the form of standardised questionnaires, gathering information directly from patients without the interpretation of a clinician about their symptoms, function, health status, health utility, and the impact of the illness on their day-to-day activities and quality of life (D’Amico et al. 2019; Nowinski et al. 2017; Wiklund 2004). Through PROs, patients can provide their own perspective on their disease. This patient-centric approach is especially important in MS, a disease marked by a wide range of symptoms and a highly individualised impact on physical, cognitive, and emotional well-being (Ziemssen 2011).

Common symptoms experienced by pwMS include but are not limited to fatigue, pain, motor, cognitive and visual impairment, and mood disturbances. These can vary greatly among individuals based on their combination (which symptoms), their intensity (how bad/how often), and their order of presentation (when and for how long). Not all these symptoms can be measured objectively. Traditional clinical measures, such as the EDSS (Kurtzke 1983), relapse rate or imaging features, do not provide a comprehensive picture of the symptoms and do not capture the extent to which these affect patients’ quality of life (Brichetto et al. 2020). For example, in the EDSS, which is the most common outcome used and favoured by regulators, cognition, vision, and upper limb function are underrepresented (Cohen et al. 2021).

This gap is where PROs play a pivotal role. Instruments such as the MS Impact Scale (MSIS-29) (Hobart et al. 2009; Hobart et al. 2001) and the MS Walking Scale (MSWS-12) (Hobart et al. 2003) are a clear example of this. They have been developed to specifically address the need for sensitive

and specific measures of general motor impairment and walking disability in MS. These tools have demonstrated reliability and validity in reflecting the patients' own assessment of their symptoms and functional limitations (McGuigan et al. 2004; Riazi et al. 2002) and have been integrated within online registries for longitudinal assessment and monitoring purposes (Jones et al. 2013; Rodgers et al. 2022). Other examples are the Hospital Anxiety and Depression Scale (HADS) (Bjelland et al. 2002; Snaith et al. 1986) and the Fatigue Severity Scale (FSS) (Krupp et al. 1989), measuring respectively mood disturbances and fatigue. Unlike the MSIS-29 and the MSWS-12, these PROs were not specifically designed for MS, but target specific symptoms and allow comparisons across different conditions.

Furthermore, the scope and utility of PROs go beyond just measuring symptoms. They provide valuable information about treatment effectiveness and side effects, rate patients' satisfaction with the care received, and their overall quality of life (Wiklund 2004). The EQ5D (EuroQol Group 1990), for example, is a 5-item questionnaire designed to give a summary measure of health-related quality of life and is a measure that allows evaluating the ecological validity of other markers of disability, as well as the real impact that treatment has on the patient's day-to-day life. This is particularly important in clinical trials and longitudinal studies where PROs can provide a more complete assessment of therapeutic efficacy from the patient's perspective (Wiklund 2004). PROs, in general, play an important role in tailoring treatment to individual needs, encouraging joint decision-making between patients and practitioners (D'Amico et al. 2019).

The potential of PROs in managing MS has been further enhanced by recent developments in digital health technologies. The accuracy and continuity of patient monitoring have been improved by the incorporation of PROs into electronic health records and the utilisation of mobile health applications for real-time symptom tracking (Brichetto et al. 2020; Lavorgna et al. 2018; Marziniak et al. 2018). Because MS is an unpredictable and variable condition, this digital evolution allows for a more dynamic and responsive approach to disease care.

However, some of these PROs were not co-created with patients and lack understanding about what matters the most to patients (The Lancet Neurology 2019; Westergaard et al. 2022). Other limitations in the use of PROs arise from the fact that they are more subjective than objective as they reflect individuals' own frame of reference, which can also be subject to change (response shift) over time as, for example, individuals adapt to their condition. This hinders objective comparisons with people with the same diagnosis or others without the diagnosis, as well as with themselves at previous points in time (Li et al. 2011; McGinley et al. 2022). Overall, these elements discourage their suitability for use as primary outcomes in trials (Strijbis et al. 2022). Nonetheless, they do offer complementary and distinct information from traditional clinical measures that can be useful to provide a comprehensive understanding of the impact of the disease on people's daily lives, to assess the ecological validity of other more objective measures, and to direct research efforts in a direction that is most relevant to pwMS. More work is needed to further validate and explore the utility of the available PROs and to develop new PROs specific to MS following a more thorough development process involving conceptual frameworks, that define and conceptualise a priori the variable to measure, patient involvement, to ensure addressing functional domains that matter most to patients, and validation and harmonisation across cultures (Bharadia et al. 2022; Zaratin et al. 2022).

In the future, the ongoing creation and improvement of PROs will be essential to maintaining patient voices at the forefront, helping to direct clinical practice and research efforts to enhance overall patient

outcomes and quality of life as MS management continues to change.

2.8 Cognition in MS

2.8.1 Impact of cognitive impairment in people with MS

Cognitive impairment is prevalent in MS and greatly affects patients' quality of life, sometimes more than physical impairment (Cutajar et al. 2000; Rao et al. 1991a). PwMS with cognitive deficits are less likely to be employed (Schultheis et al. 2010), have more functional impairments (Goverover et al. 2007), i.e. difficulties completing routine household tasks, self-care, driving (Honarmand et al. 2011)/taking public transport, and shopping independently, engage in fewer social and vocation activities and are more exposed to psychiatric illnesses, such as depression and anxiety, than pwMS with a purely physical disability (Rao et al. 1991a). The employment issue is particularly important because between 50% and 80% of pwMS are reported to be unemployed within 10 years of disease onset (Chiaravalloti et al. 2008), which happens early in life in MS as disease onset is typically between 20 and 40 years. Cognitive impairment is found to contribute largely to the reported rates of unemployment (Beatty et al. 1995; Rao et al. 1991a), explaining more variability in the rates than physical disability and demographic factors (Julian et al. 2008; Larocca et al. 1985). Overall, cognitive impairment contributes to isolation and reduces the physical and economic independence of pwMS, greatly affecting their quality of life (Chiaravalloti et al. 2008; Rao et al. 1991a).

Cognitive impairment, initially disregarded as less evident and thus less easily tested compared to physical impairment, is a core symptom of MS and, for all the reasons mentioned above, has been largely researched over the past two decades in its prevalence, expression, and neural basis (Sumowski et al. 2018). Despite this, it is still rarely assessed as an outcome in clinical trials or routinely evaluated in clinical practice (Sumowski et al. 2018), it is not accounted for in the most used clinical scale, the EDSS (Saccà et al. 2017), and it is not even considered for the definition of clinical phenotypes, as it is for example the case for 17% to 47% of the people with a Benign course, defined as such solely based on preservation of motor function, but who have been shown to have cognitive impairment (Benedict et al. 2020; Ton et al. 2017).

2.8.2 Prevalence of cognitive impairment in MS

Reported prevalence rates vary greatly across studies but are usually between 40% and 70% (Chiaravalloti et al. 2008; Rao et al. 1991b). The variability in the reported rates can be attributed to several reasons, including the research settings, the representativeness of the studied population, the type and number of tests being used to assess cognition, and the applied criteria used to define cognitive impairment (Amato et al. 2018; Fischer et al. 2014). In fact, pwMS are classified as impaired vs non-impaired following different criteria, which vary in their stringency, resulting in higher prevalence rates when using more liberal criteria and lower prevalence rates for more stringent criteria (Fischer et al. 2014). The most commonly used criterium classifies pwMS as impaired if they perform below the fifth percentile on about 20% of the evaluated test parameters or tests (Benedict 2009). This is a dichotomic approach which does not provide details on the domains that are impaired, does not differentiate across individuals impaired in different domains, and does not capture changes in cogni-

tion that can occur over disease time for both pwMS classified as impaired and not impaired (Pitteri et al. 2019; Sumowski et al. 2018). Other approaches have been used that group parameters or tests that measure the same underlying cognitive domain together and provide a more detailed picture of impairment across different aspects of cognition (Fischer et al. 2014). Nonetheless, these approaches don't solve all the problems, as they still are "impairment-based" dichotomic approaches that do not capture the progressive decline of cognition over time on either side of the cut-off threshold, resulting from both gradual progression and acute disease activity. Pitteri et al. 2019 highlight the benefits of moving from an "impairment-based" approach to a "function-based" approach, which looks at cognitive function as a continuous variable that changes over time.

In this context, key priorities in the field are to define what is a clinically meaningful change in cognition in MS and how to measure it (Benedict et al. 2012) and provide a baseline measure of cognition for all newly diagnosed pwMS, to use as a reference to evaluate future decrements in cognitive function over time and potentially treatment efficacy (Sumowski et al. 2018). Moreover, baseline cognitive assessment allows for the early detection and monitoring of cognitive dysfunction. This is particularly important because cognitive impairment can appear early in the disease, before or in the absence of other symptoms, including physical disability (Achiron et al. 2006; Amato et al. 2010; Glanz et al. 2007; Pelosi et al. 1997; Sumowski et al. 2018). Additionally, early cognitive impairment is predictive of CIS conversion to MS (Zipoli et al. 2009) and of future accumulation of physical disability (DeLoire et al. 2010). Cognitive impairment in general is predictive of evolution in Benign MS (Portaccio et al. 2009). Additionally, pwMS can have cognitive relapses, which can go on unnoticed if cognition is not readily and periodically assessed (Benedict et al. 2020; Pardini 2021; Pardini et al. 2014; Weinstock et al. 2022).

More at the general level, cognitive impairment is present at all stages and for all MS subtypes even if to a different extent (Benedict et al. 2020), and it is observed also at the extremes, e.g. in CIS (Khalil et al. 2011; Pelosi et al. 1997), RIS (Amato et al. 2012), paediatric-onset MS (Amato et al. 2014; McKay et al. 2019; Till et al. 2013), and late-onset MS (Butler Pagnotti et al. 2022). Cognitive impairment is more severe and more prevalent in progressive versus relapsing-remitting forms of MS (Huijbregts et al. 2004; Johnen et al. 2017; Planche et al. 2016; Ruano et al. 2017). Nonetheless, the pattern of cognitive domains impaired is in general similar between relapsing-remitting and progressive forms, with a higher number of domains being affected in the latter (Gois et al. 2021; Ruano et al. 2017).

Since cognitive impairment is present early on and in all stages of the disease, but longitudinal studies over a few years have often failed to show significant cognitive decline (Jennekens-Schinkel et al. 1990; Langdon 2011; Patti et al. 1998; Rao et al. 1991a), it might appear to be independent of disease duration. However, measured across a sufficiently long follow-up time, cognitive impairment does emerge and progress in a sizeable proportion of patients (Amato et al. 2001). Moreover, since the clinical phenotypes are dependent on disease duration, with, for example, SPMS individuals having greater disease duration compared to RRMS individuals, the relation between cognitive impairment and disease duration is inherent, confounded by the relation with disease course and yet to be fully understood. The major limitation in this sense is the lack of large longitudinal studies comprising cognitive assessment and covering disease course since onset with baseline assessment of cognition (Ruano et al. 2017; Sumowski et al. 2018). In fact, the relation between cognitive impairment and disease duration has been mostly measured using cross-sectional designs or short follow-up periods,

which make it difficult to detect individual changes in cognition over disease duration (Chiaravalloti et al. 2008).

Analogous considerations can be made when looking at the relation between cognitive impairment and physical disability, which have been shown to be mildly associated (Lynch et al. 2005; Ruano et al. 2017). In this regard, a recent study has shown that the underlying MS pathology, i.e. brain damage and grey-matter atrophy, contributes to the observed correlation between cognitive impairment and physical disability (Gois et al. 2021). Moreover, a link has been found between cognitive functioning and depression, with depressed MS patients performing worse on information-processing speed and working memory tasks (Arnett et al. 1999; Gilchrist et al. 1994; Thornton et al. 1997). Finally, pwMS often report impaired cognitive function because of fatigue (Krupp et al. 1988). Nonetheless, fatigue is often measured through self-report scales, which have not shown consistent associations with decreased cognitive performance (Johnson et al. 1997). However, when efforts have been made to measure fatigue more objectively, for example by means of tasks requiring sustained attention or following cognitively demanding tasks, decrements in cognitive performance over time have been found, suggesting an effect of fatigue on cognitive functioning (Krupp et al. 2000; Schwid et al. 2003).

Overall, the association of cognitive impairment with demographic and clinical features is yet to be fully understood as the studies available in the relevant literature present inconsistent results due to methodological limitations or do not account for all potential confounding factors which might explain the associations.

2.8.3 Cognitive domains affected in MS

Information-processing speed is reported to be the most commonly affected cognitive domain in MS and the one declining more rapidly (Bergendal et al. 2007; Costa et al. 2017; DeLuca et al. 2004; Denney et al. 2008; Forn et al. 2008). Working memory is also reported to be affected, often concomitantly with, but less frequently than, information-processing speed (DeLuca et al. 2004; Janculjak et al. 2002; Lengenfelder et al. 2003). When these deficits are present, they can confound the results on tasks measuring higher-order cognition as these often involve an information-processing component too (Macniven et al. 2008). Therefore, when measuring other higher-order cognitive abilities it is crucial to approach the interpretation of results with caution in the presence of information-processing deficits. Deficits in information-processing also result in difficulties in learning, e.g., a longer time is necessary to encode new information. Consequently, memory is another consistently impaired domain in MS (40-65% prevalence rate (Rao et al. 1993)), with the primary problem being the initial encoding of new information (DeLuca et al. 1994, 1998).

Other deficits that are reported but less frequently involve executive functions and visuo-spatial abilities (Benedict et al. 2006; Rao et al. 1991b). Executive functions refer to the cognitive abilities needed for goal-direct behaviour, problem-solving, decision-making, and adaptive responses, and deficits in this domain have been reported to have a prevalence of 15-28% (Benedict et al. 2006; Drew et al. 2008). Moreover, they have been shown to remain unchanged even after controlling for information-processing (Ruano et al. 2017). Visuo-spatial abilities, instead, can be affected in up to a quarter of pwMS (22% prevalence rate (Benedict et al. 2006)), and can occur even in the absence of primary visual processing abnormalities (Vleugels et al. 2000).

Finally, language function and basic attention are reported to be relatively preserved in MS (Benedict et al. 2006), but deficits in semantic memory (Laatu et al. 1999) and sustained or divided attention (McCarthy et al., 2005) can be present.

Despite having identified cognitive domains that are commonly affected in MS, there are still many questions that remain unanswered. In this regard, Sumowski et al. 2018 highlight the followings as key priorities of the field: a) the need to investigate patient-reported real-world multitasking deficits using cognitive-motor and cognitive-cognitive dual-task paradigms; b) the need to uncover the dependency links across cognitive domains; c) the need to overcome heterogenous groups of cognitively impaired pwMS and derive a more precise cognitive phenotyping of pwMS; and d) the need to derive composite or latent variables that are purer measures of targeted cognitive functions.

2.9 Traditional in-person cognitive assessments in MS

Traditionally, cognition function in MS has been measured by means of one-on-one face-to-face neuropsychological assessments. This created the need to have a standardised, efficient, and comprehensive tool for neuropsychological assessment to deploy in MS centres around the world. In this sense, a few batteries have been developed over the past decades specifically to measure cognitive domains that are impaired in MS. The two most commonly utilised are the Brief Repeatable Battery of Neuropsychological Tests (BRB-N) (Boringa et al. 2001), a battery lasting 45 minutes and including five tests developed by the Cognitive Function Study Group of the National Multiple Sclerosis Society, and the Minimal Assessment of Cognitive Function in MS (MACFIMS) (Benedict et al. 2006; Migliore et al. 2016), a battery lasting 90-minutes and including seven tests developed based on the findings of an international conference held in 2001 in collaboration with the Consortium of MS Centers. Both batteries include tests that were selected to have good psychometric properties (reliability, discriminative validity, and sensitivity) while covering domains that are commonly impaired in pwMS. Some of the tests administered in the two batteries overlap, while others are different. Both batteries measure information-processing speed, working memory, verbal memory, and visuo-spatial memory. However, the MACFIMS measures two domains that are not directly measured in the BRB-N, namely executive function and visuo-spatial processing, but it is consequently a longer battery (two times the duration with two additional tests). The two batteries were shown to be comparable in their overall sensitivity to MS (Strober et al. 2009).

To further reduce the duration of neuropsychological assessment in MS and encourage its use in everyday clinical practice and its international validation and deployment, the Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) was recommended in 2012 by an expert international committee of twelve members (Langdon et al. 2012). This battery lasts 15 minutes and includes three tests that were selected based on their psychometric properties, their acceptability by pwMS, their international application, and their ease of administration. These tests are the Symbol Digit Modalities Test (SDMT) (Benedict et al. 2017), measuring information-processing speed, the California Verbal Learning Test – Second Edition (CVLT – II) (Woods et al. 2006), measuring verbal memory, and the Brief Visuospatial Memory Test – Revised (BVMT – R) (Benedict et al. 1996), measuring visuospatial memory. Moreover, this battery does not require any specialist equipment or specialist expertise in cognitive assessment to be administered. For this reason, as of 2023, the BICAMS has been largely

validated, translated, and culturally adapted in 26 countries, has produced data for 150 published studies, and has been used in national and international trials (Potticary et al. 2023). The 26 international validation studies reported in the review from Potticary et al. 2023 confirmed the validity of the BICAMS, with pwMS performing worse than controls on all three tests included in the battery, and the feasibility of standardised international cognitive assessment.

However, batteries that have been selected for brevity rather than comprehensiveness often fail to identify all the cognitive deficits that can present across pwMS, especially considering that MS is a heterogeneous disease. Additionally, even if these batteries are optimised for brevity and can last as little as 15 minutes, the need to have even just 15 minutes of one-on-one face-to-face testing for every patient over time (ideally once a year), is not practical as it would require substantial time and resources (Sumowski et al. 2018). This hinders the integration of neuropsychological assessment into routine clinical practice. Limitations include the finite number of individuals who can be tested within a given timeframe, sampling bias favouring those with access to testing, typically conducted in specialised centres and university hospitals catering to specific sociodemographic populations (Potticary et al. 2023), and the reluctance of many individuals to return to testing centres in person for multiple assessments over time. As a result, there is a lack of availability of cross-sectional data from large representative samples and of longitudinal data with long follow-ups. However, such data are key for advancing our understanding of cognitive function in MS, in terms of prevalence, time course, risk factors, disease-related cognitive decline, and efficacy and targeting of treatment (Sumowski et al. 2018).

2.10 Online cognitive assessments

Online unsupervised computerised cognitive assessments offer a potential solution to these challenges. There has been a growing interest in these technologies over the past decade, with applications in both population (Hampshire 2020; Hampshire et al. 2020) and clinical research (Brooker et al. 2020; Del Giovane et al. 2023b; Hampshire et al. 2021b). While they carry substantial benefits compared to traditional in-person cognitive assessments, they come with a new set of challenges and specific problems that need to be addressed before they can be incorporated into clinical practice (Feenstra et al. 2017; Germine et al. 2019; Kessels 2019).

2.10.1 Benefits

The main benefits of online cognitive assessments are accessibility, standardisation, richness of measurement and efficiency (Feenstra et al. 2017). They are accessible to everyone in possession of a smartphone, a tablet, or a computer from the comfort of their home or any other convenient location at whatever time of the day they are available or prefer to take the assessment. Therefore, they are inclusive of people who are generally excluded from traditional neuropsychological testing because living in remote/rural locations or having mobility problems, low income or a busy lifestyle (Barak et al. 2004; Birnbaum 2004; Caine et al. 2012; Darby et al. 2014; Naglieri et al. 2004). They allow for greater standardisation in the way the test is presented, including the instructions, the stimuli, and the feedback received, in the scoring of the results, and in the transfer and storage of the data. This reduces the experimenter effects as well as the risk of human errors, enabling more objective and

precise assessments (Barak et al. 2002; Bilder 2011; Parsey et al. 2013; Reips 2000). Additionally, all users are guaranteed to use the most recent version of the online testing platform due to the central implementation of software upgrades, bug patches, and test adaptations (Barak et al. 2004). Online cognitive assessments are efficient as they are automatically deployed and scored via any personal device in an unsupervised setting, greatly reducing the resources needed to perform each assessment, in terms of personnel, time, and equipment, and therefore greatly reducing the costs (Bauer et al. 2012; Birnbaum 2004; Caine et al. 2012; Naglieri et al. 2004)). Moreover, online cognitive tests are often adaptations of standard paradigms or traditional neuropsychological tests but gamified in order to promote engagement and with the possibility of automatically generating unique stimuli to reduce practice effects in repeated assessments (Del Giovane et al. 2023b). Finally, online cognitive tests allow the collection of more detailed behaviour logs, including trial-by-trial information on the type and location of the stimulus presented, the accuracy of the response, and the reaction time, therefore, expanding the range and type of outcomes derivable for each test at levels and with precision unattainable for human observers (Bauer et al. 2012; Schatz et al. 2002).

Overall, this enables both normative and patient data collection at a large scale and repeated normative and patient data collection over longer periods of time. This results in more representative population samples (larger and less biased samples) with an enhanced possibility to account for variability in variables outside the studied factors which are known or expected to affect performance on some or all the tests included in the assessment (Feenstra et al. 2017). Examples of these are sociodemographic characteristics, such as age, gender, education, and ethnicity, which are known to affect performance on cognitive tests and could produce a false significant variability in other factors if not accounted for (Hampshire et al. 2012). Moreover, this allows for greater generalisability of the results (Birnbaum 2004) and greater ecological validity (Smit et al. 2013), as people are tested in their everyday context and not in a controlled environment.

2.10.2 Challenges

However, online unsupervised computerised cognitive assessments come with limitations and challenges, some new, some shared with traditional in-person cognitive assessments. Three main challenges are shared with traditional in-person cognitive assessments and include: i) the fact that performance on the tests covaries with various variables in the normal population, and these must be accounted for in some way to ensure a “culture fair” assessment, ii) the need to optimise the assessment in terms of cognitive domains being measured and duration, to make it relevant and accessible to a specific target population, and iii) the requirement to collect a large normative sample to be able to study and account for the aforementioned confounding variables as well as transform individual performances into standardised outcome measures.

The challenges that are specific to online cognitive assessments include the influence of testing context, device variability, and test adaptation from traditional tests (Germine et al. 2019). Testing context can negatively contribute to performance, but it is difficult to account for it as it is not feasible to measure all environmental distractions present at the time of testing (Feenstra et al. 2017). When transforming a traditional test into a computerised test, the changes in the perceptual, cognitive and/or motor complexity of the test or in the response modality might also affect test validity, which needs to be re-evaluated along with the other psychometric properties (Bailey et al. 2018;

Barak et al. 2002; Bauer et al. 2012). Moreover, the performance scores using the computerised tests need to be compared with those using the corresponding traditional tests to ensure that they are measuring the same cognitive construct and that they are similar or superior in their sensitivity and precision in discriminating affected domains, especially in applications targeting specific populations. Additionally, ad-hoc normative data needs to be collected for the computerised version of the tests (Bauer et al. 2012; Feenstra et al. 2017). Device variability both among different device categories (desktop/laptop vs tablet vs smartphone) and among devices in the same category but with different technical specifications (hardware and software) can greatly confound the results when, for example, evaluating changes in cognitive performance over time and using different devices at each assessment, or the effect of a certain variable on cognitive performance, while that variables also affect the type of device being used (Germine et al. 2019). This is particularly relevant when reaction time is the main outcome measure of the test (Kessels 2019). Using one device for everyone is not a good solution because, other than being impractical and costly, it introduces familiarity bias, for which people who are more familiar with the device being used will perform better. There is a trade-off between minimising the contribution of device variability and familiarity. Nonetheless, the challenges related to device variability can be minimised by collecting large and comprehensive normative data on people using a variety of devices with different specifications and explicitly accounting for this variability in the same way as sociodemographic variables are accounted for. Additionally, they can be minimised by being careful in designing and favouring tests and outcome measures that are robust to device variability (for example, using measures of accuracy over response time or measures with longer reaction times, or using differences with baseline measures taken using the same device) (Germine et al. 2019; Kessels 2019). However, the always-changing technological landscape poses further problems as it results in always new devices with better technical specifications and the consequent need to constantly adjust and re-validate the tests as well as collect new normative data for these devices, which are anyway bound to soon become obsolete in an endless cycle. It is important to treat online cognitive tests as software which needs to be continually updated, adjusted and validated through cycles of modification and evaluation at a rate dictated by technological progress (Germine et al. 2019). This is not always feasible in practice as it is prohibitive in terms of time and cost, especially in the case of cognitive assessment software resulting from research efforts carried out in universities and often not supported by a business model (Kessels 2019).

Additional general guidelines for the development of online unsupervised computerised cognitive assessment are aimed at minimising the impact of context and psychological factors, such as distractions, lack of motivation, fatigue, and anxiety, thoroughly designing the user interface such that it is clear (in terms of instructions, design and feedback), accessible for the target population (both in terms of computer skills and in case of visuo-motor impairments or other disabilities) and engaging (to optimise for motivation and compliance) (Feenstra et al. 2017).

Overall collaborations among cognitive neuroscientists, clinical neuropsychologists, software developers and neurologists are necessary to create, adapt and validate online unsupervised computerised cognitive assessments that are valid and reliable across platforms, sensitive and straightforward to employ in clinical practice, and user-friendly.

2.10.3 Cognitron

A successful effort in this direction is represented by the Cognitron, an online platform which encompasses a large library of online tasks that measure a comprehensive range of cognitive domains and are deployable via any device (desktop computers, laptops, tablets, and smartphones) (H2 Cognitive Designs [n.d.](#)). The tasks are designed to be straightforward and user-friendly, as, for example, instructions are complemented by visual animations using gifs, and more complex tasks include practice trials. Some tasks are adaptations of traditional pen-and-paper tasks; others instantiate novel cognitive designs. The platform was designed to enable easy set-up of custom sites for deploying different sequences of cognitive tasks during different studies. The sites are set up by creating a battery of available tasks, with the possibility to add questionnaires at whichever point of the battery asking, for example, about potential confounding variables or feedback or investigating things such as metacognitive abilities, which can provide complementary information to task performance. Moreover, participants can choose when to start each task so that they can read the instructions in their own time or take a short break if needed to avoid things such as fatigue or distractions to alter the results. Additionally, the sites are customisable so that whoever is setting up the study can choose to display other web pages at the beginning or at the end of the assessment to ask consent, give information on how the assessment should be carried out (addressing for example issues regarding the testing context or the format of the assessment) or thank the participants for taking part in the study and show them their results. Finally, Cognitron collects very detailed information both on the task performance as well as on the general behaviour during testing (e.g. overall duration, breaks taken between parts of the assessments, times the focus is lost from the testing web page, ...), which allows to implement compliance checks and ensure data quality.

Overall, Cognitron is a platform that enables the creation of customisable sites to perform efficient large-scale cognitive data collection. To date, it and its precursor platform have already been used successfully for large-scale citizen science research, epidemiological studies, birth cohorts, clinical cohorts and clinical trials (Bourke et al. [2023](#); Cheetham et al. [2023](#); Del Giovane et al. [2023b](#); Hampshire et al. [2022](#), [2012](#), [2021b](#); Soreq et al. [2019](#), [2021](#); Zhao et al. [2022](#)). Moreover, some of the Cognitron tasks have been related to underlying brain structures and functions thereby validating that they were measuring the construct they were developed to measure (Hampshire et al. [2012](#); Owen et al. [2010](#)).

In the context of population studies, good examples are the GBIT and REACT studies, in which Cognitron has been used to successfully collect data from hundreds of thousands of people from the general UK population, over a period of a few months, and in the former case, at repeated follow up timepoints (Hampshire [2020](#); Hampshire et al. [2020](#)). These population studies not only allowed investigating trends in cognitive abilities in the general population while also validating the included tasks, but they also provided extensive normative datasets to be used in the context of clinical studies to derive measures of cognitive function and impairment accounting for a wide range of sociodemographic variables which have been shown to affect cognition and were investigated as a part of these same studies. The GBIT study included a relatively large and representative sample of the UK population as it was initially promoted through a BBC2 Horizon documentary (Mosley [2020](#)). Additionally, these studies, which were coincidentally carried out in part during the pandemic, allowed investigation of the effects of COVID-19 on cognition and mental health in the general population and how they differed based on disease severity (Hampshire et al. [2021a,b](#)).

In the context of clinical studies, the Cognitron is being used to derive optimal batteries of online cognitive tasks to measure and monitor cognitive impairment in different clinical populations. This is a longer process as it involves first the selection and then the validation of a subset of tasks specific to each clinical population in terms of both domains being measured (which should reflect characteristic deficits) as well as accessibility and compatibility with any characteristic form of disability which might affect the group (e.g., visuo-motor impairment). Moreover, it involves carefully designing the site itself in order to make it accessible and user-friendly for the target population. While there have been many efforts in parallel to develop optimal and validated batteries for Parkinson's, Stroke (Gruia et al. 2023a; Gruia et al. 2023c), Alzheimer (Brooker et al. 2020; Del Giovane et al. 2023a), and Schizophrenia, it is worth mentioning the already published work on Traumatic Brain Injury, which details this process of development starting from tasks selection, continuing with tasks validation and comparison with established traditional neuropsychological tests, and finishing with the creation of a clinician-facing report based on the assessment to be used to inform clinical decision-making or target a more comprehensive cognitive assessment (Del Giovane et al. 2023b). In this thesis I will detail a similar process but for MS.

2.11 Online cognitive assessments in MS

The development and use of online cognitive assessments have been increasing in MS over the past decades. Both computerised and smartphone versions of classic neuropsychological tests have been developed for MS or adopted in MS research with promising results.

A recent systematic review from Wojcik et al. 2019, including 11 batteries and 33 individual tests, identified some computerised neuropsychological tests with acceptable psychometric properties and sensitivity to MS-related cognitive impairment, which could be valid screening and monitoring tools. The analysed and required psychometric properties were: test-retest reliability, discriminant/know groups validity, ecological validity and concurrent validity. The batteries containing tests with acceptable psychometric properties were four: the Cognitive Drug Research (CDR) battery (Edgar et al. 2011), the NeuroTrax battery (Achiron et al. 2007; Bergmann et al. 2023; Golan et al. 2019), the CogState Brief Battery (CBB) (Darby et al. 2002; De Meijer et al. 2018; Eilam-Stock et al. 2021), and the Central Nervous System-Vital Signs (CNSVS) battery (Gualtieri et al. 2006). The single best tests identified to perform quick screening were the Computerised Speed Cognitive Test (CSCT) (Ruet et al. 2013), the Processing Speed Test (PST) (Rao et al. 2017), and the Computerised Symbol Digit Modalities Test (C-SDMT) (Akbar et al. 2011). They were all computerised versions of the SDMT, designed to measure information-processing speed.

To be included in the review, all of the batteries had to have been used in at least one study including an MS sample. Nonetheless, narrowing the focus only on the acceptable batteries, they were developed in the context of other neurological populations or for the general neurological community and were not specifically tailored to MS. This on one side could be seen as an advantage: if the batteries show consistent reliability and validity across different populations, it is further proof of their psychometric properties. However, some of these batteries have been used in MS-related publications without being specifically validated for this population, or being validated only on small, non-representative samples of pwMS (e.g., including only relapsing-remitting individuals). Moreover, having been developed for

other populations, they might miss domains relevant for MS, or, vice versa, put too much focus on less relevant domains. Of the four acceptable batteries only one, the NeuroTrax, was used in a considerable number of MS-related publications – 23. The CBB, the CNSVS, and the CDR were used in respectively 5, 3, and 1 MS-related publications. Additionally, despite there being large normative data (> 1000 people) from at least one country available for all these batteries, which allowed evaluating discriminant validity, there are no normative datasets from the UK population, which restrict the use of these batteries in the country. Other limitations shared between some, or all the batteries involved their deployability – they are not deployable on any device –, their efficiency – they require a technician for administration or were not validated in the absence of a technician –, and their comparative validity – which was evaluated against single tasks (usually the PASAT and the SDMT) rather than against complete batteries (e.g., the BICAMS or the MACFIMS).

A systematic review from Foong et al. [2023](#) instead focused on smartphone monitoring of cognitive function. They identified six smartphone apps used for monitoring in MS, five of which were based on the SDMT. Despite the apps showing good psychometric properties, sometimes outperforming established neuropsychological tests, and high acceptability, their focus is too narrow as they exclusively measure information-processing speed, neglecting all other cognitive domains, which might also be affected in the absence of information-processing deficits. Nonetheless, the review offers proof that online monitoring via smartphone is not only feasible but can help to efficiently collect valid and reliable data on cognitive function in MS.

Overall, there is a need for a battery of online cognitive tasks which: i) is developed specifically for MS, ii) is deployable on any device, iii) includes tests that comprehensively measure the domains most commonly affected in MS, and not only information-processing speed, iv) is validated on a large representative sample of pwMS, v) has large normative data available from various different countries including the UK, and vi) is affordable and accessible to anyone with MS. The work presented in this thesis aims to set some steps in the direction of the development of such assessment.

Chapter 3

General Methods

In this chapter, I outline the overarching methods employed throughout my thesis. Specific methods are instead detailed separately within each results chapter. Specifically, I describe the patient-reported outcomes used to assess different aspects of MS-related disability from the patient's perspective and the online cognitive tasks used to objectively test cognitive function in the MS population. Additionally, I provide an overview of the pre-processing pipeline developed for the collected online cognitive data.

3.1 Patient-reported outcomes

In this thesis, I analysed data on the following patient-reported outcomes (PROs), which had been previously collected via the UK MS Register (UKMSR).

3.1.1 Multiple Sclerosis Impact Scale - MSIS-29

The Multiple Sclerosis Impact Scale (MSIS-29) is a validated scale developed to assess various aspects of MS-related disability, including physical, psychological, and social functioning. The MSIS-29 consists of 29 items, with 20 items assessing physical disability (MSIS-29 motor) and 9 items assessing psychological wellbeing (MSIS-29 psychological). The scale uses a 5-point Likert-type response format in its first version (Hobart et al. 2001) and a 4-point format in its second version (Hobart et al. 2009). Both versions have been used by the UKMSR at different points in time. The first version has been used prior to April 2012 and the second subsequently. Individuals rate the extent to which each item reflects their experience of MS-related impact on a scale from 1 (no impact) to 4/5 (maximum impact). The MSIS-29 is commonly used in clinical research and practice to evaluate the effectiveness of interventions, monitor disease progression, and assess the overall impact of MS on an individual's quality of life. In this thesis, I analysed data only on the MSIS-29 motor component.

3.1.2 Multiple Sclerosis Walking Scale - MSWS-12

The Multiple Sclerosis Walking Scale (MSWS-12) is a validated scale developed to assess various aspects of walking ability, including perceived limitations, difficulties, and impact on daily activities related to walking, in people with MS (pwMS) (Hobart et al. 2003). The MSWS-12 consists of 12 items. The scale uses a 5-point Likert-type response format, where individuals rate the extent to which each item reflects their experience of walking ability on a scale from 1 (no limitation) to 5 (severe limitation).

3.1.3 Hospital Anxiety and Depression Scale - HADS

The Hospital Anxiety and Depression Scale (HADS) is a validated scale specifically designed to measure anxiety and depression in individuals with physical health conditions who are receiving medical care (Snaith et al. 1986; Snaith 2003). The HADS consists of 14 items, with 7 items assessing symptoms of anxiety (HADS-A) and 7 items assessing symptoms of depression (HADS-D). Each item is rated on a 4-point Likert-type scale, with scores ranging from 0 to 3, reflecting the extent to which each statement describes individual's feelings over the past week. The total scores for HADS-A and HADS-D can range from 0 to 21, with higher scores indicating higher levels of anxiety or depression.

3.1.4 Fatigue Severity Scale - FSS

The Fatigue Severity Scale (FSS) is a validated scale developed to assess the severity of fatigue experienced by individuals in various conditions, including MS, and its impact on daily functioning (Krupp et al. 1989). The FSS consists of 9 items which assess the severity of fatigue symptoms over the past

week. The scale uses a 6-point Likert-type response format, where individuals rate their agreement with each statement on a scale, ranging from 1 to 7, with higher scores indicating greater fatigue severity. There is no established cutoff score for clinically significant fatigue, so interpretation of the scores may vary depending on the context and population being assessed. The scale is widely used in clinical research and practice to monitor changes in fatigue severity over time and evaluate the effectiveness of interventions targeting fatigue.

3.1.5 EuroQoL-5 Dimensions - EQ5D

The EuroQoL-5 Dimensions (EQ5D) is a standardised scale used to assess the impact of health conditions and interventions on an individual's overall health-related quality of life (EuroQol Group 1990). It can be used for comparisons across different populations, conditions, and interventions. The EQ5D consists of two parts. The first part is a descriptive system that assesses health status across five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels (no problems, some problems, or severe problems), resulting in a total of 243 possible health states. The second part is a visual analog scale (VAS) where individuals rate their overall health status on a scale from 0 (worst imaginable health) to 100 (best imaginable health).

3.1.6 Web Expanded Disability Status Scale - webEDSS

The web Expanded Disability Status Scale (webEDSS) is an online version of the EDSS (Kurtzke 1983), a widely used clinical tool to assess disability status and monitor disease progression in pwMS. The EDSS is a scale that ranges from 0 to 10 in half-point increments, with higher scores indicating higher levels of disability. It is based on an evaluation of various functional systems, including motor function, sensory function, coordination and ataxia, visual function, bowel and bladder function, cerebral function, and ambulation. The webEDSS can be administered and scored electronically, typically through a web-based interface; therefore, it offers advantages such as improved accessibility, remote assessments, standardised administration procedures, and the ability to store and analyse data electronically. The webEDSS demonstrated high agreement with the traditional in-person EDSS assessment, suggesting its reliability and validity as a tool for assessing disability severity in pwMS through a web-based platform (Leddy et al. 2013).

3.2 Online cognitive assessment using Cognitron

3.2.1 The online cognitive tasks

In this thesis, I analysed data collected using a set of 24 online cognitive tasks from the Cognitron web-based testing platform (Section 2.10.3).

The tasks were selected for the following reasons:

- They were engaging, brief (run in under 5 minutes) and suitable for online unsupervised deployment.

- Extensive normative data were available on them.
- They comprehensively measured different aspects of cognition, spanning a broad range of cognitive domains.
- They had already been deployed in the context of other neurological populations, enabling cross-clinical group comparisons.

Here I report a list of the tasks selected and for each of them, a brief description and the underlying cognitive domain/s that it was designed to measure.

1. Bees

- **Description:** The participant is presented with a set of “hives”. Each hive contains two sets of bees with different colours. The participant must guess which of the sets of bees is most numerous by uncovering one bee at a time at random. The more bees they uncover the more certain they can be, but the longer it will take.
- **Cognitive domain/s measured:** Impulsivity

2. Blocks

- **Description:** The participant is presented with a set of coloured blocks in a grid and a target configuration made of a subset of the blocks. The participant must remove coloured blocks to make it match the target configuration. Blocks fall with gravity.
- **Cognitive domain/s measured:** 2D spatial planning, visual problem-solving

3. Card Pairs

- **Description:** This is a variant on the classic Card Pairs game. The participant is shown a set of pairs of cards. After a set time period, the cards are turned over and the participant must select cards and match them with their corresponding pair.
- **Cognitive domain/s measured:** Associative working memory

4. Complex Reaction Time (CRT)

- **Description:** The participant is presented with a string of arrows pointing left or right, they must click on the side of the screen indicated by the arrow as quickly as possible.
- **Cognitive domain/s measured:** Information-processing speed

5. Digit Span

- **Description:** The participant is presented with a string of digits that they must recall. After the participant correctly remembers a string of digits, the next string increases in length by one. The game ends when the participant answers incorrectly three times for a given string length.
- **Cognitive domain/s measured:** Working memory

6. Emotion Discrimination

- **Description:** The participant is presented with a string of pairs of faces expressing emotions. They must discern whether the emotions on the faces are the same or different to each other.
- **Cognitive domain/s measured:** Emotional intelligence

7. Four Towers

- **Description:** The participant is given a set of four 3D perspective images with buildings and trees. Three of the images are the same but rotated of 90, 180 or 270 degrees, one is different. The participant must select which image is different.
- **Cognitive domain/s measured:** 3D spatial reasoning, visual problem-solving

8. 2D Manipulations

- **Description:** The participant is presented with a grid of coloured squares and 4 probe grids. One of the probe grids is the same as the target but has been rotated of 90, 180 or 270 degrees, the other probes have been subtly changed from the target grid. The participant must identify the probe that is the same configuration as the target.
- **Cognitive domain/s measured:** 2D spatial reasoning, visual problem-solving

9. Motor Control

- **Description:** The participant is presented with a string of small targets that appear all over the screen. They must click on the targets as quickly and accurately as possible.
- **Cognitive domain/s measured:** Hand-eye processing, motor abilities, information-processing speed

10. Objects Memory Immediate

- **Description:** The participant is presented with a string of object silhouettes that they must remember. They are then presented with a string of grids each containing one of the original objects that were displayed previously. Along with the target object, there are also 7 other distractor objects that range in similarity to the target object over three parameters (orientation, pose and item). The participant must select the target object from the distractor objects, for each grid. This task is performed at the beginning of the battery and is paired with a follow up task at the end of the battery.
- **Cognitive domain/s measured:** Short-term object recognition memory

11. Objects Memory Delayed

- **Description:** This task is paired with the Objects Memory Immediate task. The participant is presented again the string of grids without being reminded of the original string of target objects and they must again pick out the targets from the grids. The difference is in the delay between objects presentation and objects recall.
- **Cognitive domain/s measured:** Medium-term object recognition memory

12. Paired Associates Learning (PAL)

- **Description:** The participant is presented with a grid in which a series of images appear one by one in different grid squares. The participant is then presented with each image they saw previously, and they must indicate which grid square it appeared in.
- **Cognitive domain/s measured:** Associative working memory

13. Picture Completion

- **Description:** The participant is presented with a set of images that have pieces missing from them, and a set of pieces on the side of the image. The pieces may have been rotated from their original orientation. The participant must select which of the missing pieces goes in which of the slots in the image. In some of the trials there are also distractor pieces.
- **Cognitive domain/s measured:** 2D spatial reasoning, visual problem-solving

14. Simple Reaction Time (SRT)

- **Description:** The participant is presented with a string of targets, they must click anywhere on the screen as soon as the target appears.
- **Cognitive domain/s measured:** Information-processing speed

15. Spatial Span

- **Description:** The participant is presented with a string of locations on a grid that they must recall. After the participant correctly remembers a string of positions, the next string increases in length by one. The game ends when the participant answers incorrectly three times for a given string length.
- **Cognitive domain/s measured:** Working memory

16. Spotter

- **Description:** The participant is presented with a string of grey static grids that flash quickly onto the screen and then off again. Some of the grids contain numbers with varying degrees of decipherability, the participant must pay close attention to the grids and click the screen every time they see a “0” in one of the grids. The task is long and tests the participants ability to maintain focus.
- **Cognitive domain/s measured:** Sustained attention

17. Switching Stroop

- **Description:** The participant is presented with a coloured block (either red or blue) and on either side the words “red” and “blue”, these words are drawn with either red or blue ink. At the top of the task is a condition, the condition reads either “ink” or “text”. If the condition reads “ink” then the participant must click on the word that has the same ink colour as the square in the middle. If the condition reads “text” then the participant must click on the word that describes the colour of the square in the middle. The condition switches pseudo-randomly as the task progresses.
- **Cognitive domain/s measured:** Cognitive control, executive functions

18. Target Detection:

- **Description:** The participant is presented with an abstract shape as a target. They are then presented with a constantly changing set of different abstract shapes and they must quickly click the target shape every time it appears, while ignoring the distractor shapes.
- **Cognitive domain/s measured:** Attention

19. Tower of London

- **Description:** This is a variant on the classic Tower of London task. The participant is presented with 2 sets of 3 pegs with coloured rings on them, they must work out the minimum number of moves it would take to get from one configuration of rings to the other.
- **Cognitive domain/s measured:** 2D spatial planning, visual problem-solving

20. Trail Making

- **Description:** The participant is presented with a set of numbered tiles, which they must click in order from 1 upwards. They are then presented with a set of tiles with numbers and letters, they must alternate between clicking on the numbers and letters, but in both cases click in ascending order (e.g. 1-A-2-B-3-C-4-D-5-E...).
- **Cognitive domain/s measured:** Information-processing speed, cognitive flexibility

21. Verbal Analogies

- **Description:** The participant is presented with a set of analogies and must decide if the analogy is true or false. For example a true analogy would be “Apple is to fruit as hammer is to tool”, whereas an incorrect analogy would be “Apple is to fruit as large is to small”.
- **Cognitive domain/s measured:** Verbal reasoning

22. Word Definitions

- **Description:** The participant is given a set of English words with varying rarity and they must select the correct definition from a choice of 4 definitions. They have 20 seconds per word.
- **Cognitive domain/s measured:** Crystallised intelligence

23. Words Memory Immediate

- **Description:** The participant is presented with a string of words. They are then presented with another set of words that may or may not have been in the original string. They must identify which words in this new set were in the original string. This task is performed at the beginning of the battery and is paired with a follow up task at the end of the battery.
- **Cognitive domain/s measured:** Short-term word recognition memory

24. Words Memory Delayed

- **Description:** This task is paired with the Words Memory Immediate task. The participant is presented with a set of words that may or may not have been in the Words Memory Immediate task at the beginning of the battery. They must identify which words in this new set were in the original string and which were not. The difference is in the delay between words presentation and words recall.
- **Cognitive domain/s measured:** Medium-term word recognition memory

3.2.2 Information collected for each task

For each task, at the time of performance, task-specific information is recorded to comprehensively describe users' performance on that task. This task-specific information includes the stimulus presented, the response given, the correctness of the response, and the response time, all recorded on a trial-by-trial basis. Essentially every on-screen activity is recorded. This trial-by-trial record contains valuable and detailed information on users' behaviour but needs to be condensed into an appropriate set of task-specific summary performance measures to facilitate analysis.

Additionally, various other aspects are recorded for each task, which are factors that could influence task performance or in general prove useful during data pre-processing. These measures are the same for every task and reported in Table 3.1 together with their scope.

Measure name	Description	Use
user_id	A unique encrypted identifier for each user	To link the different tasks performed by the same user
device, os and browser	Device, operating system and browser on which the task was performed	To account for the effect of using different devices with different delays on the performance and to favour tasks which are not sensitive to device
timeStamp	When the task was performed	To see whether task was performed on the same day as another task (in case of dependency between tasks) or to account for things such as seasonality
startTime, endTime, duration	How long it took to perform the task	To identify outliers if the task was left open for too long, or to automatically identify tasks with a fixed duration during pre-processing
assessment_duration	If the task is part of a battery of tasks, how long it took to perform the whole battery	To see whether the assessment was performed in a single session or on multiple sessions/days
completed_assessment	If the task is part of a battery of tasks, if the battery was completed or not	Important in case of tasks dependency, or to identify the users for whom complete data are available
rawdata	Trial-by-trial information collected on the task	It is from the rawdata that task-specific summary performance measures are evaluated
focusLossCount	How many times focus was lost while performing the task	To discard people who are not compliant, e.g. distracted or cheating
timeOffScreen	For how long focus was lost while performing the task	To discard people who are not compliant, e.g. distracted or cheating

Table 3.1: Measures collected at the time of task performance, including a description of what they are and of the way in which they are used.

3.2.3 Normative data

In this thesis, I utilised normative data collected as part of a large-scale online longitudinal population study, called the Great British Intelligence Test (GBIT), conducted between December 2019 and February 2022 (Hampshire 2020; Hampshire et al. 2020). The study involved 2 baseline timepoints and 4 follow-up timepoints. The baseline timepoints were collected between December 2019 and January 2020, and between May 2020 and July 2020, respectively. The follow-up timepoints were collected between December 2020 and January 2021, June 2021 and August 2021, January 2022 and February 2022, and December 2022 and January 2023. Over 300,000 people participated in the study, which was promoted in a documentary aired on BBC Two in May 2020 as part of the “Horizon” science series (Mosley 2020).

The study aimed to gather large-scale population data on some of the online cognitive tasks available on the Cognitron platform, including the 24 tasks evaluated in this thesis. The main objective was to investigate patterns in cognitive performance across different populations and explore potential links between cognitive function and various factors, such as sociodemographic, environmental, and lifestyle factors, technology use, mental health, and neurological conditions.

Each data collection timepoint comprised an online questionnaire and a set of tasks that could be completed from any device, including mobile phones, desktop computers, laptops, and tablets. New questions were introduced at each timepoint while old ones were removed if they pertained to unchanging aspects. For this thesis, only sociodemographic data (e.g., age, gender, dominant hand, first language, ethnicity, country of residence, education, and occupation) and information on whether participants had any neurological or psychiatric conditions from the collected questionnaires were utilised. Participants were presented with a different set of tasks at each timepoint, comprising a fixed subset administered consistently and a variable subset changing across timepoints to enable evaluating additional tasks. The fixed subset was administered to all participants, whereas tasks in the variable subset were assigned using a random sampling to prevent cognitive fatigue arising from prolonged testing sessions. Consequently, the number of observations available varies widely across tasks depending on when it was introduced (at which timepoint) and at how many timepoints it was collected. While limited observations are available for tasks collected via random sampling, there are at least 3,500 observations available for each task, providing a substantial normative dataset for this thesis.

3.3 Pre-processing pipeline for online-collected large-scale cognitive data

The quality and reliability of online-collected data can be compromised by various factors, including missing or unreliable information, outliers arising from not properly engaging with the tool, and incomplete assessments, posing challenges during data analysis. These issues are exacerbated in online cognitive assessments as tasks often involve complex cognitive processes that can be challenging for some participants and require a higher level of attention and commitment compared to completing simple questionnaires. Participants may misunderstand the instructions, experience cognitive fatigue, or become distracted, leading to unreliable data. Additionally, participants may stop and restart the assessment midway due to various reasons, such as technical issues or interruptions, leading to

incomplete assessments and data that inaccurately reflect the participant’s cognitive abilities.

Since online cognitive tasks are typically administered without supervision, there is no mechanism in place to ensure that participants complete them correctly. Consequently, it becomes essential to retrospectively examine the collected data to perform quality checks and identify outliers or problems that may have arisen due to participant error or software malfunction. While manual inspection of data is feasible for small-scale studies, it is impractical and inefficient for large-scale datasets.

To address these challenges, I have collaborated with another member of my lab to develop a suite of algorithms to automatically clean and pre-process online-collected large-scale cognitive data. These algorithms ensure data reliability and accuracy by detecting and removing unreliable performances, system errors, and incomplete assessments, while preserving the integrity of the data.

In the following sections, I outline the pre-processing pipeline for online-collected large-scale cognitive data developed ad-hoc for this thesis, which incorporates some of these cleaning algorithms. Additionally, I go through the pre-processing steps applied to the normative data as a practical example. Finally, I explain how some parts of the pipeline were adapted to pre-process patient data.

3.3.1 Pre-processing pipeline overview

The pre-processing pipeline developed for the cognitive data analysed in this thesis comprises four main components, each addressing distinct aspects of the data:

- **Extraction of summary performance metrics.** The first component summarises the detailed trial-by-trial log collected for each task into meaningful summary performance metrics, facilitating subsequent analysis.
- **Data quality checks.** The second component identifies and removes duplicates, server errors, and corrupted data, which are present in the data but are invalid for analysis.
- **Distributions’ tails cleaning.** The third component identifies and removes outliers resulting from non-compliance with the task by leveraging the shape of the distribution of the extracted summary performance metrics.
- **Cleaning based on sociodemographics.** The fourth component merges the task data with the corresponding questionnaire data required for analysis, excluding observations from individuals lacking valid questionnaire data.

The following sections go through each of the components and the problems they are addressing to explain the proposed solutions in detail.

3.3.2 Extraction of summary performance metrics

The recording of all on-screen activities during online cognitive tasks generates extensive performance logs, which are challenging to analyse, particularly with large-scale data involving thousands of individuals. To address this challenge, it is essential to summarise these detailed performance logs into

meaningful summary performance metrics. This section outlines the process of deriving summary performance metrics for each employed task.

Overall, response accuracy and reaction latency are pivotal features of any cognitive task, providing valuable insights into an individual's cognitive abilities and processing speed. Consequently, a summary measure of accuracy and response latency were derived for each task, serving as comprehensive indicators of task performance.

3.3.2.1 Summary accuracy metrics

A summary measure of accuracy can be derived by summarising response correctness across trials. For tasks with a fixed number of trials, accuracy was quantified by calculating the percentage of correct responses. However, in tasks with a variable number of trials, relying solely on the percentage of correct responses may be insufficient, as two individuals with the same percentage could have completed a significantly different numbers of trials. Therefore, for timed tasks such as Manipulations 2D and Verbal Analogies, the total number of correct responses was utilised as a more representative measure of accuracy. Additionally, for tasks involving the detection of a target (e.g., Target Detection and Spotter), the percentage of correct clicks was computed alongside the number of targets detected, as the latter could be inflated by continuously clicking on the screen.

3.3.2.2 Summary latency metrics

A summary measure of latency can be obtained by calculating the mean, median, or sum of reaction times across trials. For tasks requiring a single response (e.g., Tower of London, Four Towers, and Switching Stroop), the median reaction time across trials was favoured as more robust to outliers compared to the mean. For tasks with timed trials (e.g., SRT, CRT, and Word Definitions), the median reaction time was computed only across trials that did not time out. In memory tasks (e.g., Words/Objects Memory Immediate and Delayed, Digit Span, Spatial Span, and PAL), the summary latency was evaluated as the median over correct trials exclusively, ensuring that the measure reflected the time taken to accurately remember information. For tasks involving multiple moves or responses per trial (e.g., Card Pairs and Trail Making), the mean or sum of the total time to complete each trial was calculated instead of the time taken for each move. The sum was preferred to the mean particularly when the number of moves varied across trials (e.g., Picture Completion). Finally, in tasks involving the detection of a target, the median reaction time was assessed only across trials where a target was detected.

3.3.2.3 Additional summary performance metrics

Some tasks necessitate additional measures to comprehensively summarise performance. For instance, the Bees task, designed to gauge impulsivity, requires more than just the percentage of correct responses to accurately capture impulsive behaviour. Additional metrics, such as the mean number of samples and the mean difference of samples of the two kinds required to make a decision, were evaluated to provide a more direct representation of this trait, as inversely proportional to impulsivity.

Certain tasks are more complex due to changing conditions or rules throughout the task, leading to

variations in difficulty across trials and requiring participants to adapt their strategies for accurate responses. This additional mental effort or resources needed to complete trials under more challenging conditions, compared to easier ones, is termed “cost”. Cost can manifest through increased response time, decreased accuracy, and heightened levels of mental fatigue or stress. To quantify the cost for these tasks, the difference between summary performance measures between the harder and easier conditions was calculated, both in terms of accuracy and latency, across trials presenting each condition. A negative accuracy cost and a positive latency cost were expected when transitioning to a harder condition. Switch costs, incongruency costs, and distractor costs were assessed across some of the tasks to understand how individuals adapted to changing task conditions and managed conflicting or distracting information.

Furthermore, six tasks were intrinsically paired: SRT with CRT, Words Memory Immediate with Words Memory Delayed, and Objects Memory Immediate with Objects Memory Delayed. CRT is a more complex version of SRT, requiring participants to choose the correct side to click on rather than simply clicking when the stimulus appears. In addition to primary measures of accuracy and latency, an additional performance measure, the “choice cost”, was introduced to capture the increased difficulty of CRT compared to SRT, which is reflected in an increased latency. Similarly, Words/Objects Memory Delayed is a delayed version of Words/Objects Memory Immediate, with the former occurring at the end of the battery of tasks and the latter at the beginning, assessing medium and short-term word/object recognition memory, respectively. To compare performance between these tasks, a new performance measure, the “percent forgetting” score, was introduced, representing the percentage of items forgotten in the transition from short-term to medium-term memory.

3.3.3 Data quality checks

Large-scale online data collection poses technical challenges due to high website traffic and server crashes, leading to issues such as duplicated, incomplete, and corrupted data. These issues are exacerbated in online cognitive assessments, where interruptions can result in task or battery restarts. Given the large number of participants and the high traffic volume, the potential for technical issues in the GBIT study was substantial. Therefore, it was necessary to implement quality checks on the collected data to detect and address the resulting issues, thereby ensuring data reliability and validity.

Performing data quality checks on the cognitive data from the GBIT study, four types of problems were found:

1. **Completely duplicated entries** - data for one individual were saved two times (same user id and same task data).
2. **Multiple task data with the same user id** - individuals performed the same task more than once in a battery (same user id but different task data).
3. **Identical task data with different user ids** - the same task data were saved two or more times and assigned different user ids at the time of saving (different user ids but same task data).
4. **Corrupted task data** - incomplete, invalid or empty task data were saved.

To address these issues, the following solutions were implemented. For problem 1) all duplicated entries

were removed and only the first occurrence of task data was retained for each user. For problem 4) all corrupted task data was discarded because it was not possible to extract valid summary performance metrics from such data. For problem 2) the approach varied based on the study’s focus. In the context of this thesis, as long as task performance was legitimate, all data instances from all users were retained. However, for longitudinal studies necessitating performance comparison across timepoints, users are expected to have completed each task the same number of times in order for learning effects not to be a confounding factor. In those cases, all task data from users who completed the task more times than they should have, should be discarded.

The solution for problem 3) was more articulated. Instances of identical data across different user ids were improbable and suggested that data from a single individual were saved multiple times with different user ids. In such cases, only the data corresponding to the original user id were retained. In some cases, it was possible to identify which of the user ids with duplicated data was the original one, if looking at the data from the other tasks, only one of the user ids figured in there. Nonetheless, some duplicated task data still remained that corresponded to user ids in other tasks. Given the scale of available data, a conservative approach was adopted, removing all data across tasks corresponding to user ids with identical data on a task where the original user id could not be identified.

3.3.4 Distributions’ tails cleaning algorithms

The purpose of distributions’ tails clearing algorithms is to exploit the shape of the distribution of certain summary performance metrics to automatically detect and remove outliers from the data arising from non-compliant task performance.

3.3.4.1 Right-tail cleaning of reaction time distribution

An example of non-compliant behaviour can be observed when participants leave tasks open for extended periods while engaging in other activities, especially in untimed tasks where responses are needed to progress. This behaviour results in disproportionately high reaction times, skewing the reaction time distribution towards higher values with a sparsely populated right-tail. Given that some degree of cognitive processing slowing is expected in certain populations, it’s vital to remove outliers from the right-tail while retaining valid observations characterised by higher reaction times.

To address this issue, an algorithm has been developed to detect isolated and clearly identifiable outliers by examining the right tail of the reaction time distribution. This algorithm utilises pairwise differences between consecutive reaction times to pinpoint the threshold at which subsequent reaction times diverge significantly. Subsequently, all observations beyond this threshold are eliminated. The algorithm is implemented only when the distribution is based on a minimum of 200 observations, to avoid excessively noisy distribution and preventing the removal of genuine data. The algorithmic steps are summarised below:

“Firstly, all RTs different from zero are extracted and sorted in ascending order. Then, the pairwise differences between each pair of consecutive RTs are evaluated, and the mean and standard deviation of the distribution of differences are calculated. Next, the 95% quantile of the RT distribution is evaluated, and this value is set as the beginning of the right-tail of the distribution. The algorithm is applied only if the standard deviation of the distribution of differences

is greater than 100 times the mean, as the presence of outliers would manifest in a higher standard deviation. The differences that are larger than the mean plus one standard deviation are then detected, and the minimum RT used to calculate each of these differences is extracted. Of the resulting RTs, only those that are higher than the 95% quantile are kept (i.e., those located in the right-tail). The minimum RT of the resulting RTs is then used as a lower threshold, above which the RTs of the observations are considered too high and subsequently removed as outliers. This algorithm is applied only if the standard deviation of the distribution of differences is greater than 100 times the mean to be conservative and drop observations only if they are clearly outliers. The number 100 was selected as the multiplier for the mean after trying different values and observing where the threshold was placed on the distribution as a result of this. This parameter was carefully manually tuned for this application, but it needs to be changed and re-tuned if this same algorithm is used for other applications.”

3.3.4.2 Left-tail cleaning of reaction time distribution

Another form of non-compliant behaviour occurs when participants systematically select random responses, rushing through trials without genuine engagement with the task. This behaviour yields consistently short reaction times, leading to a bimodal reaction time distribution. While a healthy reaction time distribution resembles a slightly right-skewed normal distribution, an unhealthy distribution exhibits a second mode in the left tail. This mode arises from the overlap of two distributions: the distribution of non-compliant fast reaction times and the left-tail of the genuine distribution. This phenomenon is illustrated in Fig. 3.1.

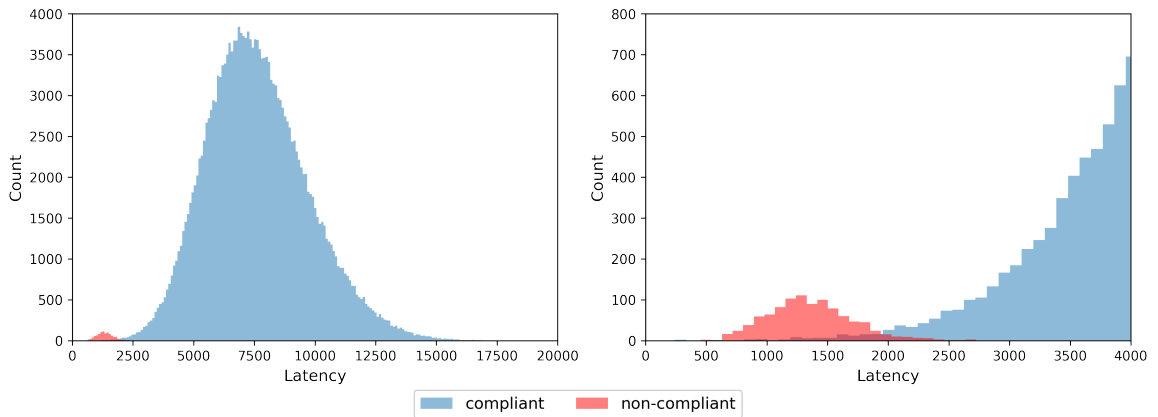


Figure 3.1: Latency distribution including non-compliant users. Illustration of typical latency distribution shape in case of non-compliant behaviours.

The degree of overlap between the two distributions varies depending on the task’s cognitive processing demands. In some instances, there is no overlap because participants cannot respond as quickly as when guessing. However, when overlap is observed, it becomes challenging to eliminate outliers without also excluding compliant participants with fast reaction times. Hence, a balance must be struck between removing outliers and retaining fast participants. The optimal threshold reaction time for discarding all observations with a smaller reaction time should ideally lie at the intersection point of the two distributions. Nonetheless, distributions derived from a sample of the population are inherently noisy, making it difficult to precisely determine this intersection point. Additionally, the chosen bin width for creating the histogram can influence the intersection point, given that reaction time is a continuous variable. Furthermore, in some cases, the number of outliers may be small enough

that they appear as isolated and sparse bins rather than forming a distinct distribution.

To address this issue, an algorithm has been developed to identify the point in the left-tail of the reaction time distribution where the distribution starts to grow monotonically, and subsequently remove all observations prior to that point. The algorithm is implemented only when the distribution is based on a minimum of 200 observations to avoid discarding genuine data because of an excessively noisy distribution. The algorithmic steps are summarised below:

“First, all RTs different from zero are extracted and sorted in ascending order. Next, the 95% quantile of the RT distribution is evaluated, divided by 100, and the resulting number is rounded to the nearest integer and set to be the bin width. The RTs are then binned using this bin width and the count of RTs in each bin is evaluated. If only one bin is present, the process is stopped. If the number of observations available to form the distribution is less than 20,000, each count is smoothed using a rolling window of 3 by replacing each count with the average of the count with the two previous counts. This was done to further reduce the effect of noise. Starting from the most left bin of the distribution, the height of the bin (corresponding to the count of RTs in that bin) is compared to the height of the following bin. If the height of the following bin is lower than the height of the previous one, then the RTs included in the previous bin are marked as unreliable. The process is repeated until the 5% quantile of the RTs distribution is reached, as only the left-tail is searched to avoid removing more than 5% of the data. The maximum RT of the last bin marked as unreliable is treated as the RT threshold below which participants were too fast to have performed the task compliantly. Finally, all observations corresponding to RTs strictly less than the RT threshold are dropped.”

3.3.4.3 Left-tail cleaning of accuracy distribution

Similar to the previously described scenario, systematically selecting random responses leads to poor performance in terms of accuracy, with participants typically performing at chance level. This behaviour is evident in tasks with a fixed number of trials that require participants to respond to proceed. Conversely, in tasks with timed trials or where trial time can expire, participants may complete the task providing minimal or no responses, resulting in very low summary accuracy.

These non-compliant behaviours manifest as a lower second mode in the left tail of the accuracy distribution. Since many of these behaviours are also associated with unhealthy reaction times (e.g., short reaction times when skimming through trials in tasks with a fixed number of trials or long reaction times when leaving the task open without responding in timed tasks), they should be detected by the previous algorithms and removed. However, some non-compliant behaviours may still persist after the initial cleaning algorithms, especially those associated with reaction times that fall within the overlap area between compliant and non-compliant distributions. Therefore, following the cleaning of the right and left tails of the reaction time distribution, a similar cleaning procedure is applied to the left tail of the accuracy distribution.

Given that reaction time is a continuous variable, selecting an appropriate bin width is essential for accurately representing its distribution. The chosen bin width impacts the resulting histogram shape and consequently the cleaning process. It involves a trade-off between a width that is too small, resulting in a noisy histogram, and one that is too large, obscuring the non-compliant distribution by including outliers in the first bin with compliant observations.

This issue was avoided in the case of accuracy, as a discrete variable was utilised to represent it for most tasks which enabled using a bin width of one. The accuracy metric chosen was the number of

correct answers for tasks for which this variable was available and the maximum achieved for span tasks (Digit Span, Spatial Span, and PAL). For tasks primarily designed to measure latency, such as Motor Control and Trail Making, no cleaning based on accuracy was performed.

The algorithm designed to clean the left tail of the accuracy distribution is similar to the one developed for the reaction time distribution, but it has been adapted to work with discrete variables and only use a bin width of one. Furthermore, due to the limited range of possible values that the accuracy variable could take for each task, the resulting accuracy histogram was less noisy compared to the reaction time histogram, rendering smoothing of the counts unnecessary. The steps applied in the algorithms are summarised below:

“First, all accuracy scores are extracted and sorted in ascending order. Next, a bin width of one is used to group the scores into bins ranging from the minimum score to the maximum score. In most cases the accuracy is a discrete variable; consequently, each bin corresponds to a single score. The count of observations in each bin is evaluated. If only one bin is present, the process is stopped here. Otherwise, starting from the leftmost bin, the height of the bin (i.e., the count of observations in that bin) is compared to the height of the following bin. If the height of the following bin is lower than the height of the previous bin, the observations in the previous bin are marked as unreliable. This process is repeated until the 5% quantile of the accuracy distribution is reached. We are only considering the left-tail of the distribution to ensure that no more than 5% of the data is discarded. If the left quantile falls in the first bin, the first bin is searched. The score corresponding to the bin after the last bin marked as unreliable is treated as the score threshold below which participants’ performance was considered too poor for them to have genuinely engaged with the task. Finally, all observations corresponding to scores strictly less than the score threshold are dropped.”

3.3.4.4 Adaptation to patient data: where to evaluate the thresholds

When applying the distributions’ tails cleaning algorithms to patient data, some adjustments are necessary. Specifically, reaction times or accuracy scores considered anomalous for controls may be valid for cognitively impaired patients.

There are advantages and disadvantages to using thresholds derived from control data for cleaning patient data or evaluating ad-hoc thresholds based on patient distributions. Using thresholds from control data benefits from a larger sample size, providing a better approximation of the true distribution. However, this approach may mistakenly identify slower reaction times or lower accuracy scores in patient data as outliers. Conversely, patient data is less abundant, leading to potentially noisier distributions with narrower value ranges.

To address this challenge, a hybrid approach has been developed. This approach involves evaluating thresholds separately for both control and patient distributions and selecting as final threshold for patient data the most conservative between the two. The term “most conservative” refers to choosing the lower value when cleaning the left tail (to retain highly impaired patients) and the higher value when cleaning the right tail (to preserve exceptionally slow performances in impaired patients). This approach combines the advantages of using large-scale control data while ensuring that cognitively impaired individuals are not erroneously classified as outliers.

3.3.5 Cleaning based on sociodemographics

To ensure a culturally unbiased assessment, it is essential to collect and account for sociodemographic factors which are known to impact cognitive performance and potentially confound results. While sociodemographic information was collected as part of the GBIT study, not all participants completed the questionnaire or provided valid responses. This component of the pipeline only retains task data from users who provided valid answers to questions pertaining to the following sociodemographic variables: age, gender, dominant hand, ethnicity, first language, country of residence, education, and occupation.

To achieve this, task data from users who did not complete the sociodemographic questionnaire at all were excluded. Subsequently, users providing at least one invalid and non-imputable response to the essential sociodemographic questions were discarded. Here, "non-imputable" refers specifically to age, as missing values for age could be imputed if age was reported at one timepoint but not at another. Additionally, users reporting ages below 16 or above 90 were excluded. Furthermore, users reporting neurological or psychiatric conditions were also discarded, as they cannot serve as appropriate controls.

Finally, the following pre-processing steps were conducted on the sociodemographic data of the users retained:

- Age was categorised into decades.
- Language was binary-encoded as "English" or "Other" due to the under representation of certain language categories.
- Residence was binary-encoded as "United Kingdom" or "Abroad" due to the under representation of some residence categories.
- Ethnicities were grouped according to categories utilised in the Census, namely: 1) Asian or Asian British, 2) Black, Black British, Caribbean, or African, 3) Mixed or multiple ethnic groups, 4) White, and 5) Other ethnic groups.

3.3.6 Conclusions

While the implemented cleaning measures and algorithms were deemed sufficient for the scope of this work, it's important to acknowledge their limitations and propose potential improvements. The primary limitation lies in the arbitrary nature of certain numerical values utilised in the distributions' tails cleaning algorithms. These values were manually tuned and tailored to the specific dataset under examination. However, the underlying rationale of these algorithms can be extended to similar types of online-collected data, including questionnaire data when reaction times are being recorded. In such cases, the parameters would need to be re-adjusted and tuned to the specific application.

To address these limitations and enhance the cleaning process, an ongoing project is being conducted in collaboration with Valentina Giunchiglia. The objective of this project is to further refine and generalise the existing pipeline and algorithms. Specifically, efforts are being made to reduce the number of parameters requiring fine-tuning and explore the feasibility of deriving them in a data-driven manner as part of the algorithms.

By addressing these aspects, the cleaning measures and algorithms can be enhanced to improve their applicability and effectiveness in various online data collection scenarios, thereby advancing the quality and reliability of the analyses conducted.

Chapter 4

Validity and utility of patient-reported outcomes to assess physical disability in Multiple Sclerosis

In this chapter, I present a prospective study where I analysed 11 years of patient-reported outcome data from the UK MS Register. I developed a novel approach to maximise data usage while coping with missing data and accounting for variable levels of engagement as a potential confounding factor. I used this approach to investigate differences in two of the motor-related patient-reported outcomes collected by the Register, the Multiple Sclerosis Impact Scale motor component and the Multiple Sclerosis Walking Scale, as a function of subtype, disease duration, and future transition to a secondary progressive course. As a result, I found robust significant associations between physical disability and disease subtype and duration, indicating progressive deterioration over time for all subtypes but Benign and significant differences among all subtypes.

4.1 Introduction

In chronic conditions such as Multiple Sclerosis (MS), the limitations of randomised controlled trials (RCTs) have been highlighted by their short length, strict entry criteria and limited ability to determine comparative effectiveness. This is compounded by an emerging view that early treatment is the optimal strategy to maximise patients’ long-term benefits (Cerqueira et al. 2018). This need for prolonged follow-up means that registries are becoming the only viable way of determining the long-term impact of treatments. Registries collect “real-life” data, without a directed research intervention but can be retrospectively analysed to compare outcomes based on a particular exposure. Successful registries at a national (Hillert et al. 2015) and international level (Kalincik et al. 2019) have delivered key insights focusing predominantly on long-term disability, a feature missing from RCTs. Traditional registries have, in turn, been extended by online registries that have widened the range and increased the scale and frequency of data that can be collected, including patient-reported outcomes (PROs) (Rodgers et al. 2022). Critically, online registries offer the opportunity for cost-effective home-based longer-term follow-up; this is of particular interest in MS, a disease with a variable outcome, which evolves for some over the short-term but for the majority over the long-term.

Registries capture what can be described as “sparse” data, as information is not collected regularly at equidistance timepoints, like in a clinical trial, nor it is completed every time by every participant. Data are usually collected in routine clinical practice, most commonly face-to-face, and submitted periodically on an obligatory or voluntary basis. This can result in attrition bias arising due to differences in follow-up and care patterns and detection bias where there are variations in the density and duration of follow-up between centres. Lastly, informed censoring occurs when subjects drop out due to surrogate results (Kalincik et al. 2019; Sormani 2020; Tur et al. 2019). A further issue unique to online MS registries is that the most common outcome used and favoured by regulators, the Expanded Disability Status Scale (EDSS), requires face-to-face contact (Barkhof et al. 2009; Kurtzke 1983). The EDSS is an ordinal scale based on a neurological examination at lower scores, measured walking distance in the middle range and finally on function at high scores (Hobart et al. 2000). The EDSS has several problems that complicate its use but ironically its major issue is the variability that arises from the requirement to be performed by a trained examiner. This issue is exacerbated in large longitudinal datasets where many examiners are involved who may not follow a standardised protocol as required in trials. Telephone, online and PRO versions have been developed but have differing characteristics to the classic version (Kuri et al. 2021; Lechner-Scott et al. 2003; Leddy et al. 2013).

Alternative outcome measures have been the focus of much work with regulators (Cohen et al. 2012; Committee for Medicinal Products for Human Use (CHMP) 2015). PROs provide one such alternative that aims to collect data entered by patients capturing a range of MS impacts. PROs are potentially a useful approach for registries as they ensure data are entered for the same person by the same person over the longer-term and can be completed conveniently over the internet. The use of PROs as disease biomarkers has been long debated (Pellegrini et al. 2019). For adoption by the wider community, measures not only need internal and external validity but also must reliably quantify change over time, e.g., capture how disability is modulated by age and disease duration, differentiate among disease subtypes, and ideally anticipate progression between subtypes (Committee for Medicinal Products for Human Use (CHMP) 2015; Food and Drug Administration 2009).

The United Kingdom Multiple Sclerosis Register (UKMSR) is a registry initiative focused on MS

that has been active since 2011 in the UK (Middleton et al. 2018). The UKMSR serves as a way for people with MS (pwMS) to connect with one another, report their personal views on their disease and engage with research. It consists of an internet portal where pwMS can take part independently of their healthcare team but also engages with 46 National Health Service (NHS) MS centres where pwMS are recruited, and their healthcare professional and objective outcomes can be recorded in an independent and partially overlapping dataset. This has resulted in a unique database, both in terms of size and longitudinal follow-up time. The UKMSR captures a wide range of disease and non-related information, including demographics of the users, characteristics of the disease at onset, symptomatology at onset and throughout the disease and PROs resulting from the periodic administration of online questionnaires.

In this chapter, I analysed large-scale registry data for two PROs that have been mainstays of the UKMSR – the Multiple Sclerosis Impact Scale (MSIS-29) motor component (Hobart et al. 2001) and the Multiple Sclerosis Walking Scale (MSWS-12) (Hobart et al. 2003) – both focused on physical disability. These PROs were developed to capture the motor element of the EDSS. The main aim was to demonstrate their validity and utility in measuring physical disability in pwMS. First, I developed a simple but flexible statistical approach based on robust permutation modelling to maximise the proportion of available data utilised in the context of sparsity. Then, I validated the measures with respect to their sensitivity to expected changes with disease duration and subtype. Finally, I explored the potential of these uniquely large and longitudinal data provided by a register to derive new insights into the nature of change across time, including patterns in the PRO-measured motor deficits that precede the development of progressive MS.

4.2 Methods

4.2.1 Data collection

The UKMSR is an online, UK-wide register supported by NHS clinical centres (Research Ethics Committee: South-West Central Bristol NRES initially as 16/SW/0194, now 21/SW/0085). The register includes independent verification of treatments and EDSS outcomes from NHS centres in a separate but overlapping population. The register also gathers PROs in the form of questionnaires administered periodically to the users regarding their physical and psychological well-being as well as disease-specific information such as about relapses, treatments, etc. Participants entered data 3-monthly from 2011-2018 and 6-monthly subsequently after receiving email reminders. Since September 2018, participants have a 28-day window in which to complete the PROs, although often they are all completed in 1 day. Demographic data collected include age, gender, and year of disease onset.

4.2.2 Patient-reported outcomes

Two of the PROs collected by the UKMSR were analysed: the MSIS-29 motor and the MSWS-12. The MSIS-29 motor component was used to assess motor abilities. The UKMSR has been collecting data using the MSIS-29 motor component version 1 (MSIS-29v1) (Hobart et al. 2001) prior to April 2012 and version 2 (MSIS-29v2) (Hobart et al. 2009) subsequently. Answers to the 20 questions that form the MSIS-29 motor component are each scored between 1 and 5 in version 1 and between 1 and

4 in version 2. As a consequence of this, the scores for MSIS-29v1 give a total ranging from 20 to 100, whereas those for MSIS-29v2 give a total from 20 to 80. To account for the changes in scales, and leverage all data available, the totals for both versions were rescaled to a value in the range of 0 to 100 using a min-max normalisation procedure fitted separately for each version.

The MSWS-12 version 2 was used to assess walking function (Hobart et al. 2003). The score was normalised as above. Participants were excluded from the MSWS-12 assessment if they indicated that they could not walk (the questions were not relevant to them).

4.2.3 Data curation and handling sparsity

Online self-reported data are by nature large but also sparse; consequently, not all individuals within the UKMSR had complete demographic and clinical records, nor did they complete the available PROs at every collection window.

A conservative approach was adopted to address this sparsity, keeping as many observations as possible for each analysis. The minimum information needed to be included in the study was to have completed at least one of the considered PROs once, to have recorded a legitimate year of birth and of symptoms onset and to have a MS subtype at completion date known or derivable as detailed below. The pre-processing, visualisation, and analysis steps included: (i) assignment of MS subtype labels at individual timepoints; (ii) pooling the observations into disease duration groups; (iii) plotting the PRO-derived trajectories over disease duration stratified by disease subtype and transitioning subtype; (iv) plotting the PROs' distributions at each disease time bin stratified by subtype; (v) evaluating the presence of a main statistical effect of disease duration and subtype on the scores as well as their interaction; (vi) when a main effect was detected, performing post-hoc pairwise comparisons to understand which groups were statistically distinct and in what direction.

4.2.4 Assignment of MS subtype labels at patient-reported outcome timepoints

Disease subtype can change with disease progression and, in some cases, can only be established over time. For this reason, users of the register are asked every 18 months to update their subtype based on what they have been told by their clinical team. Nonetheless, the dates when these updates are made do not always coincide with the dates when users complete the PROs. Consequently, it was necessary to determine the subtype at any given PRO timepoint according to the time course of recorded classifications for each user (Fig. 4.1).

Specifically, there are four main MS subtypes: relapsing-remitting (RRMS), secondary progressive (SPMS), primary progressive (PPMS) and benign (BN). At first diagnosis, pwMS are labelled as RRMS, SPMS, PPMS or do not know (DK). If there are no further episodes, then patients are subsequently classified as BN (it takes at least 15 years from onset to confirm a BN label) (Tallantyre et al. 2019). People with RRMS may transition to SPMS (the estimated median time for transitions to happen is about 19 years from onset) (Rovaris Marco et al. 2006).

For analysis, class labels at any given PRO timepoint were determined based on the time course of subtype answers in the following manner. (i) People with DK classifications early in the time course had these classifications replaced with the first subsequent valid label. (ii) People who were classified

later in the time course as either PPMS or BN had these labels applied to all preceding timepoints retrospectively, as by definition neither subtype can transition from another. (iii) People with SPMS labels at baseline were assumed to remain SPMS for all subsequent timepoints as this subtype cannot progress to another. (iv) People who were designated as RRMS at all timepoints were labelled as such. (v) Timepoints for people who were designated as RRMS at baseline but who transitioned to SPMS were assigned the temporally closest recorded label. Timepoints where the offset between questionnaire completion date and label registration was greater than 2 years were excluded. This cut-off was chosen because transitions occur over time, and it can take over 1 year to confirm a transition, for example, from RRMS to SPMS. Follow-up was for a maximum of 11 years. Consequently, observations labelled as BN and included between 0 and 10 years from onset were so based on a maximum of 11 years follow-up and are indicated with a dotted line in the plots.

Of particular interest is the question of whether there are differences early in the time course of MS for RRMS people who do vs do not subsequently transition to SPMS. In the first part of the analyses, the PROs of individuals who transitioned before or during the follow-up period were grouped together, regardless of whether they were collected during the individuals' RR or SP phase. These observations were retrospectively labelled as SPMS. Consequently, records across four subtypes were analysed and compared, namely: BN, RRMS, SPMS and PPMS. In the second part of the analyses, only PRO timepoints in the RR phase were considered, and they were divided based on whether the respective patients transitioned during the follow-up period. This resulted in two transitioning subtypes: RRNoTrans (which coincides with RRMS from above) and RRTrans (a subset of SPMS from above). A visual representation of the nested grouping is reported in Fig. 4.1. The number of unique users and the total number of MSIS-29 and MSWS-12 records for each group are reported in Fig. 4.2.

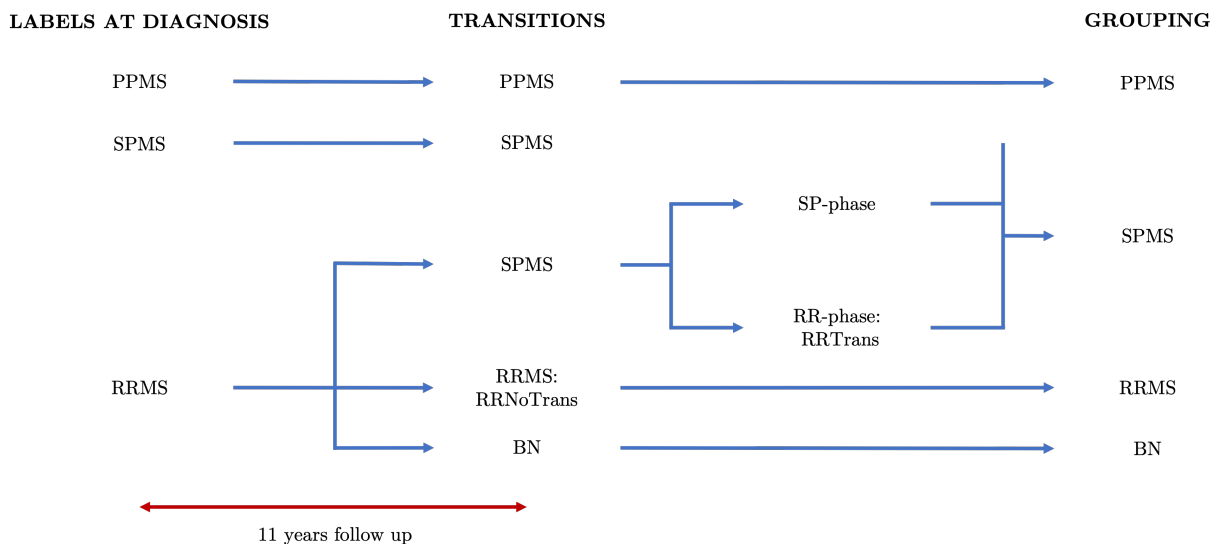


Figure 4.1: Schematic of disease subtype labels assignment. Illustration of how MS subtype labels at questionnaire completion dates were derived from the time course of recorded labels for each user and of how the records were subsequently grouped based on these labels.

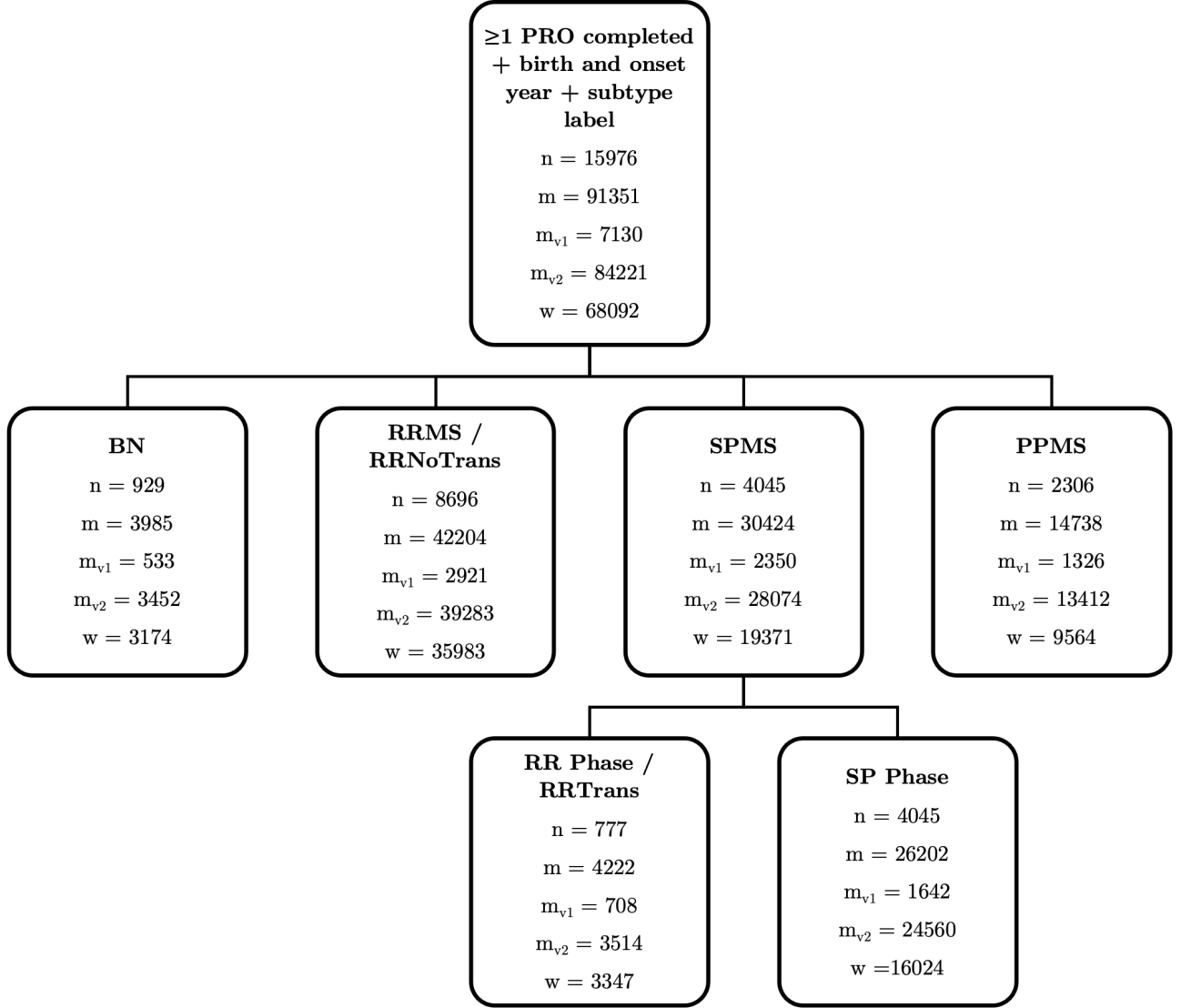


Figure 4.2: Flowchart illustrating how the studied subgroups were derived from the over-all UKMSR population. For each subgroup the number of unique users (n) and the total number of MSIS-29 motor (m) and MSWS-12 (w) questionnaires completed by the users are reported. The flowchart also reports how many of the MSIS-29 motor questionnaires completed were, respectively, version 1 (mv1) and version 2 (mv2) of the questionnaire. The minimal requirements to be included in the study consist of having completed at least one PRO, having reported a legitimate year of birth and of symptoms onset and having disease subtype label at completion date known or derivable.

4.2.5 Patient-reported outcome trajectories with disease duration

The evolution of the mean PRO-derived scores for the different subtypes was examined as a function of disease duration. Observations were grouped into 5-year blocks starting from 0 to 50 years with disease. Disease duration was defined as time from first symptom onset to time of PRO observation. Observations at disease duration > 50 years were formed into one further group due to low frequency.

To prevent repeated observations within time bins, all observations from a single individual were averaged within a time bin. This allowed leveraging all available data towards more representative measurements within each timeframe whilst meeting the assumption of group independence needed for subsequent analyses.

4.2.6 Statistical analysis

Data curation, pre-processing, and statistical analysis were conducted in Python 3.7.3.

Although UKMSR PRO data are powerful in terms of population size and timescale, they present challenges for statistical analysis; specifically, (1) groups' sizes are imbalanced (both in terms of disease time bins and subtypes), (2) data are sparse, (3) the number of observations varies across individuals, and (4) distributions of scores are non-gaussian. A robust pipeline based on Monte Carlo permutation analysis was developed to address these challenges. Permutation analysis is a natural choice because it quantifies the probability of the observed cross-condition or cross-participant differences given random permutations of the exact data, thereby coping with imbalanced class sizes and non-normally distributed data whilst enabling complexities, such as the variable number of observations per subject, to be controlled for when generating the null distribution (Nichols et al. 2002; Pesarin et al. 2010).

Steps applied in the pipeline were as follows (and are also illustrated in Fig. 4.3). (i) Dependent on whether group size was two groups or multiple groups/factors, the real observed across groups t or F statistic was calculated from the non-permuted data (t_{real} or F_{real} respectively). (ii) Data labels were permuted across observations with randomisation constrained to account for potentially confounding factors (i.e., by switching labels across subjects with the same number of observations thereby preserving the level of engagement). (iii) The t or F value was recalculated for the permuted data (t_{perm} and F_{perm}). (iv) Steps ii-iii were repeated 10,000 times, producing the null distribution. (v) P values were calculated according to where the real observed statistic ranked within the permutation null distribution.

When a main effect was detected, and if more than two groups were being compared, pairwise permutation analyses were performed to further characterise the basis of the observed main effect. Unless otherwise stated, a two-tailed alpha threshold of $p < 0.05$ was applied for rejecting the null hypothesis.

Specifics of individual permutation analyses are detailed in the results section and population sample sizes per analysis are reported in Fig. 4.2 and Table 4.1.

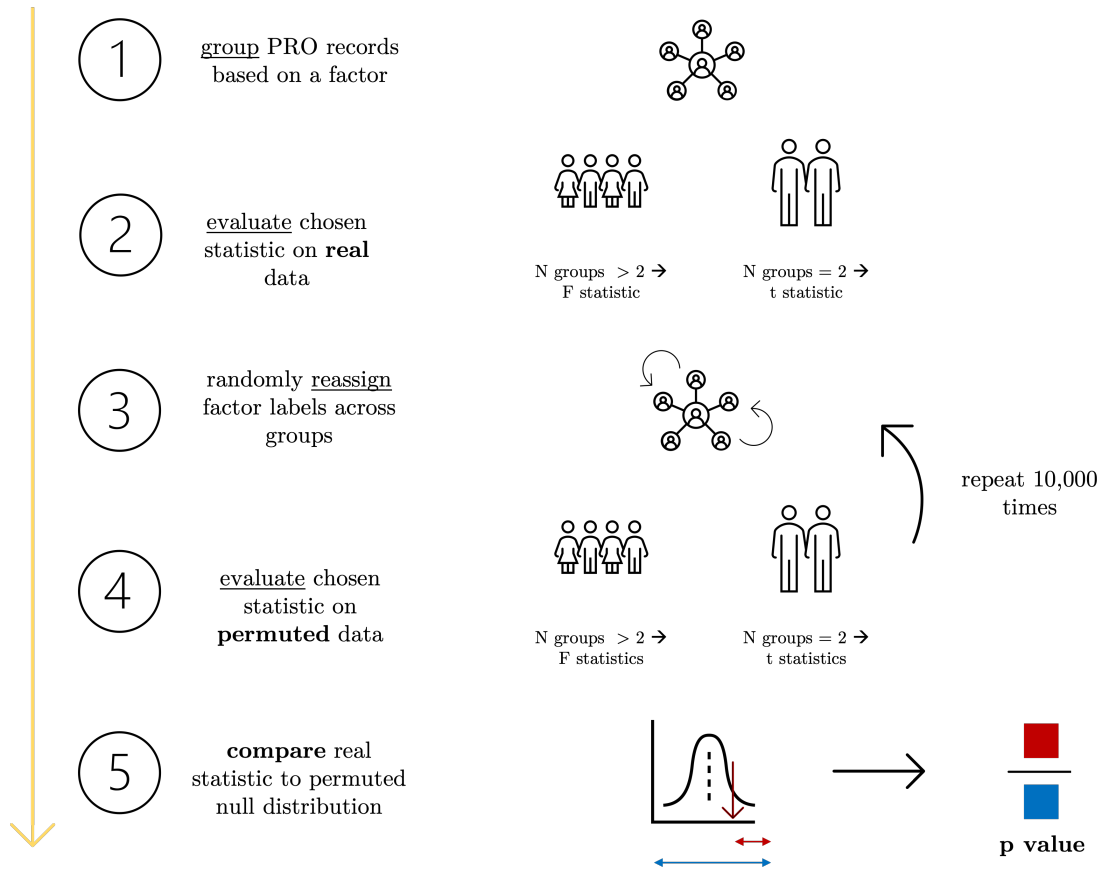


Figure 4.3: Schematic explaining how permutation testing was carried out.

Disease duration [yrs]	Total N	Disease subtype					Mean age [yrs, range]	Gender [female, missing]
		BN	RRMS	SPMS		PPMS		
				RR Phase	SP Phase			
0-5	9339 / 7934	265 / 219	7288 / 6314	294 / 224	306 / 259	1186 / 918	42.2, 18-82	7477, 3
5-10	16703 / 13772	401 / 319	10818 / 9329	724 / 557	1570 / 1124	3190 / 2443	47.4, 18-84	12987, 29
10-15	16272 / 12603	497 / 388	8807 / 7557	696 / 565	2940 / 1954	3332 / 2139	50.8, 19-86	12296, 10
15-20	13839 / 10227	610 / 501	6094 / 5216	702 / 546	3797 / 2289	2636 / 1675	53.7, 25-84	10398, 17
20-25	11187 / 7747	539 / 423	3906 / 3263	583 / 463	4339 / 2593	1820 / 1005	55.8, 26-86	8386, 9
25-30	8660 / 5858	504 / 410	2309 / 1883	523 / 433	4170 / 2464	1154 / 668	58.1, 32-87	6575, 3
30-35	6322 / 4072	443 / 354	1403 / 1128	319 / 261	3553 / 2038	604 / 291	60.3, 37-86	4983, 0
35-40	4077 / 2685	255 / 203	812 / 663	179 / 131	2477 / 1515	354 / 173	62.6, 43-91	3095, 1
40-45	2639 / 1714	223 / 167	493 / 409	94 / 70	1602 / 938	227 / 130	65.3, 46-88	1938, 0
45-50	1393 / 941	141 / 118	180 / 148	72 / 68	835 / 507	165 / 100	68.0, 51-84	1088, 0
>50	920 / 539	107 / 72	94 / 73	36 / 29	613 / 343	70 / 22	70.9, 58-91	737, 0

Table 4.1: Characterisation of PRO responses of the UKMSR population with the minimal requirements at all timepoints. MSIS-29 / MSWS-12 record counts are reported.

4.3 Results

4.3.1 Descriptive analysis of user engagement with the UKMSR

Data were collected over 132 months, from May 2011 to April 2022. Users joined the UKMSR in a rolling manner at different points in time and in the disease. They could choose to engage with any of the PROs available during the collection windows. As a result, users completed the PROs a different number of times, at a variety of timepoints over the collection period, with 62% of the users for MSIS-29 motor and 64% for MSWS-12 having completed between 1 and 4 questionnaires thus far (Fig. 4.4A). The distribution of the time intervals elapsed between subsequent questionnaire completion dates across all users is displayed in Fig. 4.4B and indicates how often users engage with the register (MSIS-29 motor: $time_{mean} = 7months$, $time_{median} = 5months$, $time_{max} = 131months$; MSWS-12: $time_{mean} = 7months$, $time_{median} = 4months$, $time_{max} = 113months$). This shows how, even though many of the users who completed the PROs in a certain collection window also completed them in the previous window (active users), others come back to engage with the UKMSR after a gap of many years. The number of MSIS-29 motor and MSWS-12 collected at different disease time bins is reported in Fig. 4.4C. The flowchart in Fig. 4.2 illustrates how each of the studied subgroups was derived from the overall UKMSR population. The characterisation of PRO responses at all time points is reported in Table 4.1; the users are accounted for as many times as the number of questionnaires they have completed at different timepoints.

4.3.2 Validity of the patient-reported outcomes

The first part of the analysis focused on assessing whether the considered PROs are valid measures of physical disability in MS, looking for statistical differences in the scores as a function of disease subtype and disease duration. Given that MS is a disease with a highly variable outcome dependent on subtype that plays out mostly over the long term, I hypothesised that a good marker of physical disability for this disease should be sensitive to both group and duration.

4.3.2.1 Patient-reported outcomes evolve differently across subtypes

First, I examined differences in PROs as a function of disease duration by disease subtype. Fig. 4.5 displays the mean score trajectories with 95% confidence intervals of the two PROs over disease duration for the different subtypes. The trajectories were obtained by pooling observations together in 5-year bins starting at onset separately for each subtype and evaluating the mean score for each bin. The trajectories were based on 91,351 observations for MSIS-29 motor and 68,092 for MSWS-12. Fig. 4.5 shows differences in trajectories by subtype from onset through the full disease course. In particular, BN and RRMS have lower average scores compared to PPMS and SPMS, since onset and throughout.

Permutation testing was performed randomly reassigning 10,000 times subtype labels while preserving any differences in level of engagement per subtype when estimating the null distribution. The F statistic was evaluated for the real and permuted data. A robust main effect of disease subtype was found on the scores (MSIS-29 motor: $F_{real} = 2698$, $p < 0.0001$; MSWS-12: $F_{real} = 3007$, $p < 0.0001$)

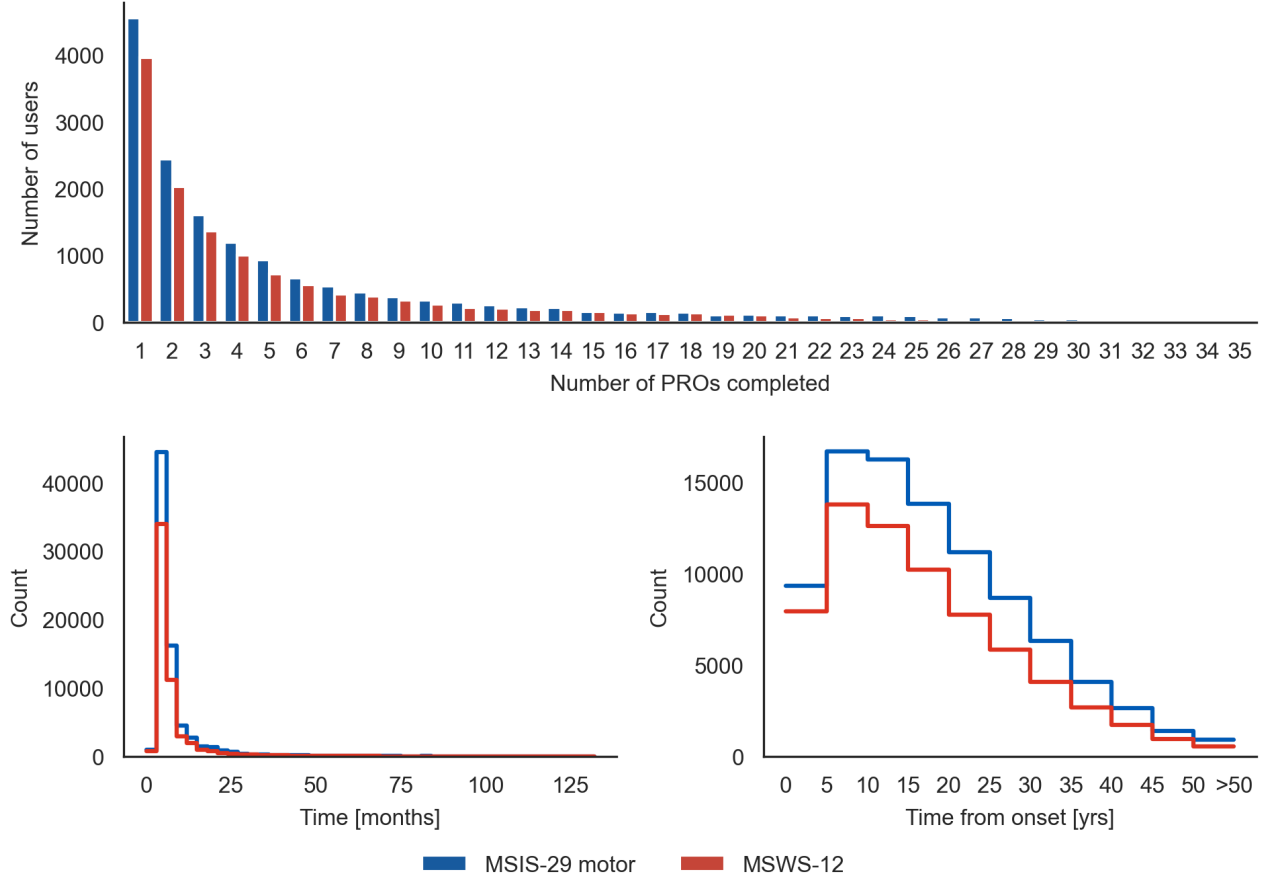


Figure 4.4: Analysis of user engagement with the UKMSR. (A) Frequency of users who completed a specific number of times MSIS-29 motor and MSWS-12. 62% of the users for MSIS-29 motor and 64% for MSWS-12 completed between 1 and 4 PROs. (B) Distribution of the time intervals elapsed between consecutive completions of MSIS-29 motor ($time_{mean} = 7months$, $time_{median} = 5months$, $time_{max} = 131months$) and MSWS-12 ($time_{mean} = 7months$, $time_{median} = 4months$, $time_{max} = 113months$) across all users. (C) Number of times MSIS-29 motor and MSWS-12 were completed at the different disease time bins.

as well as an interaction of disease subtype and duration (MSIS-29 motor: $F_{real} = 3.9$, $p < 0.0001$; MSWS-12: $F_{real} = 3.5$, $p < 0.0001$).

4.3.2.2 Disease duration influences patient-reported outcomes for all subtypes but Benign

Given that an interaction between disease duration and subtype was detected, I next looked for statistical differences in PROs as a function of disease duration, separately for each subtype. Permutation testing was performed randomly reassigning 10,000 times disease time bins for each subtype while preserving the level of engagement. Disease duration was found to have an association with the scores for RRMS (MSIS-29 motor: $F_{real} = 41.4$, $p < 0.0001$; MSWS-12: $F_{real} = 46.1$, $p < 0.0001$), PPMS (MSIS-29 motor: $F_{real} = 10.0$, $p < 0.0001$; MSWS-12: $F_{real} = 8.3$, $p < 0.0001$) and SPMS (MSIS-29 motor: $F_{real} = 7.0$, $p < 0.0001$; MSWS-12: $F_{real} = 6.5$, $p < 0.0001$), but not for BN (MSIS-29 motor: $F_{real} = 0.86$, $p = 0.796$; MSWS-12: $F_{real} = 0.37$, $p = 0.983$). This makes intuitive sense as BN individuals have a milder form of MS and experience less if any, physical disability. This can be

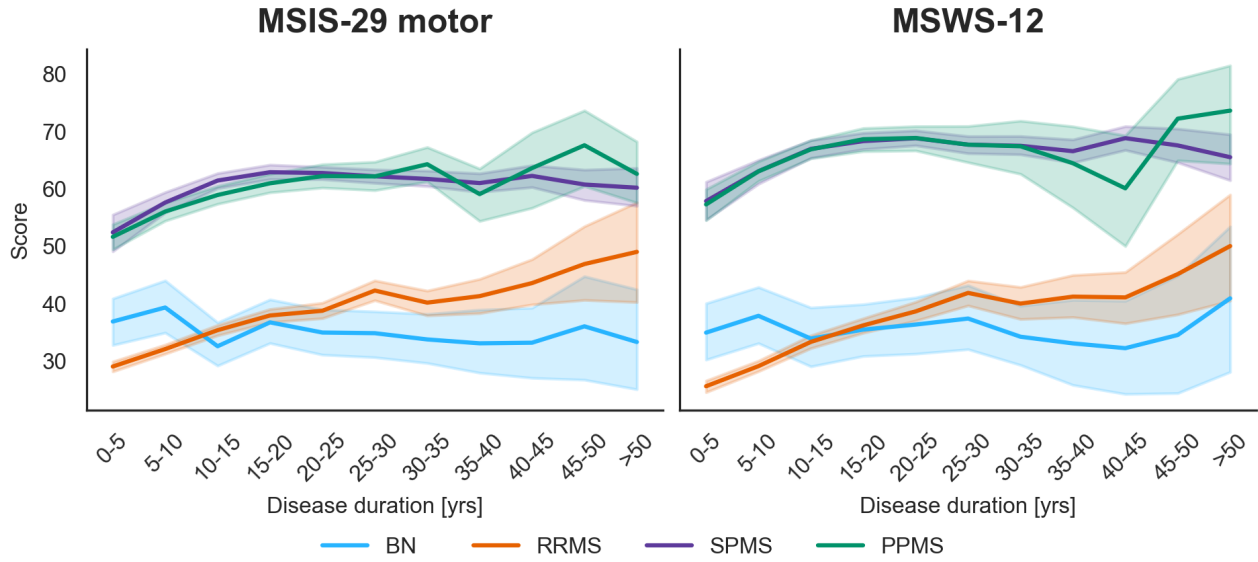


Figure 4.5: Mean score trajectories with 95% CI over disease duration for the different disease subtypes for MSIS-29 motor and MSWS-12. Performing permutational multivariate analysis of variance ($n = 10000$ permutations), a robust main effect of disease subtype was found on the scores (MSIS-29 motor: $F_{real} = 2698$, $p < 0.0001$; MSWS-12: $F_{real} = 3007$, $p < 0.0001$) as well as an interaction of disease subtype and duration (MSIS-29 motor: $F_{real} = 3.9$, $p < 0.0001$; MSWS-12: $F_{real} = 3.5$, $p < 0.0001$).

visually inspected in Fig. 4.5 where the BN trajectory fluctuates around the same range of values, while the other trajectories weakly (SPMS: MSIS-29 $m = 0.31$, MSWS-12 $m = 0.46$; PPMS: MSIS-29 $m = 1.50$, MSWS-12 $m = 1.28$; slope coefficients obtained fitting a line through the data separately for each subtype) or strongly (RRMS: MSIS-29 $m = 2.11$, MSWS-12 $m = 2.65$) increase over disease time. The SPMS trajectory is characterised by a lower gradient coefficient when fitting a line through the data among the subtypes that show an association with disease duration. This can be attributed to the fact that SPMS individuals show an increase in their scores for the first 20 years with disease but appear to plateau afterwards.

4.3.2.3 Patient-reported outcomes are sensitive to subtype within all disease time bins

Motivated by these findings, I tested whether PROs differed across subtypes within all disease time bins. First, I plotted the probability density functions of the PROs (Fig. 4.6) stratified by disease subtype for each time bin. Mean, standard deviation and skew of the distributions are reported in Table 4.2. SPMS and PPMS presented with similar distributions and centred around higher values compared to RRMS and BN at all intervals. Relapsing and progressive subtypes had distinct patterns at each disease duration with progressive distributions consistently being left-skewed and relapsing distribution mostly presenting with a right-skew. This pattern was more evident for MSWS-12. RRMS distributions were similar to BN distributions early on in the disease but, with increasing time from onset, they departed from these, losing the positive skew, shifting towards higher values and approaching the progressive distributions (skew goes from 0.8 to -0.1 over 50 years with disease for MSIS-29 and from 0.8 to -0.3 for MSWS-12). Overall, the mean of the RRMS distribution was the one that changed the most with increasing disease duration (mean from 29 to 49 for MSIS-29 motor and from

26 to 50 for MSWS-12). This makes sense intuitively as the RRMS group comprises individuals who have not reported a transition to SPMS during the current follow-up but might still transition later in the disease. These transitions, which occur gradually over time, were represented in the plots as the loss of positive skew of the RRMS distribution. On the contrary, the BN distribution remained right-skewed (skew always above 0.2).

Permutation testing by randomly reassigning 10,000 times disease subtype labels separately for each time bin showed a robust association of PROs with subtype at disease onset and throughout the time bins (MSIS-29 motor: all $p < 0.0001$; MSWS-12: all $p < 0.0001$). A complete record of F statistics and p values can be found in Table 4.3.

			Disease duration [yrs]										
			0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	>50
MSIS-29 motor	Mean	BN	36.9	39.3	32.6	36.8	35.0	34.9	33.8	33.1	33.2	36.1	33.3
		RR	29.1	32.1	35.4	38.0	38.8	42.3	40.2	41.3	43.6	46.9	49.0
		SP	52.4	57.6	61.5	62.9	62.8	62.2	61.7	61.0	62.3	60.8	60.2
		PP	51.7	56.1	58.9	61.0	62.2	62.2	64.3	59.1	63.6	67.6	62.6
	Standard deviation	BN	24.6	26.1	25.5	26.5	26.2	25.0	23.3	24.5	25.3	24.9	22.4
		RR	22.5	23.5	24.0	24.3	23.7	24.1	22.6	23.2	22.5	24.0	25.0
		SP	21.3	20.5	20.3	20.1	20.1	20.5	20.8	21.0	20.3	19.4	20.4
		PP	23.6	22.9	21.7	21.0	21.0	20.9	19.5	23.6	25.5	19.9	12.4
	Skew	BN	0.2	0.3	0.6	0.5	0.5	0.6	0.5	0.6	0.7	0.5	0.3
		RR	0.8	0.6	0.4	0.3	0.3	0.2	0.4	0.2	0.1	-0.1	-0.1
		SP	-0.3	-0.3	-0.5	-0.5	-0.5	-0.6	-0.5	-0.4	-0.4	-0.2	-0.3
		PP	-0.0	-0.2	-0.4	-0.4	-0.4	-0.6	-0.7	-0.4	-0.6	-0.9	-0.4
MSWS-12	Mean	BN	35.0	37.9	33.9	35.5	36.4	37.4	34.3	33.1	32.3	34.6	40.9
		RR	25.6	29.1	33.4	36.2	38.7	41.9	40.0	41.2	41.1	45.2	50.0
		SP	57.9	63.1	67.0	68.3	68.8	67.7	67.5	66.6	68.8	67.6	65.5
		PP	57.3	63.1	66.9	68.6	68.9	67.7	67.4	64.5	60.1	72.2	73.6
	Standard deviation	BN	27.2	28.0	29.6	28.9	28.9	29.6	27.2	28.9	28.7	26.8	28.6
		RR	24.3	25.7	26.4	26.5	26.4	26.5	24.8	26.3	23.1	26.0	24.4
		SP	21.3	20.5	19.6	18.5	18.7	20.2	19.9	20.7	17.3	17.5	18.1
		PP	24.1	21.0	19.9	19.6	18.6	19.2	21.3	25.4	26.0	17.9	13.5
	Skew	BN	0.2	0.2	0.4	0.3	0.3	0.3	0.5	0.5	0.7	0.6	0.3
		RR	0.8	0.6	0.4	0.2	0.1	0.0	0.2	0.0	-0.1	-0.1	-0.3
		SP	-0.6	-0.8	-1.1	-1.2	-1.4	-1.3	-1.3	-1.2	-1.1	-0.9	-0.9
		PP	-0.6	-0.9	-1.2	-1.4	-1.2	-1.0	-1.1	-1.2	-1.0	-1.3	-0.6

Table 4.2: Mean, standard deviation, and skew of the distributions of MSIS-29 motor and MSWS-12 across subtypes and disease time bins.

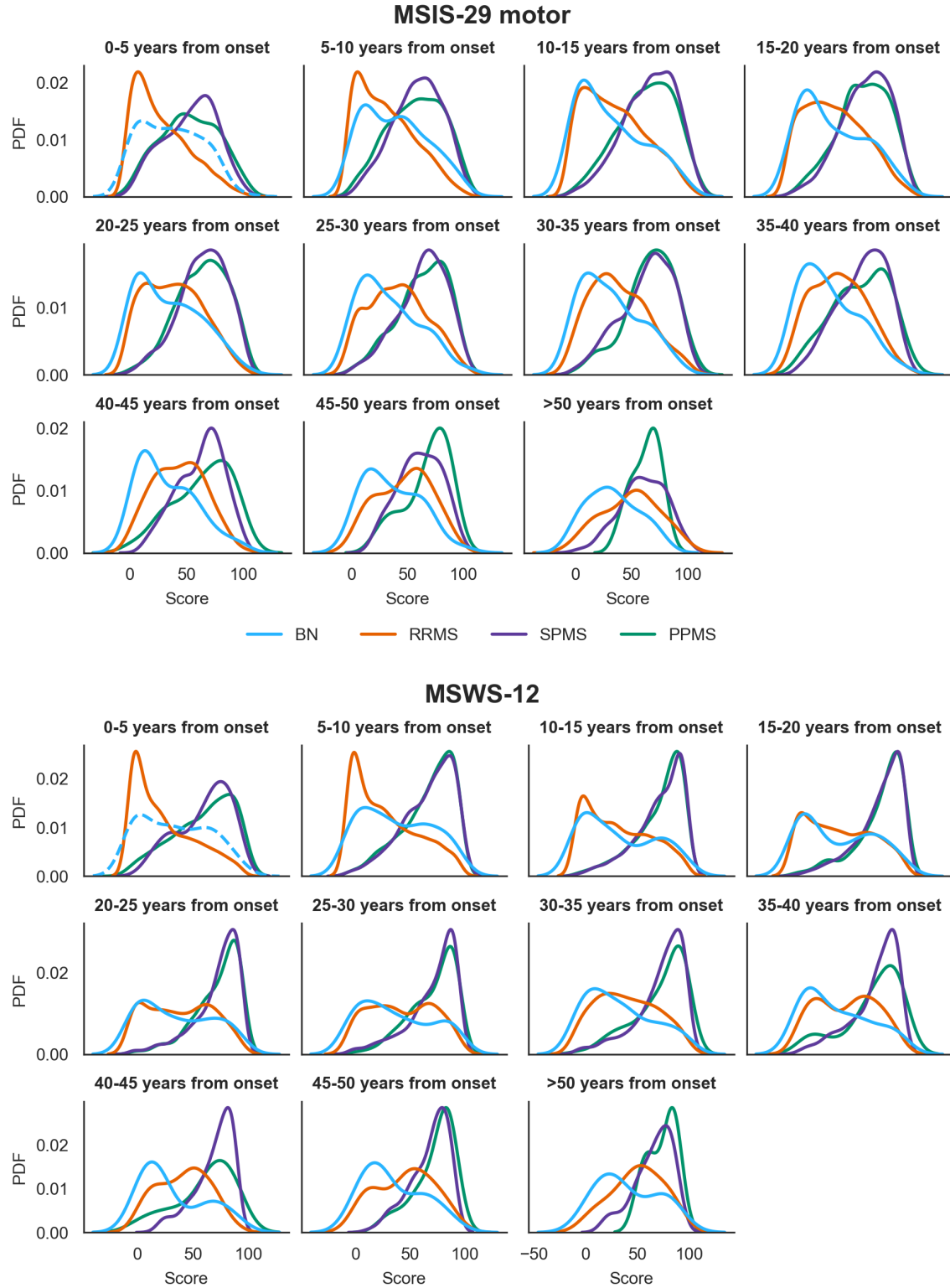


Figure 4.6: MSIS-29 motor (top panel) and MSWS-12 (bottom panel) probability density functions stratified by subtype for the different disease time bins considered. The probability density curve for the BN subgroup between 0 and 5 years from onset is dotted as this label was given retrospectively and was based on a maximum of 11-year follow-up. Performing permutational multivariate analysis of variance ($n=10000$ permutations), a robust main effect of disease subtype was found on the scores at disease onset and throughout all time bins considered (MSIS-29 motor: all $p < 0.0001$; MSWS-12: all $p < 0.0001$).

		MSIS-29 motor		MSWS-12	
		F_{real}	p	F_{real}	p
Disease duration [yrs]	0-5	169.6	<0.0001	224.5	<0.0001
	5-10	354.6	<0.0001	475.2	<0.0001
	10-15	415.1	<0.0001	481.8	<0.0001
	15-20	348.5	<0.0001	403.2	<0.0001
	20-25	294.7	<0.0001	305.2	<0.0001
	25-30	161.2	<0.0001	160.2	<0.0001
	30-35	140.5	<0.0001	139.9	<0.0001
	35-40	70.1	<0.0001	71.6	<0.0001
	40-45	49.5	<0.0001	67.4	<0.0001
	45-50	19.7	<0.0001	29.2	<0.0001
	>50	12.8	<0.0001	9.6	<0.0001

Table 4.3: Real F statistics and p values obtained randomly reassigning the disease subtype labels of the MSIS-29 motor and MSWS-12 separately at each time bin.

4.3.2.4 Pairwise comparisons across subtypes

Finally, I performed post-hoc pairwise comparisons to investigate which subtypes differed from each other. To do so, subtype labels were randomly permuted 10,000 times in pairs, first overall while preserving the level of engagement, and then separately within each disease time bin. I found that overall, the scores of all subtypes were different from the others (MSIS-29 motor: all $p < 0.0001$; MSWS-12: all $p < 0.05$) except when comparing the MSIS-29 motor scores of BN and RRMS individuals ($p = 0.51$). Regarding pairwise comparisons within each time bin, respectively 91% and 41% of the p values were not significant when comparing SPMS with PPMS or RRMS with BN (complete t statistics and p values are reported in Table 4.4). This can be visually validated in Fig. 4.6 where the distributions of SPMS and PPMS appear similar in most of the time bins. The same holds true for the distributions of BN and RRMS, especially early on. The scores for all the other subtypes' combinations were, instead, significantly different at all time bins (MSIS-29 motor: at least $p < 0.05$; MSWS-12: at least $p < 0.01$).

		MSIS-29 motor						MSWS-12						
		SP	SP	SP	PP	PP	RR	SP	SP	SP	PP	PP	RR	
		RR	BN	PP	RR	BN	BN	RR	BN	PP	RR	BN	BN	
overall	t _{real}	79.7	40.3	6.0	55.4	31.5	-0.9	81.6	42.3	3.7	57.8	33.7	-3.5	
	p	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.5094	<0.0001	<0.0001	0.0213	<0.0001	<0.0001	0.0069	
Disease duration [yrs]	0-5	t _{real}	13.9	6.2	0.4	19.0	6.4	-4.0	16.1	7.8	0.3	21.7	8.3	-4.0
		p	<0.0001	<0.0001	0.6947	<0.0001	<0.0001	<0.0001	<0.001	<0.001	0.8043	<0.0001	<0.0001	<0.001
	5-10	t _{real}	23.8	9.1	1.3	25.7	8.0	-3.7	25.8	10.9	0.0	30.3	11.4	-3.7
		p	<0.0001	<0.0001	0.1925	<0.0001	<0.0001	0.0003	<0.001	<0.001	0.9833	<0.0001	<0.0001	0.0004
	10-15	t _{real}	28.7	16.1	2.5	25.0	13.8	1.5	29.2	15.7	0.1	27.7	15.3	-0.2
		p	<0.0001	<0.0001	0.0135	<0.0001	<0.0001	0.141	<0.001	<0.001	0.9426	<0.0001	<0.0001	0.817
	15-20	t _{real}	28.3	15.6	1.9	20.9	12.9	0.6	29.1	17.6	-0.3	22.7	15.3	0.3
		p	<0.0001	<0.0001	0.0671	<0.0001	<0.0001	0.5327	<0.001	<0.001	0.7893	<0.0001	<0.0001	0.755
	20-25	t _{real}	26.1	16.4	0.5	18.5	13.7	1.9	26.7	16.9	-0.0	17.3	13.5	0.9
		p	<0.0001	<0.0001	0.6388	<0.0001	<0.0001	0.0488	<0.001	<0.001	0.9735	<0.0001	<0.0001	0.3465
	25-30	t _{real}	18.5	14.5	0.0	11.9	11.8	3.3	19.5	13.6	0.0	11.0	9.9	1.6
		p	<0.0001	<0.0001	0.9903	<0.0001	<0.0001	0.0012	<0.001	<0.001	0.9985	<0.0001	<0.0001	0.1118
	30-35	t _{real}	16.6	13.4	-1.5	12.0	11.9	2.7	18.0	13.9	0.0	9.2	8.9	1.9
		p	<0.0001	<0.0001	0.1456	<0.0001	<0.0001	0.0081	<0.001	<0.001	0.9723	<0.0001	<0.0001	0.0576
	35-40	t _{real}	11.7	10.7	0.8	6.3	7.0	2.6	12.6	10.7	0.7	5.6	5.9	2.0
		p	<0.0001	<0.0001	0.4144	<0.0001	<0.0001	0.0088	<0.001	<0.001	0.5214	<0.0001	<0.0001	0.0458
	40-45	t _{real}	8.6	10.2	-0.5	5.3	6.6	2.8	12.2	11.6	2.4	3.7	4.2	2.0
		p	<0.0001	<0.0001	0.645	<0.0001	<0.0001	0.0055	<0.0001	<0.0001	0.0161	0.0003	0.0003	0.0492
	45-50	t _{real}	4.5	6.3	-1.9	4.1	5.5	2.0	6.6	7.9	-1.2	4.4	5.6	1.6
		p	<0.0001	<0.0001	0.0664	0.0002	<0.0001	0.0544	<0.0001	<0.0001	0.2513	0.0001	<0.0001	0.1086
	>50	t _{real}	2.7	5.8	-0.5	2.1	4.9	2.4	3.3	4.5	-1.3	2.7	3.2	1.1
		p	0.0071	<0.0001	0.6325	0.0452	0.0001	0.018	0.0001	<0.0001	0.1947	0.0084	0.0038	0.2778

Table 4.4: Real t statistics and p values obtained from pairwise comparisons of the MSIS-29 motor and MSWS-12 of subtypes overall and at each disease time bin.

4.3.3 Deriving new insights using the patient-reported outcomes

In the previous sections, the validity of the PROs as measures of physical disability in MS was advocated, as they were shown to be sensitive to the diverse levels of physical impairment that individuals belonging to different subtypes experience in the various stages of the disease. Motivated by these findings, I investigated further the utility of this type of data in providing novel insights into the developmental time course of disease subtypes.

A pressing matter in MS involves defining among the relapsing individuals who are going to become progressive and ideally being able to do so early on to treat them accordingly. Since significant differences were shown in post-hoc pairwise comparisons between PROs of RRMS and SPMS patients at every time bin, I extracted from the SPMS group only the observations in the RR phase (RRTrans) and compared those to the RRMS observations (RRNoTrans) to investigate whether the PROs in the relapsing phase vary as a function of whether the patients will transition or not. Fig. 4.7 shows the mean score trajectories with 95% confidence intervals of the PROs over disease duration for the two relapsing groups, namely RRNoTrans and RRTrans. The trajectories were obtained from 46,426 observations for MSIS-29 motor and 39,330 for MSWS-12. The trajectories of the other subtypes (BN, PPMS and SPMS in the secondary progressive phase) are displayed only for comparison.

The RRNoTrans and RRTrans trajectories look different since onset and up to around 45 years with disease: RRTrans have, on average, higher scores compared to RRNoTrans, especially at onset. To validate this, permutation testing was performed randomly reassigning the subtype labels between these two groups 10,000 times both overall, preserving the level of engagement, and separately within each time bin. The t statistic was evaluated for the real and permuted data. I found that the PROs of the two groups were significantly different overall (MSIS-29 motor: $t_{real} = -23.5$, $p < 0.0001$; MSWS-12: $t_{real} = -26.2$, $p < 0.0001$) and at each disease time bin up to 45 years with disease for MSIS-29 motor (at least $p < 0.05$) and to 35 years for MSWS-12 (at least $p < 0.001$) but they became non-significant above this, where notably, data were more sparsely sampled (complete t statistics and p values are reported in Table 4.5). This means that the PROs of relapsing individuals who transitioned to progressive forms of MS were, on average, higher since disease onset and subsequently than the PROs of relapsing individuals who did not transition during the considered follow-up timeframe.

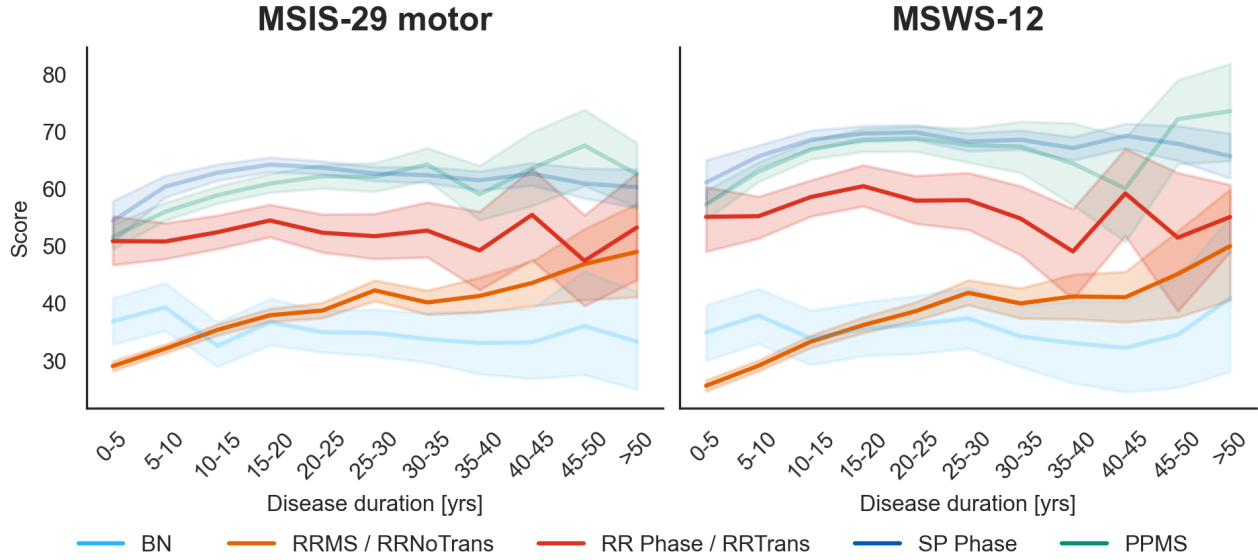


Figure 4.7: Mean score trajectories with 95% CI over disease duration for transitioning and non-transitioning individuals in the UKMSR cohort taken in their relapsing phase for MSIS-29 motor (left panel) and MSWS-12 (right panel). Being RRTrans was associated with higher scores on average in both MSIS-29 motor ($t_{real} = -23.5$, $p < 0.0001$) and MSWS-12 ($t_{real} = -26.2$, $p < 0.0001$) when performing permutation testing with t statistic and $n=10000$ permutations. Mean score trajectories with 95% CI for BN, PPMS and SPMS individuals in their secondary progressive phase are also displayed here in transparency for comparison and were not included in the statistical tests.

		MSIS-29 motor		MSWS-12	
		t_{real}	p	t_{real}	p
Disease duration [yrs]	0-5	-8.9	<0.0001	-9.9	<0.0001
	5-10	-10.6	<0.0001	-11.9	<0.0001
	10-15	-9.6	<0.0001	-11.6	<0.0001
	15-20	-9.0	<0.0001	-10.1	<0.0001
	20-25	-6.8	<0.0001	-7.2	<0.0001
	25-30	-4.0	0.0001	-5.5	<0.0001
	30-35	-4.4	<0.0001	-4.1	0.0001
	35-40	-2.1	0.0387	-1.6	0.1125
	40-45	-2.5	0.0135	-3.2	0.0014
	45-50	-0.1	0.9371	-0.8	0.4361
	>50	-0.4	0.7145	-0.4	0.6901

Table 4.5: Real t statistics and p values obtained from pairwise comparisons of the MSIS-29 motor and MSWS-12 of relapsing individuals who did vs did not transition during the follow-up, overall and at each time bin.

4.4 Discussion

The study presented in this chapter provided further validation of two PROs commonly used to measure physical disability in MS and introduced a novel permutation-based analysis approach to analyse registry data. This approach enabled handling the complexity and sparsity inherent to registers, thereby retaining the highest proportion of records possible from 11 years of prospective UKMSR PRO data whilst drawing robust statistical inferences regarding the long-term trajectories of decline that should generalise beyond just the most engaged participants. Using this approach, it was demonstrated that both PROs are viable monitoring tools capable of capturing at a large population scale the distinct patterns of physical worsening occurring across different subtypes. In further support of the utility of register-based PROs, the analysis detected subtle differences across MS subtypes early in the disease that anticipates subsequent changes in clinical labels. This highlights the potential of online monitoring technologies to estimate progression risk early and support clinical decision-making.

There are characteristic strengths and weaknesses ubiquitous to registry data that are essential traits of the data collection method. Most notably, the ideal scenario, to have all scales completed by all participants at all collection windows since disease onset, is fundamentally not feasible due to the requirement to repeatedly engage participants remotely, over long timeframes and at large scale. Consequently, data are inherently sparse and complex with engagement level being a prominent confounder, characteristics that present challenges from the analysis perspective.

Using a permutation-based approach that both handles and explicitly accounts for engagement, facilitated as much of the available data as possible to be leveraged. Specifically, 91% of the total records for MSIS-29 and 93% for MSWS-12 were included, thereby summarising over 91k records for MSIS-29 motor and 68k records for MSWS-12. This produced substantial power, enabling the derivation of robust statistical conclusions regarding the four trajectories representative of the differential evolution of physical disability across subtypes.

Notably, these trajectories together with the distributions of the scores at different time bins, provide a strong validation of the utility of PROs as large-scale monitoring tools. These plots frame the sensitivity of these PROs to both subtype and duration, as supported by the presented statistical findings. For example, the distributions for BN individuals span lower values compared to those of the other subtypes, and their trajectory does not worsen over time. This is expected as BN patients are characterised by minimal physical disability and sporadic, if not completely absent, relapses (Tallantyre et al. 2019). Conversely, RRMS PROs change the most, presenting with the steepest mean trajectory, and are characterised by a distribution that loses the positive skew as the disease evolves. Again, this pattern of results has clear validity because RRMS patients are periodically affected by relapses, which can become increasingly debilitating, and they can potentially transition to a progressive phenotype, thereby showing a more continuous silent decline (Cree et al. 2019). Consequently, the PROs analysed here provide valid alternatives to established clinical-observational scales such as the EDSS for monitoring physical disability in MS, allowing patients to be monitored at scale, remotely, more often and for longer follow-ups.

Analysis of these PROs at the timescale and population scale of the register produced new insights into the process by which some individuals phenotypically convert between subtypes. Indeed, current knowledge in MS advocates for an individualised approach to treatment with a growing emphasis on

early intervention to limit later damage (Kavaliunas et al. 2017). In this study, both analysed PROs demonstrated sensitivity to cross-group differences in patterns of early disability. Specifically, not only was a statistical difference observed in the earliest time bins between relapsing and progressive subtypes, but a robust statistical difference was also observed at those same early timepoints between the PROs of relapsing individuals who did versus did not phenoconvert to the secondary progressive subtype during the considered follow-up period. Therefore, relapsing-remitting patients at higher risk of becoming progressive may be identifiable early in the disease course. These results are consistent with insights from the pathology and spinal fluid inflammatory profile (Magliozzi et al. 2018), cortical MRI lesion development (Beck et al. 2022), and evidence of cognitive dysfunction at presentation (Pitteri et al. 2022). It is recognised that early symptoms can be vague and ignored by patients and healthcare professionals, but increasingly, there is evidence of an MS prodrome (Yusuf et al. 2020), as in the presence of early disease symptoms and impacts years prior to diagnosis.

The potential strengths in scope and scale of registry data have challenges that extend beyond sparsity and complexity (Thygesen et al. 2014). For example, a related challenge is rooted in the voluntary nature of engagement in registries, which could lead to inferences that generalise poorly to the broader MS population due to sampling bias. While bias is inherent in any sampling method (Ferreira-González et al. 2009; Tripepi et al. 2010), an examination of the relative proportions of subtypes in the studied sample revealed close concordance with those known from epidemiological studies (Cottrell et al. 1999; Klineova et al. 2018). Specifically, the UKMSR sample is representative of the MS population extending prior comparisons (Middleton et al. 2018) with 14% of the patients being PPMS and 80% being RRMS at disease onset (Cottrell et al. 1999; Klineova et al. 2018). In turn, consistent with this, approximately one-third of RRMS patients transition to SPMS (Weinshenker et al. 1989b). Notably, 6% were defined as BN, which is 3% higher than recent estimates (Tallantyre et al. 2019). Overall, while the dataset inevitably has some bias, it appears to be small and is mitigated by the adopted inclusive approach to analysis.

Regarding sampling bias over time, the number of observations collected varied with disease duration, reaching a peak at 5-10 years from onset and then steadily declining. This trend may suggest that patients engage less as MS progresses and are more likely to be recruited in the early stages. This bias can lead to complications in interpreting certain aspects of the data. For example, it appears that the PRO trajectories for SPMS plateau after 15-20 years from onset (Fig. 4.5), suggesting a stabilisation of physical disability over time. However, time-dependent sampling bias could also be a contributing factor. If participants with more severe physical disability at the outset are more likely to drop out, this could manifest as a plateau. To test this possibility, I split the SPMS individuals from the 10-20 years from onset groups into those that did and did not return in the 20-30 years from onset groups. Then, using a permutation model, I tested for a significant difference between the PROs of these two subgroups. The results showed that SPMS individuals who returned had significantly lower scores on average (mean MSIS-29 motor = 59.3; mean MSWS-12 = 67.3) than those who did not (mean MSIS-29 = 68.1; mean MSWS-12 = 74.1) with p values less than 0.0001. While this finding could explain the observed plateau, it does not rule out the possibility that physical disability stabilises to some degree over time.

In combination with other causes of sampling bias, these results further highlight the challenges of working with registry data and the importance of both coping with and accounting for engagement bias when evaluating between-subjects effects during the analysis. The presented permutation-based

approach achieves this by permuting the PROs of individuals after matching them based on their level of engagement, thereby controlling for it.

Another notable limitation is that the UKMSR currently has a maximum 11-year follow-up since its inception. This implies that individuals who joined the register at onset and are currently labelled as BN could still phenoconvert to more aggressive disease subtypes. This explains why the proportion of BN individuals in the studied sample slightly exceeds the population estimate. Similarly, some patients within the RRMS subgroup will transition to SPMS as this typically happens within 15-20 years from onset (Rovaris Marco et al. 2006), and the considered follow-up has not allowed enough time for all transitions to happen. As a result, the significant cross-group differences observed between MS subtypes likely underestimate the true underlying differences in motor impairment trajectories that can be expected as longer follow-up data become available.

4.4.1 Conclusions

Overall, the study presented in this chapter supports the value of PRO data and online registry initiatives. The devised approach allows to maximise data usage in a sparse dataset and can be easily applied in the context of other registries and extended to answer different questions. Most importantly, the findings demonstrate that the more pronounced disability of those who will become progressive is evident early in the natural history of the disease, that is, years before the clinical label is changed. Register-based PROs enable detecting this gap as they provide data at a larger population scale, spanning longer timespans, and at a more affordable cost than traditional cohort or clinical trial studies. Moving forward, it would be interesting to explore whether this early disability is also evident in mood and sleep scales or when objectively assessing cognitive symptoms. The combination of these findings would enable evaluating the potential for assigning individualised progression risk quotients based on the multivariate profile of symptoms at the early stage.

Chapter 5

Feasibility and sensitivity of online cognitive assessment in people with Multiple Sclerosis

In this chapter I present an exploratory study done in collaboration with the UK MS Register to evaluate the feasibility and sensitivity of online cognitive assessment in people with MS. I used a random-sampling strategy to evaluate 22 candidate online cognitive tasks from the Cognitron for an MS optimised battery without fatiguing the participants. I proved the feasibility of online cognitive testing in MS, identified affected cognitive domains in alignment with the literature, and derived a concise battery of tasks to optimally measure and monitor those domains remotely without supervision.

5.1 Introduction

In the previous chapter I focused on online tools for evaluating physical disability in MS, using data from the UK MS Register (UKMSR). In this chapter instead I will be focusing on online tools for assessing cognitive impairment in MS, an understudied symptom of the disease, overshadowed by physical disability that is more visually evident, and thus more easily tested. Online tools to measure physical disability, such as MS specific patient-reported outcomes (PROs), have been extensively developed and validated (Hobart et al. 2003; Hobart et al. 2001; Riazi et al. 2002). The UKMSR has been collecting data using PROs for more than 10 years (Knowles et al. 2023; Rodgers et al. 2022). The study presented in the previous chapter analysed these data longitudinally at large population scale to further validate two commonly used motor PROs and provide new insights into disease progression (Chapter 4).

In comparison, cognitive impairment has been relatively overlooked despite the high prevalence rates reported in the literature (Chiaravalloti et al. 2008). This is partly because cognitive deficits can appear in absence of other deficits and persist silently for years (Achiron et al. 2006; Glanz et al. 2007; Sumowski et al. 2018). Over the past two decades increased attention and efforts have been directed towards researching the prevalence, expression, and neural basis of MS cognitive dysfunction (Sumowski et al. 2018). Nonetheless, the development of effective and efficient tools to monitor cognitive function in pwMS remains a priority in the field. Recent research efforts have led to the development of optimised neuropsychological batteries for MS (Benedict et al. 2006; Boringa et al. 2001; Langdon et al. 2012). However, these batteries require in-person one-on-one assessment for each patient, which is impractical, costly, and inefficient. Additionally, they restrict the scale of testing, resulting in research studies on cognition in MS being characterised by small sample sizes and data on a limited number of tasks, and make it impossible to periodically screen patients throughout the course of their disease. These factors inhibit the understanding of the mechanisms underpinning cognitive impairment in MS, the interdependence of deficits across cognitive domains, the time course of disease-related decline, and the development of ad-hoc treatments for cognition. Due to the aforementioned hurdles measures of cognitive impairment have not been routinely collected as part of the UKMSR initiative, unlike for physical disability.

Online tools for assessing cognition could offer a solution to overcome these challenges: they are easily accessible, cost-effective, and efficient, allowing for wide-reaching participation, including for individuals with mobility limitations. These remote assessments require no specialised personnel or equipment, enabling large-scale periodic testing and longitudinal data collection. They provide standardised testing environments, reducing administration biases, and offer more sensitive and dynamic measures of cognitive function. By capturing detailed time courses of task performance behaviour they enable the extraction of diverse metrics, offering insights into various cognitive processes (Feenstra et al. 2017). Furthermore, they facilitate testing across a broader range of cognitive domains and include novel tasks that align more closely with real-life cognitive challenges (Kessels 2019).

Nonetheless, their feasibility, validity, and clinical utility in the context of MS still require thorough evaluation. Studies to date have employed small sample sizes and assessed only a limited range of tasks, primarily focusing on information-processing speed (Foong et al. 2023; Wojcik et al. 2019). It is important to determine whether pwMS are willing and able to engage periodically with such technology and whether these assessments are reliable and valid tools for monitoring cognitive function in this

population. Moreover, it is essential to investigate the relative sensitivity of various online cognitive tasks for detecting cognitive impairment in pwMS and select the best combination to form a battery. Additionally, extensive normative data on these tasks are necessary to enable a fair assessment by controlling for sociodemographic factors and device variability.

The Cognitron platform for online cognitive testing is a suitable candidate for this. Its reliability and validity have been demonstrated in the context of large-population studies (Hampshire et al. 2020, 2021b) and of other neurological populations (Del Giovane et al. 2023a,b; Gruia et al. 2023c). It also has a comprehensive suite of tasks that measure many cognitive domains as well as extensive normative data with detailed sociodemographic profiles, and it is deployable on any device.

In this chapter, the starting point differs from the previous one, which had the advantage of a decade's worth of data from established tools. Here, the aim is to develop and validate an online cognitive assessment battery for MS. Using Cognitron integrated into the UKMSR website, I gathered data from a large cohort of pwMS on a comprehensive set of online tasks spanning a broad range of cognitive domains. Given the impracticality of administering a lengthy task battery due to potential cognitive fatigue in pwMS (Krupp et al. 2000; Schwid et al. 2003), a randomised study design was adopted. This chapter aimed at evaluating the feasibility of online cognitive assessment in MS using Cognitron, determine the sensitivity of the evaluated tasks to MS-related cognitive deficits, and select a subset of tasks optimal for large-scale longitudinal monitoring of cognitive function in MS.

5.2 Methods

5.2.1 Study design

This is a population-based, cross-sectional observational study aimed at (i) assessing the feasibility of online cognitive assessment in pwMS and (ii) deriving a reduced battery from a superset of online tasks optimal to measure cognitive deficits in MS. Data collection was conducted between October 22, 2021 and May 18, 2022, in collaboration with the UKMSR. PwMS from the UKMSR completed different combinations of online cognitive tasks (randomised battery), resulting in a sparse dataset. The engagement of pwMS with the assessment and the sensitivity of the online cognitive tasks to MS were determined. Additionally, the study explored the correlations between cognitive function and other disability parameters, such as motor dysfunction and mood disorder, obtained from PROs. The tasks for the optimal battery were selected to maximise sensitivity to MS, minimise sensitivity to the device, and comprehensively cover the latent variable space of cognitive abilities.

Prior to data collection, a Patient and Public Involvement group provided feedback on the wording and the assessment after conducting end-to-end testing. At the study's inception, all UKMSR registrants received an email invitation to participate in the research and help shape an online tool to assess their cognition. Throughout the study, two reminder emails were sent to participants who either had initiated but not completed the assessment or had not yet started it. The resulting cognitive performance measures were linked with data already available on the UKMSR, including PROs. The following analyses excluded data from participants aged below 16 or above 90.

5.2.2 Procedures

5.2.2.1 Online cognitive assessment: the software and the tasks

The online cognitive assessment was sequenced from the Cognitron website but displayed within the UKMSR website, which cohort members are familiar with. Cognitron boasts an extensive array of internet-deliverable tasks developed to comprehensively measure different aspects of cognition. A subset of 22 tasks from Cognitron was selected for this study based on factors such as: availability of extensive normative data, alignment with other studies for cross-clinical group comparisons (e.g., traumatic brain injury (Del Giovane et al. 2023b) and Covid-19 (Hampshire et al. 2021b) studies), and comprehensive assessment of different cognitive domains (Section 3.2.1). Additionally, the Motor Control task was included to assess hand-eye processing.

For each task, all on-screen activities were recorded at the time of performance, capturing trial-by-trial information on the type of stimulus presented, response accuracy, and response time. The device, browser and operating system being used were also recorded as nuisance variables. The primary outcome metrics for each task amalgamate the trial-by-trial data to produce concise performance measures, predominantly focusing on accuracy (e.g., percentage of accurate responses) and latency (e.g., median response time across trials). Some tasks, such as Switching Stroop, Trail Making, Emotion Discrimination, Picture Completion, and CRT yield additional nuanced metrics derived from contrasts between conditions, revealing the "cost" in accuracy and response time transitioning from simpler to more challenging conditions. Notably, four tasks are intrinsically paired: SRT with CRT and Words Memory Immediate with Words Memory Delayed, and additional performance measures are integrated to discern performance disparities across these task pairs. Lastly, Bees assesses impulsivity and yields additional metrics to reflect this.

5.2.2.2 Sampling of online cognitive tasks

To minimise testing time, but gather data on all the selected tasks, a stratified random sampling methodology was employed (Fig. 5.1). Each participant was presented with a truncated battery consisting of 12 tasks, consisting of three fixed tasks and nine tasks randomly sampled from the remaining 19. With a sufficient sample size, this approach affords robust power to discern not only the sensitivities of isolated tasks but also inter-task correlations and underlying latent structures. Three tasks were kept fixed because they either gathered essential information (Motor Control) or needed to be presented at a specific point in the battery (Words Memory Immediate and Delayed).

5.2.2.3 Questionnaires

In addition to the cognitive tasks, participants completed a sociodemographic questionnaire, which included potential confounding variables such as age, gender, dominant hand, country of residence, language, ethnicity, education, and occupation. At the end of the assessment, participants also filled out a feedback questionnaire.

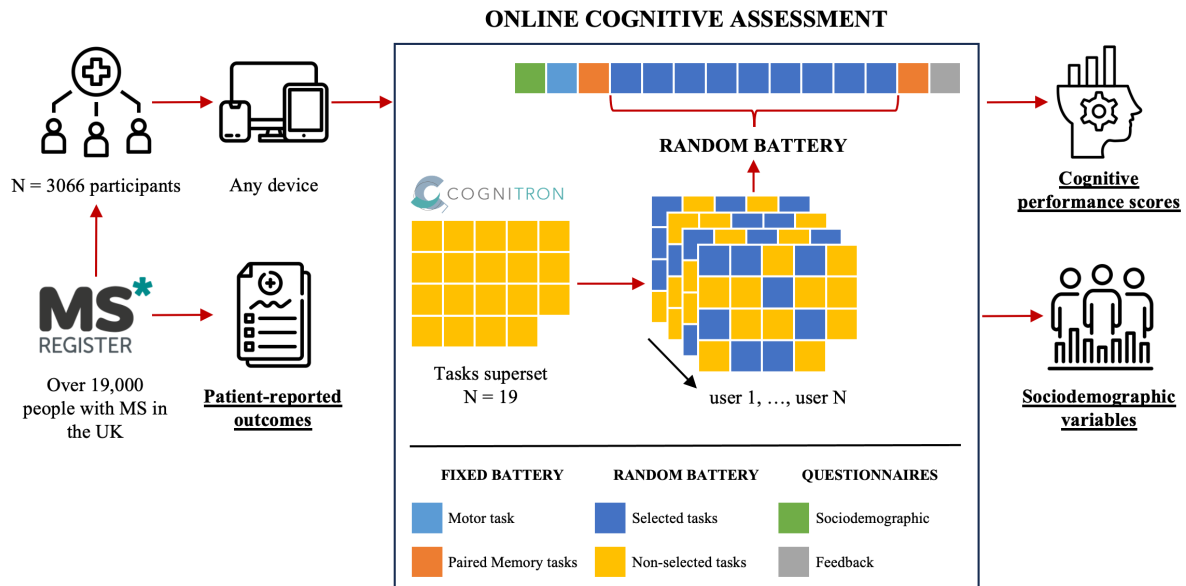


Figure 5.1: Diagram illustrating the study design. Participants were recruited via email from the UK MS Register. They could perform the assessment on any device of their choice. The online cognitive assessment consisted of 12 tasks: 3 fixed tasks and 9 tasks random sampled from a superset of 19 tasks. Overall, 22 tasks were evaluated. Participants were also asked to complete a sociodemographic questionnaire at the beginning, and a feedback questionnaire at the end of the assessment. The resulting cognitive performance scores and sociodemographic variables were linked to data already available on the Register (e.g. patient-reported outcomes) for further analysis.

5.2.2.4 Patient-reported outcomes

The UKMSR gathers PROs through periodic administration of questionnaires. Since September 2018, participants are prompted to enter data 6-monthly and have a 28-day window to complete the PROs. Collected PROs include the MSIS-29, assessing motor function and psychological well-being, the MSWS-12, measuring walking abilities, the HADS, evaluating mood disturbances, the EQ5D, assessing quality of life, the FSS, measuring fatigue, and the webEDSS, an online version of the commonly used clinical scale. The collected online cognitive data were linked with clinical information such as disease duration and subtype, as well as with PRO data available on the Register. However, for PRO data, this linkage was done only for pwMS who completed each PRO during the closest collection windows to when the cognitive assessment was conducted, i.e., either in Autumn 2021 or Spring 2022. If data for both collection windows were available, only data from the closest window were taken for each participant.

5.2.2.5 Normative data

Normative data from the UK population were collected as part of the Great British Intelligence Test (GBIT), a large-scale online longitudinal population study conducted between December 2019 and February 2023 across 2 baseline timepoints and 4 follow-up timepoints (Hampshire 2020; Hampshire et al. 2020; Mosley 2020). Data from the final follow-up timepoint were excluded due to modifications made to the tasks during this period, which had not been implemented at the time of patient data collection. For this study, only data from the 22 tasks being evaluated were retained, and individuals

meeting the following criteria were included:

- No reported neurological or psychiatric conditions (healthy controls).
- Complete sociodemographic data available (controls completed the same sociodemographic questionnaire as pwMS).
- Aged between 16 and 90 years old.
- Performing the tasks for the first time (to ensure a fair assessment, as some tasks may involve a degree of learning).

The total number of controls who completed each task is presented in Table 5.1, together with the number remaining after applying the selection criteria and the pre-processing procedures (as detailed in Section 3.3). In total, normative data from 389774 controls were used in this study, with their summary sociodemographic characteristics presented in Table 5.2.

Tasks	N observations (before selection criteria and pre-processing procedures)	N participants (before selection criteria and pre-processing procedures)	N observations = N participants (after selection criteria and pre-processing procedures)
Bees	5127	5104	4054
Blocks	899396	799128	389247
Card Pairs	8368	8000	5987
CRT	7135	6912	5210
Digit Span	625923	563712	387412
Emotion Discrimination	456628	455226	387382
Four Towers	143767	130228	101967
2D Manipulations	501905	444427	388382
Motor Control	38246	38231	25548
PAL	3763	3763	3232
Picture Completion	3529	3527	2200
Spatial Span	563740	504122	388653
Spotter	4725	4725	3879
SRT	7313	7082	5430
Switching Stroop	8363	8030	5985
Target Detection	522715	464539	387417
TOL	684354	621508	379227
Trail Making	8488	8129	6890
Verbal Analogies	546488	486951	387231
Word Definitions	510415	452419	388323
Words Memory Delayed	124823	120485	92625
Words Memory Immediate	200639	195909	98887

Table 5.1: Number of observations available for each task and number of controls who completed each task, before and after selection criteria and pre-processing procedures were applied.

Socio-demographic characteristics		Controls
Total N		389774
Age (min-max, mean)		16-90, 44.9
Gender	Female	207308 (53.2%)
	Male	178783 (45.9%)
	Other	3683 (0.9%)
Dominant hand	Right	336712 (86.4%)
	Left	42533 (10.9%)
	Ambidextrous	10529 (2.7%)
First language	English	350530 (89.9%)
	Other	39244 (10.1%)
Ethnicity	White	330900 (84.9%)
	Unknown	5213 (1.3%)
	Mixed or multiple ethnic groups	9973 (2.6%)
	Asian or Asian British	31072 (8.0%)
	Black, Black British, Caribbean or African	3552 (0.9%)
	American Hispanic	9023 (2.3%)
	Other ethnic group	41 (0.01%)
Residence	United Kingdom	339725 (87.2%)
	Abroad	50049 (12.8%)
Education	preGCSE	10725 (2.8%)
	School	129757 (33.3%)
	Degree	230173 (59.1%)
	PhD	19119 (4.9%)
Occupation	Worker	266588 (68.4%)
	Retired	56566 (14.5%)
	Disabled / Not applicable / Sheltered employment	3988 (1.0%)
	Homemaker	12105 (3.1%)
	Unemployed / Looking for work	11970 (3.1%)
	Student	36089 (9.3%)
	Unknown	2468 (0.6%)

Table 5.2: Sociodemographic characteristics of participants from the GBIT study retained as controls for this study.

5.2.3 Statistical analysis

Data curation, pre-processing, and statistical analysis were conducted in Python 3.8.5. Data from each task were analysed separately due to incomplete/sparse dataset resulting from the adopted sampling strategy. I excluded data from participants with inaccurate sociodemographic input or absent questionnaire responses. Furthermore, data from the Words Memory Delayed task, completed more than 2 hours after the Words Memory Immediate task were considered invalid and removed as this

task involves the delayed recall of a set of words introduced in the latter, and it should be completed within the same session.

5.2.3.1 Summary performance measures and pre-processing

Data pre-processing steps are outlined in detail in Section 3.3. In brief, summary performance measures were extracted for each task. Observations presenting null summary accuracy and/or latency scores were excluded. The shape of the accuracy and latency distributions from the normative and patient data were compared to establish thresholds for non-compliance and filter outliers out. Summary performance measures from each task were transformed utilising normative data, generating Deviation from Expected (DfE) scores. These scores served as a proxy for measuring the level of cognitive impairment in pwMS attributable to MS rather than other co-varying sociodemographic variable. To obtain DfE scores for a given metric, a multiple linear regression model was trained using normative data to predict scores on that metric using sociodemographic and device labels. The model was then applied to predict scores for patient data. The residuals obtained by subtracting the expected from the observed scores were then normalised dividing by the standard deviation of normative data for that particular metric, yielding sociodemographically corrected scores expressed in standard deviation units. Finally, the cognitive scores were linked with clinical information available on the Register, including disease subtype, disease duration, and PROs. For disease subtype, the linkage was done only for people who had reported on this no more than two years before the online cognitive assessment.

5.2.3.2 Task sensitivity to MS

The mean and median values of the DfE scores for a given metric serve as indicators of effect size and were utilised to gauge task sensitivity to MS, employing Cohen’s notation of effect size (where effects with values of 0.2, 0.5, and 0.8 denote small, medium, and large effects, respectively). The median was preferred over the mean due to its stability in the presence of outliers and skewed distributions, as is the case here.

To evaluate whether each task, as described by its summary performance measures, was sensitive to MS, two-tailed one sample Wilcoxon signed-rank tests were conducted. Specifically, the DfE scores for a given performance metric were tested against a hypothetical median of zero to determine whether pwMS consistently differed from controls on that metric. A significance threshold of 0.05 was applied.

5.2.3.3 Correlations with patient-reported outcomes

Pairwise correlations between DfE scores and PROs across all task and questionnaire combinations were evaluated. Kendall’s Tau was selected as the correlation metric due to the non-normal distribution of cognitive data. Each correlation was evaluated including only pwMS who had completed the task during the online cognitive assessment and the questionnaire in the Register collection window either prior to or subsequent to the administration of the cognitive assessment. If the user had completed the questionnaire at both collection windows, the one completed during the window closest to the online cognitive assessment performance was retained. Using the extended version of Cohen’s notion of effect sizes, correlations with an absolute value of 0.1 or below were considered very small and negligible,

between 0.1 and 0.3 in the small range, between 0.3 and 0.5 in the medium range, and equal to or greater than 0.5 in the large range.

5.2.3.4 Effect of device variability

To evaluate the impact of device variability on task performance, multiple linear regression models were fitted to predict the primary outcome metrics across task. The models included sociodemographic variables and device label as regressors. Subsequently, ANOVA was conducted on the regression models to determine the significance of device impact and calculate eta squared, a measure used to quantify the effect size. In Cohen's notation, an eta squared value of 0.01 denotes a small effect, 0.06 indicates a medium effect, and values of 0.14 or higher represent large effects. Effects lower than 0.01 were considered negligible.

This analysis was specifically conducted on control data to identify tasks and performance metrics robust to device variability without introducing other potential confounding factors that might affect patient data, such as visuo-motor impairment. By focusing on control data and accounting for sociodemographic variables, the aim was to isolate the effect of device variability on task performance, ensuring its accurate and unbiased assessment.

5.2.3.5 Factor analysis

Factor analysis with varimax rotation was conducted on the pairwise correlation matrix obtained by correlating the DfE scores of pwMS across the different online cognitive tasks. For each task, one primary outcome metric was selected to construct the correlation matrix. In the majority of tasks, accuracy was preferred over latency as more representative of the cognitive ability being measured as well as more robust to visuo-motor impairment. However, for tasks specifically designed to assess visuo-motor abilities or information-processing speed (namely, Motor Control, Trail Making, SRT, and CRT), latency was chosen over accuracy.

In the absence of complete data, and since values were missing completely at random for the tasks that were randomly sampled, the pairwise deletion method was used to deal with missing values. This approach involved calculating each pairwise correlation coefficient using only cases with complete data for the pair of tasks being correlated. Given that the cognitive data violated the assumption of normality, Kendall's Tau was chosen as the correlation metric.

The Kaiser Criterion, which selects factors with eigenvalues greater than 1, was applied to determine the optimal number of factors. This criterion led to the retention of six factors. Varimax rotation was then applied to produced a more parsimonious set of loadings, enabling interpretation of the latent cognitive domains represented by each factor.

5.2.3.6 Task selection for the concise optimal battery

The selection of tasks to form a concise optimal battery for MS was guided by several criteria.

1. Sensitivity to MS: tasks needed to be sensitive to MS, either in terms of accuracy, latency, or

both, to effectively measure cognitive deficits related to this condition.

2. Low device sensitivity: to facilitate remote, unsupervised administration, tasks needed to exhibit minimal sensitivity to device variability, ensuring accessibility on any device.
3. Cognitive domains coverage: the selected tasks needed to span a broad range of cognitive domains while being decorrelated in their outputs. This ensures that each task assesses distinct cognitive domains, rather than all tasks loading onto one general intelligence factor (G factor (Hampshire et al. 2012)).
4. Comparability: some degree of overlap with tasks selected for other studies was preferable, to allow for comparisons across different clinical groups.
5. Availability of control data: sufficient control data needed to be available for each task to support robust normative modelling.
6. Battery length: while there was no strict requirement for battery length, the final battery needed to be concise enough to eliminate the need for random sampling across tasks and to facilitate having more complete data in subsequent timepoints/studies.

In summary, for each identified factor, one to three tasks exhibiting high and distinct loadings (i.e., strong influence on a specific factor with minimal influence on others) were retained. In cases where multiple tasks had high loadings on a factor, preference was given to those with higher MS sensitivity and lower device sensitivity. Some tasks were included regardless of their moderate to high device sensitivity. These tasks measure aspects of cognition that are typically challenging to assess without confounding factors, such as visuo-motor impairment, but are crucial for a comprehensive understanding of cognitive function in MS, i.e., where such confounding factors would therefore have to be more carefully controlled for during analysis.

5.3 Results

5.3.1 Feasibility of online cognitive assessment in people with MS

First, I evaluated the feasibility of online cognitive assessment in pwMS by investigating user engagement in the study.

5.3.1.1 Participation rates

A total of 3066 users of the UK MS Register participated in the assessment. People who engaged did not differ significantly from the overall Register population in terms of age and disease duration, but did differ based on gender, with a slightly higher proportion of women engaging (3.2% more women, Table 5.3).

5.3.1.2 Completion rates

Not all participants completed the entire assessment, which consisted of the 12 online cognitive tasks assigned to them (Fig. 5.2A). Nonetheless, at least 1000 participants completed each task (Table 5.4). As detailed in Fig. 5.2A, among the total participants, 78.4% completed the whole assessment, with 93.4% having done so in a single session. The remaining portion of participants either completed the assessment with breaks within a single day or completed it on a different day, likely prompted by the email reminders.

5.3.1.3 Dropout rates

For those who either completed the assessment on different days or did not complete it (24.6%), I looked at data only from the first day they engaged to investigate whether there was a particular task or number of tasks where they stopped at. The percentage of dropouts was evenly distributed across tasks (Fig. 5.2C), ranging from 6.30% to 2.4%, with only the first two fixed tasks showing higher dropout rates (7.7% Motor Control; 16.2% Words Memory Immediate). 37% of the participants who dropped out, did so during the first 3 tasks.

To further investigate the motives behind participants dropping out, I compared the demographics, clinical phenotype, disease duration, and disease severity of those who completed the assessment in one session, completed in more than one session, and did not complete (Table 5.3). The analysis revealed that the three groups did not differ significantly based on age, disease duration and clinical phenotype. Nonetheless, they did differ significantly based on gender and disease severity, with non-completers having a higher proportion of males (a difference of approximately 10% to the other two groups) and both non-completers and completers in more than one session presenting with higher MSIS-29 and webEDSS scores (a difference of respectively around 5 and 0.5 respectively compared to the other group).

5.3.1.4 Engagement of active users

Finally, I looked at whether the pwMS who engaged with the online cognitive assessment were the same that periodically engage with other Register functionalities (e.g. with PROs data collection). As depicted in Fig. 5.2B, most of the pwMS who participated in the assessment were also “active users” of the UKMSR, with over 80% of them having completed four of the administered PROs during the data collection window either prior to or subsequent to the administration of the cognitive assessment. The completion rate was lower (60.1%) for the MSWS-12, primarily due to its inapplicability to people with limited or no ambulatory capacity. In fact, among pwMS who engaged with the assessment but did not complete the MSWS-12, 69.3% (constituting 27.7% of the total) reported that they could not complete it precisely because they cannot walk. Nevertheless, this outcome is positive as it indicates that the assessment was accessible to people in this category as well. Additionally, a substantial proportion of pwMS who were not active users (17%) successfully participated, resulting in an overall increase in engagement with the UKMSR.

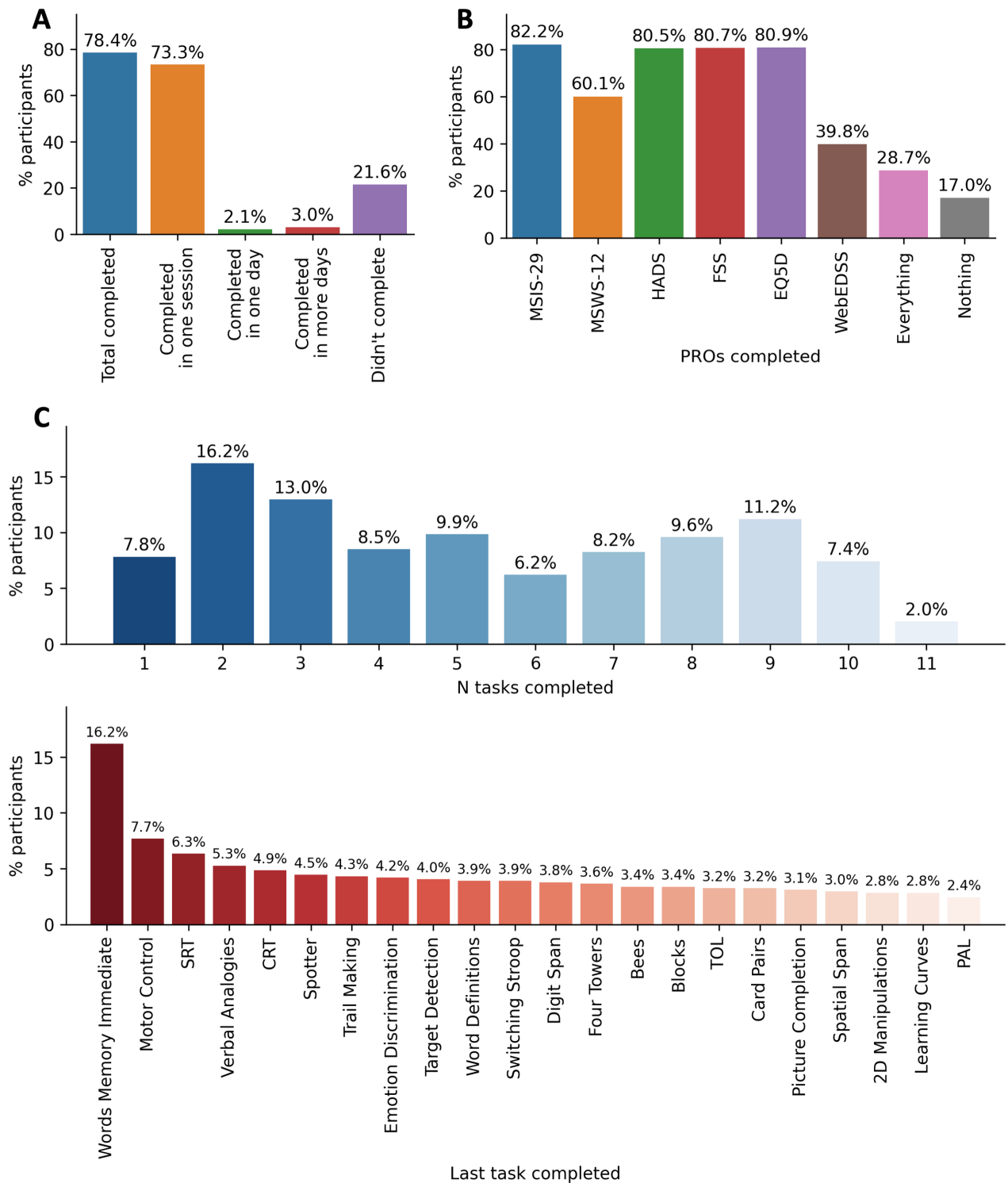


Figure 5.2: Engagement of people with MS with online cognitive assessment. (A) Percent of participants who completed, overall, in one session, in one day, or in more days, and who did not complete. (B) Percent of participants who also completed the different PROs on the UK MS Register website at the data collection windows either prior to or subsequent to the cognitive data collection. (C) Percent of participants who dropped out after completing a certain number of tasks (above) or a particular task (below).

Groups	N	Age (min-max, mean)	Gender (%)	Disease subtype (%)	Disease duration (min-max, mean)	Disease severity (mean MSIS-29 motor norm / webEDSS)
Completed in one session	2247	16-87, 54.4 (missing = 2)	Female = 79.0 Male = 20.8 Other = 0.2 (missing = 2)	RR = 54.9 SP = 29.2 PP = 13.5 BN = 2.4 (missing = 512)	1-67, 20.3 (missing = 512)	40.5 / 4.8 (missing = 358 / 1320)
Completed in more than one session	158	25-79, 54.9	Female = 82.9 Male = 17.1 Other = 0.0	RR = 50.4 SP = 32.3 PP = 15.7 BN = 1.6 (missing = 31)	0-46, 21.3 (missing = 31)	44.2 / 5.4 (missing = 24 / 96)
Did not complete	661	20-82, 54.6 (missing = 2)	Female = 70.7 Male = 29.0 Other = 0.3 (missing = 2)	RR = 50.6 SP = 32.3 PP = 15.1 BN = 2 (missing = 159)	1-55, 20.9 (missing = 159)	45.7 / 5.0 (missing = 163 / 431)
P values from groups comparison	-	0.8225	0.0001	0.6045	0.4679	0.0004 / 0.0282
Engaged with the assessment	3066	16-87, 54.4 (missing = 4)	Female = 77.4 Male = 22.4 Other = 0.2 (missing = 4)	RR = 53.7 SP = 30 PP = 14 BN = 2.2 (missing = 702)	0-67, 20.5 (missing = 702)	41.7 / 4.8 (missing = 545 / 1847)
Overall UK MS Register population (up to Spring 2022)	19188	18-90, 54.3	Female = 74.2 Male = 25.8	-	0-73, 20.1	-
P values from groups comparison	-	0.5196	0.000	-	0.1546	-

Table 5.3: Demographics, clinical phenotype, disease duration, and disease severity across participants who completed, in one or more sessions, participants who did not complete, participants overall and overall UK MS Register population. P-values were obtained using t-tests, ANOVA or chi-squared tests where appropriate.

Tasks	N participants (before pre-processing)	N participants (after pre-processing)
Bees	1111	960
Blocks	1223	1213
Card Pairs	1235	1227
CRT	1264	1180
Digit Span	1173	1136
Emotion Discrimination	1210	1200
Four Towers	1198	1157
2D Manipulations	1187	1169
Motor Control	3065	3029
PAL	1206	1196
Picture Completion	1242	1234
Spatial Span	1151	1130
Spotter	1110	1071
SRT	1233	1194
Switching Stroop	1252	1229
Target Detection	1268	1246
TOL	1212	1152
Trail Making	1207	1197
Verbal Analogies	1242	1225
Word Definitions	1241	1223
Words Memory Delayed	2412	2222
Words Memory Immediate	3021	2980

Table 5.4: Number of participants who completed each task, before and after pre-processing procedures were applied.

5.3.2 User feedback on testing time

At the conclusion of the assessment, participants who completed were asked to provide feedback on the appropriateness of the testing time. The results revealed that 55.7% of the participants reported that the testing time was either appropriate or could have been longer, while 44.3% felt that it was too long (Fig. 5.3). Given that some participants did not complete the assessment in a single session, I further examined whether there were differences in feedback on testing time between those who completed in one session vs those who completed in more sessions. The analysis indicated that the group of participants who did not complete the assessment in one session had a higher prevalence of individuals (13.3% higher) who reported that the testing time was too long. Nevertheless, 43.2% of the participants in this group still perceived the assessment time as appropriate.

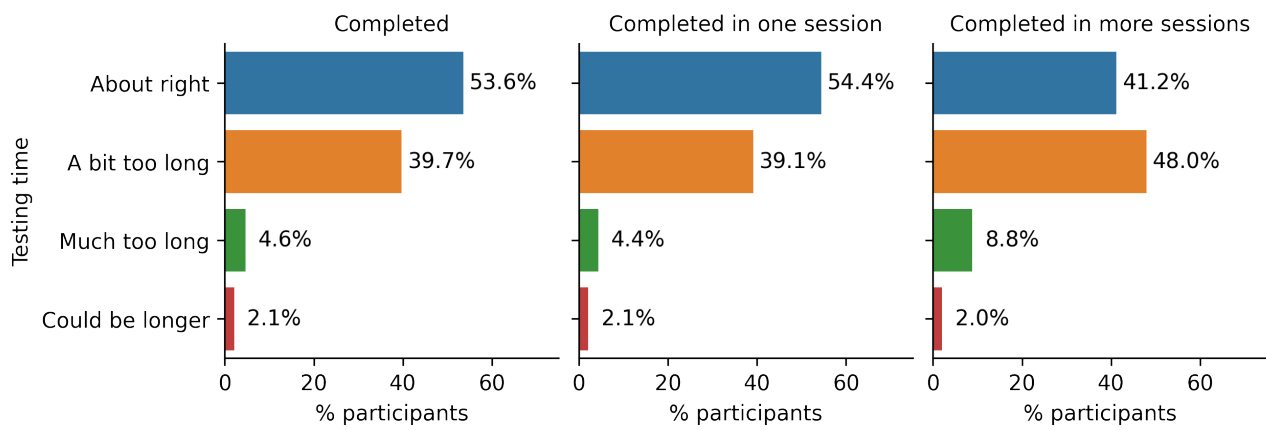


Figure 5.3: Feedback on online cognitive assessment duration. Feedback on testing time and how it varied across three groups: everyone who completed, participants who completed in one session and participants who completed in more sessions.

5.3.3 Participants

Overall, 3048 of the original 3066 participants were included in the analysis after pre-processing procedures were applied. Table 5.5 presents their sociodemographic characteristics. Valid and recent clinical labels were available for 2356 participants.

Socio-demographic characteristics		People with MS
Total N		3048
Age (min-max, mean)		16-87, 54.4
Gender	Female	2359 (77.4%)
	Male	682 (22.4%)
	Other	7 (0.2%)
Dominant hand	Right	2653 (87%)
	Left	320 (10.5%)
	Ambidextrous	75 (2.5%)
First language	English	2934 (96.3%)
	Other	114 (3.7%)
Ethnicity	White	2888 (94.8%)
	Unknown	59 (1.9%)
	Mixed or multiple ethnic groups	42 (1.4%)
	Asian or Asian British	35 (1.1%)
	Black, Black British, Caribbean or African	22 (0.7%)
	American Hispanic	2 (0.1%)
Residence	United Kingdom	3025 (99.2%)
	Abroad	23 (0.8%)
Education	preGCSE	44 (1.4%)
	School	1394 (45.7%)
	Degree	1515 (49.7%)
	PhD	95 (3.1%)
Occupation	Worker	1232 (40.4%)
	Retired	1117 (36.6%)
	Disabled / Not applicable / Sheltered employment	492 (16.1%)
	Homemaker	125 (4.1%)
	Unemployed / Looking for work	53 (1.7%)
	Student	21 (0.7%)
	Unknown	8 (0.3%)
Disease Subtype (692 missing)	Benign	53 (2.2%)
	Relapsing-remitting	1268 (53.8%)
	Secondary progressive	704 (29.9%)
	Primary progressive	331 (14.0%)
Disease Duration (min-max, mean)		0-67, 20.3 (692 missing)

Table 5.5: Sociodemographic characteristics of pwMS from the UK MS Register who participated in the online cognitive assessment and were retained for analysis after pre-processing procedures were applied.

5.3.4 Sensitivity to MS of the evaluated online cognitive tasks

Next, I assessed the sensitivity of the evaluated online cognitive tasks to MS by comparing task performance of pwMS and matched controls. This analysis was aimed at determining whether each task could be utilised to differentiate between the two groups and identify the cognitive domains most affected in MS.

Median DfE scores with 95% confidence intervals across all performance metrics for each task are presented in Fig. 5.4A including p-values obtained performing Wilcoxon signed-rank tests against zero.

5.3.4.1 Performance accuracy scores

When analysing accuracy scores, pwMS were significantly worse than controls across all tasks except for SRT, CRT, Bees, Trail Making, Motor Controls, and Four Towers. They exhibited negative median DfE accuracy scores across all remaining tasks, representing very small to large effect sizes (Fig. 5.4A). The cognitive tasks most sensitive to MS in terms of accuracy were Spotter, Switching Stroop, Picture Completion, and Card Pairs, all of which exhibited a medium-large effect size. These tasks respectively measured sustained attention, cognitive control, visuo-spatial problem-solving, and associative working memory, indicating the presence of cognitive deficits spanning multiple domains.

5.3.4.2 Performance latency scores

When analysing latencies, pwMS were significantly slower than controls across all tasks except for TOL, Verbal Analogies, Bees, and Blocks. They exhibited positive median DfE latency scores across all remaining tasks, with medium-large effect sizes for Trail Making, SRT, Card Pairs, Switching Stroop, and Picture Completion (Fig. 5.4A). This indicated that information-processing speed was affected together with the domains previously identified.

5.3.4.3 Performance costs in accuracy and latency

It is difficult to determine the extent to which the observed deficits, particularly in terms of latency, were due to visuo-motor or cognitive impairment or a combination of both. This issue does not apply when examining costs, as the contribution of visuo-motor impairment it is nullified in the subtraction process, as it affects both conditions or tasks in a similar manner as long as the visuo-motor demands are balanced across trial types, which is a fundamental principle of task design.

Fig. 5.4B presents the DfE scores for tasks for which some type of costs in terms of accuracy and/or latency could be evaluated. These typically focus around cognitive control, as control processes operate through the processes they control and must be measured subtractively. Switching Stroop was overall confirmed as the most sensitive task to MS, as it showed significant deviations in accuracy and latency for pwMS compared to matched controls also when examining costs. Particularly interesting was the median DfE incongruency cost for accuracy which was negative, indicating a greater performance drop compared to controls, and exhibited a medium-large effect size. The median DfE switch and incongruency costs for latency were instead positive, indicating a greater cognitive control process

time cost compared to controls, but exhibited only very small effect sizes. Similar results were found for the CRT choice cost and Trail Making switch cost.

The incongruency costs for Emotion Discrimination were non-significant, whereas the distractor costs for Picture Completion were significant but characterised by very small effect sizes. The percent forgetting for Words Memory Delayed was also a form of cost, with a positive median DfE score for this metric indicating that pwMS forget more items than controls in the time between Words Memory Immediate and Delayed. In this case, the percent forgetting score was non-significant, indicating that there was no increased failure of consolidation when comparing pwMS to controls.

The Bees task provided insights into pwMS's decision impulsivity, as measured by the two additional performance metrics presented in Fig. 5.4B. PwMS significantly differed from controls on both metrics, indicating a less impulsive behaviour, but exhibited very small effect sizes.

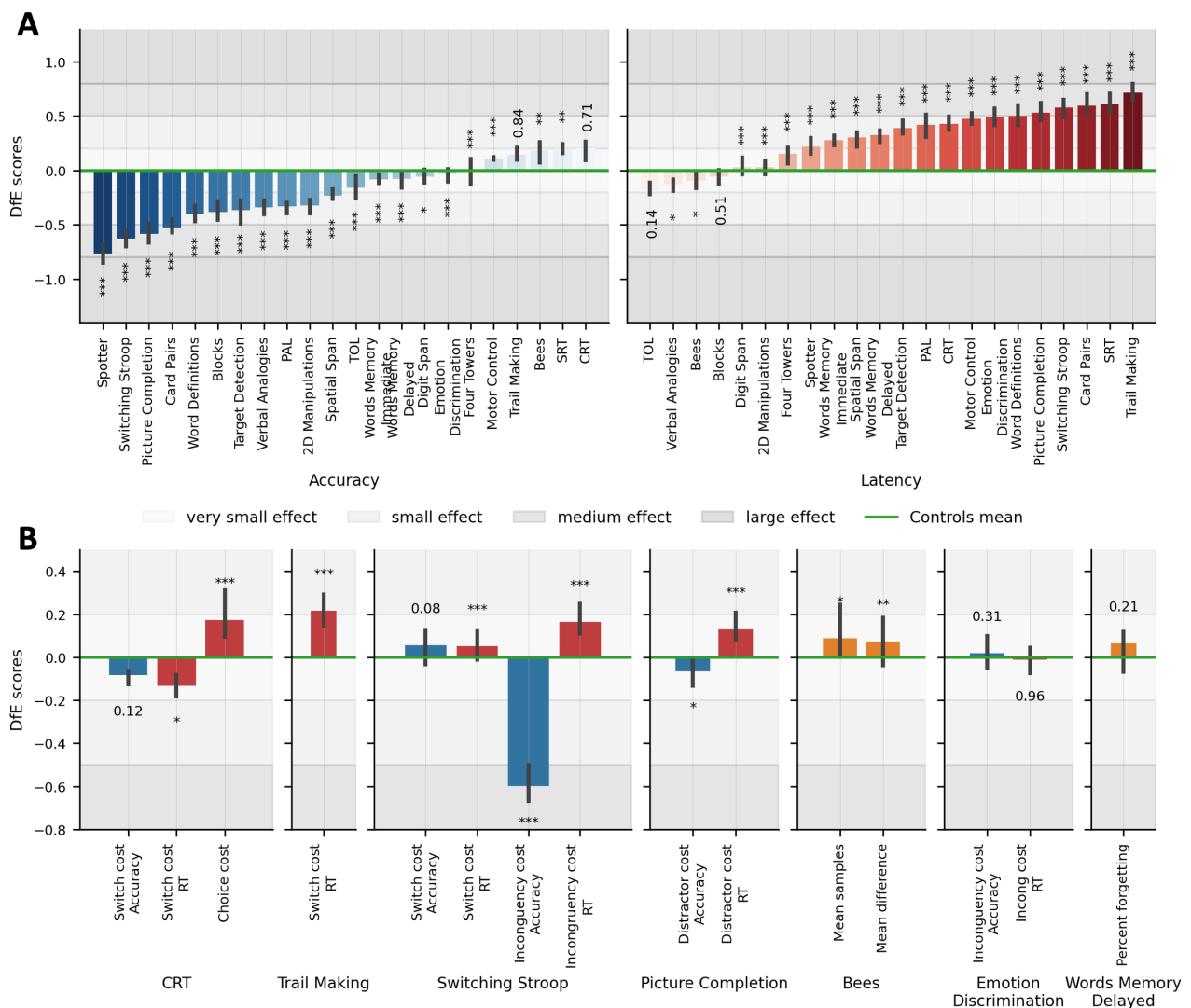


Figure 5.4: Sensitivity to MS across tasks. Median DfE score for accuracy, latency (A), and extra performance metrics (B) across tasks with 95% confidence intervals. P values are obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.

5.3.5 Correlations with patient-reported outcomes

After examining the sensitivity of online cognitive tasks to MS, the next step was to investigate how performance measures, particularly accuracy and latency, correlated with other disease-related factors such as physical disability, mental health, fatigue, and quality of life. PROs collected periodically through the UK MS Register were utilised for this purpose. The included PROs were: MSIS-29 motor and MSWS-12 to measure motor and walking abilities, HADS (anxiety and depression component) to assess mental health, FSS to measure fatigue, and EQ5D to evaluate health-related quality of life. Additionally, I compared cognitive function with disability severity using the webEDSS as a proxy.

Fig. 5.5 shows Kendall's rank correlations between performance measures across tasks and PROs. The number of observations available for each task-questionnaire combination is reported in Table 5.6.

5.3.5.1 Correlations between accuracy scores and patient-reported outcomes

The correlations between accuracy measures and PROs were consistently negative, indicating that higher PROs, reflecting higher disability, corresponded to lower DfE scores, i.e. larger deficits in accuracy when comparing pwMS to controls (Fig. 5.5A). Nonetheless, most correlations were statistically non-significant or very small ($|\tau| < 0.1$), indicating only mild associations with PROs. Medium ($0.1 \leq |\tau| < 0.3$) negative correlations were observed between the PROs measuring motor impairment and disability severity (MSIS-29 motor and webEDSS) and the accuracy measures on the Spotter, 2D Manipulations, Verbal Analogies, Word Definitions, and Target Detection tasks.

The largest correlations were observed for the Spotter task. PwMS that exhibited a greater deficit on the Spotter task when compared to controls appeared to have more severe physical disability as indicated by higher MSIS-29 motor and webEDSS scores, and lower quality of life as indicated by higher EQ5D summary score.

5.3.5.2 Correlations between latency scores and patient-reported outcomes

In contrast, the correlations between latency measures and PROs were predominantly positive, consistent with the expectation that higher PROs would be associated with increased latencies in task performance compared to controls (Fig. 5.5B). For latency measures there were more correlations in the medium range ($0.1 \leq |\tau| < 0.3$) compared to accuracy measures. The largest correlations were observed when comparing latency measures with PROs measuring motor impairment. Only two tasks, namely TOL and Four Towers, exhibited non-significant correlations with all the motor-related PROs in terms of latency. Medium correlations were also observed with the EQ5D summary score.

Remarkably, correlations with anxiety were all non-significant or negligibly small, suggesting that general anxiety did not significantly impact performance on cognitive tasks.

5.3.5.3 Correlations between costs and patient-reported outcomes

Correlations between costs, estimated through contrasts, and PROs are not reported because predominantly non-significant and all negligibly small ($|\tau| < 0.1$).

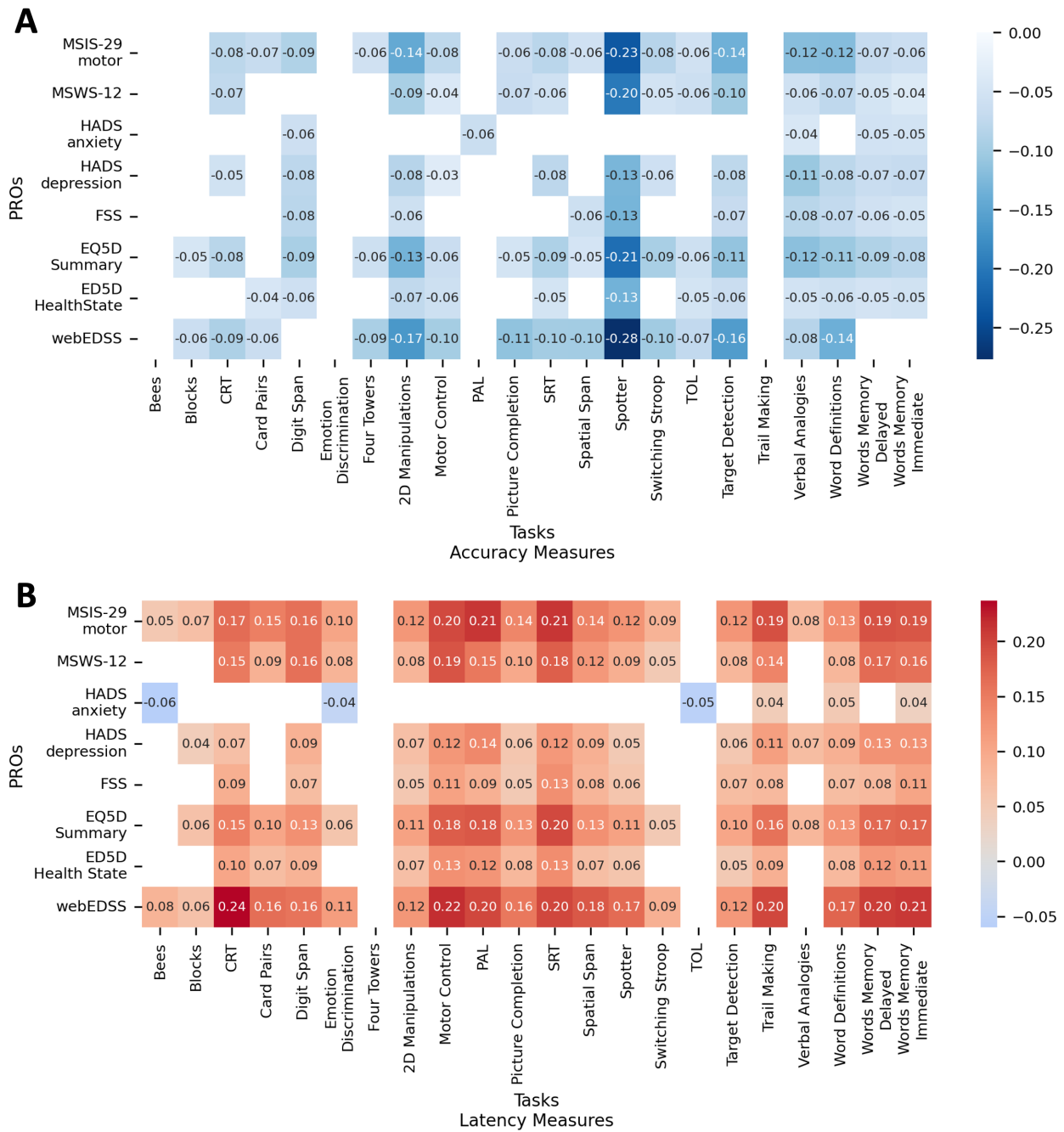


Figure 5.5: Correlations between summary performance measures and patient-reported outcomes. Kendall's rank correlations of accuracy (A) and latency (B) DfE scores across tasks with patient-reported outcomes. Only significant correlations ($p \leq 0.05$) are reported to increase readability.

Tasks	MSIS-29	MSWS-12	HADS	FSS	EQ5D	webEDSS
Bees	800	602	780	778	787	409
Blocks	1011	740	998	999	1002	508
Card Pairs	1020	759	995	1000	999	505
CRT	981	710	957	959	966	460
Digit Span	959	725	942	944	946	459
Emotion Discrimination	998	732	979	982	989	466
Four Towers	965	711	942	940	944	489
2D Manipulations	984	723	967	966	970	496
Motor Control	2511	1839	2459	2464	2470	1212
PAL	1007	744	989	989	994	499
Picture Completion	1048	781	1030	1028	1033	501
Spatial Span	950	726	931	931	933	443
Spotter	896	653	880	880	884	427
SRT	991	724	970	972	979	474
Switching Stroop	1019	750	1008	1002	1010	493
Target Detection	1040	757	1018	1015	1022	506
TOL	976	724	965	967	966	486
Trail Making	996	741	974	977	981	469
Verbal Analogies	1032	764	1015	1015	1017	503
Word Definitions	1021	756	1005	1005	1009	510
Words Memory Delayed	1874	1395	1840	1838	1847	917
Words Memory Immediate	2478	1822	2428	2431	2440	1196

Table 5.6: Number of observations available for each task-questionnaire combination, on which the pairwise correlation calculation is based.

5.3.6 Selection of online cognitive tasks for the optimal battery

The final step of this chapter involved selecting a subset from the initially evaluated 22 online cognitive tasks to form a reduced battery, specifically tailored for assessing cognitive deficits in pwMS and monitoring changes in cognitive function over time. This selection process was based on an in-depth analysis of three key aspects: task sensitivity to MS, task sensitivity to device being used, and outcomes from factor analysis.

5.3.6.1 Task sensitivity to MS

Table 5.7 provides a detailed overview of each task, highlighting for each the summary outcome metric that was selected as most representative of performance on the task. The median DfE score for that metric, and a comparative ranking based on its value are also reported. This analysis enabled the identification of tasks with higher sensitivity to MS.

5.3.6.2 Task sensitivity to device

Fig. 5.6D illustrates device sensitivity of accuracy and latency measures across tasks, evaluated using control data. This sensitivity, quantified by eta squared for device, was generally negligible or small for most accuracy measures, except in the Motor Control and Trail Making tasks. For latency measures, instead, sensitivity to device was found to be higher in tasks designed to measure visuo-motor abilities (Motor Control), information-processing speed (SRT and CRT), attention (Spotter), and different aspects of working memory (Card Pairs, Spatial Span, PAL, and Digit Span). Overall, this meant that, for example, screen size did not have a large effect on how accurately users responded, but just on how long it took them to respond. In fact, device sensitivity was higher in tasks for which users were required to click on different points of the screen, e.g. the working memory tasks, and to make more clicks on a comparable period of time, e.g. the motor, information-processing speed, and attention tasks. It makes sense that Motor Control and Trail Making were the only two tasks affected in terms of accuracy as Motor Control was measuring how spatially accurate users were when clicking on the target and Trail Making involved detecting the right element to click among other similar elements: for both tasks it was easier to make errors on smaller screens.

5.3.6.3 Factor analysis

The factor analysis, applied to the most representative metrics from the 22 tasks, identified six significant factors (*eigenvalues* > 1), as shown in Fig. 5.6B. The first factor accounted for 15.9% of the variance, with all six factors collectively explaining 31.6% of the variance. The loadings of the features onto the factors after varimax rotation are reported in Fig. 5.6C. Each factor had loadings that were interpretable as relating to a distinct cognitive domain. These were: information-processing speed, visual problem-solving, working memory, verbal abilities, memory, and attention.

5.3.6.4 Task selection

For each factor, a maximum of three tasks were selected based on their high loading, sensitivity to MS, and possibly low sensitivity to device ($\eta^2 < 0.06$). In the information-processing speed domain, Motor Control, CRT, SRT, and Trail Making had the highest loadings. Motor Control was specifically chosen for its relevance to hand-eye processing, crucial in a population with notable visuo-motor impairment. SRT and Trail Making were preferred over CRT due to their heightened sensitivity to MS. For visual problem-solving, Blocks and Switching Stroop were selected. Blocks was preferred over Picture Completion, which had a higher loading and a superior sensitivity to MS, to allow for better deployment (lower sensitivity to device) and comparison with other studies (Blocks is already present in many other studies using Cognitron to measure cognitive function in other neurological populations). Switching Stroop was selected as it ranked third in sensitivity to MS and allowed measuring cognitive control through costs, evaluated by contrasting trials with different conditions.

In the working memory domain, Card Pairs and 2D Manipulations were chosen for their high loadings and MS sensitivity, coupled with their low device sensitivity. For the same reasons Verbal Analogies and Words Definitions represented the verbal abilities domain. For the memory factor, Words Memory Immediate and Delayed were selected, despite not being particularly sensitive to MS, to maintain

comparability with other population studies and because they were the only evaluated tasks measuring memory. Finally, Spotter was retained for the attention domain, as it was the task with the highest loading on this factor and the most sensitive to MS in terms of accuracy.

Task	Main Score	Median DfE score	Rank	Sensitive to device	Factor
Spotter	Accuracy	-0.76	1	Small	Attention
Trail Making	Latency	0.72	2	No	Information processing speed
Switching Stroop	Accuracy	-0.63	3	No	Visual problem-solving
SRT	Latency	0.61	4	Yes	Information processing speed
Picture Completion	Accuracy	-0.58	5	Small	Visual problem-solving
Card Pairs	Accuracy	-0.53	6	No	Working memory
Motor Control	Latency	0.48	7	Yes	Information processing speed
CRT	Latency	0.43	8	Yes	Information processing speed
Word Definitions	Accuracy	-0.40	9	No	Verbal abilities
Blocks	Accuracy	-0.38	10	No	Visual problem-solving
Target Detection	Accuracy	-0.37	11	No	Information processing speed
Verbal Analogies	Accuracy	-0.34	12	No	Verbal abilities
PAL	Accuracy	-0.33	13	No	Working memory
2D Manipulations	Accuracy	-0.32	14	No	Working memory
Spatial Span	Accuracy	-0.23	15	No	Working memory
TOL	Accuracy	-0.16	16	No	Visual problem-solving
Words Memory Immediate	Accuracy	-0.08	17	No	Memory
Words Memory Delayed	Accuracy	-0.08	18	No	Memory
Digit Span	Accuracy	-0.05	19	No	Verbal abilities
Emotion Discrimination	Accuracy	-0.03	20	No	Visual problem-solving
Four Towers	Accuracy	0.00	21	No	Visual problem-solving
Bees	Accuracy	0.19	22	No	Visual problem-solving

Table 5.7: List of the evaluated online cognitive tasks with the metric selected as most representative for each. Additionally, the median DfE score for this metric, along with its comparative ranking based on its value are reported. Moreover, for each task it is indicated whether its most representative metric is sensitive to device and the underlying cognitive domain onto which it primarily loads. Selected tasks are highlighted in grey.

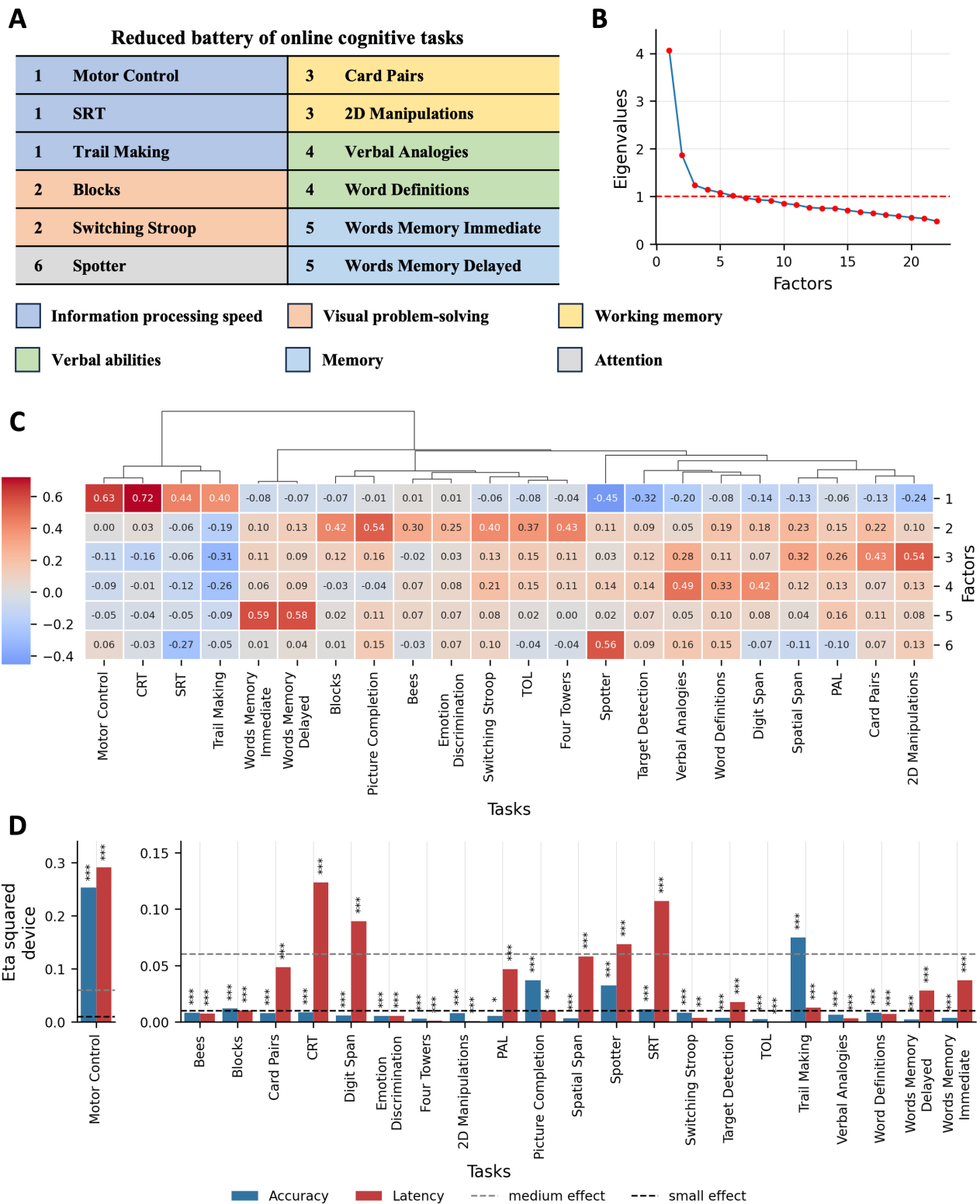


Figure 5.6: Task selection for the optimal battery. (A) Tasks selected for the optimal MS reduced battery with underlying cognitive domain they correspond to. (B) Scree plot of eigenvalues derived from factor analysis of most representative performance metrics across tasks. (C) Loadings of the tasks onto the different rotated factors, which represent underlying cognitive domains. (D) The device effect size for accuracy and latency measures across tasks, measured in terms of eta squared. Values ≤ 0.06 are small and values ≤ 0.01 are negligible.

5.4 Discussion

5.4.1 Summary of findings

The study presented in this chapter demonstrated the feasibility of online cognitive assessment in pwMS through the collection of an unprecedented amount of objective task performance data over a short span of time. The presented online cognitive assessment showed good discriminative validity, with pwMS performing worse than control on all the evaluated tasks. The cognitive domains measured by the tasks showing the highest sensitivity to MS align with the domains found to be most affected in MS literature (Benedict et al. 2006; Chiaravalloti et al. 2008; Rao et al. 1991b), providing a further validation of the assessment. Finally, a reduced battery of online tasks was identified through the study to assess multiple aspects of cognition whilst being specifically tailored for detecting impairment in MS and deployable in a home setting. Together, these results support the further use of online cognitive assessment in MS and foresee the transformative effect that could have in the field by enabling large-scale efficient data collection and longer follow-ups.

5.4.2 Evidence in favour of the feasibility of online cognitive assessment in pwMS

With over 3,000 pwMS having taken part, this study ranks high among the biggest studies ever conducted on cognition in MS. While the benefits of online cognitive assessment are widely known (Feenstra et al. 2017), it was not clear before this study how pwMS would respond to it, and if it would prove to be accessible for this population. This study provided evidence of the feasibility of online cognitive assessment in pwMS.

First, the high participation: 16% of the total users in the UK MS Register participated, with minimal prompting and advertising being done (each user only received one invitation and two reminder emails). Moreover, 17% of the participants were not active users indicating that the assessment was able to spark interest in people who do not normally engage with the Register and was accessible to people who are potentially less accustomed to using the Register online portal.

Second, the completion rate was as high as 78.4%: this was satisfactory and within the range observed for other studies carried out using the Cognitron on the normal population. The high completion rate together with the lack of statistical differences between completers and non-completers in terms of age, disease subtype, and disease duration, and the significant but minimal differences in disease severity indicated that the assessment was feasible for pwMS and if people did not complete it, it was unlikely because of disease-related factors.

A factor contributing to non-completion could be inferred looking at dropout rates: 37% of dropouts occurred during the first three tasks, likely representing participants who were curious to explore the assessment but lacked time or motivation to fully engage with it. Moreover, dropout rates oscillated but remained consistent in the rest of the battery, indicating that increased cognitive fatigue caused by battery duration was not the main driver of dropouts. Additionally, the dropout rate was fairly uniform (between 6.3% and 2.4%) across the randomly sampled tasks indicating that there were no specific tasks that participants particularly struggled with. Finally, the fact that 93.4% of the people who completed did so in a single session, generally indicated that participants could complete the assessment without needing breaks.

Nonetheless, dropouts did occur in 24.6% of cases, and assessment duration and cognitive fatigue experienced by participants, even if not the main drivers, could be contributing factors. To investigate this, I examined responses to the feedback question regarding the appropriateness of the testing time. While most completers found the testing time appropriate, non-completers did not have the opportunity to provide such feedback. Moreover, an indicator of testing time being too long was given by the increased prevalence of participants expressing concerns about assessment duration observed among those who completed in multiple sessions. Therefore, shortening the assessment duration to ensure completion both in general and within a single session could be beneficial in the future.

Further evidence in favour of the feasibility of online cognitive assessment in pwMS is the representativeness of the sample being tested: the people who engaged with the assessment were not statically distinct from the overall Register population in terms of age and disease duration, and whilst they had a higher proportion of females, it was only by 3%. Additionally, the relative prevalence of MS subtypes in the sample being tests is consistent with estimates from epidemiological studies with 83.7% being RRMS at onset (Cottrell et al. 1999; Klineova et al. 2018), 30% transitioning to SPMS (Weinshenker et al. 1989a), 14% being PPMS (Cottrell et al. 1999; Klineova et al. 2018), and 2.2% being Benign (Tallantyre et al. 2019). Overall, this bodes well for the generalisability of these results to the wider MS population and indicates a good accessibility of the tool to pwMS from any subtype and stage.

Taken together, these results demonstrate that a large and representative sample of pwMS in the UK was interested in testing their cognition using an online tool and successfully managed to do so opening the possibility to have such tool permanently available as part of the UKMSR portal.

5.4.3 Benefits of online cognitive assessment

Such scale and representativeness of the sample were achieved within a short timeframe partly owing to the unique features of online cognitive assessment: efficiency, accessibility, and engagement through gamification (Barak et al. 2004; Barak et al. 2002; Bauer et al. 2012; Del Giovane et al. 2023b; Feenstra et al. 2017). The assessment was in fact widely accessible to anyone in possess of any device, be it a smartphone, tablet, laptop, or computer, at any time of the day and from any location of their choice, fostering inclusivity and practicality. With no personnel and equipment requirements, it was substantially more time- and cost-effective compared to traditional neuropsychological assessments, not only in administration but also in scoring and data storage.

A further strength was that the standardised nature of the online assessment ensured that all participants were presented with identical information, resulting in greater consistency and comparability of results. Finally, the availability of normative data from over 300,000 people from the general UK population (Hampshire 2020; Hampshire et al. 2020; Mosley 2020) allowed ensuring a culture fair assessment achieved systematically controlling for sociodemographic variables and device variability which are factors known to affect cognitive performance (Del Giovane et al. 2023b; Hampshire et al. 2021b).

5.4.4 Validity of the evaluated assessment and cognitive domains affected in MS

Having ascertained its accessibility to pwMS, the next step was to evaluate the validity of this particular online cognitive assessment in detecting cognitive deficits in pwMS. Even if to a different extent, pwMS performed worse than controls on all the evaluated cognitive tasks and these differences remained even after accounting for sociodemographic variables and device variability, thereby proving that the tool had good discriminative validity.

The cognitive domains identified as affected, as measured by the tasks showing the highest sensitivity to MS, were in close concordance with those known to be affected from MS literature: information-processing speed (Costa et al. 2017; DeLuca et al. 2004; Forn et al. 2008), attention (McCarthy et al. 2005), cognitive control (Benedict et al. 2006; Drew et al. 2008), visuo-spatial processing (Benedict et al. 2006; Vleugels et al. 2000), and working memory (DeLuca et al. 2004; Lengenfelder et al. 2003; Pelosi et al. 1997). This on the one hand confirmed previous findings holistically, with a different assessment method, and using a larger and less biased sample. On the other hand, it provided a further validation for the technology which identified the same set of domains as affected, without previous knowledge from the literature being used to pre-select the set of tasks to be evaluated.

Additionally, using this technology, it is possible to measure different aspects of the affected domains and provide more detailed information on these aspects, offering more comprehensive performance metrics. A good example of this is the Switching Stroop task, for which six performance metrics were extracted, offering an overview of the different aspects of the executive function domain that could be affected. Measures extracted for this task include accuracy and response time, and how they change when participants are presented with conflicting vs coherent information on the screen, or when the rule participants have to follow to respond changes vs stays the same across subsequent trials.

Some of the results from this study were unsurprising, like the identification of information-processing speed and working memory among the most affected domains. Others were less expected and highlighted how domains such as attention and cognitive control which were reported to be less affected in the literature (Benedict et al. 2020; Chiaravalloti et al. 2008), appeared here to be equally affected. A notable example was the Spotter task which measures sustained attention and showed the highest sensitivity to MS in terms of accuracy. However, accuracy on this task also showed the largest negative correlation ($r = -0.28$) with the webEDSS. This could potentially represent different aspects of disability affecting performance on the Spotter task, such as visuo-motor impairment or trouble concentrating (commonly referred to as "brain fog"), which is the construct that this task is designed to measure. Considering that the correlation between the accuracy on this task and the MSIS-29 motor ($r = -.23$) was very similar to that with the webEDSS, while the correlation with the FSS was smaller ($r = -.13$), it is more likely that the high correlation between the webEDSS and the Spotter task was being driven by the visuo-motor component of disability. Further evidence for this interpretation can be found by correlating the webEDSS and the MSIS-29 motor directly, for which $r = 0.758$. Nonetheless, other evidence from the literature suggests that even if decreased cognitive performance does not correlate with self-reports of fatigue (Johnson et al. 1997), as it is also the case for this study, fatigue, when measured more objectively, has an effect on cognitive performance (Krupp et al. 2000). This is especially observable through decrements in performance over time in tasks where sustained attention is required (Schwid et al. 2003). This is a clear example of a case where the possibility to collect richer and more detailed trial-by-trial data offered by computerised assessments could make

the difference: this type of data is available for the Spotter task and is being analysed as part of a separate study to determine whether the observed association with disease severity is attributable to visuo-motor impairment, fatigue, or a combination of both.

Another surprising finding was the minimal sensitivity to MS exhibited by the two memory tasks, Words Memory Immediate and Words Memory Delayed, considering that memory is reported to be widely affected (Rao et al. 1993) and measured in the most common neuropsychological batteries in MS (Benedict et al. 2006; Boringa et al. 2001; Langdon et al. 2012). A possible explanation is that these two tasks measure recognition memory rather than recall memory, which is the standard construct measured by neuropsychological tests (Edgar et al. 2011). However, aligning with notions from the literature, the percent of items forgotten between the immediate and the delayed recall is not significantly different between patients and controls, and both tasks show similar effect sizes, indicating that memory problems experienced by pwMS are dependent on difficulties initially acquiring and encoding rather than consolidating information (DeLuca et al. 1994, 1998).

5.4.5 Comparison with traditional neuropsychological batteries

Traditional neuropsychological batteries commonly used in MS tend to prioritise information-processing speed and memory leaving aside domains such as attention (not included in the BICAMS, in the BNB-N, nor in the MACFIMS), cognitive control, and visuo-spatial processing (not included in the BICAMS nor in the BNB-N), which are highlighted as important in this study (Sumowski et al. 2018). The main motivation behind this exclusion is the need to shorten and simplify the batteries to promote efficient and widespread administration, possibly without the requirement of specialist personnel and equipment (Langdon et al. 2012). This need is mitigated when assessments are computerised and can be administered unsupervised.

With computerised assessment, the overall duration should be tailored to avoid participant fatigue. However, beyond that, there is more flexibility to administer comprehensive assessments. PwMS can take as much time as needed to complete the battery, pausing and taking breaks as necessary, without time constraints arising from supervisors managing multiple scheduled patient appointments. This is particularly helpful for patients with mobility problems who may struggle to attend in-person assessments, and in cases like the recent Covid-19 pandemic, during which the risks to those on immunosuppressive therapies going to hospital visits were high (Evangelou et al. 2021).

5.4.6 Identification of the optimal battery of online cognitive tasks

This study identified a 12-task battery that assesses six distinct cognitive domains, including the ones commonly measured in traditional MS batteries, without redundancy. In fact, the battery was selected maximising the number of domains being measured over a sustainable assessment duration (can be completed in less than an hour). This resulted in a battery tailored to measure impairment in MS and deployable in a predetermined, fixed sequence, eliminating the requirement for a random sampling approach in future administrations. Additionally, the battery was optimised for execution in unsupervised settings as most of the selected tasks demonstrated robustness to device variability. SRT and Motor Control were the only two tasks included despite their sensitivity to device, for their critical role in measuring rapid response times, an aspect inevitably influenced by things such as screen

size.

While this preliminary study selected the best tasks to include in battery, their psychometric properties still need thorough evaluation in the context of MS. Moreover, while the memory tasks currently included in the battery just focus on verbal memory, traditional batteries also measure visuo-spatial memory (Sumowski et al. 2018). Computerised object memory tasks are available on the Cognitron and have been shown to be more sensitive in a number of clinical populations. Therefore, they should be evaluated in the future for inclusion in and completion of the battery.

5.4.7 The confounding effect of visuo-motor impairment on cognitive performance measures

An open question and possible limitation of this study pertains to whether visuo-motor impairment acted as a confounder of observed performance decrements on the online cognitive tasks. While the accuracy measures were, by design, robust to visuo-motor impairment, the same does not hold for latency measures. For instance, pwMS might be slower at responding because of visual or upper limbs impairment and this could be mistakenly attributed to information-processing deficits. To investigate this, I looked at the correlation of cognitive performance measures with PROs measuring physical disability, such as the MSIS-29 motor and the MSWS-12, as well as with measures of disease severity, such as the webEDSS. These correlations were largely negligible or non-significant for accuracy measures confirming their robustness to motor impairment apart from accuracy measures for the Spotter, Manipulations 2D, Verbal Analogies, Word Definitions, and Target Detection tasks. These are all tasks for which response time can affect performance accuracy. They are therefore more prone to be affected by motor slowing.

Correlations with latency measures were instead in the medium range for the majority of the tasks indicating that these might be affected by motor impairment. Nonetheless, it is not clear from the literature whether there is a legitimate association between physical disability and cognitive impairment and to what extent (Gois et al. 2021; Lynch et al. 2005). Therefore, the observed correlations might be capturing the co-development of motor and cognitive impairment.

One way to assess this is to analyse costs. As costs are evaluated by subtraction, factors such as visuo-motor impairment, which are consistent across trials, are factored out, and the resulting measures are only dependent on the cognitive component of impairment, which varies with the differential cognitive demands of the trials being compared. The correlation between costs and PROs measuring physical disability were largely non-significant and all negligibly small suggesting that patient-reported motor impairment was independent of cognitive impairment. If that is the case, motor impairment could be a contributing factor to the observed performance latency increase in pwMS compared to controls. This would also explain why TOL and Four Tower exhibited non-significant correlation with the motor-related PROs: these are tasks characterised by long processing times for which latency is mostly determined by cognitive processing rather than motor function.

However, there are ways to model, evaluate, and account for the effect of visuo-motor impairment on the cognitive performance measures. Research efforts are being carried out in this sense by other members of my lab: a model is being developed to disentangle the visuo-motor, device, and cognitive component of performance applicable to all the Cognitron tasks (Giunchiglia et al. 2024, 2023). While

this model is being developed and becomes available for use, it is important to note that it would not be possible to develop such method for traditional pen-and-paper neuropsychological tests, as this heavily relies on the detailed trial-by-trial performance data that can be collected only using computerised tasks. Moreover, in the meantime, the observed performance deficits can be confirmed, when possible, by looking at costs and accuracy measures, which are more robust to the potential confounding effect of visuo-motor impairment. Good examples of this are the costs for the Trail Making and the Switching Stroop tasks, which confirmed the deficits observed in the cognitive control domain. Overall, despite physical disability does not seem to affect completeness rates, it might affect cognitive performance scores, in particular latency scores on tasks involving quick responses, and caution should be taken when interpreting their results.

5.4.8 Conclusions

In summary, this study demonstrated the feasibility and showcased the strengths of online cognitive assessment in MS by successfully testing a significantly larger number of participants compared to previous investigations and proving to be accessible to patients with diverse disease characteristics. This approach not only overcomes the limitations of smaller, non-representative sample sizes and of non-comprehensive, inefficient assessments, but also offers numerous advantages, including cost-effectiveness, remote administration, standardised procedures, accessibility, inclusivity, and enhanced measurement precision. By leveraging the accessibility of the MS Register, it allowed not only to confirm trends observed in smaller, potentially biased samples, but also to uncover new trends, highlighting the potential of widespread implementation of online cognitive assessment in the wider MS community. Finally, the study allowed identifying a reduced optimal battery of tasks, tailored to detect impairments in MS, whilst being comprehensive, non-redundant, engaging, and efficiently deployable remotely on any device.

The next chapter will focus on validating this battery on a separate sample of pwMS and on investigating the associations of deficits across different cognitive domains and disease-related factors such as subtype and duration.

Chapter 6

Optimal battery for online cognitive assessment in MS: validation, application and insights on disease-related associations

In this chapter I validated the optimal battery of online cognitive tasks on a partially overlapping sample of people with MS using a fixed study design and replicated the results from the previous chapter on the independent component of the sample. With the available complete data, I further evaluated the prevalence of cognitive impairments in people with MS and investigated their associations with disease-related factors.

6.1 Introduction

In the previous chapter I assessed the sensitivity to MS of 22 online cognitive tasks and selected a subset to form an optimised battery to measure cognitive impairment in pwMS. Of the initial 22 tasks 13 were selected based on high sensitivity to MS, low sensitivity to device, and low redundancy. This selection aimed to create a battery that could specifically gauge cognitive domains affected in MS while maximising the diversity of measured domains.

While the implemented random sampling approach allowed for an efficient assessment of numerous tasks without introducing participant fatigue, it yielded a sparse dataset, imposing constraints on subsequent analysis. This data sparsity meant that each task could be analysed in isolation, providing insights into task-specific sensitivity to MS. However, it hindered the examination of global trends, such as evaluating the overall prevalence of cognitive impairment in pwMS or assessing associations between cognitive function and disease-related factors. These are all critical aspects that merit further exploration in the field.

Previous studies have reported prevalence rates ranging from 40% to 70% (Chiaravalloti et al. 2008; Langdon 2011), with variations attributed to differences in samples, neuropsychological assessments, and criteria for classifying impaired pwMS across studies (Amato et al. 2018; Benedict 2009; Fischer et al. 2014). Additionally, the samples are typically small and non-representative of the broader MS population, and the chosen neuropsychological assessments often lack comprehensive coverage of all potentially affected domains. This is caused by short and non-comprehensive assessments being preferred due to time and resources constraints (Potticary et al. 2023). In terms of classification criteria there is a broad array of cut-offs being applied to cognitive performance scores that influence prevalence rates. Prevalence has been shown to vary from 4% to 81% in late and from 0% to 68% in early MS, depending on the threshold used (Fischer et al. 2014). While it might be sensible to use more liberal thresholds for earlier disease stages to enable detection of subtle impairments, concerns have been raised on the merits of employing a dichotomous classification (unimpaired vs impaired) at all. There is growing consensus on the need to approach cognition as a continuous function (Pitteri et al. 2019). Therefore, it is crucial to both assess the impact of different thresholds on the employed cognition markers and derive continuous measures of cognitive function to allow the exploration of changes over time at either side of the threshold.

Cognitive impairment has been observed at all stages of MS and in all possible subtypes (Benedict et al. 2020), sometimes even in absence of other symptoms (Sumowski et al. 2018). However, associations with other disease factors (e.g., disease subtype, disease duration, physical disability, depression, and fatigue) remain unclear due to contradictory results in the literature (Chiaravalloti et al. 2008). For example, secondary progressive individuals consistently show more significant impacts than relapsing remitting individuals across studies (Huijbregts et al. 2004; Johnen et al. 2017; Planche et al. 2016; Ruano et al. 2017). This might be attributed to progressive subtypes susceptibility to cognitive impairment, or to the increased disease duration associated with being secondary progressive, or to a combination of both. Factors contributing to the observed differences are challenging to determine as some longitudinal studies reveal no associations between disease duration and cognitive decline (Jennekens-Schinkel et al. 1990; Patti et al. 1998; Rao et al. 1991b), while others show significant associations (Amato et al. 2001; Smestad et al. 2010). These contradictory outcomes stem from studies investigating the evolution of cognitive impairment with disease duration being based on cross-sectional

rather than longitudinal designs or being characterised by short follow-ups (Chiaravalloti et al. 2008).

The use of the optimised battery of online cognitive tasks enables efficient large-scale periodic data collection and has the potential to generate longitudinal datasets that can address some of these open questions and uncover associations between cognitive impairment and other aspects of the disease.

The current chapter presents a study conducted one year after the study presented in Chapter 5. The primary aim was to validate and further improve the optimal battery previously identified. The secondary aim was to showcase the potential of the data collectable using this battery in addressing some of the outstanding questions in the field. This time, a fixed study design was employed, with all users being presented with all tasks in the same order. Two additional tasks measuring memory were incorporated for evaluation, as the two memory tasks evaluated in the previous study did not show satisfactory sensitivity to MS and were included in the battery simply because they were the sole representatives of that domain.

Similarly to the previous study, Cognitron was integrated into the UK MS Register website and all registrants were invited to participate. This resulted in a partially overlapping sample of pwMS participating in the study, including a subset of users performing the tasks for the first time, and another subset with available longitudinal data. This allowed the replication of the previous chapter’s results on an independent sample of pwMS, further validating the feasibility of online cognitive assessment and the sensitivity of selected tasks to MS. Moreover, conducting the study approximately 1 year after the previous one enabled controlling for the period of administration, with the bulk of data being collected in both studies between the end of the Autumn and the beginning of the Winter. This timing also enabled the examination of changes in cognitive function over a 1-year period for pwMS who participated in both studies. Complete data were available for users who completed the assessment, enabling the investigation of global trends in cognitive impairment in pwMS. These included the overall prevalence of cognitive impairment, its variation using different classification strategies, as well as associations of cognitive function with disease-related factors such as disease subtype, disease duration, and early-stage MS.

6.2 Methods

6.2.1 Study design

This is a population-based cross-sectional observational study, designed to validate the optimal battery of online cognitive tasks found in Chapter 5 on a partially overlapping sample of pwMS. Data collection was conducted from November 8, 2022 to January 13, 2023 in collaboration with the UK MS Register. PwMS from the UKMSR completed a fixed battery of 14 online cognitive tasks. Some of the users had already completed some of the tasks during the previous data collection. This allowed to sort the observations across tasks in two groups: a group of observations from participants performing the assessment for the first time, that constituted an independent sample from the previous timepoint, unaffected by confounders such as learning effects, and a group of observations from participants performing the assessment for the second time, for whom data at both timepoints were available on some of the tasks (overlapping sample). In this chapter I will refer to the previous data collection as timepoint 1 (TP1) and to the current data collection as timepoint 2 (TP2). Data from

the independent sample from timepoint 2 were used to replicate the findings from timepoint 1 and validate the sensitivity to MS of the tasks in the battery, consequently substantiating their selection for the final battery. Data from the overlapping sample were used to examine longitudinal changes in cognitive function in pwMS over a 1-year period, and how these related to changes in other aspects of the disease (e.g., changes in motor function). Data from both groups were used to evaluate prevalence rates of impairment in the MS sample and investigate how cognitive function is associated with disease factors, including subtype and duration. Fig. 6.1 reports a schematic of how observations were grouped, and for each group, how many pwMS were included and for which analysis they were utilised.

At the study's inception, all registrants of the UK MS Register were sent an invitation via email to take part in the research and engage with the online tool designed to assess their cognition. Throughout the study duration, two reminder emails were sent to participants who either had initiated but not completed the assessment or had not yet started it. The cognitive scores were then linked with clinical and PRO data already available on the Register. The following analyses excluded data from participants aged below 16 or above 90.

6.2.2 Procedures

6.2.2.1 Online cognitive assessment: the software and the tasks

The online cognitive assessment was delivered via the UKMSR website by means of an iframe containing the Cognitron website, which sequenced the tasks. Out of the 22 Cognitron tasks evaluated in timepoint 1 (Chapter 5), 12 tasks were selected for timepoint 2 optimising to have high sensitivity to MS, low sensitivity to device, and low redundancy.

Two of the tasks, namely Verbal Analogies and Spotter, were slightly modified to improve their design based on observations from the previous study and to align with other large cohort projects. For Verbal Analogies and 2D Manipulations, which are both timed tasks, task duration was reduced respectively from 3 to 2.5 minutes and from 3 to 2 minutes. Moreover, for all of the tasks, the orientation in which they were presented was changed from horizontal to vertical to ensure better compatibility and deployability via smartphones given recent updates to iOS. Finally, given that the Words Memory Immediate and Delayed tasks exhibited low sensitivity to MS and were retained primarily as the sole memory tasks evaluated in timepoint 1 (Section 5.4.6), two additional memory tasks were introduced in the battery that had not been previously evaluated. These tasks, Objects Memory Immediate and Delayed, measure visuospatial rather than verbal recall and have been shown elsewhere to have greater sensitivity to subtle changes in memory functions and provide more information on the nature of memory errors.

This resulted in a battery of 14 tasks, which was administered to all participants in a predetermined, fixed order. As compared to the previous study, the design was intended to produce complete data on all tasks for participants who completed the assessment and helped eliminate any potential confounders arising from varying the tasks as well as the order of the tasks across users. The battery was composed of the following tasks, presented in the following order:

1. Objects Memory Immediate

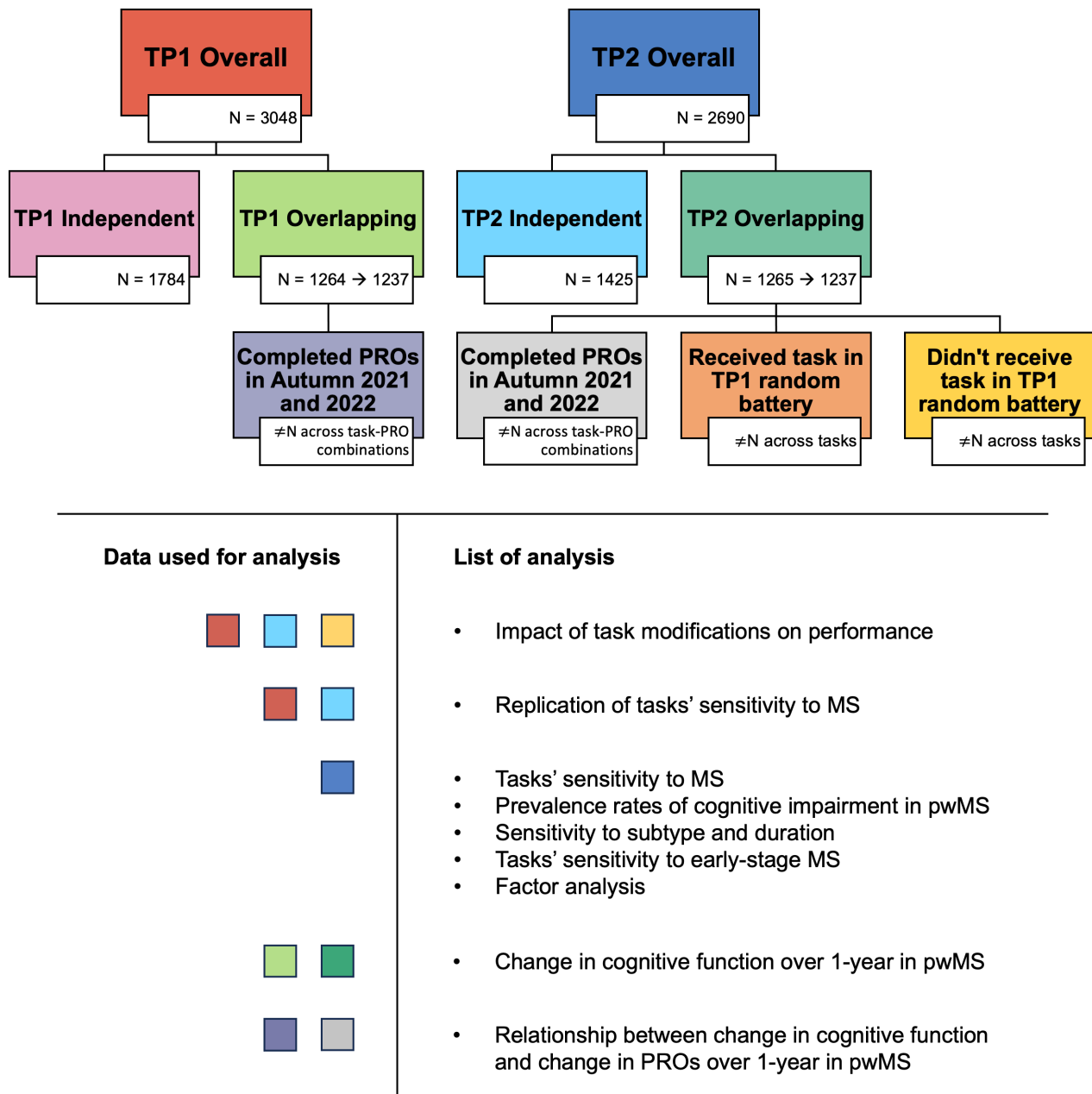


Figure 6.1: Diagram illustrating the total number of people with MS having participated in timepoint 1 and 2, how they were grouped, how many users were in each group, and which combination of groups was used for each analysis presented in this chapter. Numbers refer to pwMS retained for analysis after pre-processing procedures were applied and represent the total number of users in each group. The number of users available for each task varies across tasks. Overall users from timepoint 2 were initially grouped based on whether they had participated or not in timepoint 1 (and vice versa), resulting in one independent and one overlapping sample. For the overlapping samples, the number of users who had valid scores at each timepoint, as well as the number of users who had valid scores at both timepoints, were reported. The overlapping sample from timepoint 2 was further divided for each task based on whether users had received or not that specific task in their random sampled battery during timepoint 1. Moreover, two further groups were derived from the overlapping samples from timepoint 1 and 2, including users who had filled in the PROs during the collection windows prior to both cognitive assessments (in Autumn 2021 and 2022). The number of users in these groups varied based on the specific task-PRO combination considered.

2. Words Memory Immediate
3. Motor Control
4. 2D Manipulations
5. Card Pairs
6. Switching Stroop
7. Blocks
8. Verbal Analogies
9. Words Definitions
10. Simple Reaction Task
11. Trail Making
12. Objects Memory Delayed
13. Words Memory Delayed
14. Spotter

For each task a primary outcome metric for accuracy (e.g., percentage of correct responses) and latency (e.g., median response time) were derived from the trial-by-trial data. For the Switching Stroop and Trail Making tasks, additional outcome metrics were calculated, reflecting the “cost” in accuracy and response time when shifting from simpler to more complex conditions. Four tasks—Object Memory Immediate and Delayed, and Words Memory Immediate and Delayed—were paired, with extra outcome metrics included to evaluate performance differences between these task pairs. When a single outcome metric needed to be identified for each task as most representative of performance, accuracy was chosen over latency for most tasks, because by design has the lowest sensitivity to confounding effects of device and motor deficits, with the exception of Motor Control, SRT, and Trail Making. For these tasks, latency was selected instead, as they were specifically designed to measure reaction times.

6.2.2.2 Questionnaires

Alongside the battery of online cognitive tasks, participants completed a questionnaire in the beginning including items for potential sociodemographic confounders (age, gender, dominant hand, country of residence, language, ethnicity, education, and occupation) and for their MS characteristics (disease subtype and duration). The sociodemographic questions overlapped with those asked for the previous timepoint.

6.2.2.3 Patient-reported outcomes

The resulting cognitive data were linked with data already available on the UKMSR for each participant. Specifically, data were linked with PROs data collected during the Register’s Autumn 2022

collection window, which was chosen for its temporal proximity to the administration of the online cognitive assessment. The PROs collected included MSIS-29, MSWS-12, HADS, FSS, EQ5D, and webEDSS, measuring a spectrum of disease manifestations, from motor symptoms to mood disturbances.

6.2.2.4 Normative data

In alignment with Chapter 5, this study utilised normative data from the GBIT study, which spanned two baseline and four follow-up timepoints. However, prior to the current study and the final follow-up timepoint of the GBIT study, some modifications were made to the tasks. As already outlined in Section 6.2.2.1, Verbal Analogies and Spotter underwent design changes, the durations of Verbal Analogies and 2D Manipulations were shortened, and all tasks were reoriented from horizontal to vertical for better smartphone compatibility given iOS updates. Consequently, all GBIT timepoints, except for the last follow-up, used the pre-modifications versions of the tasks. However, not every task for which normative data were required in this chapter was included in this final GBIT timepoint, leading to a lack of normative data for the revised versions of some tasks.

To address this, the impact of task modifications on performance was evaluated across tasks. This was tested using patient data, after merging the two available timepoints, because it was the only way to have a dataset containing both versions of each task (before and after modifications). It was assumed that the control data would show a similar effect. For the tasks where only orientation was modified, the impact on performance was either non-significant or very small (Section 6.3.3.1); hence it was deemed negligible, allowing the use of control data obtained performing horizontal tasks as normative for patient data obtained performing vertical tasks. This approach was chosen also to: i) maintain consistency with what was done in the Chapter 5 (same normative data where possible), allowing for a fair comparison between patient timepoints; ii) have larger normative data as more observations were available for the horizontal version of the tasks which were used in most GBIT timepoints.

For Verbal Analogies, significant performance differences were observed due to task modifications related to how analogies of varying difficulty were sampled. However, the modified version of this task was included in the final GBIT timepoint, providing relevant control data. In contrast, Spotter underwent major modifications and was not included in the final GBIT timepoint, leading to its exclusion from the analysis presented in this chapter pending future control data availability.

For 2D Manipulations, trial-by-trial control data from the original version of the task were cut to match the modified task duration. Primary outcome metrics were recalculated for this shorter version of the original task and used as control data. Despite these adjustments, the task modifications still had a significant, albeit small, impact on performance, necessitating caution in results interpretation.

Finally, for Objects Memory Immediate and Delayed, which were introduced in the second patient timepoint and directly in their vertical versions, vertical control data from the final GBIT timepoint were utilised as normative.

Table 6.1 details the version of each task from which the selected normative data were derived. Similarly to Chapter 5, data were retained from the 13 tasks being evaluated (excluding Spotter) for individuals who were healthy controls without neurological or psychiatric conditions, who had complete sociodemographic data, were aged 16-90, and were first-time task performers to ensure unbiased

assessment. Table 6.1 summarises the total number of controls who completed the selected version of each task and the number of controls remaining after applying the selection criteria and the pre-processing procedures (as detailed in Section 3.3). In total, normative data from 389,800 controls were used in this study, with their summary sociodemographic characteristics presented in Table 6.2.

Tasks	Selected version of the task	N observations (before selection criteria and pre-processing)	N participants (before selection criteria and pre-processing)	N observations = N participants (after selection criteria and pre-processing)
Objects Memory Immediate	Vertical	25257	25233	17969
Words Memory Immediate	Horizontal	200639	195909	98887
Motor Control	Horizontal	38246	38231	25548
2D Manipulations	Cut Horizontal	501905	444427	388394
Card Pairs	Horizontal	8368	8000	5987
Switching Stroop	Horizontal	8363	8030	5985
Blocks	Horizontal	899396	799128	389247
Verbal Analogies	Modified Vertical	23138	23118	16503
Word Definitions	Horizontal	510415	452419	388323
SRT	Horizontal	7313	7082	5430
Trail Making	Horizontal	8488	8129	6890
Objects Memory Delayed	Vertical	21631	21627	14637
Words Memory Delayed	Horizontal	124823	120485	92625
Spotter	-	-	-	-

Table 6.1: Version of each task from which the selected normative data were derived. Number of observations available and number of controls for the selected version of each task, before and after selection criteria and pre-processing procedures were applied.

6.2.2.5 Timepoint 1 data

Some of the analysis presented in this chapter involved comparing data from timepoint 2 with data from timepoint 1 presented in Chapter 5. In timepoint 1, data from 3066 pwMS on 22 different online cognitive tasks were collected using a random sampling approach. In this chapter, only timepoint 1 data from 3048 pwMS on the 12 tasks that were selected for the optimal battery (except for Spotter) were retained for analysis. Data were pre-processed as detailed in the previous chapter (Section 5.2.3.1) and DfE scores were derived using the normative data there presented (Section 5.2.2.5). Moreover, the resulting cognitive data from timepoint 1 were linked with PROs data collected during the Register’s Autumn 2021 collection window. This window was chosen because it was temporally closest to the timepoint 1 cognitive assessment and collected 1-year before the PROs data linked with the timepoint 2 data. This enabled examination of changes over a 1-year period in cognitive and PRO-derived measures and their relationship.

Socio-demographic characteristics		Controls
Total N		389800
Age (min-max, mean)		16-90, 44.9
Gender	Female	207309 (53.2%)
	Male	178809 (45.9%)
	Other	3682 (0.9%)
Dominant hand	Right	336736 (86.4%)
	Left	42536 (10.9%)
	Ambidextrous	10528 (2.7%)
First language	English	350556 (89.9%)
	Other	39244 (10.1%)
Ethnicity	White	330927 (84.9%)
	Unknown	5211 (1.3%)
	Mixed or multiple ethnic groups	9975 (2.6%)
	Asian or Asian British	31073 (8.0%)
	Black, Black British, Caribbean or African	3552 (0.9%)
	American Hispanic	9020 (2.3%)
	Other ethnic group	42 (0.01%)
Residence	United Kingdom	339747 (87.2%)
	Abroad	50053 (12.8%)
Education	preGCSE	10709 (2.7%)
	School	129580 (33.2%)
	Degree	230402 (59.1%)
	PhD	19109 (4.9%)
Occupation	Worker	26661 (68.4%)
	Retired	56793 (14.6%)
	Disabled / Not applicable / Sheltered employment	3987 (1.0%)
	Homemaker	12065 (3.1%)
	Unemployed / Looking for work	11924 (3.1%)
	Student	35940 (9.2%)
	Unknown	2470 (0.6%)

Table 6.2: Sociodemographic characteristics of participants from the GBIT study retained as controls for this study.

6.2.3 Statistical analysis

Data curation, pre-processing, and statistical analysis were conducted in Python 3.8.5.

6.2.3.1 Summary performance measures and pre-processing

Data pre-processing steps were the same as for timepoint 1 (Section 5.2.3.1). They are detailed in the Chapter 3 and are here just briefly summarised. Data from participants who provided incorrect sociodemographic information or did not complete the questionnaire were omitted. Additionally, data from the Words/Objects Memory Delayed task were excluded if this was completed over two hours after the Words/Objects Memory Immediate task. This exclusion was due to the requirement for the Delayed task to be conducted within a relatively set timeframe following the Immediate task.

Summary performance metrics were derived for each task, with records showing null values for summary accuracy and/or latency being discarded. The distributions of accuracy and latency in both normative and patient datasets were compared to establish thresholds for non-compliance and to eliminate outliers. These performance metrics were then transformed using normative data in Deviation from Expected (DfE) scores, which are scores adjusted for sociodemographic factors and expressed in standard deviation units, that act as indicators of the level of cognitive impairment in pwMS.

6.2.3.2 Impact of task modifications on performance

To evaluate the impact of task modifications on performance, data from timepoint 1, collected utilising original task versions, were compared with data from timepoint 2, collected utilising modified task versions. For timepoint 1 data, all users were included. For timepoint 2, only users performing each task for the first time were included, to avoid capturing learning effects across timepoints. The datasets from both timepoints were combined, and a label was added to each observation to indicate whether it was generated from a task in its original version (timepoint 1) or in its modified version (timepoint 2). The analysis was conducted on raw performance scores, before filtering outliers and evaluating DfE scores, to prevent any confounding of the results.

A multiple linear regression model was fitted to predict the scores on each performance metric using sociodemographic variables, device label, and task modifications label as predictors. Sociodemographic variables and device label were included in the regression models as potential confounding factors. The reference category for task modifications was set to be the original version of the task. The effect size of task modifications was quantified as the beta coefficient obtained from the regression model divided by the standard deviation of the predicted performance metric. Two separate regression models were fitted for each cognitive task, one for accuracy and one for latency. The significance of the resulting standardised regression coefficients was assessed using a series of t-tests with an alpha threshold of 0.05.

6.2.3.3 Replication of timepoint 1 results

In this analysis, data from timepoint 2 were compared with data from timepoint 1 to investigate the replicability of the findings from the previous chapter. For timepoint 1, all users were included, while for timepoint 2 only users who had not engaged at all with data collection at timepoint 1 were retained. This approach ensured working with a completely independent sample, thereby enhancing the robustness of the replication analysis. To prevent data leakage from users who participated in timepoint 1, pre-processing procedures, like outliers' removal, were performed separately on the

independent sample from timepoint 2, specifically for this analysis. Moreover, to allow for a fair comparison, trial-by-trial data for 2D Manipulations and Verbal Analogies from timepoint 1 were cut to match tasks duration in timepoint 2, and performance measures were recalculated for these shorter versions of the tasks.

For both timepoints, median DfE scores across tasks were calculated for both accuracy and latency measures as indicators of task sensitivity to MS. The median DfE scores across tasks from timepoint 1 were then correlated with those from timepoint 2 using Pearson’s correlation, with an alpha threshold of 0.05 set to determine statistical significance. A high significant correlation would imply the replicability of the findings from timepoint 1 in the independent sample from timepoint 2.

As a further validation, this process was repeated using mean DfE scores across tasks, as these could also serve as indicators of task sensitivity to MS. However, the median was preferred over the mean as the primary statistic due to its greater stability in the presence of outliers and skewed distributions, which were prevalent in these data.

6.2.3.4 Definition of global cognitive impairment

To classify cognitively impaired individuals among pwMS at timepoint 2 and to estimate the prevalence of cognitive impairment in this cohort, three classification criteria, representing different levels of stringency, were adopted from existing literature (Fischer et al. 2014). Three criteria were selected to examine how prevalence rates fluctuated with changing stringency. Each criterion was based on an abnormality threshold — defined as a SD cut-off over which performance was deemed abnormal — and a critical number of abnormal parameters. Thus, a person was classified as cognitively impaired if they scored abnormally on a predefined number of performance metrics.

PwMS who completed the assessment in timepoint 2 (excluding Spotter) yielded 33 scores encompassing accuracy, latency, costs, and additional performance metrics. To maintain consistency in the total number of scores across pwMS, only those who completed the full assessment, and had valid scores for all tasks were included in the analysis. For accuracy measures and accuracy costs, a negative SD cut-off was applied to pwMS’s DfE scores to classify the scores as abnormal. Conversely, for latency measures, latency costs, and percent forgetting, a positive SD cut-off was applied.

The selected criteria were named ‘liberal’, ‘fair’, and ‘conservative’, reflecting their respective stringency levels. The liberal criteria classified a score as abnormal if it fell below -1SD in case of accuracy, and above 1SD in case of latency, from the expected score for at least 13% of the parameters (at least 5 scores). The fair criteria (the one most commonly used in literature (Benedict 2009)) used a 1.5SD cut-off for at least 20% of the parameters (at least 7 scores), and the conservative criteria a 2SD cut-off for the same proportion. Prevalence rates of cognitive impairment were then derived using these criteria. Additionally, age, gender, clinical phenotype, and disease duration were compared between impaired and non-impaired pwMS for each classification criterion, with statistical differences being assessed using t-tests and chi-squared tests at a significance threshold of 0.05.

6.2.3.5 Definition of task-level cognitive impairment

In a similar manner, to classify whether performance of pwMS from timepoint 2 on a particular task was impaired, three different abnormality thresholds were established at 1SD, 1.5SD, and 2SD. These thresholds corresponded to a liberal, fair, and conservative definition of impairment, respectively. For each task, these thresholds were applied to the most representative outcome metric. In tasks where accuracy was the primary outcome metric, performance was considered impaired if the DfE accuracy was less than the negative SD cut-off. Conversely, for tasks focused on latency, impairment was defined if the DfE latency exceeded the positive SD cut-off. The proportion of pwMS classified as impaired for each task using these different thresholds was then calculated. The normative data was utilised to evaluate the analogous proportions of controls, which provided baseline levels of impairment that could be considered normal or expected in the general population.

6.2.3.6 Sensitivity to MS and to early-stage MS

To evaluate whether each task, as described by its respective performance metrics, was sensitive to MS and to what extent, a series of two-tailed Wilcoxon signed-rank tests were conducted. Specifically, the DfE scores for a given performance metric were tested against a hypothetical median of zero to determine whether pwMS significantly differed from controls on that metric. A significance threshold of 0.05 was applied. The median DfE score for each performance metric was used as an indicator of effect size because it is more stable than the mean in presence of outliers and skewed distributions. Cohen's notation of effect size was used to quantify the effects with values of 0.2, 0.5, and 0.8 denoting small, medium, and large effects, respectively.

A similar approach was employed to evaluate the sensitivity of each task to early-stage MS. For this analysis, only pwMS within the first 5 years since disease onset at the time of cognitive assessment were included. Moreover, early-stage sensitivity was evaluated only for the primary performance metric of each task.

6.2.3.7 Sensitivity to disease subtype and duration

The influence of disease-related factors, including disease subtype and duration, on task performance was examined using non-parametric two-way permutational ANOVA. This analysis focused solely on the primary performance metric of each task. Disease subtype and duration were treated as categorical variables. Disease subtype had six possible categories: Benign, Relapsing-Remitting (RRMS), Transitioning to Secondary Progressive (TrnSPMS), Secondary Progressive (SPMS), Primary Progressive (PPMS), and do not know (DK). Disease duration, originally recorded in years since onset, was divided into five categories based on 10-year intervals, ranging from 0 to 40 years, with an additional category for durations exceeding 40 years, due to their lower frequency.

Two-way permutational ANOVA was selected over other statistical techniques because it does not assume normality or homogeneity of variances, which was appropriate given the skewed nature of some of the performance metrics. A separate model was evaluated for the primary performance metric of each task. For each factor — subtype, duration, and their interaction — the F statistic was calculated on the original, non-permuted data. The scores on the performance metric were then

permuted across users to disrupt the relationship between the factors and the performance metric while preserving the data's structure. The F statistics were recalculated on the permuted data, repeating the process 10,000 times to establish a null distribution for the F statistic of each factor. The observed F statistics were then compared against these null distributions to assess the significance of each factor, applying a significance threshold of 0.05. For tasks for which one or more factors were significantly associated with performance, post-hoc pairwise comparisons were conducted using the Mann-Whitney U test. The Bonferroni method was employed to adjust for multiple comparisons, ensuring a rigorous evaluation of the pairwise differences.

6.2.3.8 Factor analysis

Factor analysis was conducted on data from pwMS who completed the full assessment at timepoint 2 and had valid scores for all tasks. Only the primary outcome metric for each task was included in the factor analysis, yielding 13 variables per user. The Kaiser Criterion was used to determine the optimal number of factors, i.e. the factors with eigenvalue greater than 1. The first unrotated factor was taken for each user as a measure representative of general intelligence (G factor), as this is the factor that captures the highest variance in the data. Varimax rotation was then applied to the resulting loading matrix to facilitate, with a more interpretable arrangement of loadings, the inference of cognitive domains underlying each factor. Rotated factor scores were evaluated for each user.

The impact of disease subtype and duration on the G factor and on the rotated factors was evaluated using multiple linear regression models. A multiple linear regression model was fitted to predict each factor using as categorical regressors disease subtype, disease duration, and their interaction (respective categories are already outlined in Section 6.2.3.7). ANOVA was conducted on the regression models to obtain the significance of the impact of disease subtype, duration, and their interaction on the factors. When a regressor had a significant impact on a factor, the standardised regression coefficients for each category of the regressor were examined, and their significance against the reference category was evaluated using t-tests with an alpha threshold of 0.05. For disease subtype, 'Benign' was set as the reference category, while for disease duration the first bin (0-10 years from onset) was selected.

6.2.3.9 Longitudinal analysis

For longitudinal analysis, only pwMS who completed each task at both timepoints, and had valid scores for each were included. The performance deltas for each user were calculated by subtracting for both accuracy and latency metrics across tasks the raw score from timepoint 1 to that from timepoint 2, and then dividing by the standard deviation of all raw scores across users and timepoints for that metric. The resulting performance deltas were t-tested against zero to evaluate whether there was a significant change in accuracy or latency metrics across tasks at the MS population level over a 1-year period. The mean standardised performance deltas for each task served as indicators of effect size, with values of 0.2, 0.5, and 0.8 denoting small, medium, and large effects, respectively (Cohen's notation).

Additionally, for users in the overlapping sample who had also completed the MSIS-29 in the Register collection windows prior to each cognitive assessment, the questionnaire deltas were evaluated in a similar way as the performance deltas. These deltas were then correlated with the performance deltas

to investigate the relationship between changes in motor and cognitive function in pwMS over a 1-year period. The same was done for MSWS-12, HADS, FSS, EQ5D, and webEDSS, providing insights into the relationship of changes in cognitive function with changes in walking abilities, depression and anxiety, fatigue, quality of life, and general disability severity, respectively.

6.3 Results

6.3.1 User engagement

6.3.1.1 Participation and return rates

A total of 2696 users of the UKMSR participated in the online assessment at timepoint 2. 47.0% of them had already participated at the previous timepoint, meaning that 41.3% of the users from timepoint 1 returned to do the assessment at timepoint 2 (Fig. 6.2B). Among pwMS from timepoint 1, those who returned did not significantly differ from those who did not based on gender, disease duration, and disability severity. However, there were significant differences in age and disease subtype, with pwMS returning being on average older (by 3.6 years), and having a higher proportion of secondary progressive (by 4%) and a lower of relapsing-remitting individuals (by 7.6%) as detailed in Table 6.3. Despite having completed some of the tasks before, most users who returned had not already seen all of the tasks in the timepoint 2 battery, as during the previous timepoint they were presented only with a random subset of 12 tasks sampled from the superset of evaluated tasks. Moreover, not all users completed the entire assessment at timepoint 2, resulting in a non-homogeneous distribution of observations across tasks. Nonetheless, at least 2197 users completed each task, and at least 1274 were doing so for the first time. Table 6.4 presents the number of observations available for each task, both overall and separately for participants who were doing each task for the first time. When comparing the demographics and clinical factors of the overall samples from timepoints 1 and 2, the only statistically significant differences were for disease duration and MSIS-29 motor scores (Table 6.3), with pwMS from timepoint 2 having on average shorter disease durations (by 2.5 years) and higher motor disability (by 2.9 points of the normalised MSIS-29 motor scale). Although statistically significant, these differences were relatively small.

6.3.1.2 Completion rates

Completion rates for timepoint 2 are reported in Fig. 6.2A. Among the total participants, 81.5% completed the whole assessment, with 86.3% of them having done so in a single session. The remaining portion of participants either completed the assessment with breaks within a single day or completed it on a different day, likely prompted by the email reminders. Completion of the assessment did not depend on age, disease severity or disease duration, but it did have an association with disease subtype, with completers in one session having a higher proportion of relapsing remitting individuals (by 10.4% compared to completers in more sessions and by 5.4% compared to non-completers) and a lower proportion of progressive individuals (by 10.8% compared to completers in more sessions and by 6.5% compared to non-completers) as reported in Table 6.3. For the subset of participants who either completed the assessment in more days or did not complete it (25.8%), I looked at data only

from the first day they engaged with the assessment to investigate whether there was a particular task or number of tasks where they most commonly stopped (Fig. 6.2D). Dropout rates were higher either after the first task, or after more complex and challenging tasks such as Switching Stroop and 2D Manipulations, while dropping in the second half of the battery.

6.3.1.3 Engagement of active users

Finally, I looked at whether the pwMS who engaged with the online cognitive assessment also engaged with the PROs data collection during the Register window prior to the assessment. As depicted in Fig. 6.2C, a substantial proportion of timepoint 2 participants were also “active users” of the UKMSR, with 50.6% of them having completed all the administered PROs, and 68.4% having completed at least one. However, a substantial proportion of pwMS who were not active users (31.6%) successfully participated, resulting in an overall increase in engagement with the Register.

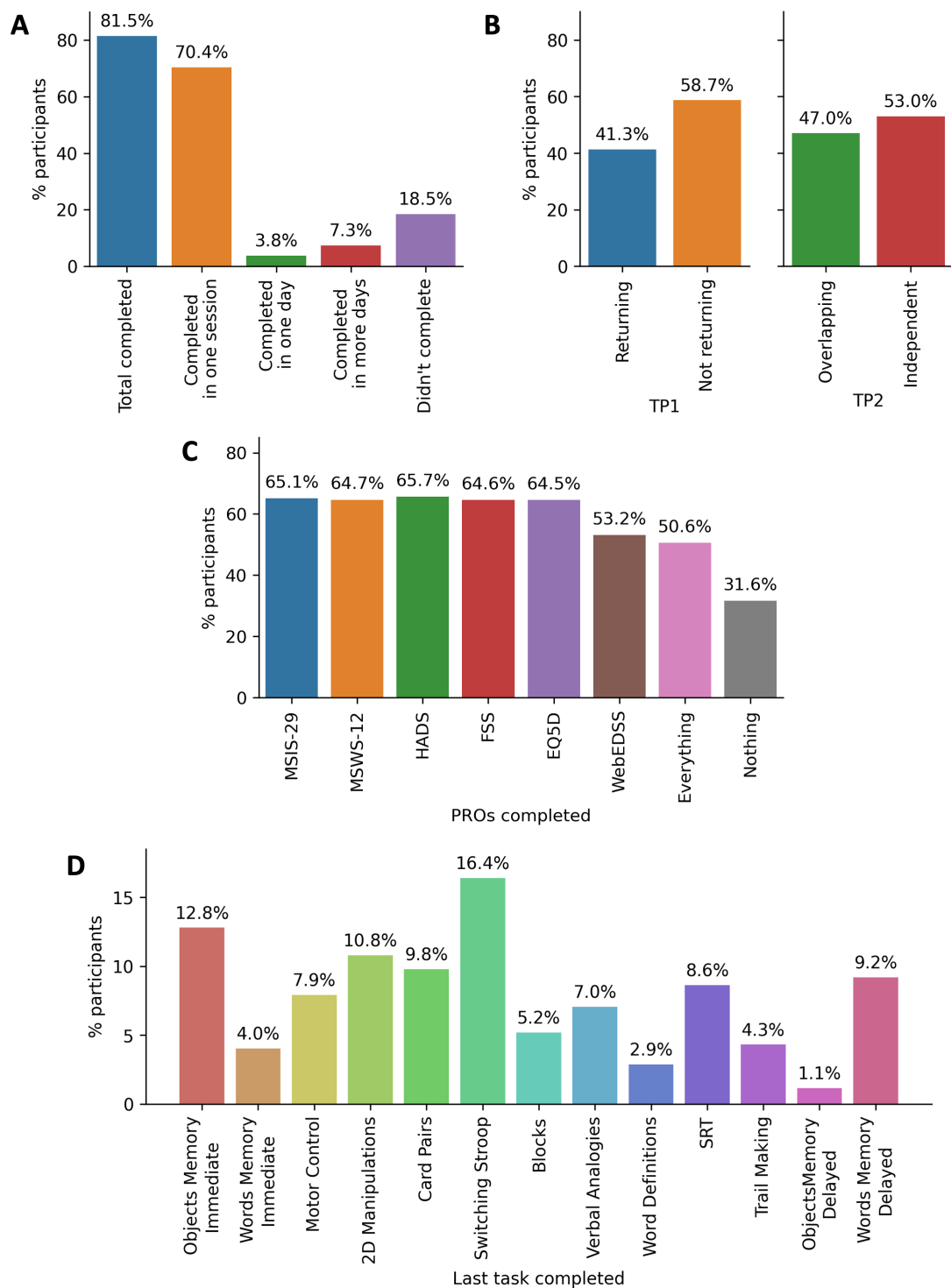


Figure 6.2: Engagement of people with MS with online cognitive assessment. (A) Percent of participants from TP2 who completed, overall, in one session, in one day, or in more days, and who did not complete. (B) Percent of participants from TP1 who did vs did not return and percent of participants from TP2 who did vs did not do TP1. (C) Percent of participants from TP2 who completed the different PROs on the UK MS Register website at the data collection window prior to the cognitive data collection. (D) Percent of participants who dropped out after completing a particular task.

Groups	N	Age (min-max, mean)	Gender (%)	Disease subtype (%)	Disease duration (min-max, mean)	Disease severity (mean MSIS-29 motor norm / webEDSS)
Completed in one session	1897	18-88, 53.8 (missing = 3)	Female = 78.8 Male = 21.1 Other = 0.1 (missing = 3)	RR = 56.1 SP = 22.5 PP = 12.1 DK = 3.9 TrnSP = 2.6 BN = 2.8 (missing = 3)	0-63, 17.8 (missing = 4)	43.2 / 4.9 (missing = 635 / 867)
Completed in more than one session	300	21-86, 55.2	Female = 78.3 Male = 21.7 Other = 0	RR = 45.7 SP = 26.7 PP = 18.7 DK = 4.7 TrnSP = 2.7 BN = 1.7	0-60, 18.9	49.9 / 5.2 (missing = 99 / 126)
Did not complete	499	20-80, 53.7 (missing = 2)	Female = 74.8 Male = 25.2 Other = 0 (missing = 2)	RR = 50.7 SP = 28 PP = 13.1 DK = 3.2 TrnSP = 2.6 BN = 2.4 (missing = 2)	0-58, 18.4 (missing = 2)	47.4 / 5.0 (missing = 207 / 270)
P-values from groups comparison	-	0.1546	0.0412	0.0099	0.2448	0.0007 / 0.0985
TP1 - returning	1267	16-87, 56.5	Female = 77.7 Male = 22 Other = 0.2	RR = 49.4 SP = 32.3 PP = 15.9 BN = 2.4 (missing = 239)	1-64, 21 (missing = 239)	42 / 4.9 (missing = 103 / 669)
TP1 - not returning	1799	20-85, 52.9 (missing = 4)	Female = 77.2 Male = 22.6 Other = 0.2 (missing = 4)	RR = 57 SP = 28.3 PP = 12.6 BN = 2.1 (missing = 463)	0-67, 20.1 (missing = 463)	41.5 / 4.8 (missing = 442 / 1178)
P-values from groups comparison	-	<0.0001	0.9242	0.0027	0.0606	0.6216 / 0.3225
TP1 - overall	3066	16-87, 54.4 (missing = 4)	Female = 77.4 Male = 22.4 Other = 0.2 (missing = 4)	RR = 53.7 SP = 30 PP = 14 BN = 2.2 (missing = 702)	0-67, 20.5 (missing = 702)	41.7 / 4.8 (missing = 545 / 1847)
TP2 - overall	2696	18-88, 53.9 (missing = 5)	Female = 78 Male = 21.9 Other = 0.1 (missing = 5)	RR = 54 SP = 24 PP = 13 DK = 3.8 TrnSP = 2.6 BN = 2.6 (missing = 5)	0-63, 18 (missing = 6)	44.6 / 4.9 (missing = 941 / 1263)
P-values from groups comparison	-	0.1184	0.3004	0.1718	<0.0001	0.0005 / 0.2121

Table 6.3: Demographics, clinical phenotype, disease duration, and disease severity across: i) participants from TP2 who completed, in one or more sessions, vs who did not complete; ii) participants from TP1 who did vs did not return to take part in TP2; iii) overall participants from TP1 vs TP2. P-values were obtained using t-tests, ANOVA or chi-squared tests where appropriate.

Tasks	N participants overall (before pre- processing)	N participants performing each task for the first time (before pre- processing)	N participants overall (after pre- processing)	N participants performing each task for the first time (after pre- processing)
Objects Memory Immediate	2696	2696	2681	2681
Words Memory Immediate	2632	1399	2621	1393
Motor Control	2611	1379	2597	1370
2D Manipulations	2576	2085	2557	2068
Card Pairs	2519	2005	2499	1989
Switching Stroop	2471	1960	2439	1932
Blocks	2393	1890	2381	1880
Verbal Analogies	2372	1878	2331	1842
Word Definitions	2339	1829	2324	1815
SRT	2327	1858	2263	1811
Trail Making	2288	1819	2275	1809
Objects Memory Delayed	2265	2265	1899	1899
Words Memory Delayed	2252	1274	1927	1085
Spotter	2197	1773	-	-

Table 6.4: Number of participants who completed each task, both overall and separately for participants who were doing each task for the first time, before and after pre-processing procedures were applied.

6.3.2 Participants

Overall, 2690 of the original 2696 participants were kept for analysis after pre-processing procedures were applied. Table 6.5 presents their sociodemographic characteristics.

Socio-demographic characteristics		People with MS
Total N		2690
Age (min-max, mean)		18-88, 53.9
Gender	Female	2099 (78%)
	Male	589 (21.9%)
	Other	2 (0.1%)
Dominant hand	Right	2315 (86.1%)
	Left	306 (11.4%)
	Ambidextrous	69 (2.6%)
First language	English	2588 (96.2%)
	Other	102 (3.8%)
Ethnicity	White	2581 (95.9%)
	Mixed or multiple ethnic groups	41 (1.5%)
	Asian or Asian British	34 (1.3%)
	Black, Black British, Caribbean or African	24 (0.9%)
	Other ethnic group	10 (0.4%)
Residence	United Kingdom	2671 (99.3%)
	Abroad	19 (0.7%)
Education	preGCSE	63 (2.3%)
	School	1269 (47.2%)
	Degree	1291 (48%)
	PhD	67 (2.5%)
Occupation	Worker	1089 (40.5%)
	Retired	916 (34.1%)
	Disabled / Not applicable / Sheltered employment	502 (18.7%)
	Homemaker	108 (4%)
	Unemployed / Looking for work	41 (1.5%)
	Student	21 (0.8%)
	Unknown	13 (0.5%)
Disease Subtype	Benign	70 (2.6%)
	Relapsing-remitting	1451 (53.9%)
	Transitioning to secondary progressive	70 (2.6%)
	Secondary progressive	645 (24%)
	Primary progressive	351 (13%)
	Do not know	103 (3.8%)
Disease Duration (min-max, mean)		0-63, 18 (missing = 1)

Table 6.5: Sociodemographic characteristics of pwMS from the UK MS Register who participated in the online cognitive assessment and were retained for analysis after pre-processing procedures were applied.

6.3.3 Analyses on independent sample

The first set of analyses focused on comparing task performance in timepoint 1 with task performance in timepoint 2. This comparison aimed to ascertain if the findings regarding task sensitivity to MS, reported in the Chapter 5, could be replicated in an independent sample of pwMS. Additionally, it sought to validate the selection of tasks for the final battery, which was optimised for monitoring cognitive function in MS.

6.3.3.1 Task orientation does not impact performance

The first step involved evaluating whether it was feasible and fair to compare between performance at timepoint 1 and 2 considering that the tasks underwent some modifications prior to timepoint 2.

As presented in Fig. 6.3, task modifications were not a significant predictor of performance across most tasks. Furthermore, where task modifications did impact performance, the effect sizes were generally very small ($\beta < 0.1$) and thus considered negligible. The exceptions were Verbal Analogies and 2D Manipulations, for which task modifications exhibited a medium and small effect on the accuracy scores, respectively. Nonetheless, these were the only tasks that underwent modifications beyond just the orientation of presentation.

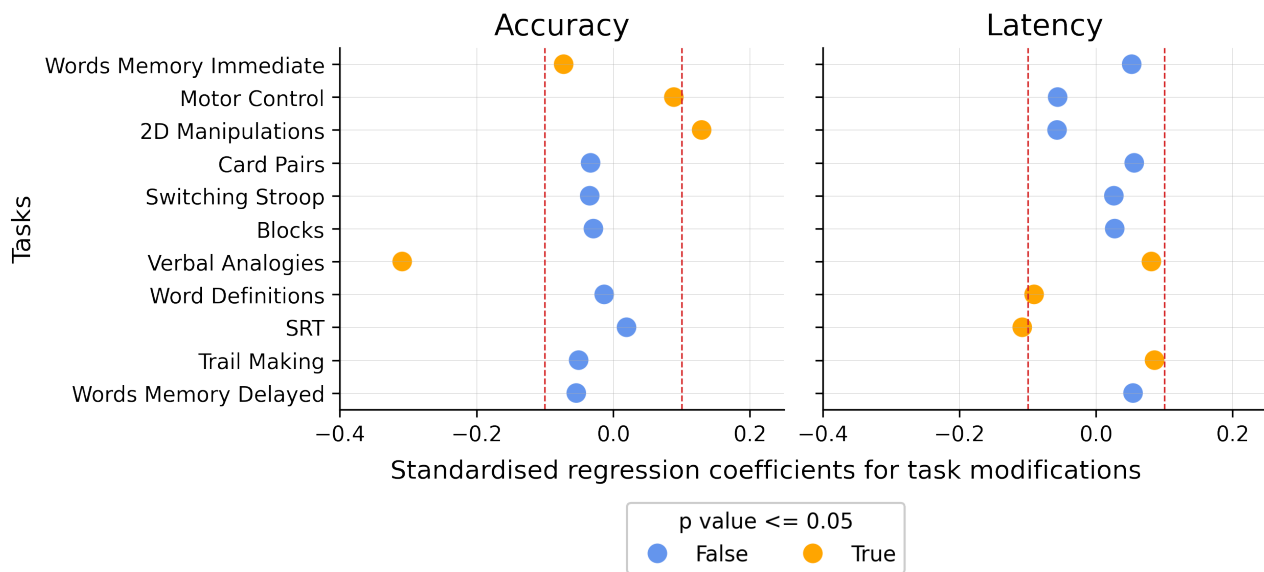


Figure 6.3: Impact of task modifications on performance. Standardised regression coefficients for task modifications resulting from multiple linear regression models predicting performance scores across tasks from sociodemographic variables, device label and task modifications.

6.3.3.2 Task sensitivity to MS replicates across timepoints

After ensuring that a fair comparison across timepoints was possible, the focus shifted towards investigating the replicability of timepoint 1 results (Chapter 5) using the independent sample from timepoint 2.

Sensitivities to MS across tasks at timepoint 1 vs at timepoint 2 are reported in Fig. 6.4 for both accuracy and latency measures. Notably, in both cases, there was a high significant correlation between the sensitivities at timepoint 1 and 2 (0.94 for accuracy and 0.89 for latency measures). The most prominent changes in sensitivity were observed in the median DfE accuracy score for Verbal Analogies, which changed from -0.34 to -0.67, indicating larger deviations of pwMS from matched controls on average with the modified task version, and the median DfE latency for Trail Making, which changed from 0.72 to 1.09, again suggesting larger deviations (Table 6.6). Therefore, even in case of changes in sensitivity, the direction of these changes was such that reaffirmed or even amplified the task sensitivity to MS.

High and significant correlations between timepoints (0.91 for both accuracy and latency measures) were also observed when comparing the mean rather than the median DfE scores across tasks, confirming the replicability of the results with an alternative indicator of task sensitivity to MS (Table 6.6).

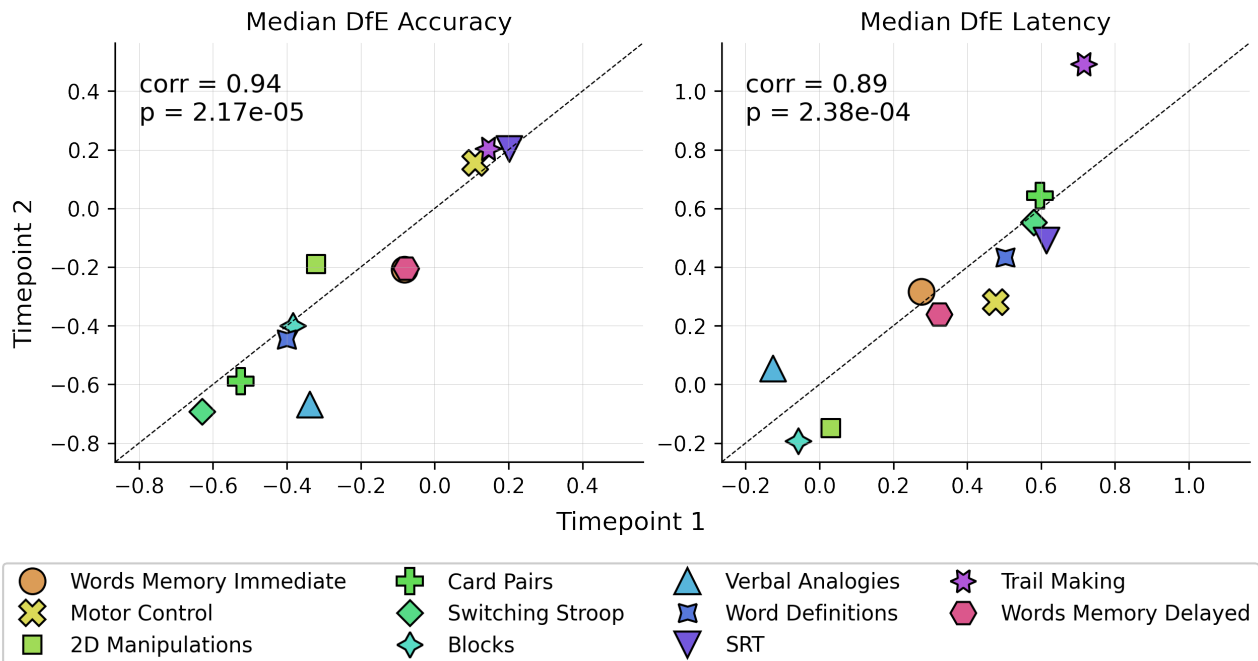


Figure 6.4: Replication of task sensitivity to MS across timepoints. Scatter plots of sensitivity to MS in terms of accuracy (left) and latency (right) across tasks for timepoint 1 (x-axis) vs timepoint 2 (y-axis). Sensitivity to MS is defined as the median DfE score for each task. Overall correlations between sensitivities at timepoint 1 vs at timepoint 2 for accuracy and latency are reported on the plots.

Tasks	Median DfE Accuracy		Median DfE Latency		Mean DfE Accuracy		Mean DfE Latency	
	TP1	TP2	TP1	TP2	TP1	TP2	TP1	TP2
Words Memory Immediate	-0.08	-0.21	0.28	0.32	-0.30	-0.46	0.67	0.70
Motor Control	0.11	0.16	0.48	0.28	0.04	0.16	0.87	0.69
2D Manipulations	-0.32	-0.19	0.03	-0.15	-0.31	-0.14	0.33	0.19
Card Pairs	-0.53	-0.59	0.60	0.64	-0.56	-0.62	0.97	1.06
Switching Stroop	-0.63	-0.69	0.58	0.55	-0.68	-0.76	0.79	0.80
Blocks	-0.38	-0.40	-0.06	-0.19	-0.46	-0.52	0.18	0.15
Verbal Analogies	-0.34	-0.67	-0.13	0.05	-0.32	-0.61	0.07	0.29
Word Definitions	-0.40	-0.45	0.50	0.43	-0.51	-0.61	0.73	0.57
SRT	0.20	0.20	0.61	0.49	-0.14	-0.14	1.00	0.89
Trail Making	0.15	0.20	0.72	1.09	-0.20	-0.23	1.27	1.76
Words Memory Delayed	-0.08	-0.21	0.32	0.24	-0.22	-0.33	0.64	0.61
Correlations	0.94		0.89		0.91		0.91	
P values	0.00002		0.0002		0.0001		0.0001	

Table 6.6: Median and mean DfE scores for accuracy and latency performance measures across tasks at timepoint 1 (TP1) and at timepoint 2 (TP2).

6.3.4 Analyses on complete sample

The second set of analyses focused on utilising the comprehensive and complete dataset from timepoint 2 to investigate patterns of cognitive impairment in MS. Specifically, the focus was on assessing the overall prevalence of cognitive impairment among pwMS, as well as its manifestation across various tasks and cognitive domains. Additionally, this analysis aimed to investigate the influence of disease-related factors, such as disease subtype and duration, on cognitive function in MS.

6.3.4.1 Sensitivity to MS of tasks in the optimal battery

First, the sensitivity to MS across all tasks was assessed, including those not evaluated at timepoint 1 (Objects Memory Immediate and Delayed). Median DfE scores with 95% confidence intervals across all performance metrics for each task are presented in Fig. 6.5 including p-values obtained performing Wilcoxon signed-rank tests against zero.

The most sensitive tasks in terms of accuracy were Switching Stroop, Verbal Analogies, Card Pairs, and Objects Memory Delayed, exhibiting medium-large effect sizes. The most sensitive tasks in terms of latency were Trail Making, Card Pairs and Switching Stroop. Among the costs, the incongruency cost in accuracy for Switching Stroop and the switch costs in latency for Trail Making were the most prominent. Overall, this indicated deficits across multiple domains, particularly cognitive control, information-processing speed, working memory, verbal reasoning, and memory. For Motor Control, Trail Making and SRT, only latency measures, including costs in latency, were negatively impacted

by MS, which aligns with the design of these tasks targeting reaction time. Conversely, for Blocks and Verbal Analogies, accuracy measures were predominantly affected, reflecting their focus on accuracy.

Interestingly, pwMS displayed larger deviations from matched controls in both Object Memory Immediate and Delayed compared to Words Memory Immediate and Delayed, for both accuracy and latency measures. Additionally, the median DfE scores on percent forgetting measures for both the Objects Memory and Words Memory tasks were either non-significant or negligibly small, indicating that the ability to consolidate information from the Immediate to the Delayed versions of the tasks was in general not impaired.

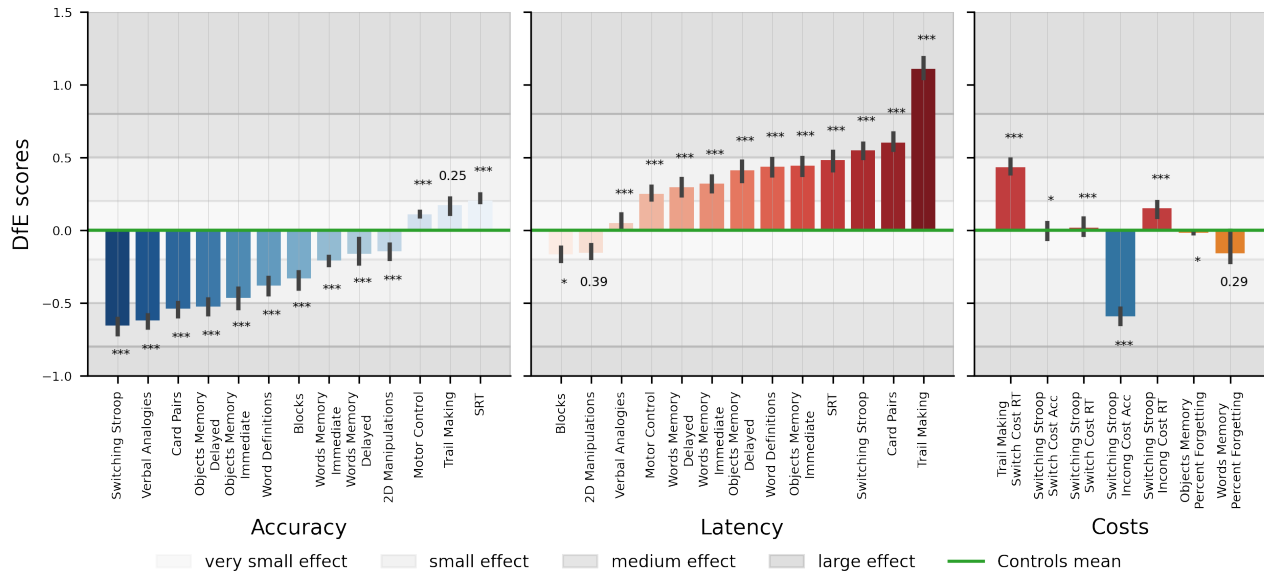


Figure 6.5: Sensitivity to MS across tasks. Median DfE score for accuracy, latency, and extra performance metrics across tasks with 95% confidence intervals. P values were obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.

6.3.4.2 Prevalence rates of cognitive impairment in pwMS: global-level

The overall prevalence of cognitive impairment in pwMS from timepoint 2 varied dramatically with the criterion being used, ranging from 72.3% with a liberal criterion, to 29.7% with a fair criterion, and down to 14.9% with a conservative criterion (as shown in Table 6.7). Irrespective of the criterion applied, pwMS identified as cognitively impaired were, on average, older and had longer disease durations. Additionally, within the impaired groups, there was a lower proportion of females and of relapsing-remitting individuals compared to the non-impaired groups.

Criterion	Groups	Prevalence	Age (min-max, mean)	Gender (%)	Disease subtype (%)	Disease duration (min-max, mean)
Liberal Abnormality threshold = 1SD Critical number of abnormal parameters ≥13%	Impaired	72.3%	18-85, 54.9	Female = 77.9 Male = 22 Other = 0.1	RR = 51.5 SP = 25.7 PP = 13.7 DK = 4.0 TrnSP = 2.3 BN = 2.8	0-63, 18.8
	Non-impaired	27.7%	22-84, 50.7	Female = 84 Male = 15.8 Other = 0.2	RR = 69.6 SP = 12.6 PP = 8.8 DK = 3.2 TrnSP = 2.6 BN = 3.2	0-50, 14.8
P-values from groups comparison			<0.0001	0.0104	<0.0001	<0.0001
Fair Abnormality threshold = 1.5SD Critical number of abnormal parameters ≥20%	Impaired	29.7%	19-85, 56.1	Female = 72.2 Male = 27.6 Other = 0.2	RR = 42.2 SP = 33.4 PP = 14.9 DK = 5 TrnSP = 2.4 BN = 2.1	0-63, 20.5
	Non-impaired	70.3%	18-84, 52.7	Female = 82.7 Male = 17.2 Other = 0.1	RR = 62.6 SP = 17.2 PP = 11.3 DK = 3.2 TrnSP = 2.4 BN = 3.3	0-61, 16.5
P-values from groups comparison			<0.0001	<0.0001	<0.0001	<0.0001
Conservative Abnormality threshold = 2SD Critical number of abnormal parameters ≥20%	Impaired	14.9%	19-85, 56.8	Female = 71.3 Male = 28.7 Other = 0	RR = 36.2 SP = 39.2 PP = 13.8 DK = 6.7 TrnSP = 2.6 BN = 1.5	0-62, 22.0
	Non-impaired	85.1%	18-84, 53.2	Female = 81 Male = 18.8 Other = 0.1	RR = 60 SP = 19 PP = 12.2 DK = 3.3 TrnSP = 2.3 BN = 3.2	0-63, 17.0
P-values from groups comparison			<0.0001	0.0009	<0.0001	<0.0001

Table 6.7: Prevalence, demographics, clinical phenotype, and disease duration of people with MS classified as having or not cognitive impairment using three criteria from the literature with different stringency. The criteria were based on a combination of different abnormality cut-offs and critical numbers of abnormal parameters. P-values were obtained using t-tests or chi-squared tests where appropriate.

6.3.4.3 Prevalence rates of cognitive impairment in pwMS: task-level

Next, I selected only one most representative outcome metric for each task and assessed the percentage of pwMS impaired at the task-level using three different abnormality thresholds: at 1SD, 1.5SD, and 2SD. These thresholds corresponded to liberal, fair, and conservative definitions of impairment, respectively. I then compared the percentage of pwMS impaired for each task to the percentage of controls resulting impaired using the same thresholds, providing a baseline level of impairment. Fig. 6.6A displays violin plots of DfE scores for pwMS (in blue) and controls (in grey) across tasks with the three impairment thresholds marked. Fig. 6.6B and Fig. 6.6C show, respectively, the percentages of pwMS and controls classified as impaired using the three thresholds and the differences in impairment rates between pwMS and controls.

The proportion of users identified as impaired was consistently higher in pwMS compared to controls across all tasks, irrespective of the threshold stringency. The proportion of pwMS classified as impaired varied across tasks, ranging from 18.5% to 53.1% with a liberal threshold, 7.3% to 40.7% with a fair threshold, and 2.2% to 31.3% with a conservative threshold (as shown in Fig. 6.6B). Thus, a less stringent threshold resulted in a higher number of pwMS being classified as impaired. This discrepancy was reduced but still present when considering baseline impairment levels (Fig. 6.6C). Tasks measuring information-processing speed (Motor Control, SRT and Trail Making) exhibited the largest differences in impairment rates between pwMS and controls, followed by the Switching Stroop task. Notably, for Trail Making, the difference in impairment rates ranged from 28% with a conservative threshold to 42.9% with a liberal threshold. The smallest differences in impairment rates were observed for 2D Manipulations and Words Memory Delayed.

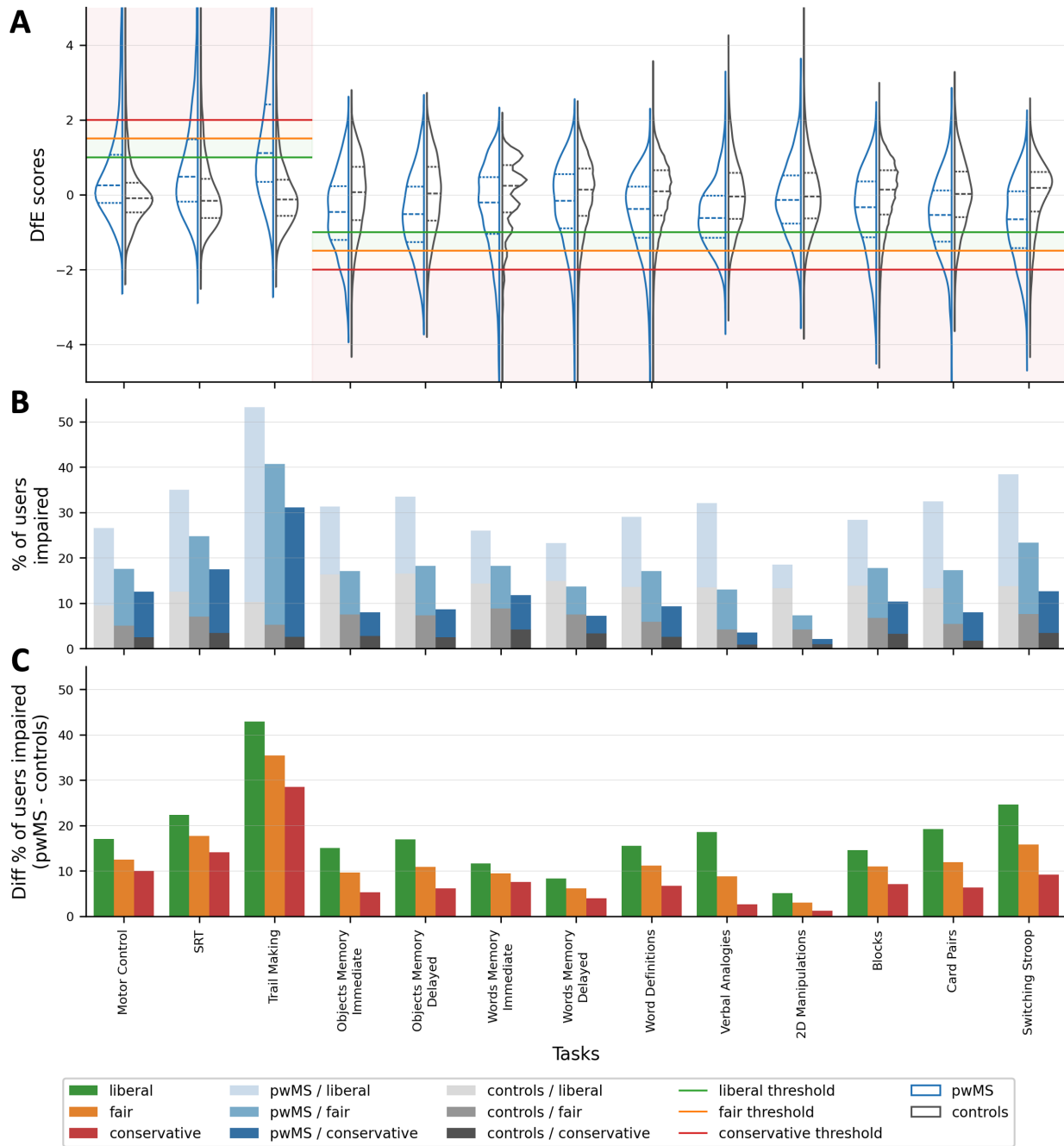


Figure 6.6: Prevalence of cognitive impairment in people with MS and controls across tasks. (A) Violin plots of DfE scores across tasks for pwMS (blue) and controls (grey) with liberal, fair, and conservative thresholds for impairment indicated on the plot. The thresholds are positive for tasks for which latency was chosen as primary outcome measure, and negative for tasks for which accuracy was chosen. (B) Barplots of the percent of users impaired for pwMS and controls, as well as (C) of the difference in percent of users impaired between the two groups, across tasks and using the different thresholds.

6.3.4.4 Sensitivity to disease subtype and duration

After assessing the prevalence of cognitive impairment in pwMS who participated in timepoint 2, I evaluated the impact of disease-related factors, specifically, subtype and duration, on task performance, focusing only on one most representative performance metric per task. Specifically, I conducted statistical analyses to investigate whether DfE scores on the selected metric for each task differed significantly based on disease subtype and duration.

Mean DfE scores with 95% confidence intervals for the primary performance metric across tasks are presented separately for each subtype and duration in Fig. 6.7A. Additionally, the associated p-values are reported. P-values were obtained performing non-parametric 2-way permutational ANOVA for each performance metric randomly shuffling 10,000 times the scores across users, including as factors disease subtype, disease duration, and their interaction.

The interaction between subtype and duration was found to be non-significant across all tasks. Subtype was found to have a significant association with performance for tasks measuring information-processing speed (Motor Control, SRT and Trail Making) and for tasks measuring delayed memory recall (Object and Words Memory Delayed). Notably, disease duration did not have a significant association with any of the tasks when looking at primary performance metrics and accounting for subtype.

Post-hoc pairwise comparisons were performed for the tasks for which a significant association with disease subtype was found. For the delayed memory recall tasks, none of the pairwise comparisons were significant after correcting for multiple comparisons using the Bonferroni method. Thus, differential patterns of cognitive impairment across subtypes were found only for tasks measuring information-processing speed. P-values for each pair of subtypes across the three significant tasks are reported in Fig. 6.7B. Regarding SRT and Motor Control, significant differences were found between relapsing subtypes (Benign, RRMS and Transitioning to SPMS) and progressive subtypes (SPMS and PPMS), with progressive subtypes presenting with higher deviations from controls on average, i.e. increased response times. For Trail Making, instead, significant differences were found between RRMS and Transitioning to SPMS, as well as between RRMS and the progressive subtypes. Additionally, for this task, PPMS significantly differed from SPMS but not from Benign. Overall, the SPMS group was the one presenting with the largest DfE scores on average.

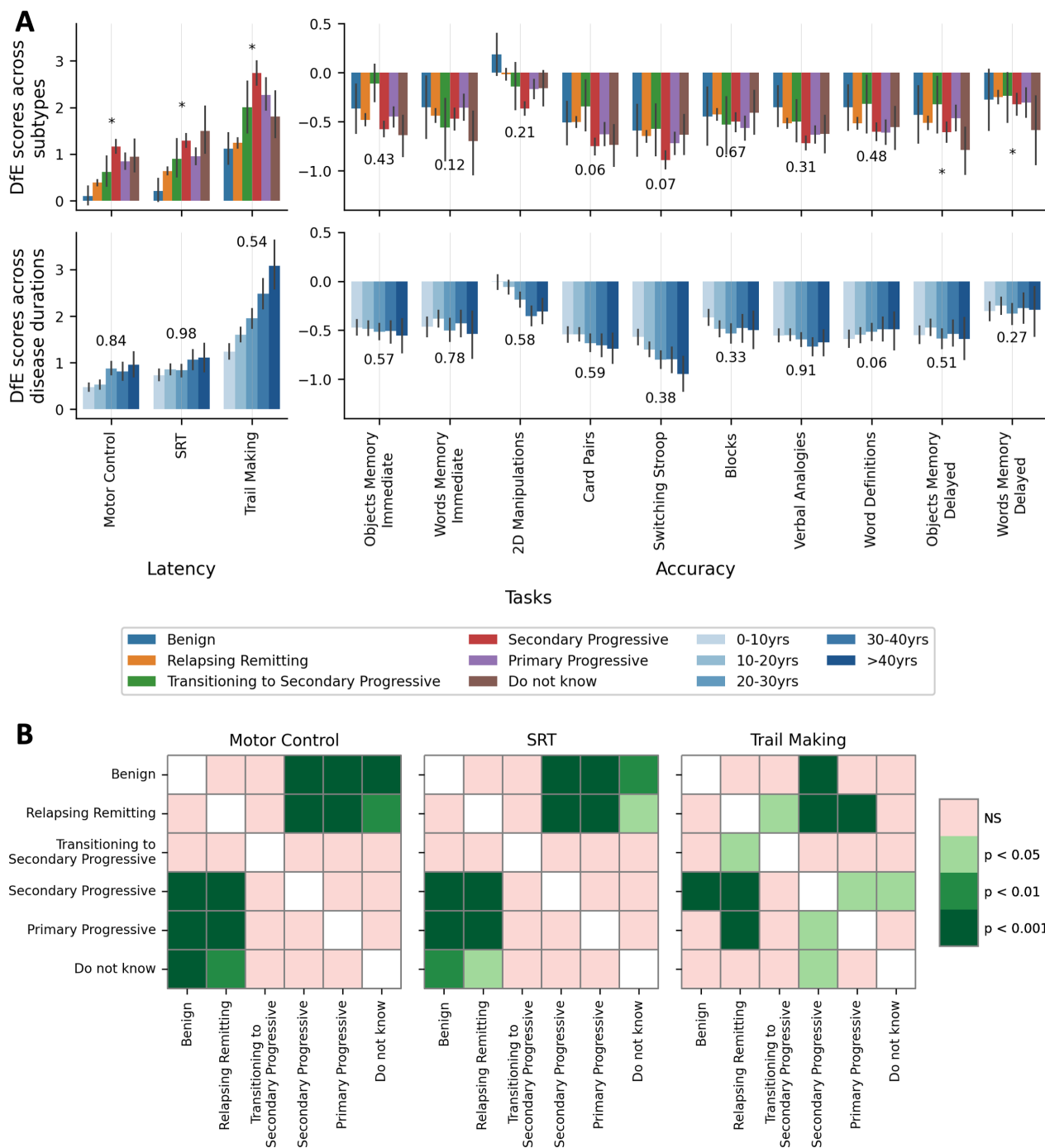


Figure 6.7: Associations of performance across tasks with disease subtype and duration. (A) Mean DfE scores with 95% confidence intervals for the primary outcome metrics for all tasks plotted across subtypes (above) and disease durations (below). 2-way permutational ANOVA was applied separately for each task to test whether the DfE scores varied significantly with disease subtype, duration, and their interaction. Resulting p values are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. (B) P-values resulting from post-hoc pairwise comparisons of DfE scores across subtypes for tasks that showed a significant association with subtype.

6.3.4.5 Sensitivity to early-stage MS

No significant associations were found between disease duration and performance across tasks when accounting for subtype, but all tasks demonstrated some degree of sensitivity to MS. For this reason, I decided to investigate task sensitivity to early-stage MS, i.e. only for pwMS between 0 and 5 years from disease onset. Among pwMS who participated in timepoint 2, there were 298 who were in the early-stage of the disease. I examined their task performance using only the primary outcome metric for each task, to define whether tasks were sensitive to early-stage MS. Median DfE scores with 95% confidence intervals across tasks for people with early-stage MS are presented in Fig. 6.8 including p-values obtained performing Wilcoxon signed-rank tests against zero.

All tasks were found to be sensitive to early-stage MS except 2D Manipulations, indicating the presence of deficits spanning multiple cognitive domains already in the first 5 years from disease onset. Trail Making, exhibiting a large effect size, followed by Verbal Analogies and Object Memory Delayed, both with medium-large effect sizes, emerged as the most sensitive tasks to early-stage MS, indicating deficits in information-processing speed, verbal reasoning, and visuo-spatial memory compared to matched controls.

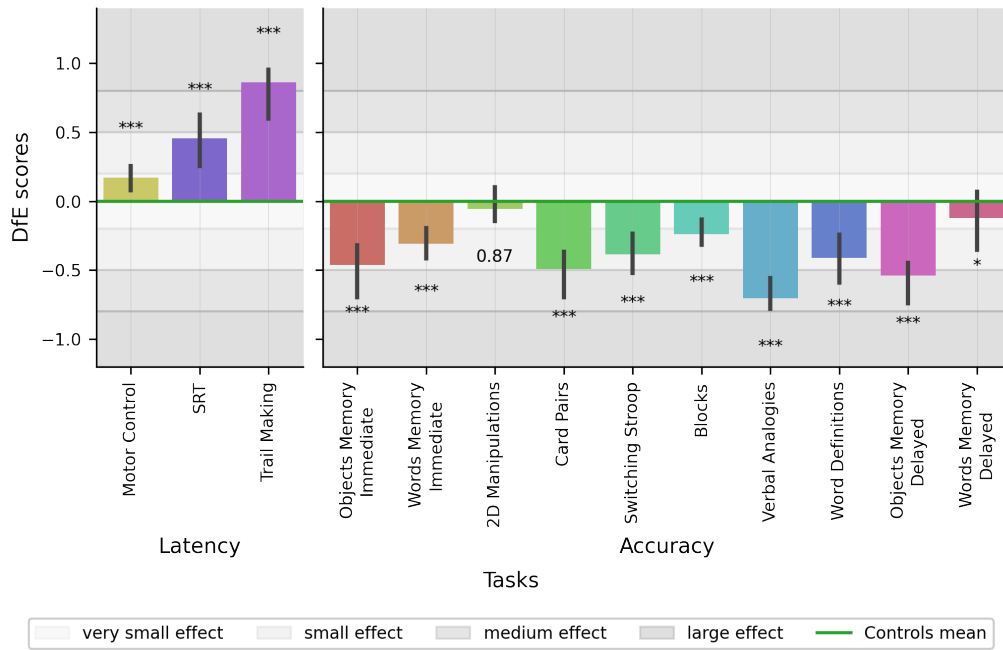


Figure 6.8: Sensitivity to early-stage MS across tasks. Median DfE scores with 95% confidence intervals for primary outcome metrics across tasks for people with early-stage MS (between 0 and 5 years from disease onset). P values were obtained performing Wilcoxon signed-rank tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.

6.3.4.6 Associations of underlying cognitive domains with disease subtype and duration

In this section, the goal was to summarise performance across tasks with fewer measures, one representing overall performance and hence general cognitive function, and the others representing performance across the different cognitive domains measured by the optimal battery.

Factor analysis, applied to the most representative outcome metrics from the 13 tasks, identified 4 factors with eigenvalues greater than 1 (Fig. 6.9D). The first unrotated factor explained 23% of the variance in the data and all features had loadings greater than 0.36 onto it. This factor was selected as representative of overall performance and labelled as the “G factor”. The four rotated factors obtained after applying varimax rotation, instead, represented the different cognitive domains evaluated by the battery, respectively Objects Memory, Information Processing Speed (IPS), Problem Solving, and Words Memory. These domains were obtained interpreting the rotated loading matrix which is presented in Fig. 6.9A. The scores on these five resulting factors (first unrotated and four rotated) were evaluated for each user, resulting in five summary performance measures.

Next, I examined how these summary measures varied with disease-related factors, such as disease subtype and duration, and if the trends were consistent with what was found at the task-level (Section 6.3.4.4). Fig. 6.9B and Fig. 6.9C present the mean scores across factors for the different disease subtypes and durations, respectively. 2-way ANOVA revealed that the interaction of disease subtype and duration was not significant across all factors. Additionally, disease subtype was significantly associated with the G factor and on the IPS factor, with progressive individuals presenting with lower mean scores compared to relapsing individuals. On the contrary, disease duration was not significantly associated with any of the factors when subtype was accounted for, aligning with what was found at the task-level.

Fig. 6.9E shows the standardised regression coefficients across the different subtypes for the significant factors, i.e. G and IPS. The significance of the coefficients was evaluated using “Benign” as the reference category. Being SPMS or PPMS, or having an unknown clinical phenotype, corresponded to having significantly lower factor scores on average compared to being Benign, indicating higher cognitive impairment both overall (G factor) as well as specifically in the IPS domain.

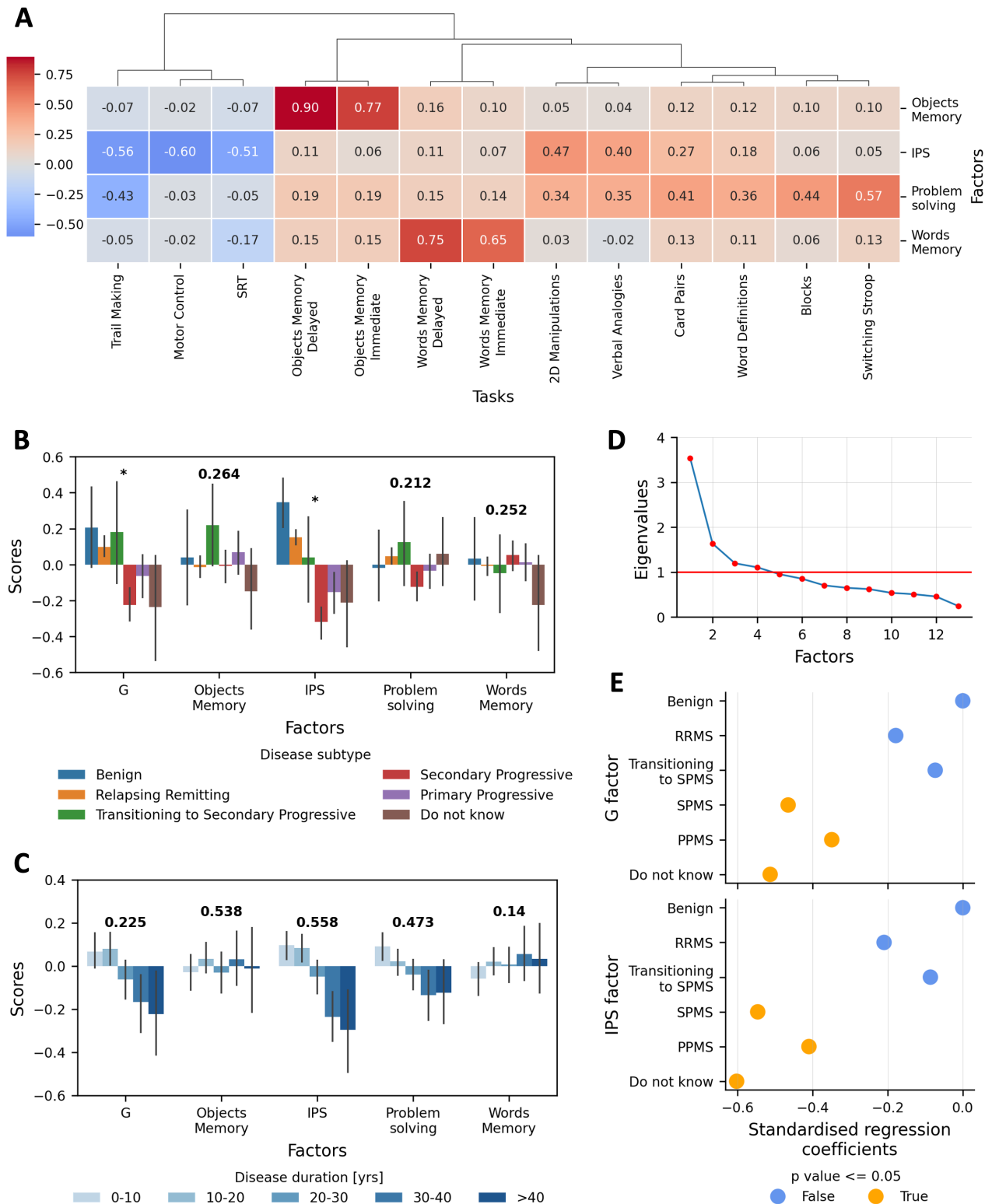


Figure 6.9: Underlying cognitive domains and their associations with disease subtype and duration. (A) Loadings of the tasks onto the different rotated factors, which represent underlying cognitive domains. (B) Mean scores for factors across subtypes and (C) disease durations. 2-way ANOVA was applied separately for each factor to test whether the scores on that factor varied significantly with disease subtype and duration. Resulting p values are reported in the figures as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. (D) Scree plot of eigenvalues derived from factor analysis of most representative outcome metrics across tasks. (E) Standardised regression coefficients across disease subtypes resulting from multiple linear regression models predicting factors' scores from them.

6.3.5 Analyses on overlapping sample

The final set of analyses focused on evaluating changes in cognitive function at the MS population level over a 1-year period. For these analyses, only users who completed each task at both timepoint 1 and 2 were included. Additionally, for users who had also completed the PROs during the Register collection windows prior to both the cognitive assessments, changes in cognitive function were related to changes in other manifestations of the disease (e.g. motor function and mood disturbances).

6.3.5.1 Longitudinal analysis

Fig. 6.10 reports the mean standardised performance deltas for accuracy and latency measures across tasks obtained subtracting performance at timepoint 1 from performance at timepoint 2 for users in the overlapping sample. Each mean delta was based on at least 444 observations. For tasks that were not randomly sampled at timepoint 1 (fixed tasks), the number of observations available reached up to 1220.

The mean deltas were non-significant or negligibly small (exhibited very small effect sizes) across most tasks indicating that cognitive function did not deteriorate over a 1-year period in the studied sample of pwMS. The only tasks showing larger (small effect sizes) and significant deltas were 2D Manipulations and Verbal Analogies in terms of accuracy and Trail Making in terms of latency. For Verbal Analogies and Trail Making the deltas indicated a worsening on average in timepoint 2 compared to timepoint 1, while for 2D Manipulations they indicated an improvement. Nonetheless, it is not possible at this stage to say whether the observed changes could be attributed to changes in cognitive function, or to other confounding factors, such as learning effects and task modifications, especially for 2D Manipulations and Verbal Analogies, for which task modifications had a significant effect on task performance (Section 6.3.3.1).

Out of the 1237 users forming the overlapping sample, 916 had completed at least one of the PROs administered in the Register during both collection windows prior to the cognitive assessments. Based on the available observations, the correlations between changes in cognitive function and changes in the PROs, were evaluated. Correlations were predominantly non-significant across all combinations of task performance measures and questionnaires, indicating that changes in cognitive function were largely independent of changes in other manifestations of the disease, such as motor function and mood disturbances. Among the significant correlation, the only ones non-negligibly small ($r > 0.1$), were found for accuracy deltas for Switching Stroop with deltas for webEDSS ($r = 0.22$), accuracy deltas for Words Memory Delayed with deltas for EQ5D ($r = -0.13$), and latency deltas for Switching Stroop ($r = 0.12$) and Word Definitions ($r = 0.12$) with deltas for HADS-anxiety component.

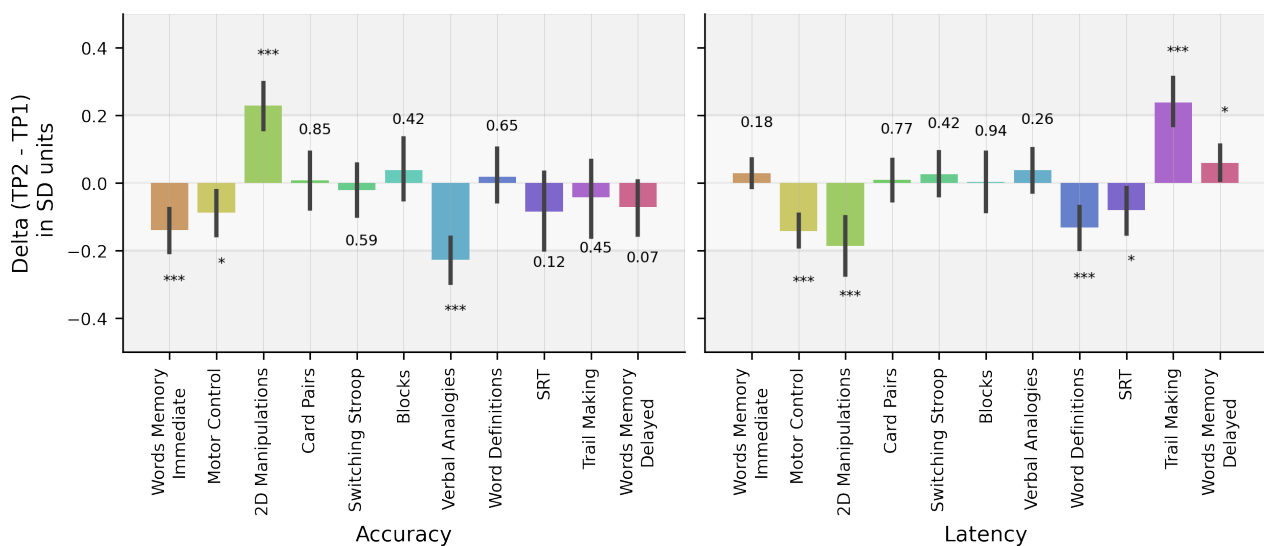


Figure 6.10: Change in cognitive function over 1-year in people with MS. Mean standardised performance deltas in terms of accuracy (left) and latency (right) scores between timepoint 1 and timepoint 2 with 95% confidence intervals. P values were obtained performing t-tests against zero and are reported in the figure as: $p > 0.05 = p$, $p \leq 0.05 = *$, $p \leq 0.001 = **$, $p \leq 0.0001 = ***$. The different effect sizes are highlighted through gradually increasing grey shaded areas.

6.4 Discussion

6.4.1 Summary of findings

The study presented in this chapter further demonstrated the feasibility and accessibility of online cognitive assessment in MS, testing over 2,000 pwMS in just two months, with almost half of the participants from the previous study returning to get tested again. Additionally, the study validated the optimal battery and substantiated task selection by replicating previous results on an independent sample of pwMS. This excluded any potential confounding factors related to changes in battery order and content across users. The optimal battery enabled the collection of complete and comprehensive cognitive data on a large and representative sample of pwMS. Using these data, it was possible to investigate associations between cognitive function and disease-related factors, particularly disease subtype and duration, with greater statistical power. Results showed that cognitive deficits spanning multiple domains were present across all subtypes and disease durations, including early in the disease. Moreover, associations of deficits across domains with increased disease duration were no longer significant when subtype, which covaries with duration, was accounted for. Finally, information-processing deficits were more pronounced in progressive compared to relapsing-remitting subtypes. Altogether these results endorsed the validity of the optimal battery and highlighted the value of the data collectable with it in confirming associations already observed in the literature and deriving new insights into cognitive function in pwMS.

6.4.2 High participation and completion rates

The study's high participation and completion rates not only emphasise the accessibility of the online cognitive assessment but also highlight the strong engagement and interest demonstrated by pwMS in

having this tool integrated into the UKMSR platform. Although fewer people participated this time (2696 compared to the 3066 of the previous study), considering the much shorter collection window (two months instead of seven months), it was an even more impressive result: 14% of the total Register users tested their cognition in just two months. Additionally, the completion rate exceeded that of the previous study (81.5% versus 78.4%), with a notable increase in the percentage of participants who were not active users of the Register (31% versus 17%). This suggests a growing interest in the technology that contributed to an overall increase in engagement with the Register.

6.4.3 Tested sample was representative of broader MS population

Furthermore, the sample of pwMS from timepoint 2 was representative of the broader MS population. The distribution of gender and subtypes in the sample closely aligned with estimates from the literature (Multiple Sclerosis International Federation 2020). Additionally, there were no significant differences in age, gender, disease subtype, and disability severity when compared to the sample from timepoint 1, which had been previously demonstrated to be representative (Section 5.4.2). The sole significant differences between the two samples pertained to disease duration and physical disability. However, these differences were minimal, with the timepoint 2 sample exhibiting, on average, a shorter disease duration by 2.5 years and a higher physical disability, as indicated by a 3-point difference on the normalised MSIS-29 motor scale. Importantly, these differences did not translate into variations in overall disability severity between the two samples (as measured by the webEDSS).

6.4.4 People with MS who participated in TP1 returned for TP2

Another positive signal was the return rate at timepoint 2, with 41.3% of participants from timepoint 1 choosing to participate again. When comparing sociodemographic and disease factors between returning and non-returning pwMS, no significant differences were found in terms of gender, disease duration, and disease severity. However, there were significant differences based on age and disease subtype, with people who returned being on average older and having a higher prevalence of progressive and a lower prevalence of relapsing subtypes. Studies from the literature indicate that cognitive impairment is associated with older age and is more severe and prevalent in progressive compared to relapsing-remitting subtypes (Johnen et al. 2017; Planche et al. 2016; Ruano et al. 2017). Therefore, given their characteristics, individuals who returned were more likely to have cognitive impairment than those who did not. If this holds true, it might imply that individuals with cognitive impairment are more inclined to undergo periodic testing, while those without may not perceive the need to be retested yearly. This would also indicate that non-returning people did not do so because they encountered difficulties with the assessment. Some may argue that this is not the case as individuals who did and did not return had no significant differences in overall disability severity, as indicated by their webEDSS. However, it is important to note that the webEDSS is predominantly focused on physical and sensory disability and could be overlooking differences in cognitive function between the two groups (Saccà et al. 2017). This hypothesis could be tested by comparing the cognitive performance scores at timepoint 1 between people who did and did not return for timepoint 2. Nonetheless, it is not possible to do so at this stage because each patient performed a different combination of tasks at timepoint 1. This limitation will be addressed in the future when an additional timepoint is collected using the optimal battery. In any case, the availability of such technology empowers pwMS who are

experiencing cognitive difficulties to easily undergo cognitive testing, identify their specific problems, and seek support, streamlining current processes.

6.4.5 Validation of the content of the optimal battery

After having further ascertained that the assessment was feasible and of interest to pwMS the optimal battery needed to be validated. The selection of included tasks was justified by the replication analysis: the sensitivities to MS across tasks evaluated on the independent sample from timepoint 2 were highly correlated with those evaluated on the sample from timepoint 1. Even in the cases in which the sensitivity to MS changed the most, for accuracy in Verbal Analogies and latency in Trial Making, the direction of the change was such that it increased the sensitivity, and consequently confirmed the inclusion of the task in the optimal battery. An exception in this sense was the 2D Manipulations task, which showed a decreased sensitivity to MS, going from a small to a very small effect size. However, it is important to keep in mind that some variations in sensitivities were to be expected given that the tasks underwent some modifications prior to timepoint 2. These modifications were shown to have a very small and therefore negligible impact on performance across most tasks, except for the Verbal Analogies and 2D Manipulations tasks. Therefore, caution must be put in interpreting the results for these two tasks: while the modifications underwent by Verbal Analogies have made the task more difficult, which is reflected in its increased sensitivity to MS, those underwent by 2D Manipulations have shortened its duration, decreasing the cognitive load on participants. This could be the reason behind the decrease in sensitivity observed for 2D Manipulations, especially considering that the normative data used for this task were obtained cutting the trial-by-trial normative data from the previous version of the task to match the modified duration. This approach could inflate patients' results as it does not account for the fact that participants might speed up when approaching the end of the task. In general, it would be sensible to collect normative data on the modified version of the 2D Manipulations task and re-assess its sensitivity to MS as well as its presence in the optimal battery.

In addition, it is essential to collect normative data for the modified version of the Spotter task. This task, although not analysed in this chapter due to the absence of appropriate normative data, is a crucial component of the optimal battery. It demonstrated the largest sensitivity to MS in the previous study (Section 5.3.4) and serves as the main representative of the attention domain.

6.4.6 Additions to the optimal battery

In regards to the memory component of the optimal battery, the Objects Memory tasks evaluated in this chapter as alternatives or additions to the Words Memory tasks, demonstrated superior sensitivity to MS both in terms of accuracy and latency, and similar results in terms of elements forgotten between the delayed and immediate versions of the task. They will therefore be included in the battery as an addition to the Words Memory tasks. The decision to include them as addition and not alternative is driven by the fact that they are measuring a different aspect of memory (visuo-spatial instead of verbal) and that participants did not seem to struggle with battery duration (given the high completion rate). However, in case the need for a shorter battery arises in the future, the recommendation is to keep the Objects Memory rather than the Words Memory tasks.

6.4.7 Design of the optimal battery

On this note, the tasks' presentation order was engineered to prevent people from dropping out and to maximise their engagement: the more challenging and cognitively demanding tasks were placed at the beginning of the battery, while the more straightforward and repetitive tasks were positioned at the end. An exception was the Spotter task, designed to fatigue the participants and measure sustained attention, which was intentionally placed at the end to prevent high dropout rates. Analysis of where participants dropped out confirmed the expectation that the rates were higher after more cognitively demanding tasks like 2D Manipulations, Card Pairs, and Switching Stroop. Notably, Switching Stroop had a particularly high dropout rate of 16%, suggesting the need for potential enhancements in task instructions.

6.4.8 Observed prevalence of cognitive impairment further validates the optimal battery

Further supporting the validity of the optimal battery are the observed prevalence rates of cognitive impairment within the sample of pwMS. Whether examining sensitivities or prevalence rates the same tasks stand out and consequently the same domains appear to be more affected. Moreover, the proportion of users identified as impaired was consistently higher in pwMS compared to controls across all tasks, irrespective of the stringency of the threshold applied.

As anticipated from existing literature (Fischer et al. 2014), the choice of threshold for classifying individuals as impaired significantly influenced the prevalence rate, both on a task-specific basis and in the overall assessment. The overall prevalence rate varied widely, ranging from 14.9% to 72.3%. Nevertheless, when comparing groups categorised as impaired vs non-impaired, the same significant differences were observed regardless of the threshold being used. The impaired groups were on average older in age, had longer disease durations, and had a lower prevalence of female and of relapsing subtypes compared to the non-impaired groups. These trends align with associations identified in the literature (Ruano et al. 2017), further reinforcing the meaningfulness of the data derived from the optimal battery. Furthermore, they indicate that the accumulation of cognitive impairment is a continuous phenomenon and the establishment of a threshold inevitably results in overlooking changes in cognitive function on either side of the classification (Pitteri et al. 2019). The optimal battery outputs continuous measures for each patient, including DfE scores and factor scores. These metrics assess performance across various domains and enable detecting changes over time while providing a more nuanced measurement of impairment severity. Importantly, these continuous measures yield the same overarching inferences as a dichotomous system while allowing for a finer-grained analysis.

6.4.9 Data from the optimal battery can be used to evaluate the impact of disease subtype and duration on cognitive function

The optimal battery allowed collecting complete and comprehensive cognitive data on a large and representative sample of pwMS. Furthermore, these data could be linked with additional information available on the Register, including PROs. Collectively these data offer a unique opportunity to investigate trends in cognitive impairment and its associations with disease-related factors as its large-

scale grants superior statistical power. In this chapter, I presented an example of this by evaluating associations with disease subtype and duration.

While previous studies consistently demonstrate the heightened severity and prevalence of cognitive impairment in progressive subtypes (Bergendal et al. 2007; Huijbregts et al. 2004; Johnen et al. 2017), the literature lacks clarity on the association between cognitive deficits and disease duration (Chiaravalloti et al. 2008). It is crucial to consider disease subtype when exploring this association, as disease duration is implicit in disease subtype; for instance, secondary progressive status typically manifests later in the disease (Weinshenker et al. 1989b).

Leveraging the extensive scale of the dataset at hand, I examined the effects of subtype, duration, and their interaction on cognitive performance across tasks. In alignment with the existing literature (Huijbregts et al. 2004; Johnen et al. 2017; Planche et al. 2016) at the task and domain levels progressive subtypes, particularly secondary progressive individuals, exhibited poorer information-processing speed and decreased general performance compared to relapsing subtypes. Notably, no significant interactions with disease duration were found, nor were there significant main effects of disease duration on any tasks or underlying cognitive domains. These findings align with some studies from the literature (Jennekens-Schinkel et al. 1990; Patti et al. 1998; Rao et al. 1991b) but are in contrast with others (Amato et al. 2001; Smestad et al. 2010). In addition, widespread cognitive deficits were shown in the early stages, specifically among individuals between 0 and 5 years from onset, further substantiating the claim that cognitive impairment is independent of disease duration.

However, employing a cross-sectional design is suboptimal for examining changes in cognitive function with disease duration (Chiaravalloti et al. 2008). To address this limitation, I analysed data from the overlapping sample to investigate potential cognitive changes over a 1-year period. The mean changes across tasks were predominantly non-significant or exhibited negligibly small effect sizes. In instances where significant yet non-negligible changes were identified, these could not be solely attributed to cognitive deterioration. This is due to potential confounding factors such as practice effects and task modifications. While the former were unlikely, given that the tasks were designed to minimise practice effects (e.g., random sampling from a wide set of stimuli), the latter may have contributed, particularly considering that the observed changes in performance were bidirectional. Even in the absence of confounding effects, the lack of associations with disease duration over a 1-year period would not conclusively negate an overall association. In fact, a study from Amato et al. 2001 has shown that if a sufficiently long follow-up time is allowed, cognitive impairment is likely to emerge and progress in a sizable proportion of pwMS.

Nevertheless, the permanent integration of this battery within the Register platform would greatly facilitate the periodic and efficient collection of cognitive data for extended follow-ups. This approach holds the potential to generate the longitudinal data necessary for addressing questions related to cognitive changes over time or similar inquiries with greater depth and thoroughness.

6.4.10 Limitations and future directions

While this study showed promising results, there are limitations that need to be addressed and additional steps required to fully validate the optimal battery and promote its widespread use across registries and in clinical practice. The first limitation pertains to the absence of complete normative

data specific to the battery. While normative data are available for each individual task within the battery, these data originate from various studies or different timepoints within the same study using different batteries. A better approach would involve the collection of normative data from controls performing an identical battery comprising all tasks presented in the same order as for patients. This would enable a fairer comparison with controls as well as the estimation of sensitivity and specificity in classifying patients from controls.

Additional limitations encompass various psychometric properties that require thorough evaluation before deeming this battery ready for practical use. These involve assessing test-retest reliability, examining potential learning effects, evaluating the impact of supervised versus unsupervised administration, and determining comparative effectiveness. Of particular significance is the need for a comprehensive comparative analysis; results obtained from the online cognitive battery should be systematically compared with outcomes from a traditional neuropsychological battery used in MS (e.g., MACFIMS (Benedict et al. 2006) or BNB-N (Boringa et al. 2001)) to establish its validity as a suitable new approach.

6.4.11 Conclusions

Despite its current limitations, the optimal battery has demonstrated promising outcomes. While it may not entirely replace traditional neuropsychological batteries, it holds potential as a valuable screening tool to identify potentially impaired patients and refer them towards in-person assessments. Notably, it possesses several advantages compared to other online cognitive assessments utilised in MS (Foong et al. 2023; Wojcik et al. 2019):

- It can be administered without supervision.
- It is specifically designed to detect cognitive deficits in the MS population and as such it offers a more targeted approach than assessments intended for the broader neurological community.
- It covers a comprehensive range of cognitive domains, not limited to information-processing speed.
- it has been successfully evaluated in a large and representative sample of pwMS across two distinct studies, yielding positive results.

The forthcoming and concluding results chapter will integrate motor data collected through PROs with cognitive data acquired through the optimal battery to cluster pwMS solely based on online-collected information about their symptoms. This synthesis aims to highlight the potential of leveraging online assessment tools to derive novel perspectives on MS.

Chapter 7

Unsupervised clustering of people with MS based on digital markers of motor and cognitive impairment

In this chapter, I present an exploratory study investigating an alternative, data-driven way to cluster people with MS solely based on disability markers obtained using online tools. This study combined data collected in collaboration with the UK MS Register using patient-reported outcomes and online cognitive assessment. As a result, four clusters were identified, each displaying distinct patterns of motor and cognitive impairment, which did not align with conventional clinical subtypes. These results further validated the utility of online tools in measuring disability, showing how they can offer a different perspective on the disease.

7.1 Introduction

Over the course of the previous chapters, I have demonstrated how online assessment tools enable the efficient collection of large-scale data from representative samples of pwMS. Using these tools, it is possible to derive digital markers of motor and cognitive impairment whose validity and utility has been demonstrated and discussed throughout. The study presented in this chapter bring together all the work presented in the previous chapters exploring an alternative way to cluster pwMS solely based on the introduced digital markers of disability.

Conventional clinical subtypes are delineated by distinct clinical manifestations arising from postulated underlying disease mechanisms (Lublin et al. 1996, 2014). For instance, individuals with a relapsing-remitting course experience alternating phases of relapses and remissions. These fluctuations are driven by underlying inflammatory activity and reflect the immune system’s periodic attacks on the central nervous system. Conversely, individuals with a progressive course undergo a gradual accumulation of disability which has a basis in underlying neurodegenerative processes. In essence, these subtypes offer insights into the diverse ways MS can manifest and progress (Klineova et al. 2018).

However, high heterogeneity exists as to how the subtypes manifest, stemming from variability in the types, combinations, and severity of symptoms, rate of disability accumulation, and outcomes (Lublin et al. 2014). This variability arises from unique individual patterns of inflammation and neurodegeneration across different brain areas. Individuals from different subtypes may exhibit similar brain damage and, therefore, similar symptoms, while individuals within the same subtype may have diverse damages leading to entirely different symptoms. Moreover, studies often indicate more similarities than differences across subtypes when examining imaging and immunological markers (Eshaghi et al. 2021). These factors result in a very heterogeneous response to treatment, both across and within subtypes, make it challenging to target the right disease-modifying therapy for each patient (Multiple Sclerosis International Federation 2013; Yang et al. 2022). Targeting treatment solely based on clinical course is not informative.

Efforts have been made to explore alternative classification methods that could provide more informative insights for selecting patients for clinical trials and guiding treatment assignment. For instance, Eshaghi et al. 2021 defined alternative subtypes using unsupervised machine learning on brain MRI data, aiming to represent different underlying pathological changes. The objective was to enable the identification of subgroups to which different treatment mechanisms did or did not apply based on their pathology. Although this approach showed promising results, its widespread application requires the availability of at least one MRI scan for each patient and some degree of standardisation in deriving MRI features for input into the model.

In this chapter, I propose an alternative method to classify pwMS based solely on digital markers of disability, i.e., measures collected through online tools. The motivation for employing digital markers arises from the need for approaches using measures that are widely accessible and efficiently collectable in a standardised manner. Additionally, compared to conventional clinical phenotypes, the focus shifted from the temporal dynamics of how symptoms progress to the types, combinations, and severity of the symptoms that individuals experience.

With this change in focus, this study aimed to identify subtypes that were more representative of underlying pathological changes while being efficient to collect. An unsupervised machine learning

technique called agglomerative hierarchical clustering was applied on the digital markers of disability to derive the alternative subtypes. As this is an exploratory study, only the measures of motor and cognitive impairment which have been extensively evaluated in this thesis were included in the clustering model. However, in case of positive results, the model could be easily extended to include additional digital markers.

The primary objective of this study was to determine whether this symptom-focused, data-driven classification provided a useful grouping of pwMS and could offer complementary information to other existing classification methods (based on relapses/progression or MRI features). Additionally, it aimed to present a novel framework for understanding MS and offer a different perspective on the disease.

7.2 Methods

7.2.1 Study design

This is a population-based, cross-sectional observational study aimed at deriving an alternative approach for clustering pwMS, solely relying on digital markers of motor and cognitive impairment. These markers were collected through PROs and online cognitive tasks, respectively. Agglomerative hierarchical clustering was employed to group pwMS based on these digital markers. The resulting clusters were characterised and labelled by examining patterns in clustering variables. Additionally, they were further defined by analysing trends in sociodemographic and clinical factors across clusters, as well as patterns in other validated PROs measuring different aspects of MS-related disability, e.g., mood disturbances and fatigue.

A key aspect of the study involved assessing the prevalence of different MS subtypes within each cluster. This was done to ascertain whether the unsupervised clustering approach yielded results consistent with the traditional clinical grouping of pwMS into subtypes or if it produced a distinct classification and, thus, a novel perspective on MS.

7.2.2 Participants

This study included pwMS from the UK MS Register who had completed the online cognitive assessment administered through the Register website between November 8, 2022, and January 13, 2023 (detailed in Chapter 6) as well as two patient-reported outcomes, namely the MSIS-29 motor and the MSWS-12, during the Register periodic collection window prior to the online cognitive assessment, i.e. between September 13, 2022, and October 24, 2022.

7.2.3 Procedures

7.2.3.1 Online cognitive data

The online cognitive assessment, powered through Cognitron and hosted on the UKMSR website via an iframe, featured a battery of 14 cognitive tasks. The battery was optimized for detecting cognitive impairment in pwMS, covered various cognitive domains, and was administered in a fixed order. A

complete list of the tasks included in the battery and specifics on their design is reported in Chapter 6. The Spotter task, the final one in the battery, was excluded from the analysis due to a lack of control data. For each of the remaining 13 tasks, one most representative outcome metric was derived summarising the trial-by-trial data. Accuracy (percentage of correct responses) was chosen for 10 of the tasks, whilst latency (median or total response time) was chosen for the remaining 3 as these were specifically designed to measure reaction time. This resulted in a set of 13 features, termed the cognitive features, for each user in the study, collectively representing their cognitive function.

In alignment with the previous chapters (chapters 5 and 6), this study utilised normative data derived from the Great British Intelligence Test (GBIT) study, a large-scale online longitudinal population study conducted between December, 2019 and February, 2023 across 2 baseline timepoints and 4 follow-up timepoints. Data from the GBIT study were retained for the 13 tasks being evaluated, for individuals who were healthy controls without neurological or psychiatric conditions, who had complete sociodemographic data, were aged 16-90, and were first-time task performers to ensure unbiased assessment. Further details on the normative data, including the number of observations available for each task, can be found in Section 6.2.2.4.

7.2.3.2 Sociodemographic data and clinical factors

Sociodemographic data and clinical factors were collected through a questionnaire administered prior to the online cognitive assessment. The questionnaire asked users information about their: age, gender, dominant hand, country of residence, language, ethnicity, education, occupation, disease onset year, and disease subtype.

7.2.3.3 Patient-reported outcome data

To be eligible for inclusion in the study, users were required to have completed two PROs routinely administered by the UKMSR during the collection window prior to the online cognitive assessment: the MSIS-29 motor and the MSWS-12. These PROs were designed to measure respectively general motor and walking abilities in pwMS. Consistent with the methodology outlined in Chapter 4, the totals obtained by summing the scores on the items of each questionnaire were rescaled to a value in the range of 0–100 using a min-max normalisation procedure fitted separately for each questionnaire. Additionally, for the MSWS-12, users who were excluded from the assessment because they indicated an inability to walk, were included as scoring 100s in the normalised scale. This resulted in 2 features, termed the motor features, for each user in the study, collectively representing their motor function.

Data from other PROs collected by the UKMSR were utilised in the study, for users for whom these data were available for the collection window prior to the online cognitive assessment. These additional PROs included the HADS, measuring mood disturbances, the FSS, measuring fatigue, the EQ5D, measuring quality of life, and the webEDSS, measuring general disability severity. For the HADS, two totals were derived by summing scores on items related to depression and anxiety separately. These totals were rescaled to a value in the range of 0–100 using a min-max normalisation procedure fitted separately for depression and anxiety totals. Moreover, users were classified as normal, borderline abnormal, or abnormal in terms of depression or anxiety based on whether they scored respectively 7 or below, between 8 and 10, or 11 and above on the respective scales. Regarding the FSS and the

EQ5D, totals were derived summing scores on the items. The total for the FSS ranged between 9 and 63, while the total for the EQ5D ranged between 5 and 15. These scores, in conjunction with the webEDSS, resulted in 5 additional variables, termed the clinical variables, collectively representing other aspects of the disease, including mental health, fatigue, quality of life, and general disability severity.

7.2.4 Statistical analysis

Data curation, pre-processing, and statistical analyses were conducted in Python 3.8.5.

7.2.4.1 Pre-processing

Data were retained for pwMS who completed the online cognitive assessment (excluding the Spotter task) at timepoint 2 (Chapter 6), along with the motor PROs (MSIS-29 motor and MSWS-12) during the UKMSR collection window prior to the online cognitive assessment. Participants were excluded if they provided invalid sociodemographic information, were not aged between 16 and 90, had null or non-compliant summary outcome measures for any of the cognitive tasks, and if they performed the Words/Objects Memory Delayed task more than two hours after the Words/Objects Memory Immediate task.

For each user, only one most representative summary outcome metric per task was retained, resulting in 13 cognitive features. These were adjusted for sociodemographic factors using the normative data and expressed in Deviation from Expected (DfE) scores. Additionally, 5 variables were derived for each user by performing a factor analysis on the cognitive features. One, referred to as the G factor, represented general intelligence and was obtained by taking the first unrotated factor. The remaining 4 were obtained by taking the optimal number of varimax rotated factors and represented Objects Memory, Information Processing Speed, Problem Solving, and Words Memory. Factors were labelled based on the interpretation of the loading matrix. Details on how these factors were derived are reported in sections 6.2.3.8 and 6.3.4.6. Summary scores derived from PROs were obtained by summing scores across items for each questionnaire, as detailed in Section 7.2.3.3. These scores were linked to the cognitive data yielding a total of 25 variables: 13 cognitive features, 5 factors representing underlying cognitive domains, 2 motor features, and 5 clinical variables.

7.2.4.2 Agglomerative hierarchical clustering

The 13 cognitive features together with the 2 motor features were used as clustering variables. Agglomerative hierarchical clustering was chosen as the main clustering algorithm because it is visually interpretable, requires no prior knowledge about how many clusters to expect in the data, and provides a multi-level view of the data by nesting the clusters, making it particularly insightful for exploratory analysis. The clustering variables were all standardised by removing mean and scaling to unit variance to ensure an equal contribution of all variables to the analysis and a fair comparison between variables of different types. The euclidean distance was used as the metric to build the pairwise distance matrix. The Ward method was employed as the linkage criterion (the method for merging clusters) as it minimises the variance within each cluster producing balanced clusters while being robust to outliers.

The optimal number of clusters was determined by comparing two possible solutions and choosing the one that yielded the most interpretable clusters when examining patterns in clustering variables (as detailed in Section 7.3.2). The three possible solutions were identified using the ‘elbow method’, i.e. looking at the points where the dendrogram showed a significant increase in linkage distance, which suggested a natural division in the data.

7.2.4.3 Associations of cluster label with clustering and clinical variables

Patterns in clustering variables across clusters were analysed to characterise and label the clusters. Additionally, patterns in factors across clusters were investigated to confirm and summarise analytically the patterns identified for the cognitive features. An evaluation of patterns in clinical variables across clusters was performed to further characterise the clusters based on unseen variables representing other aspects of disability.

The associations of cluster label with clustering variables, factors, and clinical variables were evaluated using a series Kruskal-Wallis H tests. This non-parametric test, an alternative to one-way ANOVA, is designed for comparing the medians of three or more groups when the dependent variable is not normally distributed, as is the case for most variables included in this study. For variables exhibiting a significant association with cluster label, post-hoc pairwise comparisons were conducted using Dunn’s test, to discern which clusters differed from each other based on these variables. The Bonferroni method was employed to correct for multiple comparisons, and an alpha threshold of 0.05 was applied to evaluate statistical significance.

7.2.4.4 Prevalence of cognitive impairment across clusters

To evaluate the prevalence of global cognitive impairment across clusters and determine its consistency with the patterns in cognitive features and factors across clusters, a classification criterion, representing a fair level of stringency, was chosen as it was the most used in the existing literature (Fischer et al. 2014). This criterion was based on an abnormality threshold — defined as a 1.5SD cut-off over which performance was deemed abnormal — coupled with a critical number of abnormal parameters. Thus, a user was classified as cognitively impaired if they scored abnormally on at least 20% of the performance metrics. Since one summary performance metric was selected for each task, resulting in 13 cognitive features, a user had to score abnormally on at least 3 tasks to be considered cognitively impaired. For tasks where accuracy was selected as the summary performance metric, a negative SD cut-off was applied to pwMS’s DfE scores to classify them as abnormal. Conversely, for tasks where latency was chosen, a positive SD cut-off was applied.

7.2.4.5 Associations of cluster label with sociodemographic variables and clinical factors

Patterns of sociodemographic variables, including age, gender, and education, and of clinical factors, such as disease subtype, disease duration, and age at onset, were investigated to enhance the characterisation of the clusters. T-tests or chi-squared tests were conducted to evaluate the associations with cluster label, depending on whether the variables were continuous or categorical. An alpha threshold of 0.05 was applied to assess statistical significance.

7.3 Results

7.3.1 Participants

Overall, 1180 of the original 2696 pwMS who participated in timepoint 2 (Section 6.3.2) were included in this study. Table 7.1 presents their sociodemographic characteristics.

Socio-demographic characteristics		People with MS
Total N		1180
Age (min-max, mean)		19-85, 55.8
Gender	Female	921 (78.1%)
	Male	258 (21.9%)
	Other	1 (0.1)
Dominant hand	Right	1011 (85.7%)
	Left	134 (11.4%)
	Ambidextrous	35 (3.0%)
First Language	English	1140 (96.6%)
	Other	40 (3.4%)
Ethnicity	White	1148 (97.3%)
	Mixed or multiple ethnic groups	11 (0.9%)
	Asian or Asian British	11 (0.9%)
	Black, Black British, Caribbean or African	6 (0.5%)
	Other ethnic group	4 (0.3%)
Residence	United Kingdom	1173 (99.4%)
	Abroad	7 (0.6%)
Education	preGCSE	27 (2.3%)
	School	531 (45%)
	Degree	585 (49.6%)
	PhD	37 (3.1%)
Occupation	Worker	433 (36.7%)
	Retired	470 (39.8%)
	Disabled / Not applicable / Sheltered employment	202 (17.1%)
	Homemaker	51 (4.3%)
	Unemployed / Looking for work	15 (1.3%)
	Student	7 (0.6%)
	Unknown	2 (0.2%)
Disease Subtype	Benign	34 (2.9%)
	Relapsing-remitting	618 (52.4%)
	Transitioning to secondary progressive	30 (2.5%)
	Secondary progressive	278 (23.6%)
	Primary progressive	181 (15.3%)
	Do not know	39 (3.3%)
Disease Duration (min-max, mean)		0-63, 18.6

Table 7.1: Sociodemographic characteristics of pwMS from the UKMSR who participated in the online cognitive assessment at timepoint 2 and were retained for this study.

7.3.2 Four clusters were selected as the optimal number of clusters

Agglomerative hierarchical clustering was employed to group pwMS in a data-driven way based on the motor and cognitive features. The resulting dendrogram is depicted in Fig. 7.1. Looking at the dendrogram and applying the elbow method on linkage distance, two potential solutions emerged: either two or four clusters. After thorough exploration of both solutions, four clusters were deemed optimal as they provided sufficient detail without being redundant. The solution with two clusters was discarded as it merely divided pwMS into two broad groups based on overall severity, with one group having a mean webEDSS of 3.8, and the other of 5.9. Moreover, since four is also the number of conventional MS subtypes (BN, RRMS, SPMS and PPMS), selecting four clusters enabled investigating whether pwMS would be grouped similarly using a data-driven method focused solely on motor and cognitive function. Overall, the four groups comprised respectively 31.9% (red), 16.6% (green), 23.2% (orange) and 28.3% (blue) of the pwMS in the study. The solution with five clusters was also explored for completeness but discarded as it simply further divided the least populated cluster (green) based on the type of cognitive deficits, without demonstrating differences in general severity (the two resulting clusters had a webEDSS of 5.3 and 5.4, respectively).

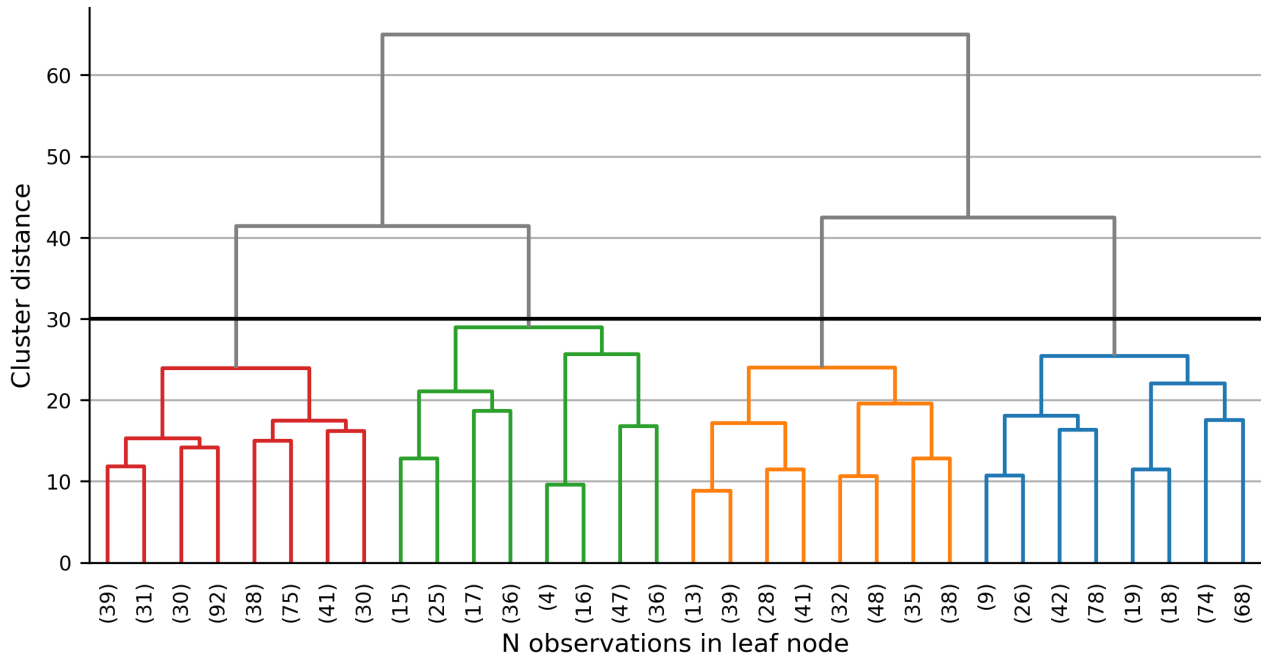


Figure 7.1: Optimal number of clusters. Agglomerative hierarchical clustering dendrogram generated using the Ward linkage method. The dendrogram provides a visual representation of how pwMS were grouped into clusters based on their similarities in the selected motor and cognitive features. Only the nodes with up to 4 merges from the final merge are shown for clarity. The x-axis represents the number of pwMS in each leaf node. The y-axis represents the distance between clusters. The vertical lines represent the fusion of clusters, with the height of each line indicating the dissimilarity at which the fusion occurred.

7.3.3 Labelling of clusters examining patterns in clustering variables

The patterns in clustering variables across groups were investigated to characterise and label them in a meaningful way. Fig. 7.2 displays median values with 95% confidence intervals across clusters

for all clustering variables (cognitive and motor features) and factors. The factors, while not used as clustering variables, were directly derived by factor-analysing the cognitive features and were here included to assist in interpreting the impaired cognitive domains across groups.

7.3.3.1 Patterns in motor function across clusters

Regarding motor function, as measured by the MSIS-29 motor and the MSWS-12, the clusters exhibited four distinct levels of impairment. The cluster encompassing 28.3% of the pwMS in the study showed no impairment (median MSIS-29 motor = 16.7 / median MSWS-12 = 14.6), the one containing 23.2% demonstrated mild impairment (median MSIS-29 motor = 38.3 / median MSWS-12 = 45.8), the one with 16.6% exhibited moderate impairment (median MSIS-29 motor = 50 / median MSWS-12 = 66.7), and the final cluster comprising 31.9% showed severe impairment (median MSIS-29 motor = 63.3 / median MSWS-12 = 100). Notably, 52.9% of the pwMS in the latter cluster were unable to walk. Cluster label demonstrated a significant association ($p < 0.0001$) with both the MSIS-29 motor and the MSWS-12. Post-hoc pairwise comparisons revealed that all clusters significantly differed from each other (all $p < 0.001$) on both variables.

7.3.3.2 Patterns in cognitive function across clusters

Regarding cognitive function, the cluster with no motor impairment showed medium to large median DfE scores on Objects Memory Delayed and Immediate, Switching Stroop, Verbal Analogies, and Trail Making, indicating problems in various cognitive domains, including memory, cognitive control, and verbal reasoning. Furthermore, 27.5% of pwMS in this cluster were classified as cognitively impaired (with 36% being the prevalence of cognitively impaired pwMS overall). This cluster was labelled as “No Motor + Mild Cognitive”.

Conversely, the cluster with mild motor impairment exhibited positive or very small median DfE scores across all tasks measuring accuracy, indicating intact cognitive function. However, this cluster showed positive but small median DfE scores across tasks measuring latency, indicating slower response times compared to controls. Overall, 3% of pwMS in this cluster were classified as impaired. This led to the cluster being labelled as “Mild Motor + No Cognitive”.

The cluster with moderate motor impairment, instead, presented large median DfE scores across all tasks, indicating widespread deficits across all measured cognitive domains. 91% of pwMS in this cluster were classified as cognitively impaired. For these reasons, this cluster was labelled as “Moderate Motor + Severe Cognitive”.

Lastly, the cluster with severe motor impairment presented medium to large median DfE scores on Objects Memory Delayed, Card Pairs, Switching Stroop, Verbal Analogies, SRT, and Trail Making, indicating problems in memory, working memory, cognitive control, verbal reasoning, and information-processing speed. The pattern and sizes of the median DfE scores were similar to the No Motor + Mild Cognitive cluster, but with additional deficits in information-processing speed. Moreover, 38% of the pwMS in this cluster were classified as impaired. Examining median scores across factors, the cluster had a lower median G factor compared to the No Motor + Mild Cognitive cluster and lower median scores on three out of the four remaining factors. Motivated by all this, this cluster was labelled as

“Severe Motor + Moderate Cognitive”.

The labelling of the clusters from the cognitive function perspective was corroborated by the median factor scores in Fig. 7.2, with the Mild Motor + No Cognitive cluster showing the highest scores and the Moderate Motor + Severe Cognitive cluster showing the lowest scores on average in the studied MS sample. The remaining two clusters were scoring in between, showing higher or lower scores compared to each other depending on the specific domain.

Overall, significant associations were found between cluster label and cognitive features and factors (all $p < 0.0001$). Post-hoc pairwise comparisons revealed that all clusters significantly differed from each other (p-values are reported in Fig. 7.3) on most variables with the exceptions of the mild and moderate cognitive impairment clusters on memory (Object Memory Delayed, and Words Memory Immediate and Delayed) and problem solving (Blocks, Verbal Analogies, and Word Definitions) tasks, and of the mild and no cognitive impairment clusters on information-processing speed tasks (SRT and Motor Control). Differences between the mild and no cognitive impairment clusters were captured in the Trail Making task, but not in the other information-processing speed tasks, highlighting that this task also provides a measure of cognitive control.

Overall, there were two clusters that exhibited only one type of deficits (either motor or cognitive impairment) and were therefore considered milder, and two clusters that manifested both types of deficits, with severity being higher in one or the other type, which were considered more severe.

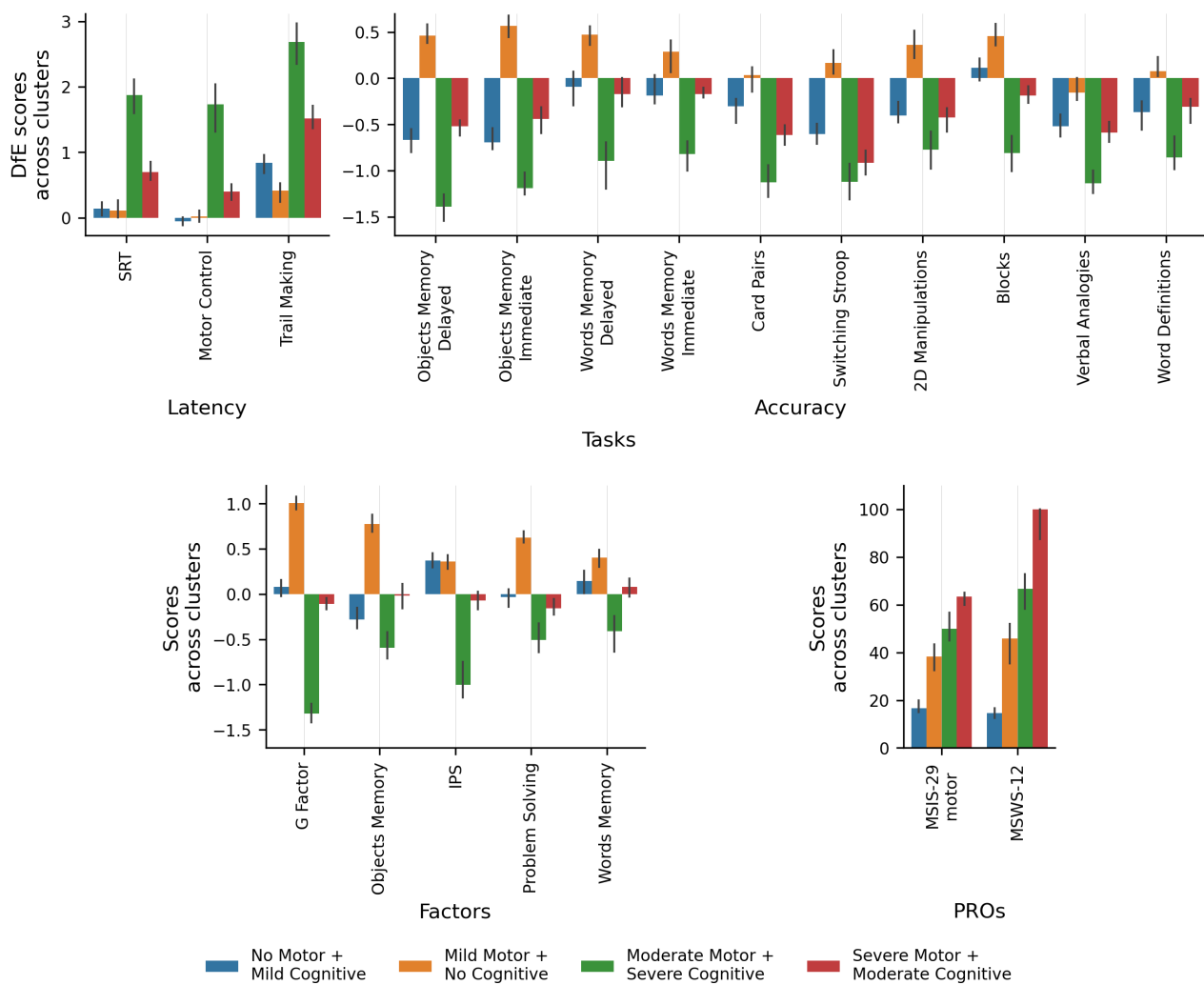


Figure 7.2: Patterns in clustering variables across clusters. Median scores with 95% confidence intervals across clusters for clustering variables (motor and cognitive features) as well as factors derived factor-analysing the cognitive features.

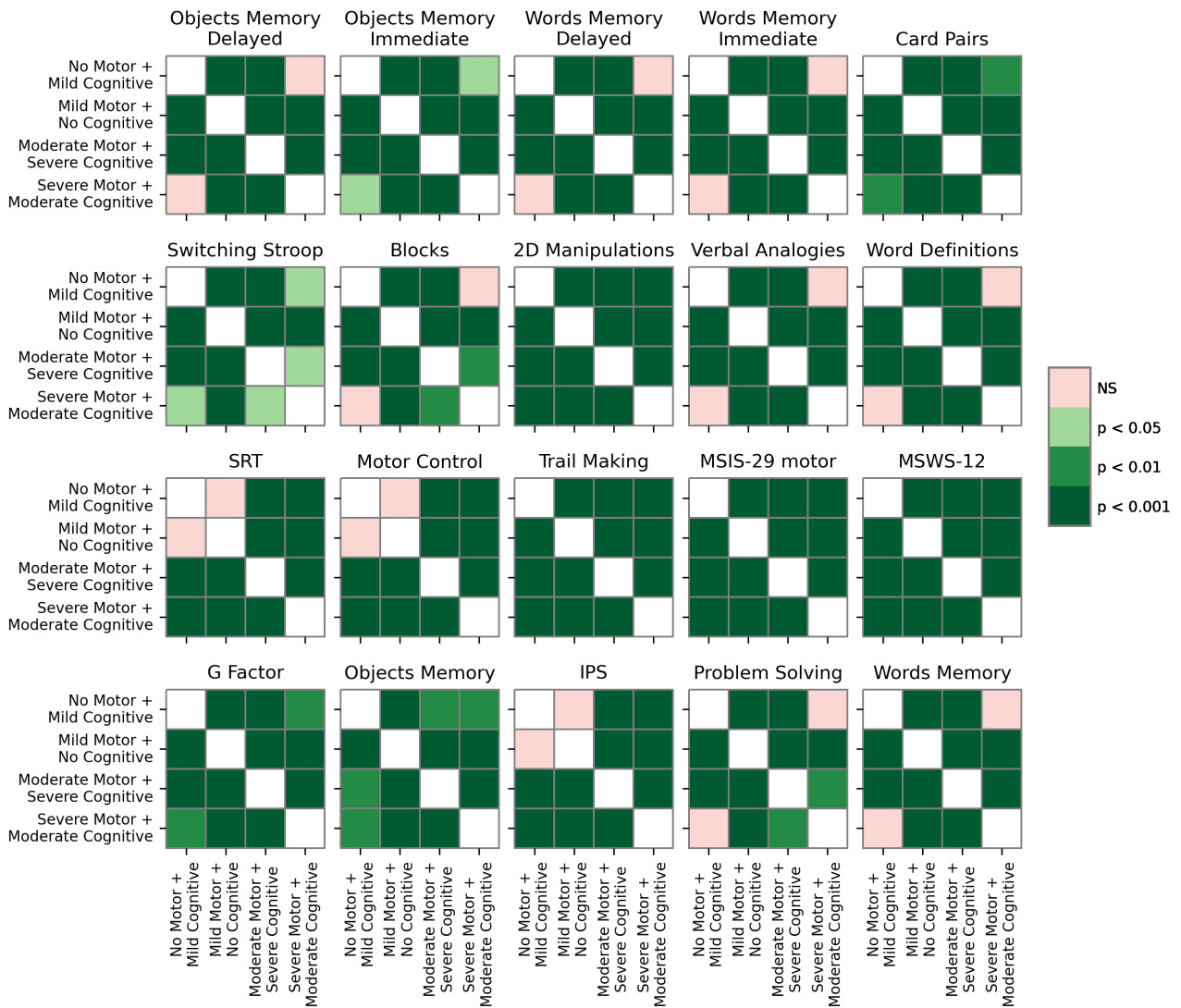


Figure 7.3: Pairwise comparisons heatmaps. P-values resulting from post-hoc pairwise comparisons of scores across clusters for motor features, cognitive features, and factors. A series of Dunn's tests was performed to obtain the p-values and the Bonferroni method was applied to correct for multiple comparisons.

7.3.4 Distribution of sociodemographics and clinical factors across clusters

To further characterise the clusters, trends in sociodemographics and clinical factors across groups were investigated. Mean age, disease duration, and age at onset across clusters are reported in Table 7.2. The distribution of gender, education levels, and subtypes across clusters is displayed in Fig. 7.4 and detailed in Table 7.2. Significant associations were found between cluster label and age, gender, education level, disease duration, and subtype. No significant associations were observed with age at onset. On average, the No Motor + Mild Cognitive cluster was the youngest, while the Severe Motor + Moderate Cognitive was the oldest. Nevertheless, the mean age difference between these clusters was only 5.6 years, and pwMS from various age groups were present in each cluster. Regarding gender distribution, it was similar in all groups except for the Moderate Motor + Severe Cognitive cluster, which had the lowest proportion of females (with a maximum difference of 14%). Notably, individuals with a degree were more prevalent in milder groups and less prevalent in both more severe groups. Concerning disease duration, it was lowest for the No Motor + Mild Cognitive cluster and highest for the more severe clusters, with a maximum difference of 5.1 years observed among cluster averages. However, individuals of any disease duration were represented in all clusters. Finally, the proportion of relapsing-remitting individuals decreased, and the proportion of progressive individuals (both SPMS and PPMS) increased in clusters with greater motor impairment. Nonetheless, individuals from any subtype were present in every cluster (Fig. 7.4).

	No Motor + Mild Cognitive	Mild Motor + No Cognitive	Moderate Motor + Severe Cognitive	Severe Motor + Moderate Cognitive	P-values from clusters comparison
Count	334	274	196	376	-
Age (min-max, mean)	19-81, 53	29-80, 55.2	19-85, 55.8	25-84, 58.6	<0.0001
Gender (% female)	81%	81%	67%	79%	0.0004
Education	preGCSE = 1% School = 40% Degree = 55% PhD = 4%	preGCSE = 2% School = 38% Degree = 55% PhD = 5%	preGCSE = 3% School = 50% Degree = 44% PhD = 3%	preGCSE = 3% School = 52% Degree = 43% PhD = 2%	0.0006
Disease Duration (min-max, mean)	0-49, 15.8	0-61, 17.7	0-62, 20.2	1-63, 20.9	<0.0001
Subtype	RR = 74% SP = 8% TrnSP = 1% PP = 7% BN = 5% DK = 4%	RR = 60% SP = 18% TrnSP = 3% PP = 14% BN = 3% DK = 2%	RR = 47% SP = 31% TrnSP = 2% PP = 15% BN = 3% DK = 4%	RR = 30% SP = 38% TrnSP = 4% PP = 24% BN = 1% DK = 3%	<0.0001
Age at onset (min-max, mean)	11-73, 37.2	13-72, 37.6	7-59, 35.6	1-70, 37.7	0.1736

Table 7.2: Sociodemographics, disease subtype, disease duration, and age at onset for pwMS across clusters. P-values were obtained using t-tests or chi-squared tests where appropriate.

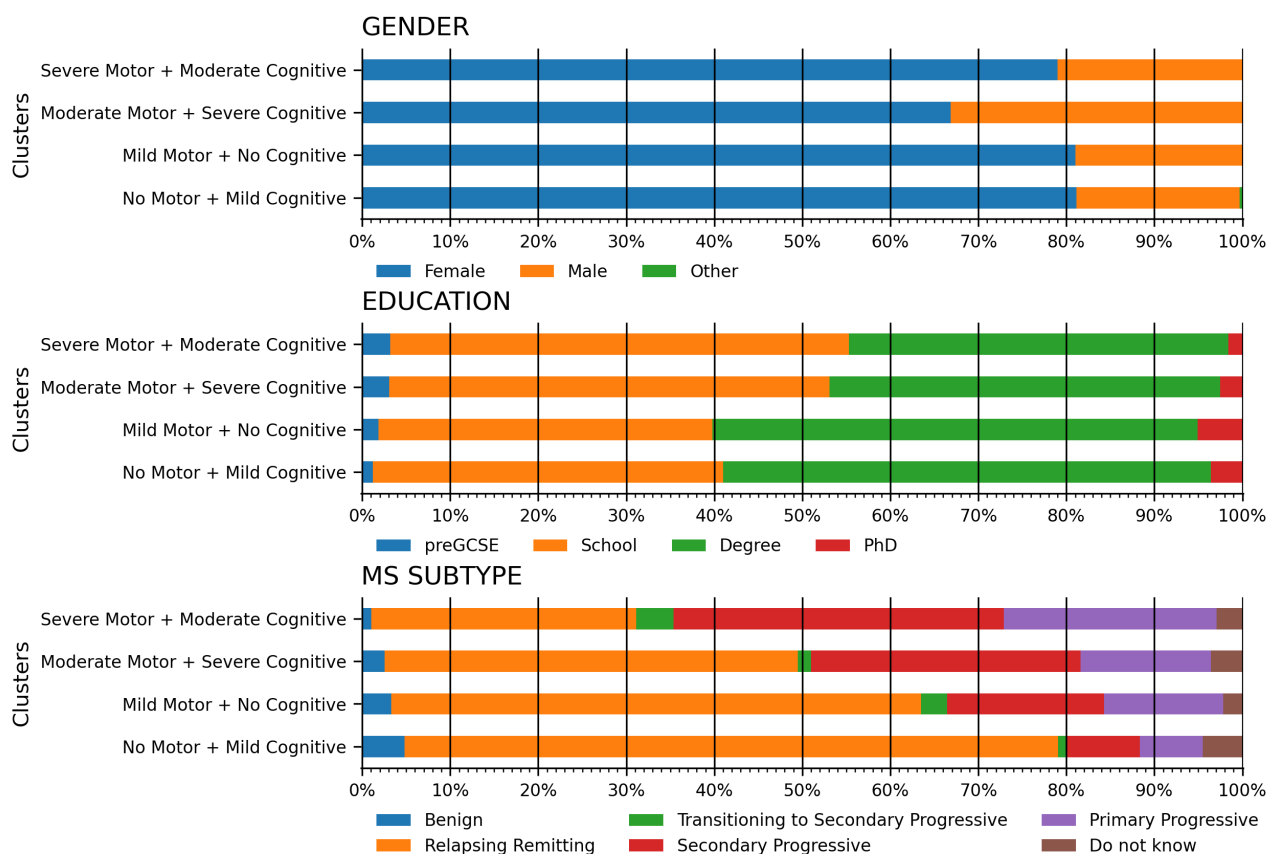


Figure 7.4: Patterns in sociodemographics and disease factors across clusters. Gender, education level, and disease subtype composition of each cluster. Bars represent the percentage of pwMS in the different categories.

7.3.5 Characterisation of clusters based on unseen clinical variables

After characterising the clusters based on the clustering variables, I examined trends in other clinical variables collected by the UKMSR through PROs. These variables included assessments of mood disturbances (HADS), quality of life (EQ5D), and fatigue (FSS). Additionally, I compared the clusters to the online version of the widely utilised clinical scale for assessing disability severity in MS, the webEDSSS. Median scores with 95% confidence intervals across clusters for all clinical variables are displayed in Fig. 7.5A.

7.3.5.1 Patterns in mental health across clusters

Regarding anxiety and depression as measured by the HADS, I examined both the summary scores as well as the prevalence of pwMS with normal, borderline abnormal, and abnormal mental health across clusters (Fig. 7.5B). Cluster label demonstrated a significant association ($p < 0.0001$) with both anxiety and depression. Median depression and anxiety scores were higher in clusters with greater motor impairment, with this pattern being more pronounced for depression. Post-hoc pairwise comparisons revealed significant differences in anxiety between the clusters No Motor + Mild Cognitive and Mild Motor + No Cognitive with the cluster Severe Motor + Moderate Cognitive. Significant differences in depression were observed between all pairs of clusters except between the Moderate

Motor + Severe Cognitive and the Severe Motor + Moderate Cognitive (Fig. 7.5C).

Analogously, the prevalence of pwMS with abnormal or borderline abnormal depression and anxiety was higher in clusters with more severe motor impairment (Fig. 7.5B). This finding suggested that mental health was more closely related to motor than to cognitive impairment. This pattern was confirmed by examining the overall correlation between the depression and anxiety summary scores with both the MSIS-29 motor and the G factor, representing general motor and cognitive impairment, respectively (Fig. 7.6). Following Cohen's notation of effect size, the correlation with the MSIS-29 motor was large for depression and medium for anxiety, with values of 0.64 and 0.32, respectively. The correlation with G, instead, was small for both depression and anxiety, with values of -0.17 and -0.11, respectively.

7.3.5.2 Patterns in self-reported fatigue across clusters

Regarding fatigue, significant associations ($p < 0.0001$) were found between cluster label and the FSS summary score. The median score was higher in clusters with greater motor impairment, indicative of heightened reported fatigue. Post-hoc pairwise comparisons revealed significant differences between the No Motor + Mild Cognitive cluster and all the other clusters, as well as between the Mild Motor + No Cognitive and the Severe Motor + Moderate Cognitive clusters (Fig. 7.5C).

7.3.5.3 Patterns in quality of life and disability severity across clusters

Finally, cluster label was also associated with quality of life ($p < 0.0001$), as measured by the EQ5D, and with general disability severity ($p < 0.0001$), as measured by the webEDSS, with all clusters being statistically distinct from each other (Fig. 7.5C). Examining median scores across clusters in Fig. 7.5A, clusters with greater motor impairment had lower quality of life and higher disability severity. In particular, the No Motor + Mild Cognitive cluster had a median webEDSS of 3, corresponding to a low disability severity, the Mild Motor + No Cognitive cluster had a median webEDSS of 4.5, corresponding to a medium disability severity, and the Moderate Motor + Severe Cognitive and the Severe Motor + Moderate Cognitive clusters had a median webEDSS of 6 and 6.5, respectively, corresponding to a high disability severity. Quality of life and webEDSS were more closely related to motor than to cognitive impairment, as it can be confirmed by the large correlations with the MSIS-29 motor (EQ5D = 0.84 / webEDSS = 0.78) and the small correlations with the G factor (EQ5D = -0.21 / webEDSS = -0.22). A complete record of the correlations between the G factor, the MSIS-29 motor, the MSWS-12, and the other clinical variables is displayed in Fig. 7.6.

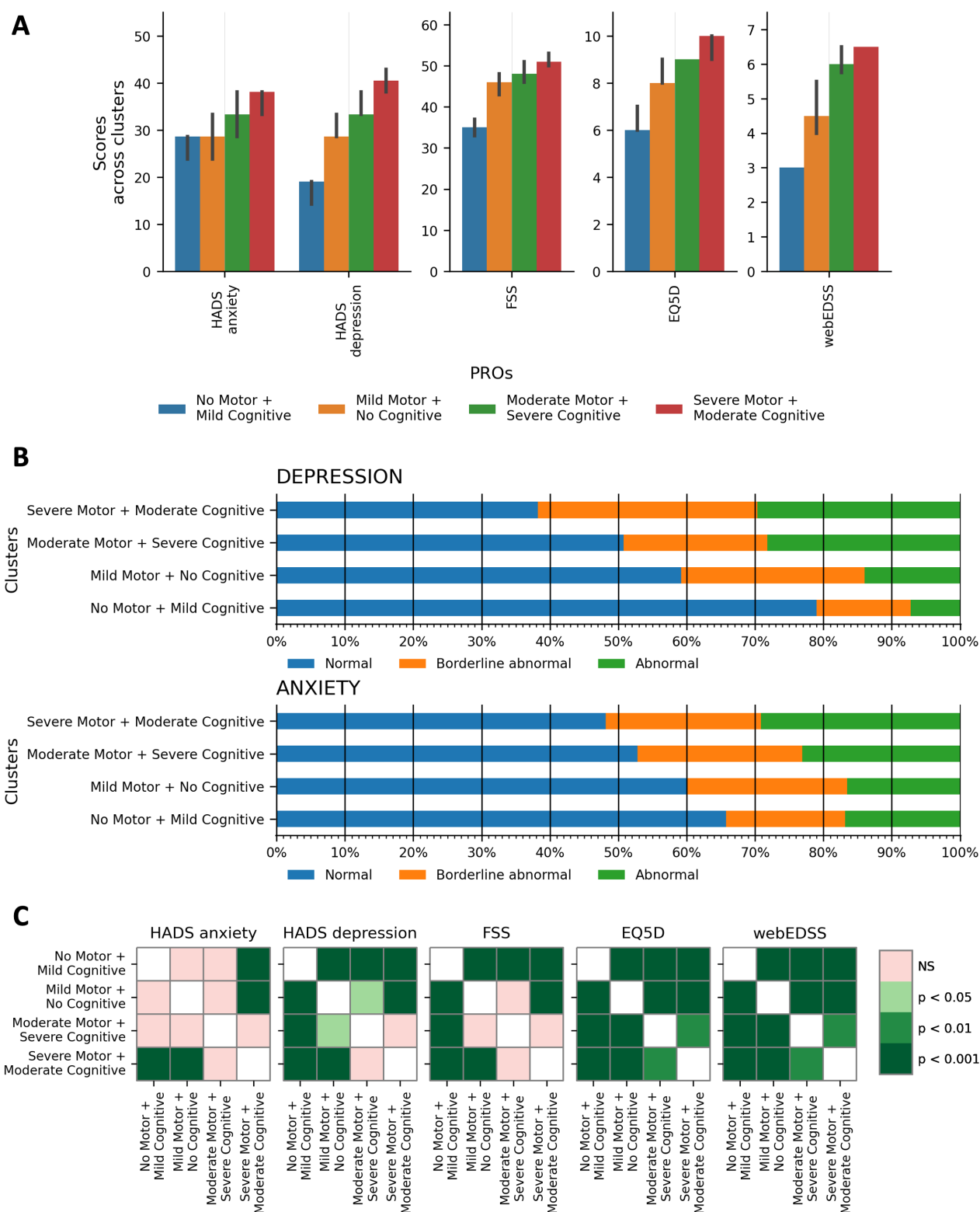


Figure 7.5: Patterns in clinical variables across clusters. (A) Median scores with 95% confidence intervals across clusters for clinical variables. (B) Prevalence of mental health problems in each cluster. Bars represent the percentage of pwMS in the different categories. (C) P-values resulting from post-hoc pairwise comparisons of scores across clusters for clinical variables. A series of Dunn's tests was performed to obtain the p-values and the Bonferroni method was applied to correct for multiple comparisons.

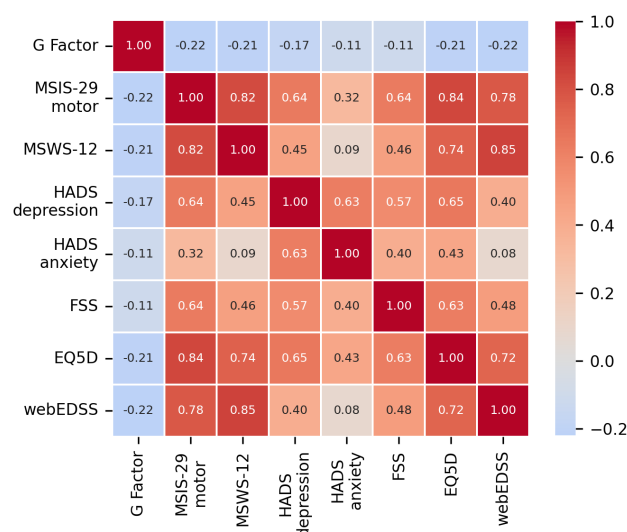


Figure 7.6: Pairwise Pearson’s correlations among the G factor, the MSIS-29 motor, the MSWS-12, the HADS-depression, the HADS-anxiety, the FSS, the EQ5D, and the webEDSS.

7.4 Discussion

7.4.1 Summary of findings

The results presented in this chapter demonstrate that pwMS, when clustered based on their manifestation of motor and cognitive symptoms, form heterogeneous groupings that did not align with conventional clinical phenotypes. Consistent with previous studies, motor and cognitive impairment occurred independently from one another, yielding clusters exhibiting diverse combinations and severity of these impairments. Despite significant variations in age and disease duration across groups, individuals of any age and disease duration were included in all identified clusters. Moreover, PROs measuring depression, fatigue, quality of life, and overall disease severity were more closely related to motor rather than cognitive function across the clusters.

Altogether, the results provided evidence that employing digital markers of impairment for clustering pwMS offered an alternative perspective on the disease, diverging from conventional clinical classifications.

7.4.2 Independence of motor and cognitive impairment

Using agglomerative hierarchical clustering, four clusters were identified as the optimal number of clusters to group pwMS based on their motor and cognitive function. The patterns of clustering variables across clusters revealed a notable dissociation between motor and cognitive impairment: there were patients with more severe cognitive deficits showing better motor function and vice versa. The clusters were heterogeneous in the combinations and severity of motor and cognitive deficits. This is unsurprising considering that the combination of symptoms patients experience arises from their affected brain areas, and each patient has a unique pattern of lesions and brain damage.

7.4.3 Insights in the No Motor + Mild Cognitive cluster

The most interesting insights were related to the No Motor + Mild Cognitive cluster which was the second largest cluster and represented pwMS with intact motor function but widespread cognitive deficits. This cluster, on average, encompassed younger patients with shorter disease duration. The formation of this cluster aligns with findings from the literature indicating that cognitive impairment can be present early on and in absence of other deficits (Achiron et al. 2006; Amato et al. 2010; Glanz et al. 2007; Sumowski et al. 2018). This highlights the importance of baseline cognitive assessment followed by regular monitoring throughout the disease to prevent cognitive deficits from progressing unnoticed. In absence of periodic assessment of cognition, there is the risk of underestimating patients' disease severity and potentially misclassify them as Benign. Indeed, a review from Ton et al. 2017 has shown that up to 47% of individuals categorised as Benign exhibit cognitive deficits, whereas a study from Saccà et al. 2017 has demonstrated that the EDSS can be miscalculated, leading to an underestimation of disability, when cognitive status is not properly measured. This is also reflected in this cluster's characteristics: it's the one with the highest prevalence of Benign cases and with the lowest webEDSS on average. Monitoring people in this cluster just from the motor perspective, would mistakenly consider them free from disability.

On another note, the cognitive deficits in this cluster encompassed domains beyond information-processing speed, including memory, cognitive control, and verbal reasoning. While deficits in information-processing speed are the most commonly reported (Costa et al. 2017; DeLuca et al. 2004; Forn et al. 2008) and prioritised during brief assessments (Langdon et al. 2012), focusing solely on this domain might lead to overlooking cognitive impairment in this cluster.

7.4.4 Insights in the other clusters

Additional interesting insights were related to the Severe Motor + Moderate Cognitive cluster. This was the largest cluster and represented pwMS with severe motor deficits alongside moderate cognitive impairment. On average, this group comprised older individuals and exhibited the highest prevalence of progressive subtypes. Additionally, it recorded the highest average MSWS-12 score and encompassed 60% of the individuals reporting an inability to walk. Furthermore, The Severe Motor + Moderate Cognitive cluster exhibited similar cognitive deficits to the No Motor + Mild Cognitive cluster, with the addition of information-processing problems.

Regarding the composition of the Moderate Motor + Severe Cognitive cluster, a higher prevalence of males was observed, resulting in a distinctive male-to-female ratio in comparison with the other clusters. This difference could not be attributed to an increased prevalence of primary progressive individuals within this cluster, as this prevalence was similar in the Mild Motor + No Cognitive cluster and lower than in the Severe Motor + Moderate Cognitive cluster. Therefore, the distinctive male-to-female ratio indicated a potential gender-related susceptibility to severe cognitive impairment, in alignment with other studies from the literature (Beatty et al. 2002; Saccà et al. 2017; Savettieri et al. 2004).

7.4.5 Do clusters represent different stages of MS?

When interpreting the clusters, the presence of significant differences in age, disease duration and disease subtype across clusters suggested that they could potentially represent different stages of MS. The milder clusters, namely the No Motor + Mild Cognitive and the Mild Motor + No Cognitive clusters, predominantly consisted of individuals who were younger or had shorter disease duration on average. These clusters also exhibited a higher prevalence of relapsing-remitting individuals. It was conceivable that these clusters represented earlier stages of MS, with individuals possibly transitioning into the other more severe clusters with disease progression. For example, individuals in the No Motor + Mild Cognitive cluster might evolve into the Moderate Motor + Severe Cognitive cluster. However, the age and disease duration ranges were spread out across all clusters and there were no obvious patterns of evolution from one cluster to another. Despite the potential for transitions between clusters, it's evident that at least a subset of individuals likely remain in the same cluster throughout their disease course. This is supported by the observation of individuals with disease durations ranging from less than 5 to more than 40 years within each cluster.

Finally, no significant differences based on age at onset were observed across clusters, indicating that the severity and type of symptoms experienced by patients are not associated with the age at which the disease manifested.

7.4.6 Clinical variables are associated with motor function

Across the clusters, patterns relating to depression, fatigue, quality of life, and disease severity revealed a closer association with motor rather than cognitive function. In fact, more pronounced motor impairment was associated with compromised mental health, heightened fatigue, diminished quality of life, and increased disability severity. These associations were evident in both the median values across clusters and the inter-correlation among variables.

While these associations might be influenced by various factors, they likely result from a combination of them. One plausible explanation is that the cognitive features represent more objective measures, collected through a standardised assessment, whereas the clinical variables and motor features rely on self-reported measures reflecting the patient-perspective on the disease. Another possibility is that the included clinical variables predominantly focus on physical disability and lack sensitivity to cognitive deficits, supported by evidence from the literature indicating that self-reports of fatigue don't show correlation with cognitive function while objective measures do (Johnen et al. 2017; Krupp et al. 2000; Schwid et al. 2003). Additionally, an examination of EQ5D items explicitly revealed a focus on mobility rather than cognition.

The rationale for emphasising motor over cognitive disability remains unclear. It is uncertain whether this emphasis stems from the greater debilitating impact of motor impairment on individuals' lives or is influenced by the inherent challenges in capturing cognitive deficits. Despite this focus on motor impairment, numerous studies underscore the debilitating nature of cognitive disability and its substantial impact on patients' quality of life, which often surpasses the impact of physical disability (Cutajar et al. 2000; Rao et al. 1991a).

Cumulatively, findings from the literature and the results presented in this chapter indicate that

cognition is inadequately represented in both clinical and patient-reported scales. Consequently, cognitive data provide additional, independent information, offering a supplementary dimension to the comprehension of the well-being of individuals with MS.

7.4.7 Limitations and future directions

One potential limitation of this study was the possibility that the observed differences in cognitive function across clusters could be attributed to normal variability in cognition within the general population. However, this scenario was ruled out by examining the distribution of education levels across clusters. Despite there were significant differences, the clusters exhibited similar prevalence of education levels in pairs while having different cognitive function and overall prevalence of cognitive impairment. Moreover, education was accounted for upstream during the evaluation of Deviation from Expected scores, which served as input to the model. The differences in education levels observed between cluster pairs instead could be explained by clusters with higher disease severity (higher webEDSS) exhibiting lower education levels. MS affects individuals early on, impacting the years when higher education is attained, with more aggressive MS and paediatric-onset MS potentially impeding such pursuits (Vargas et al. 2020). Moreover, it has been shown that higher-education levels protect from cognitive decline, at least in the early-stages (Rimkus et al. 2018). Therefore, clusters predominantly capture differences in cognitive impairment rather than individual differences or normal cognitive variability in the population.

In subsequent investigations, it would be interesting to relate the identified clusters to other more objective disease biomarkers. Comparison with imaging biomarkers would enable the exploration of specific patterns of brain lesions and damage underlying each cluster. Associations with clinical EDSS would corroborate what was observed for the webEDSS more objectively. Additionally, exploring connections with relapse occurrences and treatment history could unveil successful treatment strategies, particularly in individuals within the milder groups who have endured MS for longer disease durations. Finally, conducting follow-up assessments over time would provide insights into the stability of cluster assignment, the occurrence of transitions between clusters, and the identification of pairs of clusters among which such transitions are more likely.

7.4.8 Conclusions

Overall, this study enabled clustering pwMS solely based on digital markers of disability, shifting the focus from disease mechanisms of progression to symptoms experienced by the patients. Four distinct clusters emerged, displaying heterogeneous combinations and severity of motor and cognitive impairment. While this classification does not claim in any way to be a replacement for conventional clinical phenotypes, it represents an interesting alternative and offers the opportunity to look at MS from a different perspective.

Chapter 8

Discussion

In this chapter, I summarise and interpret the main findings of my thesis. I discuss the limitations and potential impact of my work and suggest possible future directions for research in this field.

8.1 Summary of findings

In this thesis, I aimed to develop and investigate the validity and utility of digital markers of disease-related disability in Multiple Sclerosis (MS) focusing on motor and cognitive impairment. The Multiple Sclerosis Impact Scale (MSIS-29) motor component and the Multiple Sclerosis Walking Scale (MSWS-12) were investigated as digital markers of motor impairment. Performance scores resulting from online assessment of cognition were investigated as digital markers of cognitive impairment.

The main aims of this thesis were:

1. To provide further evidence of the validity of the MSIS-29 motor and the MSWS-12 in measuring motor impairment in people with MS (pwMS) as well as investigate their utility in offering novel insights in the developmental time course of the disease.
2. To assess the feasibility and accessibility of online cognitive assessment in pwMS.
3. To identify an optimal battery of online cognitive tasks to monitor cognitive function in pwMS.
4. To validate the optimal battery of online cognitive tasks.
5. To showcase the utility of the cognitive data collectable using the optimal battery.
6. To derive symptoms-driven MS subtypes solely based on digital markers of motor and cognitive impairment.

8.1.1 Chapter 4

In Chapter 4, I further demonstrated the validity and utility of two patient-reported outcomes (PROs), the MSIS-29 motor and the MSWS-12, as measures of physical disability in MS. This retrospective and prospective cohort study over 11 years involved 15,976 pwMS from the UK MS Register. I found a robust main effect of disease subtype on the PROs and interactions of disease subtype with duration. Specifically, PROs worsen with disease duration for all subtypes apart from BN. Furthermore, the PROs of each subtype are statistically different from those of the other subtypes at all time bins except when comparing RRMS with BN and PPMS with SPMS. Notably, the patient-reported outcomes are sensitive to future transitions to progressive subtypes, with individuals who transitioned in the follow-up presenting with higher PROs in their relapsing-remitting phase compared to individuals who did not transition since onset. PROs capture the different patterns of physical worsening over disease duration and across subtypes; therefore, they are a valid tool to measure the physical impact of MS over the long term and cost-effectively.

8.1.2 Chapter 5

In Chapter 5, I demonstrated the feasibility and accessibility of online cognitive assessment in pwMS and curated a battery of tasks optimal to measure cognitive deficits in this population. Employing a random-sampling approach, I evaluated a set of 23 tasks spanning diverse cognitive domains on a sample of pwMS from the UK MS Register without inducing fatigue. High participation and completion rates across various age groups, disease durations, and subtypes underscored the tool's

feasibility and accessibility to the broader MS population. Comparing task performance with controls matched by sociodemographic factors revealed cognitive deficits across multiple domains and enabled the evaluation of task-specific sensitivity to MS. The identified impaired domains aligned with existing literature. A subset of tasks was selected to form a concise battery optimal to measure MS-related cognitive impairment while being comprehensive, non-redundant, engaging, and efficiently deployable remotely on any device.

8.1.3 Chapter 6

In Chapter 6, I validated the optimal battery derived in Chapter 5 and demonstrated its utility by testing a partially overlapping sample of pwMS from the UK MS Register. Two additional memory tasks were included in the battery for evaluation. The high participation, completion, and return rates confirmed the accessibility and appeal of the tool to the broader MS population.

A replication analysis confirmed the sensitivities to MS across tasks evaluated in Chapter 5 using an independent sample of pwMS, thereby validating the selection of tasks for the optimal battery. The two additional memory tasks demonstrated higher sensitivity to MS compared to the memory tasks already included in the battery and were consequently permanently integrated. The optimal battery enabled the collection of complete and comprehensive cognitive data on a large and representative sample of pwMS. This dataset proved valuable in confirming associations observed in existing literature with greater statistical power. On average, pwMS classified as impaired were older, had longer disease duration, and had a higher prevalence of males and of progressive subtypes compared to pwMS classified as non-impaired. However, cognitive deficits were observed across all subtypes and disease durations, including in the early stages. General performance and information-processing speed showed associations with the disease subtype. No tasks or domains showed significant associations or interactions with disease duration when considering subtype. While these findings align with some evidence in the literature, it's important to acknowledge the limitations of evaluating longitudinal trends in a cross-sectional study design. Longitudinal studies with longer follow-up periods are needed to conclusively negate associations between disease duration and cognitive impairment. Nonetheless, integrating the optimal battery into the MS Register website for periodic monitoring would greatly facilitate the collection of large-scale longitudinal data suitable to answer these types of questions.

8.1.4 Chapter 7

In Chapter 7, I explored an alternative approach to classify pwMS solely based on digital markers of motor and cognitive impairment. The PROs (the MSIS-29 motor and the MSWS-12) validated in Chapter 4 were used as makers of motor impairment. Performance scores derived from cognitive assessment using the optimal battery presented in Chapter 6 were employed as markers of cognitive impairment. PwMS who completed both PROs and the optimal battery during the Autumn-Winter 2022 were included in this study.

Four clusters emerged from agglomerative hierarchical clustering, presenting different combinations and severities of motor and cognitive deficits. This suggests that these impairments can occur independently of each other, consistent with evidence from the literature indicating that cognitive impairment can appear early on and in the absence of other deficits. Notably, one subtype had no impact on

motor function but mild cognitive impairment, highlighting that neglecting assessment of cognition may lead to underestimating disease severity and misclassifying individuals as Benign.

While associations were found between cluster labels and age, disease duration, and subtypes, each cluster contained pwMS with diverse characteristics. Although this did not exclude the possibility of transitions between clusters, it indicated that some individuals likely remain in the same cluster through their disease. Finally, deteriorated mental health, heightened fatigue, increased quality of life, and higher disease severity were more closely related to worsened motor rather than cognitive function. Altogether, these findings demonstrate that alternative approaches for grouping pwMS based on digital markers of disability can offer a novel perspective on the disease.

8.2 Interpretation

The advent of online registries in MS has created opportunities for remote patient monitoring, facilitating the collection of extensive data over longer follow-up periods for research purposes. However, the development and rigorous validation of online assessment tools are crucial for realising registers' full potential.

The UK MS Register (UKMSR) stands out as a successful registry initiative, establishing a community of over 19,000 pwMS and routinely collecting patient-reported outcome (PRO) data from them since 2011 (Middleton et al. [2021](#), [2018](#); Rodgers et al. [2022](#)). This wealth of real-world data holds promises for addressing questions related to long-term disability, a perspective often absent in randomised control trials. Nonetheless, it comes with complexities, arising from the ongoing challenge of engaging participants repeatedly over time, as well as the fact that participants join the Register at different points in their disease, resulting in a sparse dataset. Additionally, PROs are sometimes criticised for their subjective nature, given their reliance on patients' perspectives of their disease (McGinley et al. [2022](#)).

In this thesis, I built upon the work initiated by the UKMSR, taking it a step further. The main overarching theme was to prove the feasibility and validity of online assessment tools for disability in MS and demonstrate their potential for making valuable contributions to both clinical decision-making and research advancement. To achieve this, on one hand, I further validated and demonstrated the utility of tools already integrated into the Register, namely two motor-related PROs. On the other hand, I have developed and validated a new online tool to incorporate into the Register to objectively measure cognition, broadening the type of data that can be collected on the platform to include something that is not typically routinely monitored in MS.

Regarding the PROs, their feasibility in an online register context has already been demonstrated by the scale of the data collected by the UKMSR over the years (Knowles et al. [2023](#); Middleton et al. [2018](#); Nicholas et al. [2020](#)). However, my main contribution was to develop an approach to fully leverage these registry-type data and derive robust statistical insights from it. This involved addressing missing data and explicitly accounting for the level of engagement as a potential confounding factor. Using this approach, I further demonstrated the validity of two of the PROs collected by the Register as measures of motor impairment and showed their sensitivity to future transitions to progressive subtypes. The developed approach can be extended to other PROs collected by the Register and to other registry-type data collected by any patient registry.

Regarding the online cognitive assessment, unlike for the PROs, I had to start from the ground up and first demonstrate its feasibility and accessibility to pwMS. This yielded very positive results, with over 4,000 pwMS participating cumulatively in the two conducted studies, including individuals from any age, gender, disease duration, subtype, and severity. Additionally, the high completion and return rates further confirmed the accessibility of the assessment and the strong interest demonstrated by pwMS in having such a tool integrated into the UKMSR platform.

The two studies on cognition conducted via the Register using Cognitron and presented in this thesis were among the largest studies ever done on the topic in MS (Benedict et al. 2006; Rao et al. 1991b). The collected data allowed identifying cognitive domains affected in MS, estimating prevalence rates of cognitive impairment overall and across domains, and evaluating associations with disease-related factors and PROs. These are all elements that have been investigated in the literature (Benedict et al. 2020; Chiaravalloti et al. 2008; Langdon 2011), whose trends were generally confirmed by my findings with greater statistical power and more representativeness in the studied MS sample. In turn, the alignment of my findings with the literature served as validation of the assessment and as confirmation that it was sensitively measuring cognitive impairment in pwMS.

Most importantly, the findings from the two studies enabled forming an optimal and concise battery of tasks for online monitoring of cognitive function in MS. This battery is comprehensive and non-redundant but specifically tailored to measure impairment in MS. Additionally, it is efficient and can be deployed without supervision and from any device. These characteristics overcome disadvantages encountered in both traditional neuropsychological batteries (Sumowski et al. 2018) as well as other online assessments (Wojcik et al. 2019) utilised in the MS literature. For example, traditional neuropsychological batteries used in MS, such as MACFIMS (Benedict et al. 2006) and BICAMS (Langdon et al. 2012; Potticary et al. 2023), are either comprehensive but impractical for widespread use, or non-comprehensive but more efficient. However, in all cases, they require patients to come in person, leading to significant costs in terms of time and resources. Other online assessments are not specifically tailored to MS; instead, they have been developed for the wider neurological community and have often been validated only for supervised administration and for use on specific devices, greatly restricting their accessibility (Bergmann et al. 2023; De Meijer et al. 2018; Edgar et al. 2011). The optimal battery developed in this thesis could be permanently integrated into the Register platform, allowing for routine monitoring of cognition alongside the other disease features tracked by the available PROs.

A final element of this thesis involved demonstrating how the use of these online tools can lead to novel insights into the disease, complementing the information collectable through traditional assessments. While there were elements of this throughout the thesis, it is in Chapter 7 that this element is at the centre. In fact, in Chapter 7, digital markers of motor and cognitive impairment derived using online tools allowed the identification of alternative data-driven subtypes, focused on symptoms rather than disease mechanisms. While the relation between these subtypes and underlying pathology is yet to be demonstrated, as is their utility for targeting treatment, the subtypes, in their characteristics, were already rich in insights. Particularly interesting was the emergence of a purely cognitive subtype, including individuals of any age and disease duration, which, given its low average webEDSS, potentially included individuals whose disability is overlooked (Saccà et al. 2017; Ton et al. 2017). The findings in Chapter 7 showed how such a way to cluster pwMS could be insightful while being practical, as solely based on markers collected through online assessments and therefore widely accessible to everyone with

MS. While this classification does not claim to be final nor a substitution to the traditional clinical phenotypes, it is an interesting alternative that can give a different and complementary perspective on the disease.

Overall, combining validated online tools with the wide reach of the UKMSR enables the collection of large-scale data on representative samples of pwMS in the UK, holding incredible potential to address some of the open questions in the field. Additionally, online tools promote inclusivity, as anyone in possession of a device can take part, enabling the formation of more representative samples of pwMS and avoiding the problem of recruiting individuals solely based on clinic visits and treatment status (Potticary et al. 2023).

This thesis aimed to offer an example of the potential and the type of achievements possible in this direction. The tools presented are not intended to fully replace traditional processes, but rather to complement and facilitate them. These tools are thought to support and streamline the monitoring of pwMS, providing information to clinicians about changes over time and specific aspects that are changing. They can aid in identifying individuals who may need further in-person assessments and those who may not. Nonetheless, much work is still needed to fully leverage and exploit the available data, as well as to develop and validate new and existing digital markers of disability. However, it is indeed a promising direction to pursue, especially as more of these tools become shared and standardised across MS registries in Europe and worldwide, enabling data pooling and comparison of MS across countries (Fink 2016; Flachenecker et al. 2014).

8.3 Limitations

One potential future limitation concerns the cost of such tools and whether it will hinder their widespread use. For instance, the elevated cost of previously developed online cognitive assessments has been a barrier to their utilisation. However, this has not been the case for Cognitron, which has so far enabled the affordable collection of data from hundreds of thousands of participants in various studies (Brooker et al. 2020; Del Giovane et al. 2023b; Hampshire et al. 2020, 2021b), with only maintenance costs being charged. Ensuring the ongoing affordability and accessibility of these tools is crucial to make a widespread impact and prevent sociodemographic inequalities.

Another potential limitation involves the increasing computational skills required to manage, store, analyse, and extract insights from these increasingly complex data. As datasets grow in size, it becomes impractical to manually examine them. Things are further complicated in registry-like contexts where missing data are inherent. Additionally, online cognitive assessments enable collecting detailed trial-by-trial performance data, offering potential richness of information but introducing additional complexity. In fact, the analysis requires extracting and comparing multiple performance metrics to identify the most sensitive ones. Additionally, the option to leverage the full trial-by-trial data using advanced machine learning models adds another layer of intricacy.

However, with the collection of big data becoming prevalent in various aspects of our lives and the continuous advancements in the field of data science, with more individuals specialising in this area, adapting to this changing landscape is crucial. It will be important to embrace these advancements and incorporate the expertise of individuals with knowledge in these areas into the field of MS.

A final limiting factor pertains to the availability of online devices in the MS population. In fact, online assessments are available only to individuals who have access to an online device and are effective only for those with at least some degree of computer literacy. However, online assessments can also be completed during clinical visits, and access to online devices and computer literacy are likely to become less of a problem going forward. This may not be the case in poorer countries, posing the risk of widening the gap between health systems. Nonetheless, affordable and efficient online technologies could potentially help bridge this gap and provide a practical means to monitor disability in poor countries.

8.4 Future directions

There are several potential future directions for the work presented in this thesis, some of which are already in progress. Here, I am going to mention just the ones that are most interesting from my perspective or essential for the future widespread use of the presented online tools.

An essential future step involves completing the validation of the optimal battery by assessing all its psychometric properties. While the available data enabled determining task sensitivity to MS, ensuring the discriminant validity of the battery, other psychometric properties, such as test-retest reliability and concurrent validity, remain to be evaluated. With this in mind, an ongoing study set up in collaboration with the UKMSR is testing the comparative effectiveness of tasks in the optimal battery with analogous neuropsychological tests used in traditional assessments for MS (e.g., tests from BRB-N (Boringa et al. 2001), MACFIMS (Benedict et al. 2006) and BICAMS (Langdon et al. 2012)). PwMS are participating in person, performing both the optimal battery and a traditional cognitive assessment, with the order of administration being varied across participants. Furthermore, this study provides data on people performing the optimal battery in the presence of a supervisor, enabling the evaluation of differences between remote and supervised administration. Additionally, another study is being set up to test a sample of pwMS repeatedly over a short period of time to estimate test-retest reliability as well as the potential presence of learning effects.

From the analysis perspective, the "Deviation From Expected" method was utilised to account for normal variations in cognitive performance in the population with sociodemographic variables. While this method enabled accounting for factors such as age, gender and education when evaluating statistical differences between patients and controls, these factors were all considered independently within the regression models. However, some of these factors can covary and interact with each other and with other disease-related factors. For example, women after a certain age reach menopause, and this has been shown to affect cognitive performance (Conde et al. 2021). Another good example of this is age which covaries with disease duration. Future work should focus on including and estimating interactions between sociodemographic and disease-related factors in the analysis using more sophisticated models. This would enable better disentangling of the contribution of different factors to the observed deficits. A first step in this direction was already done in the context of the motor PROs. Preliminary results not presented in this thesis have shown that the correlation between age and MSIS-29 motor/MSWS-12 is different if evaluated at different disease durations (Lerede et al. 2022). In particular, the correlation appears to be higher in the earlier stages of MS and closer to zero in the later stages. This suggests that age may modulate the impact of physical disability differently

at different disease durations and that physical disability may tend to stabilise over time. While these findings are preliminary and further work is necessary to confirm them, they highlight the importance of accounting for interactions both in the context of PROs and cognitive performance measures.

Another interesting future direction, more related to the longer term, entails the periodic collection of cognitive data on the UKMSR, potentially as part of their regular collection windows. In this sense, the goal would be to administer the same battery as in Chapter 6. This battery, which consists of 14 tasks and can be completed in under 1 hour, comprehensively and sensitively measures different aspects of cognition. Given that no significant changes in cognitive function were observed over a one-year period (Section 6.3.5.1), the current recommendation is to administer this battery at intervals of at least one year. However, this timeframe should be adjusted based on future findings as new timepoints are collected and additional longitudinal data become available. A flexible approach would be to adapt the monitoring interval based on assessment results: if no changes are detected for a user, a longer interval (e.g. two years) could be used; instead, if changes are detected for a user, a shorter interval (e.g. six months) should be recommended, possibly focusing on the tasks where the performance changes were detected. This approach would enable the examination of longitudinal changes in cognition over an extended period of time and the evaluation of their associations with changes in other aspects of the disease (e.g., physical disability and mental health) as well as treatment and relapses. Additionally, through optimisation procedures, an even shorter battery could be developed for scenarios requiring quicker assessments. Based on current results, promising tasks for such a battery include Switching Stroop, Spotter, and Trail Making tasks. These tasks have demonstrated high sensitivity to MS and provide additional metrics through contrasts, enabling the measurement of multiple cognitive domains from a single task. This shorter battery could be useful to perform population screening for MS.

In the shorter term, instead, it would be intriguing to explore the relationship between the cognitive performance metrics collected in this thesis and imaging biomarkers. From the findings presented in Chapter 5 and 6, it is clear how cognitive deficits are present already early on and involve multiple cognitive domains. However, these deficits depend on the disruption of functional brain processes, which can have a basis in a number of different aspects of brain pathology, e.g., disrupted brain structural connectivity, brain atrophy, ongoing brain inflammatory processes, or changes to brain metabolism, especially of the cortical and subcortical regions associated with cognition. The extent to which the measured deficits depend on all these different aspects remains unclear. It is likely a combination of all. There may be differential associations between types and locations of brain pathology and specific cognitive tasks - but the exact mappings of these remain to be determined at present. Relatedly, the fact that progressive subtypes show more severe information processing deficits (Section 6.3.4.6) and have usually higher and more extensive brain atrophy, suggests that these deficits may be more closely related to this aspect. In contrast, early deficits which are present across all subtypes may be more related to inflammatory or metabolic changes as well as individual variations in the brain's ability to compensate for these damages through neural plasticity and reorganisation, referred to as "cognitive reserve" (Sumowski et al. 2013). To address this, a baseline assessment of cognitive function combined with MRI imaging could enable predicting cognitive preservation over time and estimating cognitive reserve. Identifying factors associated with better or worse cognitive reserve may also lead to targeted interventions to preserve cognitive function in MS. Additionally, longitudinal studies with repeated cognitive and MRI assessments with protocols to measure distinct aspects of brain pathology at different stages could clarify the contributions of these aspects to cognitive decline

over time.

Similarly, it would be intriguing to explore the relationship between the subtypes derived in Chapter 7 and imaging biomarkers. This investigation would reveal whether there is a distinct pattern of brain lesions and damage underlying each subtype, leading to more aggressive and milder manifestations of the disease, or to specific types and combinations of symptoms. Furthermore, these symptoms-driven subtypes could be further refined in the future by the inclusion of additional digital markers measuring other aspects of disability as clustering variables. Finally, the subtypes could be validated on a separate cohort of pwMS and a sample of individuals from each cluster could be followed up to investigate the potential for transitions among clusters over disease course.

Another potentially interesting application of the clustering algorithm on these data would involve deriving cognitive phenotypes based on all the cognitive scores derivable from the optimal battery. Such an approach would provide a more in-depth analysis of cognitive deficits, revealing patterns of occurrence and associations among them in specific patients. There was evidence of this already in the symptoms-driven subtypes, where the mild and moderate cognitively impaired groups were distinguishable based on the presence or absence of information-processing deficits.

My work could not conclusively determine if observed deficits in performance latency across tasks were partially attributable to the visuo-motor rather than cognitive component of impairment. However, a colleague in my lab has developed a model to disentangle the visuo-motor, device, and cognitive components of performance, applicable to any computerised cognitive task, yielding promising results (Giunchiglia et al. 2024, 2023). Future applications of this model to the MS dataset could ascertain the impact of visuo-motor impairment on performance scores. Additionally, the model could provide an objective estimate of visuo-motor impairment directly from the cognitive assessment. This measure could be validated using existing data on the motor-related PROs, and potentially serve as a digital marker of visuo-motor impairment.

Lastly, a critical future step involves the development of automatically generated patient- and clinician-facing reports after the completion of both PROs and cognitive assessment. These reports should detail the profile of deficits observed during the assessment as well as changes compared to previous assessments. This approach would incentivise patient engagement by providing tangible results. Additionally, it would also offer clinicians clear outputs, fostering their understanding of the tools' significance and encouraging their active involvement. This integration would effectively close the loop between technological advancements and clinical care.

8.5 Concluding remarks

In conclusion, this thesis demonstrated the feasibility of online large-scale collection of cognitive data and its potential to provide new insights into a previously uncaptured element of the disease. The integration of online assessment tools into the UKMSR platform enables the collection of a unique dataset that can support clinical decision-making and generate the necessary data to address some of the open questions in the field.

Bibliography

- Achiron, A. and Barak, Y. (2006). Cognitive changes in early MS: A call for a common framework. *Journal of the Neurological Sciences* 245:47–51.
- Achiron, A., Doniger, G. M., Harel, Y., Appleboim-Gavish, N., Lavie, M., and Simon, E. S. (Oct. 2007). Prolonged response times characterize cognitive performance in multiple sclerosis. *European Journal of Neurology* 14:1102–1108.
- Akbar, N., Honarmand, K., Kou, N., and Feinstein, A. (Mar. 2011). Validity of a computerized version of the Symbol Digit Modalities Test in multiple sclerosis. *Journal of Neurology* 258:373–379.
- Amato, M. P., Hakiki, B., Goretti, B., Rossi, F., Stromillo, M. L., Giorgio, A., Roscio, M., Ghezzi, A., Guidi, L., Bartolozzi, M. L., Portaccio, E., and De Stefano, N. (Jan. 2012). Association of MRI metrics and cognitive impairment in radiologically isolated syndromes. *Neurology* 78:309–314.
- Amato, M. P., Goretti, B., Ghezzi, A., Hakiki, B., Niccolai, C., Lori, S., Moiola, L., Falautano, M., Viterbo, R. G., Patti, F., Cilia, S., Pozzilli, C., Bianchi, V., Roscio, M., Martinelli, V., Comi, G., Portaccio, E., and Trojano, M. (Oct. 2014). Neuropsychological features in childhood and juvenile multiple sclerosis: five-year follow-up. *Neurology* 83:1432–1438.
- Amato, M. P., Portaccio, E., Goretti, B., Zipoli, V., Iudice, A., Pina, D. D., Malentacchi, G., Sabatini, S., Annunziata, P., Falcini, M., Mazzoni, M., Mortilla, M., Fonda, C., and De Stefano, N. (Aug. 2010). Relevance of cognitive deterioration in early relapsing-remitting MS: a 3-year follow-up study. *Multiple Sclerosis Journal* 16:1474–1482.
- Amato, M. P., Morra, V. B., Falautano, M., Ghezzi, A., Goretti, B., Patti, F., Riccardi, A., and Mattioli, F. (Aug. 2018). Cognitive assessment in multiple sclerosis—an Italian consensus. *Neurological Sciences* 39:1317–1324.
- Amato, M. P., Ponziani, G., Siracusa, G., and Sorbi, S. (Oct. 2001). Cognitive Dysfunction in Early-Onset Multiple Sclerosis A Reappraisal After 10 Years. *Archives of neurology* 58:1602–1606.
- Andersen, O. (Dec. 2012). From the Gothenburg cohort to the Swedish multiple sclerosis registry. *Acta Neurologica Scandinavica* 126:13–19.
- Arnett, P. A., Higginson, C. I., Voss, W. D., Wright, B., Bender, W. I., Wurst, J. M., and Tippin, J. M. (1999). Depressed mood in multiple sclerosis: relationship to capacity-demanding memory and attentional functioning. *Neuropsychology* 13:434–446.
- Bailey, S. K., Neigel, A. R., Dhanani, L. Y., and Sims, V. K. (May 2018). Establishing Measurement Equivalence Across Computer- and Paper-Based Tests of Spatial Cognition. *Human Factors* 60:340–350.
- Barak, A. and Buchanan, T. A. (2004). Internet-based psychological testing and assessment. *Online counseling: A handbook for mental health professionals*. Elsevier Inc., 217–239.
- Barak, A. and English, N. (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services*. Vol. 19. 2-3, 65–89.
- Barkhof, F., Calabresi, P. A., Miller, D. H., and Reingold, S. C. (May 2009). Imaging outcomes for neuroprotection and repair in multiple sclerosis trials. *Nature Reviews Neurology* 5:256–266.

- Bass, A. D., Van Wijmeersch, B., Mayer, L., Mäurer, M., Boster, A., Mandel, M., Mitchell, C., Sharrock, K., and Singer, B. (July 2020). Effect of Multiple Sclerosis on Daily Activities, Emotional Well-being, and Relationships: The Global vsMS Survey. *International Journal of MS Care* 22:158–164.
- Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., and Naugle, R. I. (Feb. 2012). Computerized neuropsychological assessment devices: Joint position paper of the American academy of clinical neuropsychology and the national academy of neuropsychology. *Clinical Neuropsychologist* 26:177–196.
- Beatty, W. W. and Aupperle, R. L. (Dec. 2002). Sex differences in cognitive impairment in multiple sclerosis. *Clinical Neuropsychologist* 16:472–480.
- Beatty, W. W., Blanco, C. R., Wilbanks, S. L., Paul, R. H., and Hames, K. A. (1995). Demographic, Clinical, and Cognitive Characteristics of Multiple Sclerosis Patients Who Continue to Work. *Journal of Neurologic Rehabilitation* 9:167–173.
- Beck, E. S., Maranzano, J., Luciano, N. J., Parvathaneni, P., Filippini, S., Morrison, M., Suto, D. J., Wu, T., van Gelderen, P., de Zwart, J. A., Antel, S., Fetco, D., Ohayon, J., Andrada, F., Mina, Y., Thomas, C., Jacobson, S., Duyn, J., Cortese, I., Narayanan, S., Nair, G., Sati, P., and Reich, D. S. (Feb. 2022). Cortical lesion hotspots and association of subpial lesions with disability in multiple sclerosis. *Multiple Sclerosis Journal* 28:1351–1363.
- Benedict, R. H. (July 2009). Standards for sample composition and impairment classification in neuropsychological studies of multiple sclerosis. *Multiple Sclerosis Journal* 15:777–778.
- Benedict, R. H., Amato, M. P., DeLuca, J., and Geurts, J. J. (2020). Cognitive impairment in multiple sclerosis: clinical management, MRI, and therapeutic avenues. *The Lancet Neurology* 19:860–71.
- Benedict, R. H., Cookfair, D., Gavett, R., Gunther, M., Munschauer, F., Garg, N., and Weinstock-Guttman, B. (2006). Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society* 12:549–558.
- Benedict, R. H., Deluca, J., Phillips, G., LaRocca, N., Hudson, L. D., and Rudick, R. (Apr. 2017). Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. *Multiple Sclerosis Journal* 23:721–733.
- Benedict, R. H., Groninger, L., Schretlen, D., Dobraski, M., and Shpritz, B. (June 1996). Revision of the brief visuospatial memory test: Studies of normal performance, reliability, and, validity. *Psychological Assessment* 8:145–153.
- Benedict, R. H. and Walton, M. K. (Dec. 2012). Evaluating cognitive outcome measures for MS clinical trials: What is a clinically meaningful change? *Multiple Sclerosis Journal* 18:1673–1679.
- Bergendal, G., Fredrikson, S., and Almkvist, O. (May 2007). Selective decline in information processing in subgroups of multiple sclerosis: an 8-year longitudinal study. *European neurology* 57:193–202.
- Bergmann, C., Becker, S., Watts, A., Sullivan, C., Wilken, J., Golan, D., Zarif, M., Bumstead, B., Buhse, M. J., Kaczmarek, O., Covey, T. J., Doniger, G. M., Penner, I. K., Hancock, L. M., Bogaardt, H., Barrera, M. A., Morrow, S., and Gudesblatt, M. (Nov. 2023). Multiple sclerosis and quality of life: The role of cognitive impairment on quality of life in people with multiple sclerosis. *Multiple Sclerosis and Related Disorders* 79:104966.
- Bharadia, T., Vandercappellen, J., Chitnis, T., Eelen, P., Bauer, B., Brichetto, G., Lloyd, A., Schmidt, H., King, M., Fitzgerald, J., Hach, T., and Hobart, J. (Apr. 2022). Patient-reported outcome measures in MS: Do development processes and patient involvement support valid quantification of clinically important variables? *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 8.
- Bilder, R. M. (Feb. 2011). Neuropsychology 3.0: Evidence-Based science and practice. *Journal of the International Neuropsychological Society* 17:7–13.

- Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology* 55:803–832.
- Bjelland, I., Dahl, A. A., Haug, T., and Neckelmann, D. (Feb. 2002). The validity of the Hospital Anxiety and Depression Scale An updated literature review. *Journal of Psychosomatic Research* 52:69–77.
- Blaschke, S. J., Ellenberger, D., Flachenecker, P., Hellwig, K., Paul, F., Pöhlau, D., Kleinschnitz, C., Rommer, P. S., Rueger, M. A., Zettl, U. K., Stahmann, A., and Warnke, C. (May 2022). Time to diagnosis in multiple sclerosis: Epidemiological data from the German Multiple Sclerosis Registry. *Multiple Sclerosis Journal* 28:865–871.
- Boringa, J. B., Lazeron, R. H., Reuling, I. E., Ade, H. J., Pfenning, L. E., Lindeboom, J., De Sonnevile, L. M., Kalkers, N. F., and Polman, C. H. (Aug. 2001). The Brief Repeatable Battery of Neuropsychological Tests: normative values allow application in multiple sclerosis clinical practice. *Multiple Sclerosis Journal* 7:263–267.
- Bourke, N. J., Trender, W., Hampshire, A., Lai, H., Demarchi, C., David, M., Hellyer, P., Sharp, D. J., and Friedland, D. (2023). Assessing prospective and retrospective metacognitive accuracy following traumatic brain injury remotely across cognitive domains. *Neuropsychological Rehabilitation* 33:574–591.
- Brichetto, G. and Zaratin, P. (June 2020). Measuring outcomes that matter most to people with multiple sclerosis: the role of patient-reported outcomes. *Current Opinion in Neurology* 33:295–299.
- Brønnum-Hansen, H., Koch-Henriksen, N., and Stenager, E. (Apr. 2004). Trends in survival and cause of death in Danish patients with multiple sclerosis. *Brain* 127:844–850.
- Brooker, H., Williams, G., Hampshire, A., Corbett, A., Aarsland, D., Cummings, J., Molinuevo, J. L., Atri, A., Ismail, Z., Creese, B., Fladby, T., Thim-Hansen, C., Wesnes, K., and Ballard, C. (Oct. 2020). FLAME: A computerized neuropsychological composite for trials in early dementia. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* 12:e12098.
- Butler Pagnotti, R., Hua, L. H., and Miller, J. B. (May 2022). Cognition and disease characteristics in adult onset versus late onset multiple sclerosis. *Multiple Sclerosis Journal* 28:933–941.
- Cadden, M. and Arnett, P. (2015). Factors associated with employment status in individuals with multiple sclerosis: Cognition, fatigue, and motor function. *International Journal of MS Care* 17:284–291.
- Caine, C., Mehta, M. P., Laack, N. N., and Gondi, V. (May 2012). Cognitive function testing in adult brain tumor trials: Lessons from a comprehensive review. *Expert Review of Anticancer Therapy* 12:655–667.
- Cerqueira, J. J., Compston, D. A. S., Geraldes, R., Rosa, M. M., Schmierer, K., Thompson, A., Tinelli, M., and Palace, J. (Aug. 2018). Time matters in multiple sclerosis: Can early treatment and long-term follow-up ensure everyone benefits from the latest advances in multiple sclerosis? *Journal of Neurology, Neurosurgery and Psychiatry* 89:844–850.
- Cheetham, N. J., Penfold, R., Giunchiglia, V., Bowyer, V., Sudre, C. H., Canas, L. S., Deng, J., Murray, B., Kerfoot, E., Antonelli, M., Rjoob, K., Molteni, E., Österdahl, M. F., Harvey, N. R., Trender, W. R., Malim, M. H., Doores, K. J., Hellyer, P. J., Modat, M., Hammers, A., Ourselin, S., Duncan, E. L., Hampshire, A., and Steves, C. J. (July 2023). The effects of COVID-19 on cognitive performance in a community-based cohort: a COVID symptom study biobank prospective cohort study. *eClinicalMedicine* 62:102086.
- Chiaravalloti, N. D. and DeLuca, J. (Dec. 2008). Cognitive impairment in multiple sclerosis. *The Lancet Neurology* 7:1139–1151.

- Chitnis, T., Glanz, B., Jaffin, S., and Healy, B. (Mar. 2009). Demographics of pediatric-onset multiple sclerosis in an MS center population from the Northeastern United States. *Multiple Sclerosis Journal* 15:627–631.
- Cohen, J. A., Reingold, S. C., Polman, C. H., and Wolinsky, J. S. (May 2012). Disability outcome measures in multiple sclerosis clinical trials: Current status and future prospects. *The Lancet Neurology* 11:467–476.
- Cohen, M., Bresch, S., Thommel Rocchi, O., Morain, E., Benoit, J., Levraut, M., Fakir, S., Landes, C., Lebrun-Frénay, C., and Mikael, C. (Sept. 2021). Should we still only rely on EDSS to evaluate disability in multiple sclerosis patients? A study of inter and intra rater reliability. *Multiple Sclerosis and Related Disorders* 54:103144.
- Committee for Medicinal Products for Human Use (CHMP) (2015). *Guideline on clinical investigation of medicinal products for the treatment of Multiple Sclerosis*. Tech. rep. London: European Medicines Agency, 1–20.
- Compston, A. and Coles, A. (2008). Multiple sclerosis. *The Lancet* 372:1502–1517.
- Conde, D. M., Verdade, R. C., Valadares, A. L. R., Mella, L. F. B., Pedro, A. O., and Costa-Paiva, L. (Sept. 2021). Menopause and cognitive impairment: A narrative review of current knowledge. *World Journal of Psychiatry* 11:412–428.
- Correale, J., Ysraelit, M. C., and Fiol, M. P. (Oct. 2012). Benign multiple sclerosis: Does it exist? *Current Neurology and Neuroscience Reports* 12:601–609.
- Costa, S. L., Genova, H. M., Deluca, J., and Chiaravalloti, N. D. (May 2017). Information processing speed in multiple sclerosis: Past, present, and future. *Multiple Sclerosis Journal* 23:772–789.
- Cottrell, D. A., Kremenchutzky, M., Rice, G. P. A., Koopman, W. J., Hader, W., Baskerville, J., and Ebers, G. C. (1999). The natural history of multiple sclerosis: a geographically based study 5. The clinical features and natural history of primary progressive multiple sclerosis. *Brain* 122:625–639.
- Cree, B. A., Hollenbach, J. A., Bove, R., Kirkish, G., Sacco, S., Caverzasi, E., Bischof, A., Gundel, T., Zhu, A. H., Papinutto, N., Stern, W. A., Bevan, C., Romeo, A., Goodin, D. S., Gelfand, J. M., Graves, J., Green, A. J., Wilson, M. R., Zamvil, S. S., Zhao, C., Gomez, R., Ragan, N. R., Rush, G. Q., Barba, P., Santaniello, A., Baranzini, S. E., Oksenberg, J. R., Henry, R. G., and Hauser, S. L. (May 2019). Silent progression in disease activity-free relapsing multiple sclerosis. *Annals of Neurology* 85:653–666.
- Cutajar, R., Ferriani, E., Scandellari, C., Sabattini, L., Trocino, C., Marchello, L. P., and Stecchi, S. (May 2000). Cognitive function and quality of life in multiple sclerosis patients. *Journal of Neurovirology* 6 Suppl 2:186–90.
- D’Amico, E., Haase, R., and Ziemssen, T. (Aug. 2019). Review: Patient-reported outcomes in multiple sclerosis care. *Multiple Sclerosis and Related Disorders* 33:61–66.
- Darby, D., Maruff, P., Collie, A., and Mcstephen, M (Oct. 2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology* 59:1042–1046.
- Darby, D. G., Fredrickson, J., Pietrzak, R. H., Maruff, P., Woodward, M., and Brodtmann, A. (2014). Reliability and usability of an internet-based computerized cognitive testing battery in community-dwelling older people. *Computers in Human Behavior* 30:199–205.
- De Meijer, L., Merlo, D., Skibina, O., Grobbee, E. J., Gale, J., Haartsen, J., Maruff, P., Darby, D., Butzkueven, H., and Van der Walt, A. (2018). Monitoring cognitive change in multiple sclerosis using a computerized cognitive battery. *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 4:2055217318815513.

- Dehghani, A., Khoramkish, M., and Isfahani, S. S. (July 2019). Challenges in the Daily Living Activities of Patients with Multiple Sclerosis: A Qualitative Content Analysis. *International journal of community based nursing and midwifery* 7:201–210.
- Del Giovane, M., Parker, T. D., David, M. C., Kolanko, M. A., Giunchiglia, V., Trender, W., Carswell, C. J., Malhotra, P. A., Sharp, D. J., and Hampshire, A. (July 2023a). Computerised cognitive testing and multi-domain structural Magnetic Resonance Imaging in patients with idiopathic Normal Pressure Hydrocephalus and Alzheimer’s disease. *Alzheimer’s Association International Conference*. ALZ.
- Del Giovane, M., Trender, W. R., Bălăeț, M., Mallas, E.-J., Jolly, A. E., Bourke, N. J., Zimmermann, K., Graham, N. S. N., Lai, H., Losty, E. J. F., Oiarbide, G. A., Hellyer, P. J., Faiman, I., Daniels, S. J. C., Batey, P., Harrison, M., Giunchiglia, V., Kolanko, M. A., David, M. C. B., Li, L. M., Demarchi, C., Friedland, D., Sharp, D. J., and Hampshire, A. (2023b). Computerised cognitive assessment in patients with traumatic brain injury: an observational study of feasibility and sensitivity relative to established clinical scales. *eClinicalMedicine* 59:101980.
- Deloire, M., Ruet, A., Hamel, D., Bonnet, M., and Brochet, B. (Mar. 2010). Early cognitive impairment in multiple sclerosis predicts disability outcome several years later. *Multiple Sclerosis Journal* 16:581–587.
- DeLuca, J., Barbieri-Berger, S., and Johnson, S. K. (Apr. 1994). The nature of memory impairments in multiple sclerosis: acquisition versus retrieval. *Journal of clinical and experimental neuropsychology* 16:183–189.
- DeLuca, J., Chelune, G. J., Tulsky, D. S., Lengenfelder, J., and Chiaravalloti, N. D. (June 2004). Is speed of processing or working memory the primary information processing deficit in multiple sclerosis? *Journal of clinical and experimental neuropsychology* 26:550–562.
- DeLuca, J., Gaudino, E. A., Diamond, B. J., Christodoulou, C., and Engel, R. A. (1998). Acquisition and storage deficits in multiple sclerosis. *Journal of clinical and experimental neuropsychology* 20:376–390.
- Denney, D. R., Lynch, S. G., and Parmenter, B. A. (Apr. 2008). A 3-year longitudinal study of cognitive impairment in patients with primary progressive multiple sclerosis: speed matters. *Journal of the neurological sciences* 267:129–136.
- Derfuss, T. (Oct. 2012). Personalized medicine in multiple sclerosis: Hope or reality? *BMC Medicine* 10:116.
- Dobson, R. and Giovannoni, G. (Jan. 2019). Multiple sclerosis – a review. *European Journal of Neurology* 26:27–40.
- Drew, M., Tippet, L. J., Starkey, N. J., and Isler, R. B. (2008). Executive dysfunction and cognitive impairment in a large community-based sample with Multiple Sclerosis from New Zealand: A descriptive study. *Archives of Clinical Neuropsychology* 23:1–19.
- Edgar, C., Jongen, P. J., Sanders, E., Sindic, C., Goffette, S., Dupuis, M., Jacquerye, P., Guillaume, D., Reznik, R., and Wesnes, K. (2011). Cognitive performance in relapsing remitting multiple sclerosis: A longitudinal study in daily practice using a brief computerized cognitive battery. *BMC Neurology* 11:68.
- Eilam-Stock, T., Shaw, M. T., Krupp, L. B., Charvet, L. E., and Charvet, L. (2021). Early Neuropsychological Markers of Cognitive Involvement in Multiple Sclerosis. *Journal of the Neurological Sciences* 423.
- Ernstsson, O., Gyllensten, H., Alexanderson, K., Tinghög, P., Friberg, E., and Norlund, A. (July 2016). Cost of illness of Multiple Sclerosis - A systematic review. *PLoS ONE* 11:e0159129.
- Eshaghi, A., Young, A. L., Wijeratne, P. A., Prados, F., Arnold, D. L., Narayanan, S., Guttmann, C. R., Barkhof, F., Alexander, D. C., Thompson, A. J., Chard, D., and Ciccirelli, O. (Dec. 2021).

- Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nature Communications* 12:2078.
- EuroQol Group (1990). EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 16:199–208.
- Evangelou, N., Garjani, A., Dasnair, R., Hunter, R., Tuite-Dalton, K. A., Craig, E. M., Rodgers, W. J., Coles, A., Dobson, R., Duddy, M., Ford, D. V., Hughes, S., Pearson, O., Middleton, L. A., Rog, D., Tallantyre, E. C., Friede, T., Middleton, R. M., and Nicholas, R. (Jan. 2021). Self-diagnosed COVID-19 in people with multiple sclerosis: A community-based cohort of the UK MS Register. *Journal of Neurology, Neurosurgery and Psychiatry* 92:107–109.
- Ezabadi, S. G., Sahraian, M. A., Maroufi, H., Shahrbafe, M. A., and Eskandarieh, S. (July 2022). Global assessment of characteristics of multiple sclerosis registries; A systematic review. *Multiple Sclerosis and Related Disorders* 63:103928.
- Feenstra, H. E., Vermeulen, I. E., Murre, J. M., and Schagen, S. B. (Jan. 2017). Online cognition: factors facilitating reliable online neuropsychological test results. *Clinical Neuropsychologist* 31:59–84.
- Ferreira-González, I., Marsal, J. R., Mitjavila, F., Parada, A., Ribera, A., Cascant, P., Soriano, N., Sánchez, P. L., Arós, F., Heras, M., Bueno, H., Marrugat, J., Cuñat, J., Civeira, E., and Permanyer-Miralda, G. (Nov. 2009). Patient registries of acute coronary syndrome assessing or biasing the clinical real world data? *Circulation: Cardiovascular Quality and Outcomes* 2:540–547.
- Fink, K. (Dec. 2016). Multiple sclerosis e-registries. *Neurodegenerative disease management* 6:23–25.
- Fischer, M., Kunkel, A., Bublak, P., Faiss, J. H., Hoffmann, F., Sailer, M., Schwab, M., Zettl, U. K., and Köhler, W. (Aug. 2014). How reliable is the classification of cognitive impairment across different criteria in early and late stages of multiple sclerosis? *Journal of the Neurological Sciences* 343:91–99.
- Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkiel, K. A., Benton, C. E., Lanyon, R., Thompson, A. J., and Miller, D. H. (Mar. 2008). Disability and T2 MRI lesions: A 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain* 131:808–817.
- Flachenecker, P. (2014). E-Mail Multiple Sclerosis Databases: Present and Future. *European Neurology* 72:29–31.
- Flachenecker, P., Buckow, K., Pugliatti, M., Kes, V. B., Battaglia, M. A., Boyko, A., Confavreux, C., Ellenberger, D., Eskic, D., Ford, D., Friede, T., Fuge, J., Glaser, A., Hillert, J., Holloway, E., Ioannidou, E., Kappos, L., Kasilingam, E., Koch-Henriksen, N., Kuhle, J., Lepore, V., Middleton, R., Myhr, K. M., Orolagos, A., Otero, S., Pitschnau-Michel, D., Rienhoff, O., Sastre-Garriga, J., Schyns-Liharska, T., Sutovic, D., Thalheim, C., Trojano, M., Vlasov, Y. V., and Yaldirli, (Oct. 2014). Multiple sclerosis registries in Europe - Results of a systematic survey. *Multiple Sclerosis Journal* 20:1523–1532.
- Flachenecker, P. and Stuke, K. (Dec. 2008). National MS registries. *Journal of Neurology* 255:102–108.
- Food and Drug Administration (2009). *Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Tech. rep.
- Foong, Y. C., Bridge, F., Merlo, D., Gresle, M., Zhu, C., Buzzard, K., Butzkueven, H., and van der Walt, A. (May 2023). Smartphone monitoring of cognition in people with multiple sclerosis: A systematic review. *Multiple Sclerosis and Related Disorders* 73:104674.
- Forn, C., Belenguer, A., Parcet-Ibars, M. A., and Avila, C. (Feb. 2008). Information-processing speed is the primary deficit underlying the poor performance of multiple sclerosis patients in the Paced Auditory Serial Addition Test (PASAT). *Journal of clinical and experimental neuropsychology* 30:789–796.

- Gafson, A., Craner, M. J., and Matthews, P. M. (Mar. 2017). Personalised medicine for multiple sclerosis care. *Multiple Sclerosis Journal* 23:362–369.
- Germine, L., Reinecke, K., and Chaytor, N. S. (Feb. 2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *Clinical Neuropsychologist* 33:271–286.
- Gholamzad, M., Ebtekar, M., Ardestani, M. S., Azimi, M., Mahmodi, Z., Mousavi, M. J., and Aslani, S. (Jan. 2019). A comprehensive review on the treatment approaches of multiple sclerosis: currently and in the future. *Inflammation research* 68:25–38.
- Gilchrist, A. C. and Creed, F. H. (1994). Depression, cognitive impairment and social stress in multiple sclerosis. *Journal of psychosomatic research* 38:193–201.
- Giovannoni, G. (Nov. 2017). Personalized medicine in multiple sclerosis. *Neurodegenerative disease management* 7:13–17.
- Giunchiglia, V., Curtis, S., Smith, S., Allen, N., and Hampshire, A. (Feb. 2024). Neural correlates of cognitive ability and visuo-motor speed: Validation of IDoCT on UK Biobank Data. *Imaging Neuroscience* 2:1–25.
- Giunchiglia, V., Gruia, D., Lerede, A., Trender, W., Hellyer, P., and Hampshire, A. (2023). Iterative decomposition of visuomotor, device and cognitive variance in large scale online cognitive test data. *Research Square*.
- Glad, S., Nyland, H., and Myhr, K. M. (May 2006). Benign multiple sclerosis. *Acta Neurologica Scandinavica* 183:55–57.
- Glanz, B. I., Holland, C. M., Gauthier, S. A., Amunwa, E. L., Liptak, Z., Houtchens, M. K., Sperling, R. A., Khoury, S. J., Guttmann, C. R., and Weiner, H. L. (Sept. 2007). Cognitive dysfunction in patients with clinically isolated syndromes or newly diagnosed multiple sclerosis. *Multiple Sclerosis Journal* 13:1004–1010.
- Gois, L. C. d. P., Pimentel-Silva, L. R., Damasceno, B. P., and Damasceno, A. (Feb. 2021). Associations between cognitive and clinical disability across MS subtypes: The role of the underlying brain damage. *Multiple Sclerosis and Related Disorders* 48:102701.
- Golan, D., Wilken, J., Doniger, G. M., Fratto, T., Kane, R., Srinivasan, J., Zarif, M., Bumstead, B., Buhse, M., Fafard, L., Topalli, I., and Gudesblatt, M. (May 2019). Validity of a multi-domain computerized cognitive assessment battery for patients with multiple sclerosis. *Multiple Sclerosis and Related Disorders* 30:154–162.
- Goverover, Y., Genova, H. M., Hillary, F. G., and DeLuca, J. (June 2007). The relationship between neuropsychological measures and the Timed Instrumental Activities of Daily Living task in multiple sclerosis. *Multiple Sclerosis Journal* 13:636–644.
- Gruia, D., Combrie, S., and Geranmayeh, F. (June 2023a). Novel unsupervised comprehensive tool for monitoring vascular cognitive impairment following stroke. *Alzheimer's & Dementia* 19:e061493.
- Gruia, D. C., Trender, W., Hellyer, P., Banerjee, S., Kwan, J., Zetterberg, H., Hampshire, A., and Geranmayeh, F. (Nov. 2023c). IC3 protocol: a longitudinal observational study of cognition after stroke using novel digital health technology. *BMJ Open* 13:e076653.
- Gualtieri, C. T. and Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology* 21:623–643.
- H2 Cognitive Designs (n.d.). *Cognitron - Home*.
- Hampshire, A. (2020). Great British Intelligence Test Protocol. *Research Square*.
- Hampshire, A., Ballard, C., and Williams, G. (Dec. 2020). Computerized neuropsychological tests undertaken on digital platforms are cost effective, achieve high engagement, distinguish and are

highly sensitive to longitudinal change: Data from the PROTECT and GBIT studies. *Alzheimer's & Dementia* 16:e041122.

- Hampshire, A., Chatfield, D. A., MPhil, A. M., Jolly, A., Trender, W., Hellyer, P. J., Giovane, M. D., Newcombe, V. F., Outtrim, J. G., Warne, B., Bhatti, J., Pointon, L., Elmer, A., Sithole, N., Bradley, J., Kingston, N., Sawcer, S. J., Bullmore, E. T., Rowe, J. B., and Menon, D. K. (May 2022). Multivariate profile and acute-phase correlates of cognitive deficits in a COVID-19 hospitalised cohort. *eClinicalMedicine* 47:101417.
- Hampshire, A., Hellyer, P. J., Soreq, E., Mehta, M. A., Ioannidis, K., Trender, W., Grant, J. E., and Chamberlain, S. R. (Dec. 2021a). Associations between dimensions of behaviour, personality traits, and mental-health during the COVID-19 pandemic in the United Kingdom. *Nature Communications* 12:4111.
- Hampshire, A., Highfield, R. R., Parkin, B. L., and Owen, A. M. (Dec. 2012). Fractionating Human Intelligence. *Neuron* 76:1225–1237.
- Hampshire, A., Trender, W., Chamberlain, S. R., Jolly, A. E., Grant, J. E., Patrick, F., Mazibuko, N., Williams, S. C., Barnby, J. M., Hellyer, P., and Mehta, M. A. (Sept. 2021b). Cognitive deficits in people who have recovered from COVID-19. *eClinicalMedicine* 39:101044.
- Heather, A., Goodwin, E., Green, C., Morrish, N., Ukoumunne, O. C., Middleton, R. M., and Hawton, A. (Apr. 2023). Multiple sclerosis health-related quality of life utility values from the UK MS register. *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 9.
- Hillert, J. and Stawiarz, L. (July 2015). The Swedish MS registry – clinical support tool and scientific resource. *Acta Neurologica Scandinavica* 132:11–19.
- Hobart, J. and Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment* 13:1–177.
- Hobart, J. C., Riazi, A., Lamping, D. L., Fitzpatrick, R., and Thompson, A. J. (Jan. 2003). Measuring the impact of MS on walking ability: the 12-Item MS Walking Scale (MSWS-12). *Neurology* 60:31–36.
- Hobart, J., Freeman, J., and Thompson, A. (May 2000). Kurtzke scales revisited: the application of psychometric methods to clinical intuition. *Brain* 123:1027–1040.
- Hobart, J., Lamping, D., Fitzpatrick, R., Riazi, A., and Thompson, A. (May 2001). The Multiple Sclerosis Impact Scale (MSIS-29): A new patient-based outcome measure. *Brain* 124:962–973.
- Honarmand, K., Akbar, N., Kou, N., and Feinstein, A. (Feb. 2011). Predicting employment status in multiple sclerosis patients: The utility of the MS functional composite. *Journal of Neurology* 258:244–249.
- Huijbregts, S. C. J., Kalkers, N. F., De Sonnevile, L. M. J., De Groot, V., Reuling, I. E. W., and Polman, C. H. (2004). Differences in cognitive impairment of relapsing remitting, secondary, and primary progressive MS. *Neurology* 63:335–339.
- Hurwitz, B. J. (2011). Registry studies of long-term multiple sclerosis outcomes: Description of key registries. *Neurology* 76:S3–S6.
- Janculjak, D., Mubrin, Z., Brinar, V., and Spilich, G. (July 2002). Changes of attention and memory in a group of patients with multiple sclerosis. *Clinical Neurology and Neurosurgery* 104:221–227.
- Jellinger, K. A. (May 2003). Multiple Sclerosis: Immunology, Pathology, and Pathophysiology. *European Journal of Neurology* 10:339–339.
- Jennekens-Schinkel, A., Laboyrie, M., Lanser, J. B. K., and Van Der Velde, E. A. (1990). Cognition in patients with multiple sclerosis After four years. *Journal of the Neurological Sciences* 99:229–247.

- Johnen, A., Landmeyer, N. C., Bürkner, P. C., Wiendl, H., Meuth, S. G., and Holling, H. (Dec. 2017). Distinct cognitive impairments in different disease courses of multiple sclerosis—A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews* 83:568–578.
- Johnson, S. K., Lange, G., DeLuca, J., Korn, L. R., and Natelson, B. (1997). The Effects of Fatigue on Neuropsychological Performance in Patients With Chronic Fatigue Syndrome, Multiple Sclerosis, and Depression. *Applied Neuropsychology* 4:145–153.
- Jones, K. H., Ford, D. V., Jones, P. A., John, A., Middleton, R. M., Lockhart-Jones, H., Peng, J., Osborne, L. A., and Noble, J. G. (Jan. 2013). The physical and psychological impact of multiple sclerosis using the MSIS-29 via the web portal of the UK MS Register. *PloS one* 8:e55422.
- Julian, L. J., Vella, L., Vollmer, T., Hadjimichael, O., and Mohr, D. C. (Sept. 2008). Employment in multiple sclerosis. Exiting and re-entering the work force. *Journal of neurology* 255:1354–1360.
- Kalincik, T. and Butzkueven, H. (Dec. 2019). The MSBase registry: Informing clinical practice. *Multiple Sclerosis Journal* 25:1828–1834.
- Kavaliunas, A., Manouchehrinia, A., Stawiarz, L., Ramanujam, R., Agholme, J., Hedström, A. K., Beiki, O., Glaser, A., and Hillert, J. (Aug. 2017). Importance of early treatment initiation in the clinical course of multiple sclerosis. *Multiple Sclerosis Journal* 23:1233–1240.
- Kessels, R. P. (Feb. 2019). Improving precision in neuropsychological assessment: Bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *Clinical Neuropsychologist* 33:357–368.
- Khalil, M., Enzinger, C., Langkammer, C., Petrovic, K., Loitfelder, M., Tscherner, M., Jehna, M., Bachmaier, G., Wallner-Blazek, M., Ropele, S., Schmidt, R., Fuchs, S., and Fazekas, F. (Oct. 2011). Cognitive impairment in relation to MRI metrics in patients with clinically isolated syndrome. *Multiple Sclerosis Journal* 17:173–180.
- Klineova, S. and Lublin, F. D. (Sept. 2018). Clinical course of multiple sclerosis. *Cold Spring Harbor Perspectives in Medicine* 8:a028928.
- Knowles, S., Middleton, R., Cooze, B., Farkas, I., Leung, Y. Y., Allen, K., Winslade, M., Owen, D. R., Magliozzi, R., Reynolds, R., Neal, J. W., Pearson, O., Nicholas, R., Pickrell, W. O., and Howell, O. W. (2023). Comparing the Pathology, Clinical, and Demographic Characteristics of Younger and Older-Onset Multiple Sclerosis. *Annals of neurology*.
- Kobelt, G., Thompson, A., Berg, J., Gannedahl, M., and Eriksson, J. (July 2017). New insights into the burden and costs of multiple sclerosis in Europe. *Multiple Sclerosis Journal* 23:1123–1136.
- Koch-Henriksen, N., Stenager, E., and Brønnum-Hansen, H. (July 2011). Studies based on the Danish Multiple Sclerosis Registry. *Scandinavian Journal of Public Health* 39:180–184.
- Krupp, L. B., Alvarez, L. A., Larocca, N. G., and Scheinberg, L. C. (1988). Fatigue in multiple sclerosis. *Archives of neurology* 45:435–437.
- Krupp, L. B. and Elkins, L. E. (Oct. 2000). Fatigue and declines in cognitive functioning in multiple sclerosis. *Neurology* 55:934–939.
- Krupp, L. B., Larocca, N. G., Muir Nash, J., and Steinberg, A. D. (Oct. 1989). The Fatigue Severity Scale: Application to Patients With Multiple Sclerosis and Systemic Lupus Erythematosus. *Archives of Neurology* 46:1121–1123.
- Kuri, A., Jacobs, B. M., Leddy, S., Schmierer, K., Turner, B., Allen-Philbey, K., Stennett, A., Giovannoni, G., Thomson, A., and Dobson, R. (Sept. 2021). Evaluation of remote assessments for multiple sclerosis in an in-home setting. *Multiple Sclerosis and Related Disorders* 54:103125.
- Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology* 33:1444–1452.

- Laatu, S., Hämäläinen, P., Revonsuo, A., Portin, R., and Ruutiainen, J. (Jan. 1999). Semantic memory deficit in multiple sclerosis; impaired understanding of conceptual meanings. *Journal of the Neurological Sciences* 162:152–161.
- Landtblom, A. M., Berntsson, S. G., Boström, I., and Iacobaeus, E. (Sept. 2021). Multiple sclerosis and COVID-19: The Swedish experience. *Acta Neurologica Scandinavica* 144:229–235.
- Langdon, D. W., Amato, M. P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H. P., Krupp, L., Penner, I. K., Reder, A. T., and Benedict, R. H. (2012). Recommendations for a brief international cognitive assessment for multiple sclerosis (BICAMS). *Multiple Sclerosis Journal* 18:891–898.
- Langdon, D. W. (June 2011). Cognition in multiple sclerosis. *Current Opinion in Neurology* 24:244–249.
- Larocca, N., Kalb, R., Scheinberg, L., and Kendall, P. (1985). Factors associated with unemployment of patients with multiple sclerosis. *Journal of chronic diseases* 38:203–210.
- Lavorgna, L., Brigo, F., Moccia, M., Leocani, L., Lanzillo, R., Clerico, M., Abbadessa, G., Schmierer, K., Solaro, C., Prosperini, L., Tedeschi, G., Giovannoni, G., and Bonavita, S. (Nov. 2018). e-Health and multiple sclerosis: An update. *Multiple Sclerosis Journal* 24:1657–1664.
- Lechner-Scott, L., Kappos, L., Hofman, M., Polman, C. H., Ronner, H., Montalban, X., Tintore, M., Frontoni, M., Buttinelli, C., Amato, M. P., Bartolozzi, M. L., Versavel, M., Dahlke, F., Kapp, J. F., and Gibberd, R. (Apr. 2003). Can the Expanded Disability Status Scale be assessed by telephone? *Multiple Sclerosis Journal* 9:154–159.
- Leddy, S., Hadavi, S., McCarren, A., Giovannoni, G., and Dobson, R. (Oct. 2013). Validating a novel web-based method to capture disease progression outcomes in multiple sclerosis. *Journal of Neurology* 260:2505–2510.
- Lengenfelder, J., Chiaravalloti, N. D., Ricker, J. H., and Deluca, D. (Mar. 2003). Deciphering Components of Impaired Working Memory in Multiple Sclerosis. *Cognitive and Behavioral Neurology* 16:28–39.
- Lerede, A., Middleton, R., Rodgers, J., Hampshire, A., and Nicholas, R. (2022). The relation between age, symptoms onset and physical disability in MS: a prospective registry study. *Poster session presented at the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS), Amsterdam, Netherlands.*
- Li, Y. and Schwartz, C. E. (2011). Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. *Quality of Life Research* 20:1543–1553.
- Lublin, F. D. and Reingold, S. C. (1996). Defining the clinical course of multiple sclerosis: Results of an international survey. *Neurology* 46:907–911.
- Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Soelberg Sørensen, P., Alan Thompson, D. J., Wolinsky, J. S., Balcer, L. J., Brenda Banwell, M., Barkhof, F., Bebo, B., Calabresi, P. A., Clanet, M., Comi, G., Fox, R. J., Freedman, M. S., Goodman, A. D., Inglese, M., Kappos, L., Kieseier, B. C., Lincoln, J. A., Lubetzki, C., Miller, A. E., Montalban, X., O, P. W., Petkau, J., Pozzilli, C., Rudick, R. A., Pia Sormani, M., Stüve, O., Waubant, E., and Polman, C. H. (2014). Defining the clinical course of multiple sclerosis: The 2013 revisions. *Neurology* 83:278–286.
- Lynch, S. G., Parmenter, B. A., and Denney, D. R. (Aug. 2005). The association between cognitive impairment and physical disability in multiple sclerosis. *Multiple Sclerosis Journal* 11:469–476.
- Macniven, J. A., Davis, C., Ho, M. Y., Bradshaw, C. M., Szabadi, E., and Constantinescu, C. S. (2008). Stroop performance in multiple sclerosis: Information processing, selective attention, or executive functioning? *Journal of the International Neuropsychological Society* 14:805–814.

- Magliozzi, R., Howell, O. W., Nicholas, R., Cruciani, C., Castellaro, M., Romualdi, C., Rossi, S., Pitteri, M., Benedetti, M. D., Gajofatto, A., Pizzini, F. B., Montemezzi, S., Rasia, S., Capra, R., Bertoldo, A., Facchiano, F., Monaco, S., Reynolds, R., and Calabrese, M. (Apr. 2018). Inflammatory intrathecal profiles and cortical damage in multiple sclerosis. *Annals of Neurology* 83:739–755.
- Magyari, M. and Sorensen, P. S. (Aug. 2020). Comorbidity in Multiple Sclerosis. *Frontiers in Neurology* 11:851.
- Marrie, R. A., Cutter, G., Tyry, T., Campagnolo, D., and Vollmer, T. (July 2007). Validation of the NARCOMS registry: Diagnosis. *Multiple Sclerosis Journal* 13:770–775.
- Marziniak, M., Brichetto, G., Feys, P., Meyding-Lamadé, U., Vernon, K., and Meuth, S. G. (Apr. 2018). The use of digital and remote communication technologies as a tool for multiple sclerosis management: Narrative review. *JMIR Rehabilitation and Assistive Technologies* 5:e5.
- Mason, K., Thygesen, L. C., Stenager, E., Brønnum-Hansen, H., and Koch-Henriksen, N. (Mar. 2012). Evaluating the use and limitations of the Danish National Patient Register in register-based research using an example of multiple sclerosis. *Acta Neurologica Scandinavica* 125:213–217.
- McCarthy, M., Beaumont, J. G., Thompson, R., and Peacock, S. (Aug. 2005). Modality-specific aspects of sustained and divided attentional performance in multiple sclerosis. *Archives of clinical neuropsychology* 20:705–718.
- McGinley, M. P. and Lapin, B. (Sept. 2022). The value of patient-reported outcome measures for multiple sclerosis. *Multiple Sclerosis Journal* 28:1489–1490.
- McGuigan, C. and Hutchinson, M. (June 2004). Confirming the validity and responsiveness of the Multiple Sclerosis Walking Scale-12 (MSWS-12). *Neurology* 62:2103–2105.
- McKay, K. A., Manouchehrinia, A., Berrigan, L., Fisk, J. D., Olsson, T., and Hillert, J. (Sept. 2019). Long-term Cognitive Outcomes in Patients with Pediatric-Onset vs Adult-Onset Multiple Sclerosis. *JAMA Neurology* 76:1028–1034.
- Middleton, R. M., Craig, E. M., Rodgers, W. J., Tuite-Dalton, K., Garjani, A., Evangelou, N., das Nair, R., Hunter, R., Tallantyre, E. C., Cauchi, M., Cairn, C., Paling, D., Fuller, S., McDonnell, G., Petheram, K., Liu, B., Nock, U., Ingram, G., Brownlee, W., Taylor, J., and Nicholas, R. (Jan. 2021). COVID-19 in Multiple Sclerosis: Clinically reported outcomes from the UK Multiple Sclerosis Register. *Multiple Sclerosis and Related Disorders* 56:103317.
- Middleton, R. M., Rodgers, W. J., Chataway, J., Schmierer, K., Rog, D., Galea, I., Akbari, A., Tuite-Dalton, K., Lockhart-Jones, H., Griffiths, D., Noble, D. G., Jones, K. H., Al-Din, A., Craner, M., Evangelou, N., Harman, P., Harrower, T., Hobart, J., Husseyin, H., Kasti, M., Kipps, C., McDonnell, G., Owen, C., Pearson, O., Rashid, W., Wilson, H., and Ford, D. V. (Aug. 2018). Validating the portal population of the United Kingdom Multiple Sclerosis Register. *Multiple Sclerosis and Related Disorders* 24:3–10.
- Migliore, S., Ghazaryan, A., Simonelli, I., Pasqualetti, P., Landi, D., Palmieri, M. G., Moffa, F., Rinaldi, P., Vernieri, F., and Filippi, M. M. (Aug. 2016). Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS) in the Italian population. *Neurological Sciences* 37:1261–1270.
- Miller, D. H., Chard, D. T., and Ciccarelli, O. (Feb. 2012). Clinically isolated syndromes. *The Lancet Neurology* 11:157–69.
- Moore, H., Nair, K. P. S., Baster, K., Middleton, R., Paling, D., and Sharrack, B. (Aug. 2022). Fatigue in multiple sclerosis: A UK MS-register based study. *Multiple Sclerosis and Related Disorders* 64:103954.

- Morales-González, J. M., Benito-León, J., Rivera-Navarro, J., and Mitchell, A. J. (Feb. 2004). A systematic approach to analyse health-related quality of life in multiple sclerosis: the GEDMA study. *Multiple Sclerosis Journal* 10:47–54.
- Mosley, M. (2020). *The Great British Intelligence Test*.
- MS Society (2022). *MS in the UK: Estimates of incidence and prevalence of MS in the UK*. Tech. rep.
- Multiple Sclerosis International Federation (2013). *Atlas of MS*. Tech. rep.
- Multiple Sclerosis International Federation (Sept. 2020). *Atlas of MS 3rd edition*. Tech. rep.
- Myhr, K. M., Grytten, N., Torkildsen, Wergeland, S., Bø, L., and Aarseth, J. H. (July 2015). The Norwegian Multiple Sclerosis Registry and Biobank. *Acta Neurologica Scandinavica* 132:24–28.
- Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., and Velasquez, R. (Apr. 2004). Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist* 59:150–162.
- Nicholas, R. S., Heaven, M. L., Middleton, R. M., Chevli, M., Pulikottil-Jacob, R., Jones, K. H., and Ford, D. V. (Jan. 2020). Personal and societal costs of multiple sclerosis in the UK: A population-based MS Registry study. *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 6.
- Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping* 15:1–25.
- Nowinski, C. J., Miller, D. M., and Cella, D. (Oct. 2017). Evolution of Patient-Reported Outcomes and Their Role in Multiple Sclerosis Clinical Trials. *Neurotherapeutics* 14:934–944.
- Okuda, D. T., Mowry, E. M., Beheshtian, A., Waubant, E., Baranzini, S. E., Goodin, D. S., Hauser, S. L., and Pelletier, D (2009). Incidental MRI anomalies suggestive of multiple sclerosis: The radiologically isolated syndrome. *Neurology* 72:800–805.
- Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., and Ballard, C. G. (June 2010). Putting brain training to the test. *Nature* 465:775–778.
- Pardini, M. (Sept. 2021). Do isolated cognitive relapses exist? Yes. *Multiple Sclerosis Journal* 27:1486–1487.
- Pardini, M., Uccelli, A., Grafman, J., Yaldizli, Mancardi, G., and Roccatagliata, L. (Sept. 2014). Isolated cognitive relapses in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* 85:1035–1037.
- Parsey, C. M. and Schmitter-Edgecombe, M. (Nov. 2013). Applications of technology in neuropsychological assessment. *Clinical Neuropsychologist* 27:1328–1361.
- Pathak, L. (Apr. 2023). Personalized Treatment for Multiple Sclerosis: The Role of Precision Medicine. *Neurology Letters* 2:30–34.
- Patti, F., Failla, G., Ciancio, M. R., L’Episcopo, M. R., and Reggio, A. (1998). Neuropsychological, neuroradiological and clinical findings in multiple sclerosis. A 3 year follow-up study. *European Journal of Neurology* 5:283–286.
- Paz-Zulueta, M., Parás-Bravo, P., Cantarero-Prieto, D., Blázquez-Fernández, C., and Oterino-Durán, A. (Aug. 2020). A literature review of cost-of-illness studies on the economic burden of multiple sclerosis. *Multiple Sclerosis and Related Disorders* 43:102162.
- Pellegrini, F., Copetti, M., Bovis, F., Cheng, D., Hyde, R., de Moor, C., Kieseier, B. C., and Sormani, M. P. (May 2019). A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. *Multiple Sclerosis Journal* 26:1064–1073.

- Pelosi, L., Geesken, J. M., Holly, M., Hayward, M., and Blumhardt, L. D. (1997). Working memory impairment in early multiple sclerosis Evidence from an event-related potential study of patients with clinically isolated myelopathy. *Brain* 120:2039–2058.
- Pesarin, F. and Salmaso, L. (Dec. 2010). The permutation testing approach: a review. *Statistica* 70:481–509.
- Pitteri, M., Magliozzi, R., Nicholas, R., Ziccardi, S., Pisani, A. I., Pezzini, F., Marastoni, D., and Calabrese, M. (Apr. 2022). Cerebrospinal fluid inflammatory profile of cognitive impairment in newly diagnosed multiple sclerosis patients. *Multiple Sclerosis Journal* 28:768–777.
- Pitteri, M., Ziccardi, S., Dapor, C., Guandalini, M., and Calabrese, M. (Nov. 2019). Lost in classification: Lower cognitive functioning in apparently cognitive normal newly diagnosed RRMS patients. *Brain Sciences* 9:321.
- Planche, V., Gibelin, M., Cregut, D., Pereira, B., and Clavelou, P. (Feb. 2016). Cognitive impairment in a population-based study of patients with multiple sclerosis: Differences between late relapsing-remitting, secondary progressive and primary progressive multiple sclerosis. *European Journal of Neurology* 23:282–289.
- Portaccio, E., Stromillo, M. L., Goretti, B., Zipoli, V., Siracusa, G., Battaglini, M., Giorgio, A., Bartolozzi, M. L., Guidi, L., Sorbi, S., Federico, A., Amato, M. P., and De Stefano, N. (Aug. 2009). Neuropsychological and MRI measures predict short-term evolution in benign multiple sclerosis. *Neurology* 73:498–503.
- Potticary, H. and Langdon, D. (Jan. 2023). A Systematic Review and Meta-Analysis of the Brief Cognitive Assessment for Multiple Sclerosis (BICAMS) International Validations. *Journal of Clinical Medicine* 12:703.
- Pugliatti, M., Rosati, G., Carton, H., Riise, T., Drulovic, J., Vécsei, L., and Milanov, I. (2006). The epidemiology of multiple sclerosis in Europe. *European Journal of Neurology* 13:700–722.
- Rao, S. M., Leo, G. J., Ellington, L., Nauertz, T., Bernardin, L., and Unverzagt, F. (1991a). Cognitive dysfunction in multiple sclerosis. II. Impact on employment and social functioning. *Neurology* 41:692–696.
- Rao, S. M., Grafman, J., DiGiulio, D., Mittenberg, W., Bernardin, L., Leo, G. J., Luchetta, T., and Unverzagt, F. (1993). Memory Dysfunction in Multiple Sclerosis Its Relation to Working Memory, Semantic Encoding, and Implicit Learning. *Neuropsychology* 7:364–374.
- Rao, S. M., Leo, G. J., Bernardin, L., and Unverzagt, F. (1991b). Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. *Neurology* 41:685–691.
- Rao, S. M., Losinski, G., Mourany, L., Schindler, D., Mamone, B., Reece, C., Kemeny, D., Narayanan, S., Miller, D. M., Bethoux, F., Bermel, R. A., Rudick, R., and Alberts, J. (Dec. 2017). Processing speed test: Validation of a self-administered, iPad®-based tool for screening cognitive dysfunction in a clinic setting. *Multiple Sclerosis Journal* 23:1929–1937.
- Reips, U.-D. (2000). The Web Experiment Method: Advantages, Disadvantages, and Solutions. *Psychological Experiments on the Internet*, 89–117.
- Riazi, A., Hobart, J. C., Lamping, D. L., Fitzpatrick, R., and Thompson, A. J. (Dec. 2002). Multiple Sclerosis Impact Scale (MSIS-29): Reliability and validity in hospital based samples. *Journal of Neurology Neurosurgery and Psychiatry* 73:701–704.
- Rimkus, C. d. M., Avolio, I. M. B., Miotto, E. C., Pereira, S. A., Mendes, M. F., Callegaro, D., and Leite, C. d. C. (May 2018). The protective effects of high-education levels on cognition in different stages of multiple sclerosis. *Multiple Sclerosis and Related Disorders* 22:41–48.
- Rodgers, J., Friede, T., Vonberg, F. W., Constantinescu, C. S., Coles, A., Chataway, J., Duddy, M., Emsley, H., Ford, H., Fisniku, L., Galea, I., Harrower, T., Hobart, J., Huseyin, H., Kipps, C. M.,

- Marta, M., McDonnell, G. V., Mclean, B., Pearson, O. R., Rog, D., Schmierer, K., Sharrack, B., Straukiene, A., Wilson, H. C., Ford, D. V., Middleton, R. M., and Nicholas, R. (Apr. 2022). The impact of smoking cessation on multiple sclerosis disease progression. *Brain* 145:1368–1378.
- Rovaris Marco, Confavreux Christian, Furlan Roberto, Kappos Ludwig, Comi Giancarlo, and Filippi Massimo (2006). Secondary progressive multiple sclerosis: current knowledge and future challenges. *The Lancet Neurology* 5:343–354.
- Ruano, L., Portaccio, E., Goretti, B., Nicolai, C., Severo, M., Patti, F., Cilia, S., Gallo, P., Grossi, P., Ghezzi, A., Roscio, M., Mattioli, F., Stampatori, C., Trojano, M., Viterbo, R. G., and Amato, M. P. (Aug. 2017). Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. *Multiple Sclerosis Journal* 23:1258–1267.
- Ruet, A., Deloire, M. S., Charré-Morin, J., Hamel, D., and Brochet, B. (Mar. 2013). A new computerised cognitive test for the detection of information processing speed impairment in multiple sclerosis. *Multiple Sclerosis Journal* 19:1665–1672.
- Saccà, F., Costabile, T., Carotenuto, A., Lanzillo, R., Moccia, M., Pane, C., Russo, C. V., Barbarulo, A. M., Casertano, S., Rossi, F., Signoriello, E., Lus, G., and Brescia Morra, V. (Aug. 2017). The EDSS integration with the Brief International Cognitive Assessment for Multiple Sclerosis and orientation tests. *Multiple Sclerosis Journal* 23:1289–1296.
- Savettieri, G., Messina, D., Andreoli, V., Bonavita, S., Caltagirone, C., Cittadella, R., Farina, D., Fazio, M. C., Girlanda, P., Le Pira, F., Liguori, M., Lugaresi, A., Nocentini, U., Reggio, A., Salemi, G., Tedeschi, G., Trojano, M., Valentino, P., and Quattrone, A. (Oct. 2004). Gender-related effect of clinical and genetic variables on the cognitive impairment in multiple sclerosis. *Journal of Neurology* 251:1208–1214.
- Schatz, P. and Browndyke, J. (2002). Applications of computer-based neuropsychological assessment. *The Journal of head trauma rehabilitation* 17:395–410.
- Schultheis, M. T., Weisser, V., Ang, J., Elovic, E., Nead, R., Sestito, N., Fleksher, C., and Millis, S. R. (Mar. 2010). Examining the Relationship Between Cognition and Driving Performance in Multiple Sclerosis. *Archives of Physical Medicine and Rehabilitation* 91:465–473.
- Schwid, S. R., Tyler, C. M., Scheid, E. A., Weinstein, A., Goodman, A. D., and Mcdermott, M. P. (2003). Cognitive fatigue during a test requiring sustained attention: a pilot study. *Multiple Sclerosis Journal* 9:503–508.
- Simpson, S., Wang, W., Otahal, P., Blizzard, L., Van Der Mei, I. A., and Taylor, B. V. (Nov. 2019). Latitude continues to be significantly associated with the prevalence of multiple sclerosis: An updated meta-analysis. *Journal of Neurology, Neurosurgery and Psychiatry* 90:1193–1200.
- Sintzel, M. B., Rametta, M., and Reder, A. T. (June 2018). Vitamin D and Multiple Sclerosis: A Comprehensive Review. *Neurology and Therapy* 7:59–85.
- Siva, A., Saip, S., Altintas, A., Jacob, A., Keegan, B. M., and Kantarci, O. H. (Aug. 2009). Multiple sclerosis risk in radiologically uncovered asymptomatic possible inflammatory-demyelinating disease. <http://dx.doi.org/10.1177/1352458509106214> 15:918–927.
- Skoog, B., Tedeholm, H., Runmarker, B., Odén, A., and Andersen, O. (2014). Continuous prediction of secondary progression in the individual course of multiple sclerosis. *Multiple Sclerosis and Related Disorders* 3:584–592.
- Smestad, C., Sandvik, L., Landrø, N. I., and Celius, E. G. (2010). Cognitive impairment after three decades of multiple sclerosis. *European journal of neurology* 17:499–505.
- Smit, M., Van der Stigchel, S., Visser-Meily, J. M., Kouwenhoven, M., Eijsackers, A. L., and Nijboer, T. C. (July 2013). The feasibility of computer-based prism adaptation to ameliorate neglect in sub-acute stroke patients admitted to a rehabilitation center. *Frontiers in Human Neuroscience* 7:48432.

- Snaith, R. P. and Zigmond, A. S. (Feb. 1986). The hospital anxiety and depression scale. *British Medical Journal (Clinical research ed.)* 292:344.
- Snaith, R. P. (Aug. 2003). The Hospital Anxiety And Depression Scale. *Health and Quality of Life Outcomes* 1:29.
- Soreq, E., Leech, R., and Hampshire, A. (Feb. 2019). Dynamic network coding of working-memory domains and working-memory processes. *Nature Communications* 2019 10:1 10:936.
- Soreq, E., Violante, I. R., Daws, R. E., and Hampshire, A. (Dec. 2021). Neuroimaging evidence for a network sampling theory of individual differences in human intelligence test performance. *Nature Communications* 12:2072.
- Sormani, M. P. (Feb. 2020). Real-world studies provide reliable comparisons of disease modifying therapies in MS - No. *Multiple Sclerosis Journal* 26:161–162.
- Strijbis, E. M., Repovic, P., Mostert, J., Bowen, J. D., Uitdehaag, B. M., Cutter, G., and Koch, M. W. (Sept. 2022). The MSIS-29 and SF-36 as outcomes in secondary progressive MS trials. *Multiple Sclerosis Journal* 28:1606–1619.
- Strober, D. L. (Nov. 2020). Determinants of unemployment in multiple sclerosis (MS): The role of disease, person-specific factors, and engagement in positive health-related behaviors. *Multiple Sclerosis and Related Disorders* 46.
- Strober, L., Englert, J., Munschauer, F., Weinstock-Guttman, B., Rao, S., and Benedict, R. H. (2009). Sensitivity of conventional memory tests in multiple sclerosis: Comparing the Rao Brief Repeatable Neuropsychological Battery and the Minimal Assessment of Cognitive Function in MS. *Multiple Sclerosis Journal* 15:1077–1084.
- Stuke, K., Peter, A. E., Ae, F., Ae, U. K. Z., Elias, W. G., Matthias, A. E., Ae, F., Haas, J., Dorothea, A. E., Ae, P.-M., Schimrigk, S., and Rieckmann, A. P. (2009). Symptomatology of MS: results from the German MS Registry. *Journal of Neurology* 256:1932–1935.
- Sumowski, J. F., Benedict, R.,ENZINGER, C., Filippi, M., Geurts, J. J., Hamalainen, P., Hulst, H., Inglese, M., Leavitt, V. M., Rocca, M. A., Rosti-Otajarvi, E. M., and Rao, S. (Feb. 2018). Cognition in multiple sclerosis. *Neurology* 90:278–288.
- Sumowski, J. F., Rocca, M. A., Leavitt, V. M., Riccitelli, G., Comi, G., Deluca, J., and Filippi, M. (June 2013). Brain reserve and cognitive reserve in multiple sclerosis: What you’ve got and how you use it. *Neurology* 80:2186.
- Tallantyre, E. C., Major, P. C., Atherton, M. J., Davies, W. A., Joseph, F., Tomassini, V., Pickersgill, T. P., Harding, K. E., Willis, M. D., Winter, M., and Robertson, N. P. (May 2019). How common is truly benign MS in a UK population? *Journal of Neurology, Neurosurgery & Psychiatry* 90:522–528.
- Tao, C., Simpson, S., Van Der Mei, I., Blizzard, L., Havrdova, E., Horakova, D., Shaygannejad, V., Lugaresi, A., Izquierdo, G., Trojano, M., Duquette, P., Girard, M., Grand’Maison, F., Grammond, P., Alroughani, R., Terzi, M., Oreja-Guevara, C., Sajedi, S. A., Iuliano, G., Sola, P., Lechner-Scott, J., Van Pesch, V., Pucci, E., Bergamaschi, R., Barnett, M., Ramo, C., Singhal, B., Spitaleri, D. L., Slee, M., Verheul, F., Fernández Bolaños, R., Amato, M. P., Cristiano, E., Granella, F., Hodgkinson, S., Fiol, M., Gray, O., McCombe, P., Saladino, M. L., Sánchez Menoyo, J. L., Shuey, N., Vucic, S., Shaw, C., Deri, N., Arruda, W. O., Butzkueven, H., Spelman, T., and Taylor, B. V. (Dec. 2016). Higher latitude is significantly associated with an earlier age of disease onset in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry* 87:1343–1349.
- The Lancet Neurology (Oct. 2011). Making EUREMS count for people with multiple sclerosis. *The Lancet Neurology* 10:865.
- The Lancet Neurology (2019). Patient-reported outcomes in the spotlight. *The Lancet Neurology* 18:981.

- Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., Mowry, E. M., Sorensen, P. S., Tintoré, M., Traboulsee, A. L., Trojano, M., Uitdehaag, B. M., Vukusic, S., Waubant, E., Weinshenker, B. G., Reingold, S. C., and Cohen, J. A. (Feb. 2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology* 17:162–173.
- Thornton, A. E. and Raz, N. (1997). Memory impairment in multiple sclerosis: A quantitative review. *Neuropsychology* 11:357–366.
- Thygesen, L. C. and Ersbøll, A. K. (2014). When the entire population is the sample: strengths and limitations in register-based epidemiology. *European journal of epidemiology* 29:551–558.
- Till, C., Racine, N., Araujo, D., Narayanan, S., Louis Collins, D., Aubert-Broche, B., Arnold, D. L., and Banwell, B. (2013). Changes in cognitive performance over a 1-year period in children and adolescents with multiple sclerosis. *Neuropsychology* 27:210–219.
- Ton, A. M. M., Vasconcelos, C. C. F., and Alvarenga, R. M. P. (June 2017). Benign multiple sclerosis: Aspects of cognition and neuroimaging. *Arquivos de Neuro-Psiquiatria* 75:394–401.
- Tripepi, G., Jager, K. J., Dekker, F. W., and Zoccali, C. (June 2010). Selection Bias and Information Bias in Clinical Research. *Nephron Clinical Practice* 115:c94–c99.
- Trisolini, M., Honeycutt, A., Wiener, J., and Lesesne, S. (2010). *Global Economic Impact of Multiple Sclerosis Global Economic Impact of Multiple Sclerosis Literature Review*. Tech. rep.
- Trojano, M. et al. (Jan. 2019). The Italian multiple sclerosis register. *Neurological Sciences* 40:155–165.
- Tur, C., Kalincik, T., Oh, J., Sormani, M. P., Tintoré, M., Butzkueven, H., and Montalban, X. (Oct. 2019). Head-to-head drug comparisons in multiple sclerosis: Urgent action needed. *Neurology* 93:793–809.
- Van Der Mei, I. A., Ponsonby, A. L., Blizzard, L., and Dwyer, T. (2001). Regional variation in multiple sclerosis prevalence in Australia and its association with ambient ultraviolet radiation. *Neuroepidemiology* 20:168–174.
- Vargas, W. S., Noble, K. G., Banwell, B., and De Jager, P. (Nov. 2020). The educational impact of childhood-onset multiple sclerosis: Why assessing academic achievement is imperative. *Multiple Sclerosis Journal* 26:1633–1637.
- Vleugels, L., Lafosse, C., Van Nunen, A., Nachtergaele, S., Ketelaer, P., Charlier, M., and Vandebussche, E. (2000). Visuo-perceptual impairment in multiple sclerosis patients diagnosed with neuropsychological tasks. *Multiple Sclerosis Journal* 6:241–254.
- Vollmer, B. L., Wolf, A. B., Sillau, S., Corboy, J. R., and Alvarez, E. (Jan. 2022). Evolution of Disease Modifying Therapy Benefits and Risks: An Argument for De-escalation as a Treatment Paradigm for Patients With Multiple Sclerosis. *Frontiers in Neurology* 12:799138.
- Wallin, M. T., Culpepper, W. J., Campbell, J. D., Nelson, L. M., Langer-Gould, A., Marrie, R. A., Cutter, G. R., Kaye, W. E., Wagner, L., Tremlett, H., Buka, S. L., Dilokthornsakul, P., Topol, B., Chen, L. H., and Larocca, N. G. (Mar. 2019). The prevalence of MS in the United States: A population-based estimate using health claims data. *Neurology* 92:E1029–E1040.
- Ward, M. and Goldman, M. D. (2022). Epidemiology and Pathophysiology of Multiple Sclerosis. *Multiple Sclerosis and Related Disorders* 8:988–1005.
- Weinshenker, B. G., Bass, B., Rice, G. P., Noseworthy, J., Carriere, W., Baskerville, J., and Ebers, G. C. (Dec. 1989a). THE NATURAL HISTORY OF MULTIPLE SCLEROSIS: A GEOGRAPHICALLY BASED STUDY: 2 PREDICTIVE VALUE OF THE EARLY CLINICAL COURSE. *Brain* 112:1419–1428.

- Weinshenker, B. G., Bass, B., Rice, G. P., Noseworthy, J., Carriere, W., Baskerville, J., and Ebers, G. C. (Feb. 1989b). THE NATURAL HISTORY OF MULTIPLE SCLEROSIS: A GEOGRAPHICALLY BASED STUDY: I. CLINICAL COURSE AND DISABILITY. *Brain* 112:133–146.
- Weinstock, Z. L. and Benedict, R. H. B. (Aug. 2022). Cognitive Relapse in Multiple Sclerosis: New Findings and Directions for Future Research. *NeuroSci* 3:510–520.
- Westergaard, K., Skovgaard, L., Magyari, M., and Kristiansen, M. (Feb. 2022). Patient perspectives on patient-reported outcomes in multiple sclerosis treatment trajectories: A qualitative study of why, what, and how? *Multiple Sclerosis and Related Disorders* 58:103475.
- Wiklund, I. (June 2004). Assessment of patient-reported outcomes in clinical trials: The example of health-related quality of life. *Fundamental and Clinical Pharmacology* 18:351–363.
- Wingerchuk, D. M. and Noseworthy, J. H. (Apr. 2002). Randomized controlled trials to assess therapies for multiple sclerosis. *Neurology* 58:S40–S48.
- Wojcik, C. M., Beier, M., Costello, K., DeLuca, J., Feinstein, A., Goverover, Y., Gudesblatt, M., Jaworski, M., Kalb, R., Kostich, L., LaRocca, N. G., Rodgers, J. D., and Benedict, R. H. (Dec. 2019). Computerized neuropsychological assessment devices in multiple sclerosis: A systematic review. *Multiple Sclerosis Journal* 25:1848–1869.
- Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., and Holdnack, J. A. (2006). The California Verbal Learning Test - second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology* 21:413–420.
- Yang, J. H., Rempe, T., Whitmire, N., Dunn-Pirio, A., and Graves, J. S. (June 2022). Therapeutic Advances in Multiple Sclerosis. *Frontiers in Neurology* 13:824926.
- Yusuf, F. L., Ng, B. C., Wijnands, J. M., Kingwell, E., Marrie, R. A., and Tremlett, H. (Aug. 2020). A systematic review of morbidities suggestive of the multiple sclerosis prodrome. *Expert Review of Neurotherapeutics* 20:799–819.
- Zaratin, P., Vermersch, P., Amato, M. P., Brichetto, G., Coetzee, T., Cutter, G., Edan, G., Giovannoni, G., Gray, E., Hartung, H. P., Hobart, J., Helme, A., Hyde, R., Khan, U., Leocani, L., Mantovani, L. G., McBurney, R., Montalban, X., Penner, I. K., Uitdehaag, B. M., Valentine, P., Weiland, H., Bertorello, D., Battaglia, M. A., Baneke, P., and Comi, G. (May 2022). The agenda of the global patient reported outcomes for multiple sclerosis (PROMS) initiative: Progresses and open questions. *Multiple Sclerosis and Related Disorders* 61:103757.
- Zhao, S., Shibata, K., Hellyer, P. J., Trender, W., Manohar, S., Hampshire, A., and Husain, M. (2022). Rapid vigilance and episodic memory decrements in COVID-19 survivors. *Brain Communications* 4:fcab295.
- Ziemssen, T. (Dec. 2011). Symptom management in patients with multiple sclerosis. *Journal of the Neurological Sciences* 311:S48–S52.
- Zipoli, V., Goretti, B., Hakiki, B., Siracusa, G., Sorbi, S., Portaccio, E., and Amato, M. P. (Dec. 2009). Cognitive impairment predicts conversion to multiple sclerosis in clinically isolated syndromes. *Multiple Sclerosis Journal* 16:62–67.