Author:
Lawton, Michael A

Title:
Prognosis of neurodegenerative diseases

methodological and empirical results for Multiple Sclerosis and Parkinson's disease

Prognosis of neurodegenerative diseases: methodological and empirical results for Multiple Sclerosis and Parkinson's disease

Michael Lawton, BSc, MSc
Population Health Sciences

Supervised by Professor Yoav Ben-Shlomo and
Professor Kate Tilling

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy by Published Work in the Faculty of Health Sciences, October 2019.

Word count = 15,297 words

Abstract
(300 words max)

Neurodegenerative diseases, like Multiple Sclerosis and Parkinson's disease, lead to disability that worsens with time. Being able to predict prognosis in these diseases is important for both patients and clinicians when making medication choices and planning for the future. My aim was to look at drug effectiveness over a ten-year period in Multiple Sclerosis, in the absence of a long-term clinical trial, and to look at prognosis in Parkinson's disease by deriving subtypes.

I developed a longitudinal model for the untreated natural history of patients with Multiple Sclerosis using multilevel models. This model was used to predict the untreated trajectories of treated MS patients over a ten year period. A comparison between the observed treated trajectories and the predicted untreated trajectories gave an estimate of long-term drug effectiveness. I carried out intention-to-treat and per-protocol approaches along with imputed analyses to adjust for missing data. The medications were found to be effective in the long-term.

I used a k-means cluster analysis on the baseline phenotype of a large cohort of recently diagnosed Parkinson's patients to attempt to derive subtypes. These subtypes were subsequently found to be associated with medication response. This approach was extended in another large inception cohort using a development and validation approach. Before combining the two cohorts in an analysis I had to harmonise the data from the cohorts as olfaction was measured using two different tests. I used Item Response Theory to convert the two tests onto the same scale. The harmonised baseline phenotypic data from both cohorts was used to estimate subtypes. This approach was relatively stable when comparing the actual and predicted subtypes in the smaller cohort. These subtypes were associated with differing rates of motor progression and to medication response.

Acknowledgements

Many thanks to Yoav who has been working with me for 8 and a half years since my very first day at the University of Bristol. He always has something insightful to add to any conversation about epidemiology.

To Kate for introducing me to multilevel models. I am now able to pass this knowledge onto others through my journal articles and teaching at the university.

Margaret, thanks for your wisdom when we worked together on the Parkinson's Disease cohorts and for lending me your PhD by publication which really helped me structure this thesis.

Thanks to Neil, Helen, Michele and Donald for giving me the opportunity to work on their world class Multiple Sclerosis and Parkinson's Disease cohorts. To all the participants of these studies and everyone involved in the collection of this data that made all these analyses possible.

Many thanks to my parents. I wouldn't be here if it wasn't for them, in more ways than one! Their kindness and compassion has kept me going through dark times and good.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Table of Contents

Contents

List of Tables and Figures

List of submitted publications

Paper P1. Lawton M, Tilling K, Robertson NP, Tremlett H, Zhu F, Harding KE et al. A longitudinal model for disease progression was developed and applied to multiple sclerosis. (2015) *Journal of Clinical Epidemiology*.;68(11):1355-1365.

Paper P2. Palace J, Duddy M, Bregenzer T, Lawton M, Zhu F, Boggild M et al. Effectiveness and cost-effectiveness of interferon beta and glatiramer acetate in the UK Multiple Sclerosis Risk Sharing Scheme at 6 years: a clinical cohort study with natural history comparator. (2015) *Lancet Neurology*, *14*(5), 497-505.

Paper P3. Palace J, Duddy M, Lawton M, Bregenzer T, Zhu F, Boggild M et al. Assessing the long-term effectiveness of interferon-beta and glatiramer acetate in multiple sclerosis: final 10-year results from the UK multiple sclerosis risk-sharing scheme. (2019) *Journal of Neurology, Neurosurgery and Psychiatry*. 90(3):251-60.

Paper P4. Lawton M, Baig F, Rolinski M, Ruffman C, Nithi K, May MT et al. Parkinson's disease subtypes in the Oxford Parkinson disease centre (OPDC) discovery cohort. (2015) *Journal of Parkinson's disease*. *5(2)*:269-279

Paper P5. Lawton MA, Hu MTM, Baig F, Ruffmann C, Barron E, Swallow DMA et al. Equating Scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in Patients with Parkinson's Disease. (2016) *Parkinsonism and Related Disorders*. 33:96-101.

Paper P6. Lawton M, Ben-Shlomo Y, May MT, Baig F, Barber TR, Klein JC et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. (2018) *Journal of Neurology, Neurosurgery and Psychiatry*. 89(12):1279-87

Please note that there is a post publication mistake on page 498 of paper P2 in the Methods chapter within the first paragraph. When referencing the Association of British Neurologist guidelines for prescribing drugs it says an EDSS score of 5.5 or lower. The correct figure is 6.5 or lower which is reported in both papers P1 and P3.

1. Introduction

"Parkinson's is a slow but inevitable process. It's hard living with it on a daily basis. The difficulty facing people with it is that they never quite know 'Can I or can't I do this today?'"

- *Helen Mirren*

1.1. Neurodegenerative disease

Neurodegenerative diseases, such as Multiple Sclerosis and Parkinson's disease, are chronic diseases that lead to disability that generally gets progressively worse with time.

Prognostic research in these diseases is important for many reasons. It can help to counsel patients and family at diagnosis as to what they can expect from the years ahead. It can help with risk stratification for treatment options - for instance, if someone had a poor prognosis they might consider medications with a worse side effects profile. It can also help in designing clinical trials or in looking at drug effectiveness in the long-term.

1.2. Objectives

The objectives for the presented work were as follows:

1. To develop a multilevel model for the natural history of untreated MS and validate this on an independent cohort.
2. Use this model as a comparator in a large cohort of treated patients followed up for 10 years to determine drug effectiveness in the long-term.
3. Derive subtypes of Parkinson's disease using cluster analysis in a large inception cohort of Parkinson's patients.
4. Harmonise the data across two large inception cohorts of Parkinson's patients
5. Use the harmonized data to derive subtypes of Parkinson's disease in both cohorts with a development and validation approach.

6. Examine how these validated subtypes relate to disease prognosis and response to medication.

1.3. Structure of the thesis

In chapter 2 I will introduce the disease Multiple Sclerosis and describe its basic epidemiology and some treatment options. The following chapter will describe the work I have done looking at prognosis in Multiple Sclerosis. In chapter 4 I will introduce Parkinson's disease and describe its basic epidemiology and the reason why clinicians believe there are subtypes of the disease. The following chapter will describe the work I have done looking at prognosis in Parkinson's disease by deriving subtypes. The sixth chapter is a discussion of the work in this thesis which includes sections on the significance of the publications, strengths and limitations of the work, any ongoing and future research, recommendations for future research, and ends with a conclusion. In the seventh chapter I will detail exactly what contribution I made to the published works. Following that, will be a list of my references and then finally all the published papers that will be referred to as P1 to P6, (see above). Along with each published paper will be any supplementary materials that are only available online.

2. Background to Multiple Sclerosis

"I realized this is what God has dealt me, and I should be thankful considering all that's happened to me in my life, but MS caused the movies to stop - stop dead - and I miss it"

- *Richard Pryor*

Multiple Sclerosis (MS) is a chronic neurodegenerative disease which is caused by damage to the nerve coating (myelin) and to the nerves themselves. This can bring about a range of symptoms which are mostly dysfunction in motor and autonomic function. The symptoms will generally worsen with time and eventually patients might need ambulatory assistance such as a walking stick or a wheelchair. In some cases the disease can cause death.

It has been estimated in the UK (2010) the prevalence of MS was 203 per 100,000 of the population and the incidence in one year was 9.64 per 100,000 [1]. Tragically it is a disease that affects people when relatively young with an average age at onset of 30 years [2]. However life expectancy is only reduced by 6-7 years [3] so people live for a long time with the disease. The presentation of disease is highly heterogenous including different clinical features and rates of accumulation of neurological disability.

It is widely accepted that there are different subtypes to the disease. The majority of patients (~85%) present with relapses ("attacks") where symptoms appear and then disappear (either partially or completely) [4]. This is called relapsing-remitting MS (RRMS). Some of these RRMS patients will convert to secondary-progressive MS (SPMS) where the frequency of relapses decreases and the accumulation of disability increases steadily [5]. Some patients (~15%) will present with a primary-progressive MS (PPMS) that gets progressively worse over time without relapses [4]. In a group of RRMS/SPMS patients the estimated median time to requiring assistance to walk was 18 years and to being bedridden was 28 years [6]. The prognosis tends to be worse in PPMS patients with an estimated median time to requiring assistance to walk being only 14 years [7].

2.1. Treatment for Multiple Sclerosis

Currently there is no cure for the disease. In 2002 there were a range of disease modifying therapies (DMTs) that were licensed for treating RRMS or SPMS. They were shown to reduce relapse rates and also the seriousness of relapses [8-11]. The four DMTs available at that time were: two forms of interferon beta-1a, interferon beta-1b and glatiramer acetate [12].

The UK's National Institute for Care Excellence decided these DMTs were not cost-effective in the long-term over a 10 or 15 year period [13]. However these conclusions were based on clinical trials that typically lasted no longer than three years and hence there was no information about the long-term effectiveness of these DMTs reducing disability accumulation. Since these drugs were already licensed to reduce relapse rates it was felt unethical to carry out a long-term placebo controlled trial, however not everyone agreed with this position [14]. To determine the long-term effectiveness of DMTs the UK MS risk-sharing scheme (MS-RSS) was established by the UK's Department of Health. The plan was to follow-up a large cohort of treated patients for a ten year period. A model for disability accumulation could then be built from historical data on untreated MS patients. This model could be applied to the MS-RSS to predict an individual's progression if they had remained untreated and compared to each individuals' observed progression on treatment thus giving an estimate of the long-term effectiveness of the DMTs.

2.2. Scales for measurement of severity of disease in Multiple Sclerosis

The most common outcome measure for measuring the severity of MS is the Expanded Disability Status Scale (EDSS) [15, 16]. It is an ordinal 20 unit scale based on a neurologist's examination. It ranges from 0 (normal) to 10 (death due to MS) in half unit increments (apart from no score of 0.5), a description of each score is in table 1. The EDSS has been used in many cohort studies and also clinical trials. An alternative scale is the MS Functional Composite [17] which is a metric scale and includes a measure of cognitive impairment which is not included within the EDSS.

A measure of quality of life, related to the EDSS, was a utility measure derived from data that reported EQ5D scores at different EDSS states, see table 1 below, where a score of 1 is perfect health and 0 death. These utility scores are reported in the appendices of papers P2 and P3. It could be argued that utility scores are more appropriate outcome measure for patients because the data are created using a patient reported outcome rather than an arbitrary scale created by clinicians to rate severity of disease.

Table 1. A description of the Expanded Disability Status Scale (EDSS) scores along with the corresponding utility scores derived using the EQ5D

EDSS SCORE	**DESCRIPTION**	**UTILITY SCORE**
0	No disability	0.9248
1	No disability, minimal signs in one functional system (FS)	0.7614
1.5	No disability, minimal signs in more than one FS	0.7614
2	Minimal disability in one FS	0.6741
2.5	Mild disability in one FS or minimal disability in two FS	0.6741
3	Moderate disability in one FS, or mild disability in three or four FS. No impairment to walking	0.5643
3.5	Moderate disability in one FS and more than minimal disability in several others. No impairment to walking	0.5643
4	Significant disability but self-sufficient and up and about some 12 hours a day. Able to walk without aid or rest for 500m	0.5643
4.5	Significant disability but up and about much of the day, able to work a full day, may otherwise have some limitation of full activity or require minimal assistance. Able to walk without aid or rest for 300m	0.5643
5	Disability severe enough to impair full daily activities and ability to work a full day without special provisions. Able to walk without aid or rest for 200m	0.4906
5.5	Disability severe enough to preclude full daily activities. Able to walk without aid or rest for 100m	0.4906

6	Requires a walking aid - cane, crutch, etc - to walk about 100m with or without resting	0.4453
6.5	Requires two walking aids - pair of canes, crutches, etc - to walk about 20m without resting	0.4453
7	Unable to walk beyond approximately 5m even with aid. Essentially restricted to wheelchair; though wheels self in standard wheelchair and transfers alone. Up and about in wheelchair some 12 hours a day	0.2686
7.5	Unable to take more than a few steps. Restricted to wheelchair and may need aid in transfering. Can wheel self but can not carry on in standard wheelchair for a full day and may require a motorised wheelchair	0.2686
8	Essentially restricted to bed or chair or pushed in wheelchair. May be out of bed itself much of the day. Retains many self-care functions. Generally has effective use of arms	0.0076
8.5	Essentially restricted to bed much of day. Has some effective use of arms retains some self care functions	0.0076
9	Confined to bed. Can still communicate and eat	-0.2304
9.5	Confined to bed and totally dependent. Unable to communicate effectively or eat/swallow	-0.2304
10	Death due to MS	0

3. Prognosis in Multiple Sclerosis

"my left side is asking for directions from a broken GPS."

Selma Blair

3.1. Multiple Sclerosis cohorts

We had access to three cohorts of individuals with Multiple Sclerosis which are detailed below. In the UK the Association of British Neurologist (ABN) criteria are used to determine whether an individual with MS can be treated with DMTs. The ABN criteria was defined as age ≥ 18 years, EDSS ≤ 6.5 and had ≥ 2 relapses during the previous two years [18]. A relapse was defined as worsening neurologic symptoms lasting > 24 hours, in the absence of fever or infection and the starting date of each relapse was recorded by an MS specialist neurologist.

3.1.1. University of Wales MS cohort

The University of Wales MS cohort (UoWMS) is based at the University Hospital of Wales which is the major tertiary referral center for neurology in Wales, United Kingdom. It serves a local population of 1.2 million and provides a network of MS clinics across South East Wales. Data were initially collected in a cross-sectional study and were updated periodically until 2002 when data were essentially collected prospectively. Sociodemographic and clinical features at disease onset are recorded in a standardized fashion, including degree of recovery and initial inter-relapse interval. Approximately 1,000 patient contacts are documented annually, and clinical data, including EDSS scores, are collected routinely at presentation and at each visit. At the time of extraction there were roughly 2,000 registered MS patients with 1,283 and 809 patients having at least 2 or 4 or more EDSS scores over time, respectively.

The UoWMS data came in four spreadsheets which identified different patient variables: one with all the EDSS data, another with all the treatment data, another with all the relapse data and the fourth with other phenotypic data. This data required time to manipulate and merge together into one dataset in a format appropriate for statistical analysis. Our aim was to create a model for the untreated disease accumulation for individuals who would have been eligible for treatment under the MS-RSS (using the ABN criteria). To select our patient population we did the following:

1. Removed any EDSS at an age of less than 18 years
2. Looked at any EDSS observations ≤ 6.5 and checked whether there were two or more relapses during the previous two years. Took the earliest of these observations for any patients and used that as the ABN eligibility start date and removed any observations made before this date.
3. Truncated the data when patients initiated any DMTs.

After data cleaning and restricting to those who were eligible for treatment under the ABN criteria we were left with 404 patients for the analysis.

3.1.2. British Columbia MS cohort

The British Columbia MS (BCMS) cohort was established in 1980 and is population based, capturing around 80% of the population of British Columbia which is a province on the western side of Canada. As of 2009, the database contained records for over 5,900 MS patients spanning 28 years of prospective follow-up from four clinics across British Columbia. The BCMS cohort has been the subject of many papers looking at the natural history of MS [19-22].

All the data cleaning and the majority of data preparation was done by the BCMS site statistician rather than by myself. Due to the original ethics approval the data could only be analysed on site in the Brain Research Centre, University of British Columbia, Vancouver, Canada. After data cleaning and restricting to those who were eligible for treatment under the ABN criteria we were left with 978 patients for the analysis.

3.1.3. United Kingdom MS risk-sharing scheme

The UK MS risk-sharing scheme (MS RSS) recruited 5,583 patients between 2002 and 2005 across 70 neurology centres in the UK[23]. These were patients who met the ABN criteria for treatment and represented about 80% of patients with MS starting treatment in the UK over this period. The high uptake of patients starting treatment in that period reduces the potential for selection bias. However, drug selection across the four DMTs was led by clinician choice and was not randomised in any way. Data were collected annually over a ten year period.

3.2. Developing a longitudinal model for untreated Multiple Sclerosis

Most previous studies that have modelled EDSS have used survival analysis [24-26], considering the time to specific milestones - for instance an EDSS of 6, which is equivalent to needing an aid to walk. However, this ignores data both before and after reaching the milestone. Also, individuals with very different disease accumulation might reach the milestone at the same time, for instance two people with annual EDSS scores of 4, 5, 6 and 0, 1, 6 would contribute the same information. Also, if an individual had reached the milestone when we started modelling they would have to be ignored.

Percentiles of EDSS scores derived at yearly intervals with data-smoothing techniques have also been used to create disability curves over time [27, 28]. However, these do not model how any given individual changes, or the relationship between patient characteristics and the centiles.

Another potential method is to use discrete Markov models which models the probability of transitioning between EDSS states as time progresses. Markov models have been used previously to relate progression in MS to baseline covariates [29, 30]. This model type was also the original model planned to be used in the MS-RSS analysis and was the only model used in the 2 year analysis of this dataset [23]. After the 2 year analysis it was decided to carry out an independent analysis using multilevel repeated measure models. This would allow the MS-RSS to test the robustness of results by using two different modelling strategies.

Multilevel repeated measures models [31] are easily able to model unbalanced data, such as we have from our observational cohorts, where individuals have different numbers of observations and the time between observations is not constant. Such models can account for both within and between patient variability. To our knowledge multilevel models had only previously been used once to model the accumulation of disability in MS [32] using a transformation of the EDSS and assuming a quadratic relationship between EDSS and time since first observation.

In my first paper, P1, we have improved upon this method by testing for different non-linear relationships (instead of the standard assumption of adding a quadratic term) using fractional polynomials, testing/adjusting for observation level variation, checking and adjusting for autocorrelation, and removing the effect of relapse so that we are modelling only the accumulation of underlying disability and not relapse related disability. As we were censoring our data when a patient started medication we also tested for indication/selection bias. We also show how we can use the baseline measurement to predict all future data, since that is how our analysis in the MS-RSS will work.

We found that a model using log and linear time terms for time since onset was optimal. Adding an additional time term to the observational level variation improved the model fit, presumably because the upper end of the EDSS scale has lower measurement error than the lower end since it is more subjective at the lower end. We adjusted for autocorrelation by grouping together observations made within a short time frame. We tested for the effect of relapse by dropping observations made within certain time frames of start of a relapse. Testing for selection bias by including ever started DMTs as a covariate provided some evidence that selection bias was not present. Although the EDSS scale is not technically continuous we felt that the QQ-plots from the residuals in our model were acceptable given our sample size.

3.3. Long-term effectiveness of Disease Modifying Therapies

The MS-RSS was to be a 10 year study, with interim analyses every two years though because of the delay in getting these data prepared by the time the 2 year data were published it was decided to work on the 6 year data as this was almost complete. Hence there were only 2 interim analyses prior to the final 10 year results, at year 2 and 6. The 2 year analysis used a longitudinal cohort from London Ontario to develop the untreated natural history model [23]. This cohort had a rule that EDSS scores could not improve from one visit to the next. In the year 2 analysis carrying out a sensitivity analysis when they used unadjusted EDSS scores at baseline gave remarkably different results, changing from treatment having a large detrimental effect in the primary analysis to treatment having a large beneficial effect in this sensitivity analysis [23]. Also it was felt this “no improvement” rule was not a reflection of reality since the EDSS will have both measurement error and also day to day variability of

patients not due to relapse which could mean that EDSS might improve at the next visit. Due to these problems after the 2 year analysis it was felt this cohort was inappropriate for modelling the untreated natural history of MS. To better reflect the reality of the EDSS the MS-RSS decided to change the natural history dataset for the year 6 and 10 analyses and the BCMS cohort was used as the historical cohort instead. Not it only was it larger, it was more contemporaneous hence reducing any bias due to secular changes in the natural history of MS. Although there were multiple drugs within the MS-RSS the plan was not to analyse any drug-specific data due to commercial confidentiality.

Using the models we developed, we were able to calculate the predicted EDSS if the patients had remained off treatment conditional on the patients baseline EDSS. Importantly for the economic analysis we also derived a predicted EDSS on treatment. To estimate predicted EDSS on treatment the pharmaceutical companies had estimated a hazard ratio, derived from their short-term clinical trials, and this hazard ratio was used to adjust the accumulation of disability in our natural history multilevel models. We used two approaches for the expected values on treatment, an intention-to-treat approach where the whole trajectory used the hazard adjusted multilevel model and a per-protocol approach where the trajectory reverted to the natural history multilevel model at the time treatment was stopped (if treatment was stopped).

At the outset of the project a deviation measure was created to inform the economic analysis as shown below

$$D(t) = \frac{D_a(t) - D_e^*(t)}{D_e(t) - D_e^*(t)} \cdot 100\%$$

$D_a(t)$ is the actual value of the outcome measure at year t ***with treatment***

$D_e^*(t)$ is the expected value of the outcome measure at year t ***with treatment*** estimated from our natural history model adjusted with the hazard ratio

$D_e(t)$ is the expected value of the outcome measure at year t ***without treatment*** estimated from our natural history model

A value of 0% would mean the drugs worked exactly as predicted, lower than 0% implies that the drugs work better than predicted and higher than 0% worse than predicted.

Although in our model development paper (P1) we used time since onset of MS as the time axis we resorted to using time since ABN eligibility as our time axis for the main analyses as requested by the scientific advisory group. This would allow us to make more robust comparisons between the multilevel model and the Markov model as they would use the same time axis. We also added age at onset as a covariate in our model [30] as this was found to be important for the Markov model. This covariate was used to adjust our intercept and also our two time terms. Table 2 below shows the validation of this new model in the UoWMS cohort where we used the first available EDSS to predict all future values. The figure below shows this same validation graphically. This validation shows good characteristics with observed and predicted EDSS being very similar across all time points with no evidence of bias.

Table 2 Observed EDSS for the UoWMS dataset, and EDSS predicted by the BCMS model relating EDSS to time since ABN eligibility and including age at onset as a binary covariate. N is the number of observations in each time frame and N* is the number of individuals contributing to that N.

Time since eligibility, years	EDSS observed			EDSS predicted		EDSS predicted – observed	
	N (N*)	Mean	SD	Mean	SD	Mean	SD
0-0.5	88 (72)	3.76	1.71	3.77	1.63	0.01	0.99
0.5-1.5	292 (153)	4.18	2.02	4.26	1.81	0.08	1.14
1.5-2.5	265 (118)	4.82	1.78	4.87	1.59	0.05	1.14
2.5-3.5	202 (101)	5.27	1.61	5.19	1.62	-0.08	1.32
3.5-4.5	182 (92)	5.61	1.77	5.58	1.67	-0.03	1.53
4.5-5.5	142 (72)	5.74	1.73	5.86	1.42	0.12	1.46
5.5-6.5	138 (54)	6.07	1.77	5.85	1.35	-0.22	1.57
6.5-7.5	75 (39)	6.09	1.79	5.98	1.38	-0.11	1.68
7.5-8.5	52 (30)	6.10	1.92	6.27	1.38	0.17	1.69
8.5-9.5	29 (14)	7.33	1.22	7.22	1.67	-0.10	1.48
9.5-10.5	30 (10)	7.48	1.16	7.45	1.80	-0.03	1.49
Total	1495 (253)	5.20	1.97	5.20	1.80	-0.001	1.35

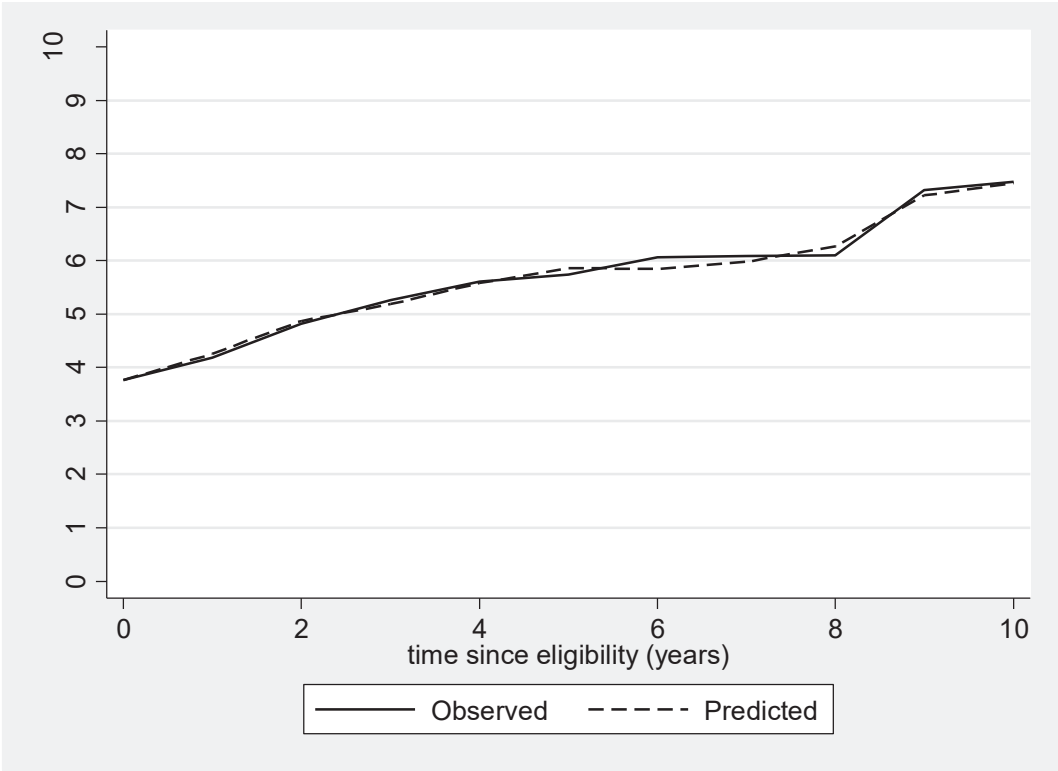


Figure 1. Observed and predicted EDSS in the UoWMS cohort using the final analysis model

EDSS was our primary outcome however we also carried out sensitivity analyses where we converted EDSS scores to utility scores. Although EDSS was treated as continuous in our models we rounded to the nearest EDSS score for all our predictions. In the Markov model, due to sparsity of some of the transition data, the Markov model was derived using only the integer values, similar to the table above. To ensure comparability between the two approaches we carried out analyses using the normal EDSS scale and another where we used only the integer values of the EDSS (called rounded EDSS in the papers).

For our primary analyses we used the last available EDSS score. However, as with any long-term study there is likely to be drop-out, so for some individuals the last available EDSS score might have been earlier, e.g. they could have dropped out after year 5. To adjust for this drop-out we carried out a sensitivity analyses where we imputed on-treatment scores at the missing years. Imputation was carried out using multilevel model fitted to the MS-RSS population where this multilevel model had the same parameterization as our natural history model. We used a single mean imputation taken from the patient specific fitted lines (accounting for individual level random effects) and also a multiple imputation where the observation level residuals were taken from the appropriate distribution to add uncertainty the mean imputed EDSS values. Imputation was carried out using different approaches

1. Intention to treat
2. Per-protocol where the trajectories were adjusted for being on or off treatment and individuals were assumed to stay on treatment for the missing years.
3. As 2. but individuals were assumed to be off treatment for the missing years

As we could not be certain that data is missing at random (missingness associated with observed data) we carried out an imputation where we added 0.5 to the EDSS scores in an attempt to account for data that is not missing at random (missingness associated with unobserved data). This would imply that the individuals missed their later visits because their EDSS was worse than predicted from their modelled trajectories.

The 95% confidence intervals for the deviation measure and relative rate of progression were derived using bootstrapping since they involved dividing multiple estimates making it hard to derive a formula for the confidence interval.

To account for the uncertainty in the BCMS model parameter that we used to derive the predicted untreated EDSS, we carried out a sensitivity analyses in the year 6 analysis where we sampled from the parameter distribution (which is a multivariate normal distribution). We did this 500 times predicting EDSS on and off treatment for each re-sampled model. The mean and variance of these 500 predictions were then combined with the use of Rubin's rules [33].

For our year 10 analysis, P3, as well as repeating the analyses from year 6 we also carried out pre-specified sub-group analyses. These included stratifying by baseline EDSS (≤3.5, 4 to 5.5 and ≥ 6), type of MS at baseline (RRMS or SPMS) and date of baseline assessment (on/before 31 August 2003 or after 31 August 2003). The last sub-group reflected some concern that the patients initially recruited into the RSS were a mixture of prevalent and incident cases and may therefore have better prognosis due to "length bias" whilst later patients recruited would only include patients who newly met the inclusion criteria. Some of these subgroup analyses were combined for instance stratifying by both baseline EDSS and date of baseline assessment. In the year 10 analysis we also explored the apparent changes in treatment effect over time by considering the difference between observed and predicted off treatment at 2, 4, 6, 8 and 10 years. This seemed to show a "waning" effect where the effect of the medication (difference between observed and predicted off treatment) was more pronounced at earlier time points.

Generally, all results show that the medications are effective in the long-term. The analyses using EDSS as the outcome had different estimates of effect of drugs between the Markov and multilevel models although both are consistent with a positive effect of medication over the long-term. However the analyses using the utility score as an outcome gave more similar results across the two modeling approaches. Imputed analyses were similar to the main analyses providing evidence that missing data and drop-out did not bias any of our estimates.

We can formalise the causal inference we are doing using the framework set out by Hernan and Robins [34]. Our conceptual question is whether treatments used within the RSS reduce progression rates in EDSS for those meeting the ABN criteria. If A= 1 is on treatment and A = 0 is not on treatment, whilst Y is our outcome (EDSS progression) the two causal effects, in essence, that we are interested in deriving are:

(i) $E[Y|A= 1] - E[Y|A= 0]$ (what we called absolute treatment effect in paper P3)

(ii) $E[Y|A= 1] / E[Y|A= 0]$ (what we called relative treatment effect in paper P3)

We have observed $E[Y|A= 1]$ as all of the patients in the RSS are on treatment, although this is still just an estimate of the target population of interest (all MS patients fulfilling the ABN criteria).

The counterfactual outcome $E[Y|A= 0]$ is unknown for all the patients in the RSS, something that we are trying to predict using the untreated natural history model we have derived, along with the covariate age at onset and baseline EDSS. For our predictions to be a valid exchangeable counterfactual then conditional on the baseline EDSS and age at onset all other risk factors for EDSS progression would need to be the same within the RSS population and the BCMS population used for the untreated model. This is often called the no unmeasured confounder assumption.

In table 1 of paper P2 we can see that sex, time since symptom onset and number of confirmed relapses in the past 2 years are relatively similar within the BCMS and RSS. We also know from, table 2 and figure 1, that conditional on the baseline EDSS and age at onset our model predicts EDSS progression in an external dataset without any bias. However we can never be certain that all risk factors for EDSS progression are the same within the RSS population and the BCMS population.

4. Background to Parkinson's Disease

"I don't have a choice whether or not I have Parkinson's but surrounding that non-choice is a million other choices that I can make"

-Michael J. Fox

Parkinson's disease is a chronic neurodegenerative disease and was first described by Dr. James Parkinson in the early 1800's in a paper called "An essay on the Shaking Palsy". It is caused by a loss of dopamine in parts of the brain and is commonly categorized by its motor symptoms bradykinesia (slowness of movement), tremor, rigidity and postural instability. The level of dopamine will continue to decline over the years leading the disease to get progressively worse.

Parkinson's disease is the second most common neurodegenerative disorder, after Alzheimer's disease [35], with an estimated crude incidence of 13.0 per 100,000 person-years [36]. Parkinson's currently has a prevalence of 210.1 per 100,000 of the population in the UK. This rises from a prevalence of 1.8 per 100,000 of those aged 20-29 to 1,696 per 100,000 in those aged 80-84 [37]. The numbers with PD is expected to rise considerably over the next 30 years due to an aging population [38]. A recent meta-analysis showed that mean disease duration at death in 10 studies ranged from 7 to 14 years [39].

Diagnosing Parkinson's disease comes with its own challenges. There are other Parkinsonian syndromes such as progressive supranuclear palsy (PSP), multiple system atrophy (MSA) and corticobasal degeneration that in the early stages of disease are often mistaken for PD. In fact the only way to be certain is following autopsy where one study showed that 24/100 patients diagnosed with PD did not have PD following postmortem brain examination [40]. Another study showed that at least 15% of patients with a diagnosis of PD do not satisfy strict clinical criteria [41].

As well as the classic motor symptoms it is now well recognized that there are many non-motor symptoms predominant in PD such as cognitive impairment, anxiety, apathy, sleep

problems, hyposmia (loss of sense of smell) and autonomic problems like constipation, urinary dysfunction and postural hypotension. A recent systematic review showed the rate of PD dementia was 24.5% [42]. Whilst another publication showed the rate varying from 2% in early-onset cases to 81% in an unselected patient population [43]. Along with the motor symptoms, dementia will clearly have large implications with respect to loss of independence and the requirement for support or a caregiver.

The motor and non-motor symptoms present at differing severities within each person with Parkinson's and prognosis will also vary considerably. This heterogeneity has led many clinicians to believe that there are subtypes of the disease [44, 45]. Unlike MS which has well defined subtypes (RRMS, SPMS and PPMS) there is no consensus within PD about how to define these subtypes.

5. Prognosis in Parkinson's Disease

"It isn't the mountains ahead to climb that wear you out: it's the pebble in your shoe"

- Muhammad Ali

5.1. Parkinson's Disease cohorts

For the papers presented in this thesis I worked on two prospective cohorts of individuals with Parkinson's disease. The Oxford Parkinson's Disease Centre Discovery cohort (hereafter referred to as the Discovery cohort) and the Tracking Parkinson's cohort (hereafter referred to as the Tracking cohort).

Within both cohorts, patients are recently diagnosed at recruitment, within 3.5 years of diagnosis, hence reducing the chance of any survival bias seen with prevalent cohorts. They must also satisfy the Queen Square brain bank criteria for PD diagnosis [46]. In short this criteria requires a patient to have bradykinesia and at least one of the following: muscular rigidity, rest tremor or postural instability. Demographic and phenotypic data were collected by questionnaires completed via a clinician, nurse or the patient themselves. Patients are followed up every 18 months and ~90% of the questionnaires are collected in both cohorts. All patients gave informed consent and full ethical approval was granted to both studies. The charity Parkinson's UK funded both of these cohorts.

The Discovery cohort [47] has roughly 1000 patients with PD that were recruited from 11 hospitals in the Thames Valley Region. Patients were recruited between September 2010 and January 2016. Exclusion criteria for participation were: non-idiopathic parkinsonism, dementia within one year of diagnosis suggestive of Dementia with Lewy bodies and cognitive impairment precluding informed consent. People at-risk of developing PD (first degree relatives of PD patients and those with REM sleep behavior disorder) and some control subjects were also recruited as part of this cohort. These at-risk and control individuals were not used for any analyses described within this thesis.

The Tracking cohort [48] has roughly 2000 patients with PD that were recruited from 72 sites providing secondary care treatment for PD patients across the entire United Kingdom. Patients were recruited between February 2012 and May 2014. Exclusion criteria were:

severe comorbidities not allowing clinic visits, other forms of parkinsonism (like PSP and MSA), and patients with drug-induced parkinsonism. A smaller dataset of ~250 individuals with early onset PD (age onset <50 years) and first degree relatives were also recruited. These early onset PD and relatives were not used for any analyses described within this thesis.

The Discovery cohort data was stored in an Open Clinica online database and the Tracking cohort data was stored in a ClinBase online database. This data required considerable time to download and manipulate into a format that would allow detailed statistical analysis. I also spent time checking for any illogical or inconsistent data and contacting admin staff at each centre so this data could be checked against the original paper CRF's.

As mentioned in the previous chapter diagnosing PD can be difficult. So in an attempt to ensure that individuals with other similar diseases were not included within any of our analysis datasets we excluded any patients with a firm alternative diagnosis or who had a probability of PD <90% as rated by a research neurologist/movement disorder specialist at their latest available visit. The numbers that were dropped in papers P4 and P6 can be seen from the flow-charts in the respective papers.

5.2. Scales for measurement of disease severity

One of the most common measures of disease severity comes from the Movement Disorder Society (MDS) Unified Parkinson's Disease Rating Scale (UPDRS) [49] which is split into four parts. We are most interested in part III which measures the severity of motor function from 33 questions on a Likert scale of 0 to 4. The MDS-UPDRS III is our main outcome measure when looking at motor prognosis with values from 0 to 132 where higher values are associated with worse motor function.

The main cognitive outcome in both cohorts was the Montreal Cognitive Assessment (MoCA) [50] which is commonly adjusted for the education years to account for a patient's intelligence whereby a more intelligent person might find the cognitive tasks easier. The MoCA is on a scale of 0 to 30 where higher values are associated with better cognition.

5.3. Parkinson's Disease subtypes

Our method to derive subtypes was first to carry out a factor analysis and then a k-means cluster analysis on the factor scores and any variables that were not loading into one of these factors. The factor analysis is important because if many variables from a similar domain (that are highly correlated) were included in a k-means analysis then that domain would be given greater importance in the k-means algorithm. We also used multiple imputation with chained equations to adjust for missing data which is more robust and less biased than excluding individuals with missing data (complete case analysis), although in some untestable situations a complete case analysis would be unbiased whilst an imputed analysis would be biased [51].

In our analysis we had variables from the following phenotypic domains:

1. Motor function from the MDS-UPDRS III split into five domains
    a. Bradykinesia
    b. Tremor
    c. Rigidity
    d. Postural stability
    e. Speech
2. Get up and go (getting up from a chair walking turning around and sitting back down)
3. Flamingo (standing on one leg)
4. Dexterity
5. Apathy
6. Fatigue
7. Pain
8. Anxiety
9. Depression
10. Impulsive- compulsive behaviours
11. Cognition
    a. MoCA
    b. Mini-mental state examination (MMSE)
    c. Phonemic and Semantic fluencies

12. Orthostatic blood pressure (difference in blood pressure when going from a lying position to a standing position)
13. Olfaction (sense of smell)
14. REM sleep behavior disorder (RBD)
15. Daytime sleepiness
16. Hallucinations
17. Constipation
18. Urinary function
19. Personality traits

There have been many other attempts at deriving PD subtypes and a systematic review [45] from 2010 reviewed all papers published on PD subtypes using cluster analysis and came up with the following recommendations for future studies on this topic:

1. Select a sample of patients with a similar disease duration.
2. Critically select a set of conceptually similar clinical variables that adequately represent the clinical spectrum of PD
3. Take the limitations of K-means cluster analysis (CA) into account or apply another CA technique that does not have these limitations
4. Critically evaluate the cluster results: Are they clinically meaningful and interpretable? Which variables discriminate best between the clusters? How do the clusters differ with respect to variables not included in the CA?
5. Validate the results in independent samples. Studies that apply a similar design in different cohorts and take into account the aforementioned recommendations will likely increase our knowledge on subtypes in PD.

To take into account these recommendations we did the following

1. As mentioned in section 5.1 our cohorts were based on individuals who were recently diagnosis (within 3.5 years of diagnosis) and hence more similar disease duration than other cross-sectional studies.
2. Our cohorts were well phenotyped across a wide range of important motor and non-motor domains such as motor, cognitive, autonomic, sleep dysfunction, and psychological well-being.

3. We took into account the limitations of k-means by:
    a. Using hierarchical cluster analysis prior to the analysis to help determine the number of clusters
    b. Standardizing all variables so that they have equal weighting within the CA
    c. Using 500 random starts to prevent the selection of local rather than global optima
4. We looked at independent associations between our clusters and response to medication.
5. In our first paper we did not validate our results in an independent sample, however we did use cross-validation techniques to look at the stability of our analysis approach. We had already established a collaboration with another cohort, Tracking, that would enable us to carry out a future cross-cohort validation.

In the systematic review the largest CA was based on 346 patients. A recent follow-on paper written by the same author of the systematic review paper was based on 344 patients and validated in another 357 [52]. Two papers looking at cluster analyses in separate cohorts have since been written by Fereshtehnejad, the first based on only 113 patients [53] and the second based on 421 patients [54]. Our analysis was based on 769 patients making it the largest ever published at the time. However in the same year another paper was published based on 1,510 PD patients [55]. This paper was based on patients with much longer disease duration (64 months on average) and was not as deeply phenotyped as our own cohort (they only had motor, cognition, depression, sleep quality and constipation variables).

In paper P4 we found three factors: a psychological well-being factor, a non-tremor motor factor and a cognitive factor. Using two different statistics on a hierarchical cluster analysis suggested that either the two or five cluster solution was optimal. After looking at the different clusters and seeing that the two cluster solution only provided a good and poor group we decided that five groups solution was optimal. The five subtypes from the k-means cluster analysis were arbitrarily called: (1) mild motor and non-motor disease, (2) poor posture and cognition, (3) severe tremor, (4) poor psychological well-being, RBD and sleep, (5) severe motor and non-motor disease with poor psychological well-being. Importantly these subtypes were related to variables not included in the k-means algorithm like age and gender which might point towards differing aetiology. Also they were related to medication

effectiveness when we used the crude Clinical Global Impression of Change Scale. In a cross-validation approach we found 73.8% of individuals were stable.

5.4. Harmonising scales for cross cohort collaborations

One of the most common non-motor features in PD is dysfunction with olfaction (sense of smell). However, the way that we measured olfaction was not carried out in the same way across and within the two cohorts. The Tracking cohort measured olfaction using the University of Pennsylvania Smell Identification Test (UPSIT) but then changed to using Sniffin' sticks when the UPSIT became difficult to source from the USA. The Discovery cohort measured olfaction only using Sniffin' sticks.

The UPSIT [56] is based on 40 scratch and sniff panels and each panel has a corresponding multiple choice question with one correct answer and three incorrect answers or "distractors". This test has a forced choice paradigm so if an individual taking the test is unsure of an answer they were asked to guess the correct answer. Hence the UPSIT is on a scale of 0 to 40 where a higher score is better olfaction.

The Sniffin' sticks identification test [57] is based on 16 sticks that look similar to a marker pen. Again the test has multiple choice questions with one correct answer and three distractors and has a forced choice paradigm. So the Sniffin' is on a scale of 0 to 16 where a higher score is better olfaction.

Since olfaction was measured using different tests on different scales we need a way to put our olfaction data onto the same scale. One crude method would be to standardize the data from each scale. However this would be based on the assumption that the average olfaction and variability in olfaction is the same for people taking the UPSIT and the Sniffin'. We already had strong evidence that some motor and non-motor characteristics were different across the two cohorts so making that assumption could potentially be incorrect. Therefore we used scale equating methods such as equipercentile equating and Item Response Theory (IRT) to convert UPSIT to Sniffin' scores.

Equipercentile equating matches score based on their percentile ranks after first smoothing the centile distributions. This method requires that the two groups are equivalent in terms of distribution usually because individuals have taken both tests or individuals are randomized to receive one test. Equipercentile equating has been applied to cognitive scales in PD by other groups and ourselves [58-60]. IRT fits a series of latent variable models (where the latent variable is olfaction for our analysis) for each item on a test. These models are then harmonized across the two tests using items that are common to the two tests. This harmonization relies on the fact that the model parameters would be equal if the latent variable was identical in the populations taking each test. These methods are described in detail elsewhere [61].

As well as the data from the Discovery and Tracking cohorts the PI from the Tracking cohort was able to collect a small dataset of 61 PD and 67 control subjects who took both the UPSIT and Sniffin' tests so that we could validate any method converting UPSIT to Sniffin' scores. This group of individuals were recruited as a convenience sample from the regional West of Scotland Movement Disorder Clinic. Our main outcome when comparing converted and true scores was to be the concordance correlation coefficient [62] whilst also considering the mean/median of the difference. Concordance is very similar to a correlation coefficient but is derived from deviation about the line of perfect agreement. When we compared our IRT converted Sniffin' scores to true Sniffin' scores within this datatset it had a high concordance of 0.80 and a mean and median difference of 0.14 and 0 respectively. Using the equipercentile equating method gave a concordance 0.79 of and a mean and median difference of 0.66 and 1 respectively, providing more evidence that the distribution of olfaction was different in those who took the UPSIT and Sniffin' tests. However our IRT conversion method had acceptable characteristics with converted and true scores being similar.

As well as allowing us to analyse olfaction in these two cohorts this conversion method will also assist future researchers doing cross-cohort collaborations or individual patient meta-analyses. For this aim we also equated scores from the UPSIT to the Brief-SIT (a smaller 12 item version of the UPSIT) and from a 12 item version of the Sniffin' to the full 16 item Sniffin'.

5.5. Parkinson's Disease subtypes using a development validation approach

Our first analysis to derive Parkinson's subtypes was based on only the Discovery cohort but now that we were able to harmonise the olfaction data we could extend our approach to also include the Tracking cohort. In this new analysis we used a development/validation approach where we developed the clusters in both cohorts and then created a model for the Tracking cohort clusters using a discriminant analysis. We could then predict these Tracking cohort clusters in the Discovery cohort and compare to the k-means clusters to determine the stability of our approach using a kappa statistic. This satisfies the criteria for point 5 in the systematic review's recommendations. To better satisfy point 4 we critically evaluated our clusters by looking at levodopa response and also motor (MDS-UPDRS III) and cognitive (MoCA) progression using multilevel random slope and intercept models. We also carried out a sensitivity analysis adjusting for patient withdrawal from the study using pattern-mixture models [63] as we were concerned that patients loss to follow-up might bias our progression estimates.

To our knowledge at the time only the two papers written by Fereshtehnejad have used cluster analysis on baseline phenotype and then looked at whether they predict subsequent prognosis [53, 54]. However we have used all the follow-up data in a multilevel random slope and intercept model which is arguably better than the approach used by Fereshtehnejad using only the baseline measurement and the latest available follow-up measure. Using all available data will give us a more precise and less biased measure of the slope. We also used pattern-mixture models in an attempt to adjust for the potential loss-to-follow up bias whilst Fereshtehnejad excluded those individuals who dropped out after their baseline assessment.

We looked at levodopa response using what is called a levodopa challenge. This involves a patient omitting their usual levodopa dose approximately 12 hours before the morning challenge test. The patient is then given their usual dose of oral levodopa and the MDS-UPDRS III performed at baseline and 1 hour later. The percentage change from the pre and post dose MDS-UPDRS III then gives a more quantitative measure of drug effectiveness than the simple Clinical Global Impression of Change Scale used in paper P4.

There were some variables that were not collected in Tracking but were collected in the Discovery cohort and used in the P4 paper. These were the get up and go, flamingo, and dexterity variables, one of the two depression measures, MMSE and the phonemic fluency. This only removed four variables from our non-tremor motor factor, one variable from the psychological well-being factor and prevented us from detecting a cognitive factor. However since the MoCA had the highest loading on the cognitive factor losing these variables did not have a large impact on any of the domains in the cluster analysis.

Using the same statistical approach as the previous paper again showed that either two or five clusters was optimal. We then examined the proportions between the k-means clusters in Discovery and those predicted by the Tracking discriminant model looking at each solution in turn. The five and four cluster solutions gave the same overall agreement of 67.9% and as the four cluster solution was more parsimonious we decided that was best.

We called the four subtypes: (1) fast motor progression with symmetrical motor disease, poor olfaction, cognition and postural hypotension; (2) mild motor and non-motor disease with intermediate motor progression; (3) severe motor disease, poor psychological well-being and poor sleep with an intermediate motor progression; (4) slow motor progression with tremor dominant, unilateral disease. The third cluster from our P6 paper was essentially a merger of the fourth and fifth cluster from the P4 paper.

Our approach was relatively stable with a kappa of 0.58 (95% confidence interval 0.54 to 0.61) between the k-means clusters in Discovery and those predicted by the Tracking discriminant model. According to accepted guidelines which are used to determine the strength of agreement[64] 0.4-0.6 is moderate agreement and 0.6-0.8 is substantial agreement. Since our kappa is close to the border of these two groups and the 95% confidence interval includes both groups we interpreted this as being moderate to substantial agreement, although a more conservative view would be to just say moderate agreement. Interesting we found 67.9% of patients stable which compares to 73.8% stable using cross-validation in the P4 analysis. This might be due to the baseline differences between the two cohorts reported in P6 rather than cross-validation not working as well as a true external validation.

The levodopa challenge showed that medication response differs significantly across the clusters. We also found that MDS-UPDRS III progression rates differ significantly across the clusters with the difference between the fastest and slowest progressors in the Tracking cohort being 2.6 points per year. Given that this is near equivalent to the primary endpoint in some clinical studies this has implications for future clinical trials in Parkinson's. A clinical trial is more likely to have a false negative finding if we fail to take into account that some individuals will not respond to medication due to their disease subtype.

When we compared our approach to a cruder but more recognised method for determining subtypes (tremor dominant, postural instability gait difficulty and mixed phenotype) we did not find significant difference in MDS-UPDRS III progression rates. Although in the Tracking cohort only we did find a suggestion that the PIGD subtype has faster cognitive decline. Our validated clusters were not strongly associated with different rates of cognitive decline, however the patterns of cognitive decline for the clusters across the two cohorts were remarkably similar. It may be that the MoCA, which was designed to be a screening test, lacks the sensitivity for us to detect differences across our four clusters or we need a longer follow-up for this to emerge.

The olfaction conversion proved very important. If we had removed this variable due to not being able to harmonise the data, or we had used a crude method we may not have been able to pick out our first cluster robustly (which had poor olfaction). This cluster turned out to have the fastest motor progression.

In paper P6 not only have we taken into account all five recommendations for future research from the systematic review but also our analysis had much far greater numbers (>2500 in both cohorts) than any previous cluster analysis in PD. Also our cohort was unselected whilst some other cohorts like the Parkinson's Progression Markers Initiative cohort are based on recently diagnosed (within two years of diagnosis) patients and also must be drug-naïve. By excluding individuals who are on dopaminergic therapy close to diagnosis we are likely to

exclude those with the most severe disease and hence add an element of selection bias into the cohort and the clusters could not be generalized to all recently diagnosed patients.

6. Discussion

"Sure, Parkinson's may be one step forward and two steps back, but I've learned that what is important is making that step count."

- *Michael J. Fox*

6.1 How our results compare with others

A latent class growth analysis has been used to model EDSS in PPMS patients to consider heterogeneity in disease progression [65]. The authors applied similar methods to our own using fractional polynomials to determine the trajectories and also found that including time since disease onset was superior to using age. However they found the best trajectory was modelled with three time terms: linear time, square root of time and square of time. Square root of time is relatively similar to log of time in that the effect diminishes as time progresses.

Other papers have used different methods to answer the same question about drug effectiveness in the long-term. One previous paper [66] looked at the association between use of interferon Beta and disease progression in RRMS patients within the BCMS cohort and found a remarkably similar hazard ratio of 0.77 when compared to our relative rate of progression, 0.72, providing more evidence that the drug slows progression in the long-term. However it should be noted that when using a contemporary control cohort instead of a historical control cohort the hazard ratio was in the opposite direction with little overlap between the 95% confidence intervals. Another paper has reviewed 10 long-term follow up studies, which were generally open-label studies continued after a short term RCT [67]. One study found evidence that interferon beta-1b reduced mortality [68] and another study found that interferon beta-1a slowed progression, gave greater independence and improved quality of life at 15 years [69]. This review concluded that early treatment gives persistent long-term benefits including conversion to clinically definite MS and time to and risk of relapse [67]. Another recent paper showed that patients initially treated with DMTs had a lower risk of converting to SPMS over a median follow-up of 7.6 years [70]. A new study that used data from the BCMS cohort has shown that use of beta interferon for >3 years was associated with increased survival [71]. Data from a new large registry cohort which examined the records of 2466 patients followed up for at least 10 years has shown that use of DMTs was associated

with a smaller increase in EDSS score [72]. A systematic review published in 2016 looking at 14 studies reporting on the long-term effectiveness of interferon or Glatiramer acetate showed a pooled hazard ratio of 0.49 (95% confidence interval 0.34-0.69) to reaching an EDSS of 6 [73]. All providing more evidence that these drugs are effective in the long-term and backing up our findings, although the problem of publication bias whereby negative studies are not published should be considered.

In the discussion of paper P6 we list qualitatively similar findings from previous cluster analysis papers in PD. In particular a previous paper that also looked at disease progression for their data derived subtypes discovered three clusters that were associated with different rates of motor progression [53]. Their diffuse/malignant cluster that progressed the fastest had higher rates of cognitive impairment and orthostatic hypotension. Our fast motor progression cluster also had worse than average cognition and orthostatic hypotension. Another recent paper found three clusters and their fastest progression cluster also had worse cognition and higher SCOPA-AUT scores [74]. The SCOPA-AUT is a questionnaire about autonomic function which includes questions about cardiovascular disfunction and one about orthostatic hypotension.

6.2 Significance of publications

We used a novel methodology to model disease progression in MS. At the time only one other MS longitudinal analysis had used multilevel models and we improved on that approach by considering the best time axis and different trajectories with fractional polynomials whilst also thinking about complex measurement error and removing the effect of relapse and autocorrelation. We have looked at drug effectiveness over a 10 year period using a large cohort of treated patients and a natural history model comparator which has never been done before. Our work on paper P2 has been highly cited with 77 citations (Google Scholar, Aug 2019). The 10 year results in paper P3 has helped inform the current NICE guidelines for treating Multiple Sclerosis with Beta interferons and glatiramer acetate [75]. In particular it was noted that all treatments in the RSS slowed disease progression and that the effect will not stay constant over time, what was called the waning effect. I was an invited plenary speaker at the 26th annual meeting of the European Charcot foundation in

Italy, November 2018 with a talk entitled "The use of multi-level modelling to produce virtual placebo groups for long term assessment of MS drug effectiveness: experience from the risk sharing scheme".

We have carried out a cluster analysis in PD using two cohorts in a development and validation framework and found clusters associated with different rates of motor progression and response to medication. Ours is not only the largest cluster analysis ever in PD but is in early PD, hence more applicability for recently diagnosed patients, and it satisfies all five criteria for future research from a systematic review of cluster analyses in PD [45]. At the time it was one of only three papers that had looked at disease progression of their data derived subtypes [53, 54]. The other papers used a cruder method to consider disease progression by only using the baseline and most recent follow-up measure, whilst our method used all available follow-up measurements and also accounted for withdrawal from the study. We used methods to convert UPSIT scores to Sniffin' scores, as well as validating the conversion, which has never been attempted before. These methods will enable individuals to harmonise data across cohorts for instance carrying out individual patient meta-analysis with cohorts who used different olfaction tests.

Our work on cluster analysis has been presented at many conferences. I created a poster about paper P4 and took this to the Young Statisticians Meeting in 2015. Since then I have created two posters about paper P6. One focused on the main results of the cluster analysis and the subsequent longitudinal analysis of motor disability. The other was more methodological in nature and compared our standard longitudinal analysis to the pattern-mixture model analysis (adjusting for drop-out). I took both posters to a centre PD meeting (a collaboration between the Oxford Parkinson's Disease Centre and two European PD Centres [76]) in March 2018 and out of three poster prizes I won two awards. One got the award for most collaborative project (presumably because the analysis was based on two cohorts) and another the award for most innovative project (presumably because it focused on a novel method in PD to account for withdrawal). The main poster I also took to the University of Bristol Population Health Sciences annual symposium where I was awarded the prize for best poster. In Sept 2018 I took my more methodological poster to the Royal Statistical Society conference in Cardiff. I was an invited plenary speaker at the West Midlands Parkinson's

Network Meeting in June 2019 with a talk entitled “Clinical subtypes in Parkinson’s: Results from the Discovery and Tracking Parkinson’s cohorts”.

6.3 Strengths and limitations

The strengths of our MS papers are that we have developed and validated our natural history model within two observational cohorts from different geographical regions. The RSS cohort is large and incorporates most individuals starting therapy within that window reducing the possibility of selection bias. Our modelling approach, multilevel models, in the RSS was compared to an alternate modelling approach, Markov models, which gave similar answers that the drugs are effective in the long-term. Using imputation and changing how we defined the patient population (for instance including or excluding observations where individuals had switched treatments) also gave us similar answers providing more robust evidence the drugs are effective in the long-term.

A major limitation was the allocation of drugs used in the RSS was not randomized in any way, instead it was decided by clinician and patient choice. If the drug had been randomized then we would now have robust evidence to decide whether any drugs within the scheme perform better. The BCMS study had an average follow-up time of only 5.8 years with only 159/978 (16.3%) of individuals contributing at least 10 years of follow-up. A placebo controlled trial would clearly give better evidence than our approach although it would be difficult to run over 10 years with non-responders likely to drop out thinking they might be on placebo. It is also harder to justify with licensed drugs that are labelled as “disease modifying therapies”, though this was based on strong evidence of a reduction in relapse rate rather than disability. The assumption being made that relapses lead to gradual decline in physical ability and hence this should translate into reduced disability. However, a counter-argument is that relapses reflect a response to primary neuronal degeneration and may be an epi-phenomenon so simply masking relapses may not result in reduced disability.

An alternative approach to looking at the causal effect of a medication would be to carry out a large observational study using inverse probability weighting or propensity score methods [77-79]. These are methods that adjust for the probability of receiving treatment and are less

biased than standard approaches in non-randomised observational studies although these also rely on many assumptions and requires untreated individuals who can be compared to treated individuals. We changed to using time since ABN eligibility after developing our model using time since onset, however this new model had no bias when validated in an external dataset.

There were many other problems and biases with the RSS that were outlined at the inception of the study [14]. One issue was that prevalent cases were included in the study which could bias estimates of drug effectiveness. We examined sensitivity of our conclusions to this by carrying out a sub-group analysis that stratified by recruitment date, thereby effectively removing those who were prevalent cases. Interestingly it was felt that prevalent cases might bias in favour of treatment but this is not what was observed in this sub-group analysis with the drugs seemingly more effective when excluding the prevalent cases. Another concern was that patients might not be followed up once they stop treatment. However, in the year 10 paper we had a mean follow-up of 8.7 years and 9 years or more data for nearly 80% of patients and we also used imputation to account for withdrawal. Two potential biases that were not possible to examine were the lack of blinded assessment of the outcome and the issue of non-randomised comparisons. It was also felt that the EDSS is not that sensitive to change but this is the most widely used measure of disability in MS research [80, 81] and was found to be differ both within and between patients.

The strengths in our PD papers are we had two cohorts with almost identical inclusion/exclusion criteria and similar data collection instruments allowing validation of our modelling approach. These are the largest numbers ever used in a PD cluster analysis. Both cohorts are well phenotyped across a range of motor and non-motor symptoms and the data was harmonised where the questionnaires were different for olfaction. The individuals within this study have more similar disease duration than some other cross-sectional studies. The patients in both cohorts are followed-up every 18 months allowing for a longitudinal analysis. In our longitudinal analyses we used pattern-mixture models to adjust for withdrawals from the cohort which provided similar results. However this method does rely on untestable assumptions about the missing data, that it is normally distributed but we could argue that this

is less relevant with the kind of sample size we were dealing with. We were able to validate our olfaction test conversion method in an external dataset.

One limitation relates to our clustering approach as alternative clustering methods can perform better than k-means in certain circumstances [82] for instance k-means tends to favour spherical clusters which could lead to misclassification if data was skewed within a cluster. It is also very difficult to choose the number of clusters and other individuals with the same data might have decided that a different number of clusters was optimal. We used multiple imputation to adjust for missing data in both cohorts. However there are particular situations, when the data is missing not at random, where an imputed analysis would be biased and a complete case analysis would be unbiased [51]. In the Discovery cohort we had a relatively small proportion completing the levodopa challenge and this test was not carried out at baseline in both cohorts so it was affected by withdrawal. Hence our associations with the levodopa challenge could have selection bias whereby individuals with worse disease who need to drop-out of the study might respond less to levodopa [83]. Some of the questionnaires that we used were designed as screening questionnaires, like the MoCA and the RBD questionnaire[84], whilst we were using them as severity questionnaires which might reduce the potential to identifying differences in these phenotypes across individuals. Also if we had a perfect diagnostic method for picking up other parkinsonian syndromes, like MSA and PSP, and other diseases often mistaken for PD, like dystonic/essential tremor, it would give us the most accurate representation of PD subtypes without enrichment from other diseases. Although the duration of disease in our two cohorts was much more similar than other cross-sectional PD cluster analyses the range of 3.5 years is still relatively large and duration of disease is a big confounding factor in terms of disease severity and could make patients phenotype seem more different than the true difference at a similar time-point. An ideal cohort would only recruit participants within 6-12 months of diagnosis. Also we found statistically significant differences in duration of disease between our clusters which could have impacted on the robustness of our clusters. However we felt this was not clinically significant as the largest difference in mean duration between clusters was only 3.5 months. An alternative approach to remove the confounding effect of disease duration would have been to carry out regressions of each cluster variable against time (with appropriate non-linear time terms if required) and then to use the residuals from this regression in the cluster analysis [85]. However, that approach does require that all of the variables you are considering

are continuous and not categorical which would have meant excluding some variables on constipation and urinary function from our analysis.

Our kappa (0.58) between the actual and predicted clusters in the Discovery cohort is moderate rather than very high. This may be due to the baseline differences between the two cohorts however one might expect this is due to different proportions in subtypes between the two cohorts which is what we observed. It could also be due to inter-rater variability because the Discovery cohort has much fewer centres and fewer raters. Then there is also the possibility that our subtyping approach does not function correctly or that subtypes do not actually exist and there are instead underlying continuous axes of phenotypic variation that contribute to the heterogeneity in PD. However reassuringly we do see very similar patterns of gender, age, motor and cognitive progression across the two cohorts post cluster analysis.

6.4 Ongoing and future research

A project has started looking at the genetics of MS progression in the UoWMS cohort. This project will use Genetic Wide Association Study (GWAS) approaches to discover single SNPs associated with progression, and also two sample Mendelian Randomisation approaches to look at causal effects of different exposures on disease progression. The work I did on developing a model for disease progression will help inform the models of genetic variables on prognosis.

Due to the interest in our work on paper P6, another British PD cohort, ICICLE [86], has asked if we could collaborate and derive our subtypes within their cohort. We are currently harmonizing our data allowing us to derive these subtypes in ICICLE and we are drawing up a formal data sharing agreement. This will allow us to look at the effect of some serum immune biomarkers [87], already assayed in this cohort, against our subtypes whilst further validating our subtypes in another cohort.

We are continuing to look at our validated PD subtypes in relation to both genetics and blood biomarkers (Apolipoprotein-A1, C-reactive protein, Uric Acid and Vitamin D). A paper with

the results of these biomarkers vs. subtypes has recently been accepted for publication by Movement Disorders. We have found evidence that Apolipoprotein-A1 is significantly reduced and C-reactive protein significantly increased within our third severe motor, poor psychological well-being and poor sleep cluster. This provides more evidence that our subtypes have a biological basis. The initial results of the genetic analyses look interesting and we are starting to write up a paper with these results. In this genetic analysis so far we have created a genetic risk score for developing PD using data from a large GWAS of PD vs. controls. When comparing our clusters against this genetic risk score, our third cluster seems to have a lower genetic risk of the disease than clusters 2 and 4. This along with the biomarker associations could suggest that the third cluster has a different etiology than the other clusters. We have also started a project looking at Mendelian Randomisation of PD progression in both the Discovery and Tracking cohorts to try and find different exposures that are causally related to disease progression.

There are plans for us to use a different methodology, growth mixture models [88], to define subtypes. Growth mixture models cluster individuals together who have similar trajectories in a longitudinal analysis. If our main aim is to find clusters associated with disease progression, then using growth mixture models and how these progression clusters relate to baseline characteristics is arguably better than clustering on baseline characteristics and subsequently looking at progression. We are also considering testing our k-means cluster analysis against other cluster analysis methods such as latent profile analysis [89] which is a model-based clustering approach trying to cluster using statistical distributions rather than an algorithm that is used to distinguish clusters. Another future plan is to look at whether we can reduce the number of variables to reliably predict our subtypes which would mean they could be used within clinical practice much faster.

I have been invited, as one of two statisticians, to work on a Movement Disorders taskforce doing a systematic review of Parkinson's disease subtypes due to my experience in this area. My job is to rate the quality of the statistical methods within each paper and to broadly categorise the methods that have been used.

We are currently writing a paper, using the Tracking Parkinson's cohort data, that compares different methods for dichotomizing UPSIT scores and looking at the agreement between these methods. Our conversion allows us to combine the UPSIT and Sniffin' data within that cohort for this analysis.

6.5 Recommendations for future research

How to define the time axis in neurodegenerative diseases is particularly difficult. Age is often very inappropriate because if you start the time at the lowest age, for instance 18 years old, and someone does not get diagnosed until 40 years old then you have 22 years of "dead" space to model. This is not so much an issue in a linear analysis (at least when considering the slope), however if you are deriving non-linear trajectories with fractional polynomials this "dead" space can absorb a lot of the non-linearity, for instance the effect of log time diminishes with time. Time since symptom onset is difficult to determine because it relies on the awareness and memory of the individual. Also in PD motor symptom onset might be easier to remember for someone whose first motor symptom was tremor compared to someone whose first motor symptom was bradykinesia. Time since diagnosis seems the best candidate, in my opinion, however there might also be heterogeneity in someone's disease severity before seeking a diagnosis from a medical professional thus introducing bias in the true time scale of the disease. An alternative is to add a latent variable to the time term in our multilevel models, allowing the heterogeneity in time to be modelled by a random effect, which has been applied to Parkinson's previously [90] and this latent time term seemed to separate well healthy controls, prodromal and PD patients.

Determining non-linearity in progression rates is important as linear models fitted to non-linear trajectories might, for example, over predict at baseline but under predict at the other extremes of time. Using fractional polynomials allows for a simpler and easier to interpret model than methods like cubic splines although at the cost of being less flexible [91]. Our research suggests that simply adding a square time term to a model to deal with non-linearity, which is quite common, may not be the most suitable option and has also been shown before in other contexts [92]. Unfortunately in our PD cohorts we do not currently have the data to look at non-linearity (needs 4 or more time-points on average). Considering complex

measurement error (heteroscedasticity in observation level variation) is also important for multilevel models and can improve model fit and change standard errors. However many standard statistical programs like STATA do not allow for multilevel models that add time terms to observation level random effects. So if researchers find heteroscedasticity of observation level residuals across time they should consider modelling this using a package like MLwiN [93]. Alternatively if researchers are modelling cohorts that have balanced data (time of measurements for each individual is the same) then observational level variation can be freely estimated at each time point using structural equation modelling approaches [94].

The best evidence for subtypes would come from cohorts where all individuals are phenotyped at the same time point, for instance at diagnosis, or using residuals from a regression model including time as cited in the strength and limitations. Better phenotyping of PD cohorts, especially with regards to cognition would assist in deriving subtypes through a cluster analysis, however to get more cognitive data in a very large cohort such as ours is costly and adds to participant burden. If measurements of MDS-UPDRS III can be taken in the "off" state after overnight withdrawal of medication then this would give measurements unconfounded by medication response. However this needs to be weighed up against the distress and discomfort this would give to patients. Newer methods such as using smartphones to measure motor disability will eventually be cheaper and should have far smaller inter-rater variability than current methods. However these smartphone apps are still in their infancy and require further validation until they can replace the current in-clinic questionnaires [95, 96]. Also to reduce the inter-rater variability to the current accuracy of smartphone sensors would require all patients using these apps to follow the protocol correctly which may be hard to achieve in practice.

It is important to validate any statistical model [97-99]. An external validation in an independent distinct cohort is clearly best, as we did with our MS natural history model, our olfaction scale harmonization method and our PD subtypes. If the sample size of a cohort was particularly high then a split-sample approach where part of the dataset is kept aside for validation would suffice. However, this kind of random split-sampling often means that the dataset for development and validation is more similar than an external dataset as the individuals in the development and validation sets are recruited and measured under the same

protocol. There are also other internal validation techniques available like bootstrapping and cross validation [100].

Parkinson's disease has many motor and non-motor features. Most of these features rely on questionnaires completed by the patient or in a consultation with a nurse or clinician. However for each motor and non-motor feature there are often a number of questionnaires to choose from which are generally measured on a scale unique to that questionnaire with different numbers of questions often rated on a different Likert scale. This means that most of the observational PD cohorts have their disease phenotyped on different scales. Data harmonization of these scales will allow more cross cohort collaborations or large individual patient meta-analyses. These type of methods [61] have been applied before to cognitive scales in PD [58-60] allowing multicenter analyses [101, 102]. It will also allow cohorts to create large collaborations giving the kind of numbers required to have decent power in genetic analyses where the effect sizes tend to be small.

One issue with determining causal effects in observational studies is the problem of confounding. It is often difficult to know the confounders in advance of a study and to measure them accurately leading to unmeasured and residual confounding [103]. A relatively new method that overcomes the problem of confounding is to use Mendelian Randomisation (MR) [104, 105]. This method uses genes as an instrumental variable that according to Mendel's Laws should be assigned randomly at birth. Thus creating a natural randomized study where individuals are randomly assigned to have genes that might increase or decrease the risk of the exposure of interest allowing us to find the causal effect of an exposure on an outcome. Two sample MR [106] takes the gene-exposure associations from a genome wide association (GWAS) study and the gene-outcome associations from another study. In the future two sample MR might allow us to determine exposures causally related to disease progression using these MS and PD cohorts. This could assist in finding new treatments that could be tested within an RCT or in counselling of patients. However in practice this does not always work as in the MR study that showed high uric acid was beneficial [107] in PD but the phase 3 clinical trial (SURE-PD3 NCT02642393) testing raising uric acid was ended early due to futility [108]. Using the two sample MR framework also means that as long as there is genetic information on the individuals within a study and there is a GWAS study on a blood

biomarker then this biomarker can be tested against prognosis without the need to assay it in participants within the study which could prove costly.

Another important question to consider is whether risk-sharing schemes should be used in the future for other neurodegenerative diseases. For instance imagine the hypothetical situation where a drug is licensed as a symptomatic medication for one of the many non-motor features of PD but there is ambiguity as to whether it reduces motor disease progression and hence may or may not be cost effective. The major argument for the MS-RSS was a question of ethics and whether there is clinical equipoise. I think there are two ways of viewing this

1. There is clinical equipoise for the question "Is this medication effective in the long-term for reducing disease progression".
2. There is not clinical equipoise for the question "Should this medication be prescribed for a person with this symptom" because it is already licensed for treating that disease.

A long-term randomised controlled trial (RCT) is clearly going to be the best evidence but a risk-sharing scheme is a good alternative for a condition that lasts many years if these ethical issues exist along with the problems of retaining participants within the study who see themselves as non-responders. Whether an RCT would use a placebo or an active comparator would depend on currently available treatment options routinely used in clinical practice. In a risk sharing scheme I think it makes logical sense to develop and validate any natural history model prior to the inception of such a scheme (unlike what happened in the MS-RSS) and to think carefully about how to reduce any of the biases that might be an issue in this scheme. Developing this model is going to be far cheaper than any risk sharing scheme or RCT and if there was a situation where this model did not validate well then you may already have spent millions on the risk sharing scheme without having a decent comparator for your treated cohort. New tools for assessing the risk of bias in non-randomised studies of interventions also exist (ROBINS-I) [109] which should be considered at the beginning of any such study. A previous paper found they could not replicate the results from an RCT using a historical comparator [110], although this study did have a very small sample size.

There would be considerable risks when developing an untreated natural history model for a risk sharing scheme. It could be difficult to obtain the data and would most likely involve obtaining a cohort initiated before any of the drugs in question were available. This would ultimately lead to a potentially biased comparison of historical untreated data with contemporary treated data. Also a major assumption (the unmeasured confounding assumption) of such a study would be that any risk factors for progression are equally distributed within the risk sharing scheme and the historical cohort. The alternative would be identifying all risk factors for progression in advance and including them as covariates within the untreated natural history model. As within any long-term study there would be difficulty in dealing with drop-outs (in both the risk sharing scheme and the observational untreated cohort) which requires using methods that adjust for missing data and/or assuming that the missing data does not have a particular pattern, such as assuming the data isn't missing not at random. Within a long-term placebo controlled trial of a licensed drug it might become difficult to maintain blinding if particular side-effect profiles are well known and documented. An advantage of having an RCT with an active comparator is it would allow indirect comparisons within a network meta-analysis framework [111] provided that active comparator has also been compared with other treatments in RCT's.

Alternatives to a risk sharing scheme or a long-term RCT would include a long-term observational study following up both treated and untreated patients. Such a study could use methods for causal inference in observational studies such as inverse probability weighting that were mentioned in chapter 6.3, although such studies require knowing and being able to measure all confounders. Another alternative would be to have an active comparator (depending on currently available treatment options) within an observational active comparator new user study design [112].

We think that failure to account for PD subtypes in an RCT could bias results towards the null especially if some of the subtypes do not respond as well to a specific medication, which is what we observed with levodopa. This raises important questions about how to design a clinical trial accounting for subtypes and can we find simpler methods to predict a patients subtype. To have the power to detect a clinically important difference within each subtype would require recruiting many more patients to a study with cost implications of potentially

needing four times the number of patients to account for each of the four subtypes. An alternative might be to specifically target a medication to only one of the subtypes, however this then limits the generalizability of the RCT to all PD patients. There is clearly a trade-off between the cost of screening for the sub-types versus the greater efficiency and hence reduced cost of the subsequent RCT if limited to a sub-type. Future work needs to model these options to determine the trade-off point whilst new technologies may result in far reduced costs than currently required for sub-type identification.

PD subtypes might eventually be useful for counselling of patients at diagnosis. However this will require longer follow-up data and firm clinical end-points such as time to developing severe disability or dementia. These kind of end-points would be more familiar to a patient than the annual change in an clinical score that a patient would not understand (such as the MDS-UPDRS III).

6.6 Conclusions

The work presented in this thesis considers prognosis in two neurodegenerative diseases, Multiple Sclerosis and Parkinson's disease. There were six objectives which we have met in the following ways:

1. We developed a multilevel model for the natural history of MS and validated this in an independent cohort. Our modelling approach considered the choice of time axis, the best fitting trajectory, complex observation level variation, and adjustments for relapses and autocorrelation. We found that a model with time since onset as the time axis with both linear and log time terms best modelled the trajectories. Adding a linear time random effect to the observation level variation improved the model fit. Grouping together observations made within a ¼ year time frame removed most of the autocorrelation. We found that removing observations made within three months of start of relapse in the UoWMS cohort (although only one month in BCMS) reduced our observation level variation giving evidence it removed any effect of relapse on the trajectories. Our final natural history model validated very well in the independent

dataset - future measurements predicted from the first measurement were very close to the observed measurements on average.

2. This natural history model was used as a comparator in a large cohort of treated patients followed up for 10 years to determine drug effectiveness in the long-term. Our modelling approach used a complete case analysis along with imputed analyses for individuals who were lost to follow-up. We looked at different outcomes, EDSS and utility scores, carried out sub-group analyses and considered whether the effectiveness changed over time. Our analyses suggest that these disease modifying treatments are effective in the long-term.
3. We derived subtypes of Parkinson's disease using a k-means cluster analysis in a large cohort of individuals who were recently diagnosed with Parkinson's. Our cohort was well phenotyped across important motor and non-motor symptoms and all of these were included within our cluster analysis. The modelling approach first carried out a factor analysis followed by a k-means cluster analysis where we took into account the limitations of the k-means method. We found three factors, one a psychological well-being factor including depression and anxiety, another a non-tremor motor factor including rigidity, bradykinesia and postural instability and a third including cognitive variables. Using clinical and statistical methods we decided that five clusters gave the optimal solution. We carried out an internal validation of our method using cross-validation which showed it was relatively stable. These clusters were associated with response to medication.
4. We harmonized the data across two cohorts of Parkinson's patients to allow us to carry out a cluster analysis using the data from both cohorts. Our modelling approach involved using item-response theory to convert scores from one olfaction test, the UPSIT, to another, the Sniffin. This conversion method performed well when it was validated in an external dataset of PD patients and controls who took both olfaction tests.
5. We used the harmonised data to derive subtypes of the disease using the same approach as before within both cohorts. The larger cohort (Tracking) was used as the development cohort and the smaller (Discovery) was used as the validation. In this analysis we decided that four clusters gave us the optimal solution. Following that we used a discriminant analysis model derived from the clusters in the larger cohort to predict the clusters in the smaller. We then compared the predicted and derived clusters and found our approach to be moderately stable with a kappa of 0.58.

6. We took our validated subtypes and used multilevel models to look at prognosis in motor and cognitive variables. Our subtypes were different in terms of their motor prognosis, and this conclusion remained after adjusting for withdrawal from the cohort using pattern-mixture models. Using data from a levodopa challenge we also found our subtypes were associated with response to levodopa.

Modelling prognosis in neurodegenerative diseases is important due to the substantial heterogeneity in patient progression. The work presented herein details some empirical and methodological results from looking at prognosis in two neurodegenerative diseases. This work adds to the existing literature in demonstrating how one can use more advanced statistical methods to inform the evaluation of observational data on effectiveness as well as how to determine prognostic sub-groups that may have implications for the design of future trials in this area.

7. Statement of contribution to published work

"Parkinson's is my toughest fight. No, it doesn't hurt. It's hard to explain. I'm being tested to see if I'll keep praying, to see if I'll keep my faith. All great people are tested by God."

- *Muhammad Ali*

For paper P1 I cleaned and checked all of the UoWMS cohort data, as noted in chapter 3, and then manipulated both the UoWMS and BCMS data ready for analysis. I helped write the analysis plan and performed all of the statistical analysis myself. I was the first author for this paper and contributed to writing the majority of this paper. For papers P2 and P3 I helped write the analysis plan and performed all of the statistical analysis that was relevant to the multilevel modelling approach. Although I was not the lead author I wrote the methods section relevant to the multilevel modelling approach, helped design the results tables and reviewed the paper. As my work on the MS-RSS was funded by the National Institute for Health Research Health Technology Assessment (NIHR-HTA) there is also a HTA report [12] that I did all the statistical analysis for and designed all the tables and figures. This report was not included as a paper within this thesis because of the duplication of most of the work.

For papers P4 to P6 I cleaned and checked all of the Discovery cohort and Tracking cohort data, as noted in chapter 5. For paper P4 I helped to write the analysis plan and performed all of the statistical analysis. I was the lead author of this paper but required some help from my clinical colleagues to write the discussion section. For P5 I helped to write the analysis plan and carried out all of the statistical analysis. I was also the lead author on this paper. Additionally these harmonization methods were not familiar to any of our collaborators or anyone within my department. Hence all the statistical analysis and programming I did was completely self-taught from reading published papers and books. For P6 I wrote the analysis plan and performed all the statistical analysis. I was the lead author of this paper but again required some clinical input to write the discussion section.

The table below gives a summary of the contributions I made to each paper

Table 3. Summary of my contributions to the published papers

Paper	Cleaned Data	Wrote Analysis Plan	Performed Analysis	Prepared 1st draft	Contribution to writing	Times cited[a]
P1	X (UoWMS cohort only)	X (some assistance)	X	X	90%	14
P2		X (For MLM)	X (For MLM)		15%	77
P3		X (For MLM)	X (For MLM)		15%	5
P4	X	X (some assistance)	X	X	80%	48
P5	X	X (some assistance)	X	X	90%	18
P6	X	X	X	X	80%	11

[a] As of Aug 2019 in Google Scholar

References

1. Mackenzie IS, Morant SV, Bloomfield GA, MacDonald TM, O'Riordan J. Incidence and prevalence of multiple sclerosis in the UK 1990-2010: a descriptive study in the General Practice Research Database. J Neurol Neurosurg Psychiatry 2014;85(1):76-84.
2. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapses and progression of disability in multiple sclerosis. N Engl J Med 2000;343(20):1430-1438.
3. Sadovnick AD, Ebers GC, Wilson RW, Paty DW. Life expectancy in patients attending multiple sclerosis clinics. Neurology 1992;42(5):991-994.
4. Hurwitz BJ. The diagnosis of multiple sclerosis and the clinical subtypes. Ann Indian Acad Neurol 2009;12(4):226-230.
5. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: Results of an international survey. Neurology 1996;46(4):907-911.
6. Scalfari A, Neuhaus A, Degenhardt A, et al. The natural history of multiple sclerosis: a geographically based study 10: relapses and long-term disability. Brain 2010;133(Pt 7):1914-1929.
7. Koch M, Kingwell E, Rieckmann P, Tremlett H. The natural history of primary progressive multiple sclerosis. Neurology 2009;73(23):1996-2002.
8. The IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. . Neurology 1993;43(4):655-661.
9. Comi G, Filippi M, Wolinsky JS. European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging--measured disease activity and burden in patients with relapsing multiple sclerosis. European/Canadian Glatiramer Acetate Study Group. Ann Neurol 2001;49(3):290-297.
10. Jacobs LD, Cookfair DL, Rudick RA, et al. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG). Ann Neurol 1996;39(3):285-294.
11. PRISMS (Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis) Study Group. Randomised double-blind placebo-controlled study of

interferon beta-1a in relapsing/remitting multiple sclerosis. . Lancet 1998;352(9139):1498-1504.

12. Tilling K, Lawton M, Robertson N, et al. Modelling disease progression in relapsing-remitting onset multiple sclerosis using multilevel models applied to longitudinal data from two natural history cohorts and one treated cohort. Health Technol Assess 2016;20(81):1-48.

13. National Institute for Health and Care Excellence (NICE). Beta Interferon and Glatiramer Acetate for the Treatment of Multiple Sclerosis. NICE Technology Appraisal Guidance (TA32). London: NICE; 2002.

14. Sudlow CL, Counsell CE. Problems with UK government's risk sharing scheme for assessing drugs for multiple sclerosis. BMJ 2003;326(7385):388-392.

15. D'Souza M, Kappos L, Czaplinski A. Reconsidering clinical outcomes in Multiple Sclerosis: Relapses, impairment, disability and beyond. Journal of the Neurological Sciences 2008;274(1-2):76-79.

16. Kurtzke JF. Rating Neurologic Impairment in Multiple-Sclerosis - an Expanded Disability Status Scale (Edss). Neurology 1983;33(11):1444-1452.

17. Cutter GR, Baier ML, Rudick RA, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999;122 (Pt 5):871-882.

18. Association of British Neurologists. Revised (2009) Association of British Neurologists Guidelines for Prescribing in Multiple Sclerosis [Available from: https://www.theabn.org/media/docs/ABN%20publications/ABN_MS_Guidelines_2009_Final(1).pdf] accessed 27/06/2016.

19. Tremlett H, Paty D, Devonshire V. The natural history of primary progressive MS in British Columbia, Canada. Neurology 2005;65(12):1919-1923.

20. Tremlett H, Yinshan Z, Devonshire V. Natural history of secondary-progressive multiple sclerosis. Mult Scler 2008;14(3):314-324.

21. Kingwell E, van der Kop M, Zhao Y, et al. Relative mortality and survival in multiple sclerosis: findings from British Columbia, Canada. J Neurol Neurosurg Psychiatry 2012;83(1):61-66.

22. Shirani A, Zhao Y, Kingwell E, Rieckmann P, Tremlett H. Temporal trends of disability progression in multiple sclerosis: findings from British Columbia, Canada (1975-2009). Mult Scler 2012;18(4):442-450.

23. Boggild M, Palace J, Barton P, et al. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. BMJ 2009;339:b4677.

24. Scalfari A, Neuhaus A, Daumer M, Ebers GC, Muraro PA. Age and disability accumulation in multiple sclerosis. Neurology 2011;77(13):1246-1252.
25. Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. Neurology 2010;74(24):2004-2015.
26. Vukusic S, Confavreux C. Natural history of multiple sclerosis: risk factors and prognostic indicators. Curr Opin Neurol 2007;20(3):269-274.
27. Achiron A. Predicting the course of relapsing-remitting MS using longitudinal disability curves. J Neurol 2004;251 Suppl 5:v65-v68.
28. Achiron A, Barak Y, Rotstein Z. Longitudinal disability curves for predicting the course of relapsing-remitting multiple sclerosis. Mult Scler 2003;9(5):486-491.
29. Gauthier SA, Mandel M, Guttmann CR, et al. Predicting short-term disability in multiple sclerosis. Neurology 2007;68(24):2059-2065.
30. Palace J, Bregenzer T, Tremlett H, et al. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. BMJ open 2014;4(1):e004073.
31. Goldstein H. Multilevel statistical models. 4th ed. Chichester, West Sussex: Wiley, 2011.
32. Di Serio C, Lamina C. Investigating Determinants of Multiple Sclerosis in Longitunal Studies: A Bayesian Approach. Journal of Probability and Statistics 2009;2009.
33. Rubin DB. Inference and Missing Data. Biometrika 1976;63(3):581-590.
34. Hernan MA, Robins J.M. Causal Inference: What If.: Boca Raton: Chapman & Hall/CRC., 2020.
35. Elbaz A, Carcaillon L, Kab S, Moisan F. Epidemiology of Parkinson's disease. Rev Neurol (Paris) 2016;172(1):14-26.
36. Evans JR, Cummins G, Breen DP, et al. Comparative epidemiology of incident Parkinson's disease in Cambridgeshire, UK. J Neurol Neurosurg Psychiatry 2016;87(9):1034-1036.
37. Parkinsons UK. The incidence and prevalence of Parkinson's in the UK. Results from the Clinical Practice Research Datalink Reference Report 2018 [Available from: https://www.parkinsons.org.uk/sites/default/files/2018-01/Prevalence%20%20Incidence%20Report%20Latest_Public_2.pdf] accessed 23/12/2018.
38. Bach JP, Ziegler U, Deuschl G, Dodel R, Doblhammer-Reiter G. Projected numbers of people with movement disorders in the years 2030 and 2050. Mov Disord 2011;26(12):2286-2290.

39. Macleod AD, Taylor KS, Counsell CE. Mortality in Parkinson's disease: a systematic review and meta-analysis. Mov Disord 2014;29(13):1615-1622.
40. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 1992;55(3):181-184.
41. Schrag A, Ben-Shlomo Y, Quinn N. How valid is the clinical diagnosis of Parkinson's disease in the community? J Neurol Neurosurg Psychiatry 2002;73(5):529-534.
42. Aarsland D, Zaccai J, Brayne C. A systematic review of prevalence studies of dementia in Parkinson's disease. Mov Disord 2005;20(10):1255-1263.
43. Emre M. Dementia associated with Parkinson's disease. Lancet Neurol 2003;2(4):229-237.
44. Graham JM, Sagar HJ. A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. Mov Disord 1999;14(1):10-20.
45. van Rooden SM, Heiser WJ, Kok JN, Verbaan D, van Hilten JJ, Marinus J. The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. Mov Disord 2010;25(8):969-978.
46. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry 1988;51(6):745-752.
47. Szewczyk-Krolikowski K, Tomlinson P, Nithi K, et al. The influence of age and gender on motor and non-motor features of early Parkinson's disease: Initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. Parkinsonism & related disorders 2013;20(1):99-105.
48. Malek N, Swallow DM, Grosset KA, et al. Tracking Parkinson's: Study Design and Baseline Patient Data. J Parkinsons Dis 2015;5(4):947-959.
49. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23(15):2129-2170.
50. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53(4):695-699.
51. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019;48(4):1294-1304.

52. van Rooden SM, Colas F, Martinez-Martin P, et al. Clinical subtypes of Parkinson's disease. Mov Disord 2011;26(1):51-58.
53. Fereshtehnejad SM, Romenets SR, Anang JB, Latreille V, Gagnon JF, Postuma RB. New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phenotypes. JAMA Neurol 2015;72(8):863-873.
54. Fereshtehnejad SM, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. Brain 2017;140(7):1959-1976.
55. Ma LY, Chan P, Gu ZQ, Li FF, Feng T. Heterogeneity among patients with Parkinson's disease: Cluster analysis and genetic association. Journal of the Neurological Sciences 2015;351(1-2):41-45.
56. Doty RL, Shaman P, Dann M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiol Behav 1984;32(3):489-502.
57. Wolfensberger M, Schnieper I, Welge-Lussen A. Sniffin'Sticks: a new olfactory test battery. Acta Otolaryngol 2000;120(2):303-306.
58. Lawton M, Kasten M, May MT, et al. Validation of conversion between mini-mental state examination and montreal cognitive assessment. Mov Disord 2016;31(4):593-596.
59. van Steenoven I, Aarsland D, Hurtig H, et al. Conversion between mini-mental state examination, montreal cognitive assessment, and dementia rating scale-2 scores in Parkinson's disease. Mov Disord 2014;29(14):1809-1815.
60. Armstrong MJ, Duff-Canning S, Psych C, Kowgier M, Marras C. Independent application of montreal cognitive assessment/mini-mental state examination conversion. Mov Disord 2015;30(12):1710-1711.
61. Kolen MJ, Brennan RL. Test equating, scaling, and linking: second edition: Springer, 2004.
62. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255-268.
63. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 1997;2(1):64-78.
64. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174.
65. Signori A, Izquierdo G, Lugaresi A, et al. Long-term disability trajectories in primary progressive MS patients: A latent class growth analysis. Mult Scler 2018;24(5):642-652.

66. Shirani A, Zhao YS, Karim ME, et al. Association Between Use of Interferon Beta and Progression of Disability in Patients With Relapsing-Remitting Multiple Sclerosis. Jama-J Am Med Assoc 2012;308(3):247-256.
67. Hartung HP, Graf J, Kremer D. Long-term follow-up of multiple sclerosis studies and outcomes from early treatment of clinically isolated syndrome in the BENEFIT 11 study. J Neurol 2019.
68. Goodin DS, Reder AT, Ebers GC, et al. Survival in MS: a randomized cohort study 21 years after the start of the pivotal IFNbeta-1b trial. Neurology 2012;78(17):1315-1322.
69. Bermel RA, Weinstock-Guttman B, Bourdette D, Foulds P, You X, Rudick RA. Intramuscular interferon beta-1a therapy in patients with relapsing-remitting multiple sclerosis: a 15-year follow-up study. Mult Scler 2010;16(5):588-596.
70. Brown JWL, Coles A, Horakova D, et al. Association of Initial Disease-Modifying Therapy With Later Conversion to Secondary Progressive Multiple Sclerosis. JAMA 2019;321(2):175-187.
71. Kingwell E, Leray E, Zhu F, et al. Multiple sclerosis: effect of beta interferon treatment on survival. Brain 2019;142(5):1324-1333.
72. Fyfe I. Multiple sclerosis: Real-world long-term benefits of disease-modifying MS therapy. Nat Rev Neurol 2016;12(7):372.
73. Signori A, Gallo F, Bovis F, Di Tullio N, Maietta I, Sormani MP. Long-term impact of interferon or Glatiramer acetate in multiple sclerosis: A systematic review and meta-analysis. Mult Scler Relat Disord 2016;6:57-63.
74. Zhang X, Chou J, Liang J, et al. Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. Sci Rep 2019;9(1):797.
75. National Institute for Health and Care Excellence (NICE). Beta Interferon and Glatiramer Acetate for the Treatment of Multiple Sclerosis. NICE Technology Appraisal Guidance (TA527). London: NICE; 2018.
76. University of Luxembourg Centre-PD 2019 [Available from: https://www.centre-pd.lu/] accessed 15/09/2019.
77. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011;46(3):399-424.
78. Austin PC, Stuart EAJSim. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. 2015;34(28):3661-3679.
79. Mansournia MA, Altman DGJB. Inverse probability weighting. 2016;352:i189.

80. Hoogervorst EL, Eikelenboom MJ, Uitdehaag BM, Polman CH. One year changes in disability in multiple sclerosis: neurological examination compared with patient self report. J Neurol Neurosurg Psychiatry 2003;74(4):439-442.
81. Multiple Sclerosis Trust. Expanded Disability Status Scale (EDSS) 2018 [Available from: https://www.mstrust.org.uk/a-z/expanded-disability-status-scale-edss] accessed 25/10/2019.
82. Raykov YP, Boukouvalas A, Baig F, Little MA. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS One 2016;11(9):e0162259.
83. Malek N, Kanavou S, Lawton MA, et al. L-dopa responsiveness in early Parkinson's disease is associated with the rate of motor progression. Parkinsonism Relat Disord 2019.
84. Stiasny-Kolster K, Mayer G, Schafer S, Moller JC, Heinzel-Gutenbrunner M, Oertel WH. The REM sleep behavior disorder screening questionnaire--a new diagnostic instrument. Mov Disord 2007;22(16):2386-2393.
85. Colas F, Meulenbelt I, Houwing-Duistermaat JJ, et al. Reliability of cluster results for different types of time adjustments in complex disease research. Conf Proc IEEE Eng Med Biol Soc 2008;2008:4601-4604.
86. Yarnall AJ, Breen DP, Duncan GW, et al. Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study. Neurology 2014;82(4):308-316.
87. Williams-Gray CH, Wijeyekoon R, Yarnall AJ, et al. Serum immune markers and disease progression in an incident Parkinson's disease cohort (ICICLE-PD). Mov Disord 2016;31(7):995-1003.
88. Ram N, Grimm KJ. Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups. Int J Behav Dev 2009;33(6):565-576.
89. Oberski D. Mixture models: Latent profile and latent class analysis. Modern statistical methods for HCI: Springer, 2016:275-287.
90. Iddi S, Li D, Aisen PS, et al. Estimating the Evolution of Disease in the Parkinson's Progression Markers Initiative. Neurodegener Dis 2018;18(4):173-190.
91. Binder H, Sauerbrei W, Royston PJSiM. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. 2013;32(13):2262-2277.
92. Long J, Ryoo JJBJoM, Psychology S. Using fractional polynomials to model non‐linear trends in longitudinal data. 2010;63(1):177-203.

93. Rasbash J, Browne W, Goldstein H, et al. A user's guide to MLwiN. 2000;286.
94. Bollen KA, Curran PJ. Latent curve models: A structural equation perspective: John Wiley & Sons, 2006.
95. Lo C, Arora S, Baig F, et al. Predicting motor, cognitive & functional impairment in Parkinson's. Ann Clin Transl Neurol 2019;6(8):1498-1509.
96. Arora S, Baig F, Lo C, et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. Neurology 2018;91(16):e1528-e1538.
97. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(4):453-473.
98. Mayer DG, Butler DG. Statistical Validation. Ecol Model 1993;68(1-2):21-32.
99. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245-247.
100. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. Brit Med J 2009;338.
101. Mestre TA, Pont‐Sunyer C, Kausar F, et al. Clustering of motor and nonmotor traits in leucine‐rich repeat kinase 2 G2019S Parkinson's disease nonparkinsonian relatives: A multicenter family study. 2018;33(6):960-965.
102. Van Steenoven I, Aarsland D, Weintraub D, et al. Cerebrospinal fluid Alzheimer's disease biomarkers across the spectrum of Lewy body diseases: results from a large multicenter cohort. 2016;54(1):287-295.
103. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol 2007;166(6):646-655.
104. Smith GD, Ebrahim S. Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. Biosocial Surveys: National Academies Press (US), 2008.
105. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601.
106. Lawlor DAJIjoe. Commentary: Two-sample Mendelian randomization: opportunities and challenges. 2016;45(3):908.
107. Simon KC, Eberly S, Gao X, et al. Mendelian randomization of serum urate and parkinson disease progression. Ann Neurol 2014;76(6):862-868.

108. The Cure Parkinson's Trust. Inosine Trial to End Early [Available from: https://www.cureparkinsons.org.uk/news/inosine-ends-early] accessed 27/10/2019.
109. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
110. Ali M, Juttler E, Lees KR, et al. Patient outcomes in historical comparators compared with randomised-controlled trials. Int J Stroke 2010;5(1):10-15.
111. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. Intern Emerg Med 2017;12(1):103-111.
112. Lund JL, Richardson DB, Sturmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. Curr Epidemiol Rep 2015;2(4):221-228.

Journal of Clinical Epidemiology 68 (2015) 1355−1365

Journal of Clinical Epidemiology

A longitudinal model for disease progression was developed and applied to multiple sclerosis

Michael Lawton[a,*], Kate Tilling[a], Neil Robertson[b], Helen Tremlett[c], Feng Zhu[c], Katharine Harding[b], Joel Oger[c], Yoav Ben-Shlomo[a]

[a]*School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, BS8 2PS, UK*
[b]*Institute of Psychological Medicine and Clinical Neuroscience, Cardiff University, University Hospital of Wales, Cardiff, CF14 4XN, UK*
[c]*Faculty of Medicine (Neurology), UBC Hospital, 2211 Wesbrook Mall, University of British Columbia, Vancouver, BC V6T 2B5 Canada*

Accepted 5 May 2015; Published online 14 May 2015

Abstract

Objectives: To develop a model of disease progression using multiple sclerosis (MS) as an exemplar.

Study Design and Settings: Two observational cohorts, the University of Wales MS (UoWMS), UK (1976), and British Columbia MS (BCMS) database, Canada (1980), with longitudinal disability data [the Expanded Disability Status Scale (EDSS)] were used; individuals potentially eligible for MS disease-modifying drugs treatments, but who were unexposed, were selected. Multilevel modeling was used to estimate the EDSS trajectory over time in one data set and validated in the other; challenges addressed included the choice and function of time axis, complex observation-level variation, adjustments for MS relapses, and autocorrelation.

Results: The best-fitting model for the UoWMS cohort (404 individuals, and 2,290 EDSS observations) included a nonlinear function of time since onset. Measurement error decreased over time and ad hoc methods reduced autocorrelation and the effect of relapse. Replication within the BCMS cohort (978 individuals and 7,335 EDSS observations) led to a model with similar time (years) coefficients, time [0.22 (95% confidence interval {CI}: 0.19, 0.26), 0.16 (95% CI: 0.10, 0.22)] and log time [−0.13 (95% CI: −0.39, 0.14), −0.15 (95% CI: −0.70, 0.40)] for BCMS and UoWMS, respectively.

Conflict of interest: M.L. had his expenses paid by the MS Trust to attend a meeting of the UK MS RSS scientific advisory group to outline the plan for these analyses and also his travel and accommodation expenses for visiting Vancouver to analyze the BCMS data set; M.L. received support from MS trust and funding from NIHR HTA. K.T. had her expenses paid by the MS Trust to attend a meeting of the UK MS RSS scientific advisory group outline the plan for these analyses. K.T. is a principal investigator on grant from NIHR HTA to develop the MS model. N.R. has declared no conflict of interests. H.T. is funded by the Multiple Sclerosis Society of Canada (Don Paty Career Development Award), is a Michael Smith Foundation for Health Research Scholar, and is the Canada Research Chair for Neuroepidemiology and Multiple Sclerosis. H.T. has received research support from the National Multiple Sclerosis Society, Canadian Institutes of Health Research, and UK MS Trust; speaker honoraria and/or travel expenses to attend conferences from the Consortium of MS Centres (2013), the MS Society of Canada, endMS Summer School (2012 and 2014), the National MS Society (2012 and 2014), Bayer Pharmaceutical (speaker, 2010, honoraria declined), Teva Pharmaceuticals (speaker 2011), ECTRIMS (2011, 2012, and 2013), UK MS Trust (2011), the Chesapeake Health Education Program, US Veterans Affairs (2012, honorarium declined), Novartis Canada (2012), Biogen Idec (2014, honorarium declined), and American Academy of Neurologists (annual meeting speaker, 2013, 2014, honorarium declined). Unless otherwise stated, all speaker honoraria are either donated to an MS charity or to an unrestricted grant for use by her research group. F.Z. has declared no conflict of interest. K.H. has declared no conflict of interest. J.O. has carried out consultancy work for and obtained grants to run clinical trials (both unrelated to this work) from Bayer, BIOGEN-IDEC, EMD Serono, Novartis, Aventis, and Teva-Neuroscience. Y.B-.S. had his expenses paid by the MS Trust to attend a meeting of the UK MS RSS scientific advisory group to outline the plan for these analyses and has a relative with MS who is currently on treatment for the disease. Y.B-.S. is a coapplicant on grant from NIHR HTA to develop the MS model.

Funding: The BCMS database has been funded through various grants over the years, including, Canadian Institutes of Health Research, the MS Society of Canada, US National MS Society, UK MS Trust, MS/MRI Research group, and unrestricted grants from Dr Donald Paty. This project was funded by the NIHR Health Technology Assessment programme (project number 10/55) and will be published in full in the Health Technology Assessment journal series. Visit the HTA program Web site for more detailswww.hta.ac.uk/link to project page. This report presents independent research commissioned by the National Institute for Health Research (NIHR). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NHS, the NIHR, MRC, CCF, NETSCC, the HTA programme or the Department of Health.

* Corresponding author. Tel.: +44-117-9287255; fax: +44-117-9287325.

E-mail address: Michael.Lawton@bristol.ac.uk (M. Lawton).

http://dx.doi.org/10.1016/j.jclinepi.2015.05.003

Conclusion: It is possible to develop robust models of disability progression for chronic disease. However, explicit validation is important given the complex methodological challenges faced.

Keywords: Multiple sclerosis; Repeated measures model; Multilevel model; Fractional polynomials; Prognosis; Observational cohorts

1. Introduction

Prognostic models for chronic diseases [1] are needed to guide management decisions and counseling of patients and their families. Such models can consider outcomes ranging from treatment response to changes in disability [2−4] and may model individual disease trajectories. This poses technical challenges because progression may be nonlinear, the outcome measure(s) may not be continuous or normally distributed, and individuals may have been observed a different number of times and at irregular intervals.

One example in which modeling long-term trajectories poses challenges is multiple sclerosis (MS). MS is a chronic inflammatory neurodegenerative disorder, with considerable interindividual variation in the disease course. Most patients present with relapsing-remitting MS (RRMS), in which symptoms appear for a varying amount of time and then disappear (either partially or completely). However, over time individuals with RRMS can progress to secondary progressive MS (SPMS), where the frequency of relapses decreases and the accumulation of disability increases steadily [5].

Disability in individuals with MS is commonly measured using the Expanded Disability Status Scale (EDSS) [6]. EDSS is an ordinal scale, based on a neurologist's examination, ranging from 0 (normal neurologic examination) to 10 (death due to MS) in half unit increments (but there is no score of 0.5). Previous studies of EDSS progression have used survival analysis [7−9], considering the time to specific milestones, for example, an EDSS score of 6, which is equivalent to needing an aid to walk. This ignores available data both before and after reaching the milestone and therefore fails to differentiate two individuals reaching a milestone at the same time but with different trajectories.

Empirical percentiles derived at yearly intervals and data-smoothing techniques have been used to create disability curves over time at different percentiles [10,11]. These methods do not model how a given individual changes over time or the relationship between the centiles and patient characteristics.

Markov models have been used to relate progression in MS to age and disease duration as well as other baseline covariates [12,13]. However, such models assume that further progression essentially depends only on the previous measurement and may be less able to cope with issues such as missing data and the need for imputation.

An alternative approach is using multilevel repeated measure models where observations are clustered within individuals [14]. We have used multilevel models to model disability after stroke [4,15], and prostate-specific antigen changes in men with localized prostate cancer [16,17]. Such models could account for both within and between patient variability of the EDSS measurements in MS. These multilevel models are ideal to analyze unbalanced data, that is, where observations are unequally spaced in time and differ in number between individuals. Multilevel models have been used to model the accumulation of disability in MS using a transformation of EDSS [18], assuming a quadratic curve for each individual and ignoring observation-level (within individual) variation over time. Our aim was to develop a generalizable model for the natural history of patients with relapsing-onset MS in two independent data sets who were not treated with any specific disease modifying therapy (DMT) for MS but who would have been eligible for a DMT. This was to facilitate future comparisons with long-term cohorts of DMT-treated patients, such as the UK MS risk sharing scheme [13,19]. Here, we report how we have approached a variety of analytical challenges and our proposed solutions for the development of our natural history (untreated) model of MS disease progression.

2. Methods

2.1. Study design and settings

We used data from the University of Wales MS (UoWMS) cohort, United Kingdom, and the British Columbia MS (BCMS) database, Canada, to develop and validate the model.

2.1.1. UoWMS cohort

The University Hospital of Wales is the major tertiary referral center for neurology in Wales, United Kingdom, serving a local population of 1.2 million and provides a network of MS clinics across South East Wales. Data were initially collected in a cross-sectional study in 1985 [20] and were updated periodically [21,22], until 2002 when data were essentially collected prospectively [23]. Sociodemographic and clinical features at disease onset are recorded in a standardized fashion, including degree of recovery and initial interrelapse interval. Approximately 1,000 patient contacts are documented annually, and clinical data, including EDSS scores, are collected routinely at presentation and at each visit. The database, at the time of extraction, had around 2,000 registered MS patients with

What is new?

- Considering different functions of time to find the best-fitting trajectory of the Expanded Disability Status Scale (EDSS).
- Accounting for nonconstant measurement error in the multilevel model of EDSS.
- Adjusting data to avoid including the potential confounding effects of short-term disease fluctuations (i.e., multiple sclerosis relapses) on an individual's background longer-term disability outcomes.
- Investigating autocorrelation and suitable approaches to minimize this.
- Replicating and cross-validating the model in an independent cohort.

1,283 and 809 patients having at least 2 or 4 or more EDSS scores over time, respectively.

2.1.2. BCMS cohort

BCMS database [8,24–30], established in 1980, is population based, estimating to capture 80% of the BCMS population. Strengths of the BCMS database include longitudinal follow-up of both DMT-treated and untreated patients, and consistent care provided by the same four core neurologists who have examined over 85% of the patients considered for this study. EDSS scores are recorded after a face-to-face consultation with an MS specialist neurologist. As of 2009, the database contained records for over 5,900 MS patients spanning 28 years (>25,000 cumulative years) of prospective follow-up, from four MS clinics in British Columbia.

2.1.3. Eligibility criteria

We included patients in either cohort (UoWMS and BCMS) if they ever became eligible for DMTs. This eligibility was according to the 2001 Association of British Neurologists (ABN) criteria for interferon beta and glatiramer acetate (IFN-β/GA) use (adapted from online supplementary appendix IV Health Service Circular 2002/2004), defined as: aged ≥18 years, EDSS ≤6.5, and had ≥2 relapses during the previous 2 years. Similar criteria are broadly adopted in other legislative areas as well as the United Kingdom and British Columbia, Canada. All EDSS observations before a patient reaching the ABN eligibility criteria were excluded.

A relapse was defined as worsening neurologic symptoms lasting >24 hours, in the absence of fever or infection. The starting date of each relapse was recorded by an MS specialist neurologist.

As we wished to model the natural history of MS, we truncated the patient profiles once a DMT was initiated. In addition, in the BCMS cohort, the data were truncated to 1995, the last year in which the DMTs were not widely available in British Columbia. This was to avoid "indication bias" whereby a patient's trajectory may influence the decision as to whether they started a DMT [27].

2.2. General multilevel model

We modeled the EDSS scores of individuals with MS using multilevel models [14]. Our model had two levels: observations (level 1) within individuals (level 2). A simple multilevel repeated measure model is a linear random intercept and random slope model. This type of model estimates a linear population mean along with a specific line for each individual. A graphical representation of the model shown below is given in Fig. 1.

$$y_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i}) \cdot t_{ij} + e_{ij}, \quad (1)$$

where y_{ij} and t_{ij} are the EDSS score and the time variable for the ith individual at the jth time point. Hence, $\beta_0 + u_{0i}$ is the ith individual's baseline EDSS, whereas β_0 is the mean baseline EDSS, and $\beta_1 + u_{1i}$ is the ith individual's slope over time, whereas β_1 is the mean slope over time.

The u_{ki} ($k = 0, 1$) is often referred to as the individual-level random effects and the e_{ij} as the observation-level

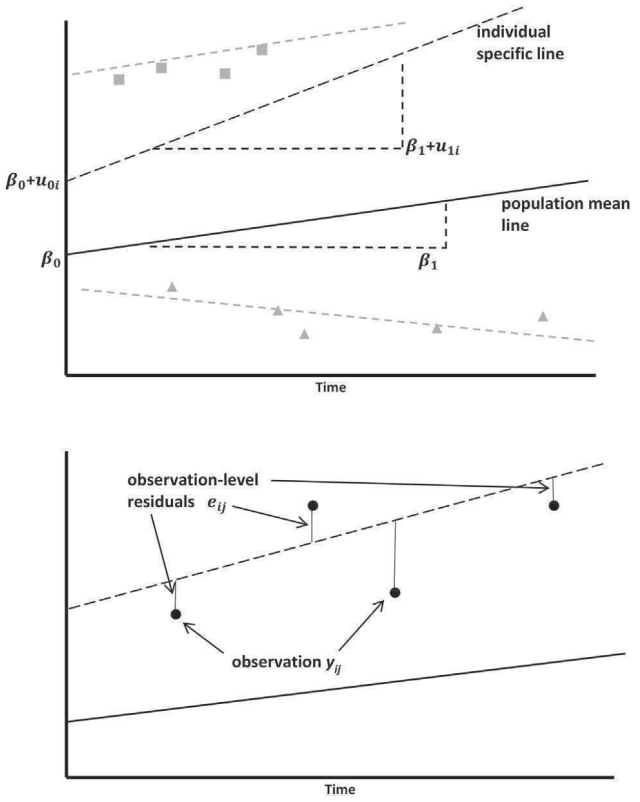


Fig. 1. Graphical representation of the simple multilevel model with linear random intercept and random slope model as shown within Equation (1). Squares are subject 1, circles subject 2, and triangles subjects 3.

random effects, whereas the β_k ($k = 0, 1$) are the fixed effects. Conceptually, the individual-level random-effects measure the deviation of the individual-specific line from the population mean line, and the observation-level random-effects measure the deviation of observations about the individual-specific line.

The e_{ij} is assumed normally distributed with mean zero and variance σ^2, and u_{0i} and u_{1i} are assumed normally distributed with mean zero and an unstructured covariance matrix D_u.

Removing u_{1i} from the equation would give us a linear random intercept model, and removing both u_{0i} and u_{1i} would give us a linear fixed effects only model.

2.3. Developing the model

We initially developed our model using the UoWMS data set and then used the BCMS data set for replication. The model was originally developed in UoWMS (the smaller of the two cohorts) because access to the BCMS data set was only possible at the University of British Columbia, Canada. We cross-validated each model using the other data set.

2.3.1. Choice of time axis

We considered modeling EDSS as a function of either the age of an individual or the time since onset of MS at each observation [8,24]. It is important to center the time axis at a meaningful time point such as the minimum age of the onset (18 years) or zero for the time since onset. Models with different types of time are nonnested but use the same data, so the Akaike information criterion (AIC) was used to compare the models, selecting the model with the lower AIC. In addition, we considered the root mean square error (RMSE) for the difference between the individual-specific predicted EDSS and observed EDSS, and the proportion of these differences that were less than 0.5 or more than or equal to 2 EDSS points.

2.3.2. Choice of function of time

The next model choice was the best-fitting trajectory of EDSS. A simple multilevel model (see Equation 1) allows a random intercept and slope for each individual. Options for more complex models include fractional polynomials to choose the best-fitting curve [4,31], fitting cubic (or linear) splines, finding a transformation of the outcome or time axis (or both) which have a linear relationship, or smoothing. We used fractional polynomials [4] to find the best-fitting trajectory because these require all data to be positive we added one to time. This procedure, see Web Appendix A at www.jclinepi.com, tests what functions of time best represent the individual trajectories of EDSS over time. Fractional polynomials have the advantage that the model has a simple algebraic form. Linear splines also have a simple form but assume biologically implausible piecewise linear growth [32]. Cubic (or other complex) splines, although more flexible, are more difficult for prediction purposes than fractional polynomials as they require estimation of the curves between each knot point rather than one single global curve. Also, all splines involve the selection of knot points, which would further complicate the multilevel model. Having selected the best-fitting fractional polynomial for each time axis, we then compared these two models using the same criteria as above.

As sensitivity analyses, we repeated the fractional polynomial procedure on a restricted data set, only including observations made within 30 and then 15 years from MS symptom onset to check to what extent outliers were influencing the choice of trajectory.

The model with both the best-fitting time axis and function of that time-axis is referred to hereafter as the ‘‘best-fitting’’ simple model.

2.3.3. Observation-level variation

Observation-level variation is the extent to which EDSS observations on a given individual at any one time are likely to differ and can be considered as a mixture of measurement error and within-person fluctuation, which will change over time, because there is greater interrater and intrarater variability for lower EDSS values [33,34]. To examine complex measurement error empirically, we plotted the observation-level residuals against time for the best-fitting simple model. Fractional polynomials were then used to obtain the best-fitting function of time for the observation-level variance, in the same way as described in the previous section, see Web Appendix B at www.jclinepi.com. The best-fitting simple model, with the addition of the selected best-fitting observation-level variance function, is referred to as the ‘‘complex’’ model.

2.3.4. Autocorrelation

Autocorrelation occurs when measures on the same individual are correlated more than would be implied by the overall within-individual correlation. We investigated autocorrelation by examining the association in lagged differences between observations and the individual-specific predictions. A large correlation coefficient for these lagged differences can indicate autocorrelation.

As an ad hoc method to reduce autocorrelation, we divided each individual’s time axis into quarter year intervals. If there was more than one observation within that interval, a new observation was created by taking the median time and the median EDSS score of all the observations within that interval.

Other possible methods to take into account autocorrelation would be autoregressive-moving average models. We could also have created a more complex model with an autocorrelation parameter that is a function of the time between observations within an individual [35]. Other methods to account for autocorrelation usually require data to be balanced, and more research is necessary to

incorporate methods for measuring autocorrelation in unbalanced repeated measure models.

2.3.5. Relapses

Our focus was to model the true accumulation of disability over time and avoid short-term disability secondary to an acute relapse. Consequently, all EDSS observations recorded within 1 month postrelapse were removed from both data sets. However, some patients may continue to improve in physical disability beyond the 1-month postrelapse window [36]. Therefore, we carried out sensitivity analyses using the complex model taking account of autocorrelation, by also removing all observations within 3 and then 6 months after a documented relapse.

2.3.6. Assessing model assumptions and fit

We assessed the normality of the residuals by using QQ plots and the fit of the model by comparing the actual and predicted EDSS values. All analyses were carried out using Stata software (Texas, USA) [37], and all multilevel models were estimated by the runmlwin command [38].

2.3.7. Indication bias

We could bias our results by censoring individual observations after they started treatment because individuals who start treatment might differ to those who never started treatment. We avoided ''indication bias'' in the BCMS cohort by truncating the data at 1995. However, because of the smaller size of the UoWMS cohort, we instead tested for indication bias by including ''starting a DMT'' as a covariate within the multilevel model.

2.3.8. Conditional predictions for cross-validation

We carried out external validation using the model from one cohort to predict the data from the other cohort, by predicting future trajectory based on the first observed EDSS score for each individual [4,39]. We used the BCMS model to predict the UoWMS data and the UoWMS model to predict the BCMS data. The EDSS scores were predicted using the complex model accounting for autocorrelation and relapses.

3. Results

Table 1 shows the characteristics of the MS patients included in the two cohorts, that is, those reaching eligibility for drug treatment. The BCMS data set included more than twice as many individuals as the UoWMS and had a larger number of EDSS observations per person. However, the patient characteristics were similar for sex and proportion ever starting a DMT. For age and disease duration at ABN eligibility and the proportion of patients with SPMS at ABN eligibility, there was some moderate evidence of a difference between data sets (*P*-values between 0.018 and 0.052) although these differences were small. BCMS patients were on average 2 years younger at the onset than UoWMS patients, although the longer disease duration at ABN eligibility meant that they were only 1.3 years younger on average at ABN eligibility. A higher proportion of the BCMS cohort reached secondary progressive disease

Table 1. Patient demographics of all those eligible[a] for disease-modifying drug treatment within the two multiple sclerosis cohorts from British Columbia, Canada, and the University of Wales, United Kingdom, with all observations made within 1 month postrelapse removed

Mean (SD; range) or *N* (%) unless otherwise stated	British Columbia, Canada	University of Wales, United Kingdom	*P*-value difference
N	978	404	
Number of EDSS observations; mean per person(range)	7,335; 7.5 (1–73)	2,290; 5.7 (1–72)	
Females	728 (74.4%)	306 (75.7%)	0.611[b]
Age at the onset, yr	29.1 (8.6; 3.4–61.1)	31.1 (8.7; 13.4–60.0)	<0.001[c]
Age at eligibility, yr	37.3 (9.3; 18.1–7.0)	38.6 (9.1; 18.8–80.1)	0.018[c]
Disease duration at eligibility, yr	8.2 (6.9; 0.2–38.9)	7.4 (7.1; 0.5–43.8)	0.052[c]
SPMS reached by eligibility date	150 (15.3%)	83 (20.5%)	0.019[b]
Ever reached SPMS[d]	563 (57.6%)	139 (34.4%)	<0.001[b]
Relapses in 2 years before eligibility: median(quartiles; range)	2.9 (1.2; 2–9)	3.5 (0.9; 2–9)	
EDSS at eligibility: median(quartiles; range)	2 (1,3.5; 0–6.5)	3.5 (2,4.5; 0–6.5)	
Year of EDSS at eligibility: range	1980–1995	1976–2011	
Year of last EDSS included in the present study: range	1981–1995	1984–2011	
Prospective follow-up time[d], yr (first eligible EDSS to last DMT-free EDSS)	5.8 (3.8, 0–15)	2.98 (3.9, 0–29.3)	<0.001[c]
Prospectively followed[d], ≥5 years	560 (57.3%)	92 (22.8%)	<0.001[b]
Prospectively followed[d], ≥10 years	159 (16.3%)	16 (4.0%)	<0.001[b]
Time between observations, yr	0.9 (1.0; 0.0, 11.3)	0.6 (1.2; 0.0, 21.3)	<0.001[c]
Ever prescribed a DMT[e]	232 (23.7%)	109 (27.0%)	0.201[b]

Abbreviations: SD, standard deviation; EDSS, Expanded Disability Status Scale; SPMS, secondary progressive MS; DMT, disease modifying therapy; BCMS, British Columbia MS.

[a] Using the Association of British Neurologists (ABN) criteria.

[b] Chi-squared test.

[c] *T*-test.

[d] For the BCMS cohort only includes observations made within the dates of the truncated data set, that is, 1980–1995.

[e] Includes DMT exposure up until 2011 in the BCMS cohort, that is, beyond the 1980–1995 window.

Table 2. Akaike information criterion (AIC), root mean square error (RMSE), and percentage of observations within 0.5 EDSS and 2 or more EDSS of all the models fitted using the difference between the observations and individual-level predictions[a]

Model	AIC	RMSE individual-level predictions	% (*N*) within ±0.5 EDSS individual-level predictions	% (*N*) ≥2 EDSS difference individual-level predictions
Linear fixed effects only (age)	9496.11	1.92	17.6 (404/2,290)	32.3 (739/2,290)
Linear random intercept (age)	6549.66	0.71	61.8 (1,416/2,290)	1.9 (43/2,290)
Linear random slope and intercept (age)	6256.66	0.59	69.7 (1,595/2,290)	0.7 (15/2,290)
Linear fixed effects only (time since onset)	9385.02	1.88	15.6 (357/2,290)	31.9 (731/2,290)
Linear random intercept (time since onset)	6525.80	0.71	62.1 (1,421/2,290)	1.7 (38/2,290)
Linear random slope and intercept (time since onset)	6203.40	0.58	70.7 (1,619/2,290)	0.7 (16/2,290)
$\sqrt{t}$ log t (time since onset)	6063.25	0.55	71.2% (1,630/2,290)	0.5 (11/2,290)
t, log t (time since onset)	6066.72	0.54	72.5 (1,660/2,290)	0.5 (11/2,290)
t, log t (time since onset), adding t to observation-level variance	6013.40	0.55	71.2 (1,630/2,290)	0.5 (11/2,290)
t, log t (time since onset) adding t to observation-level variance with restriction Var(t) = 0	6018.23	0.55	72.1 (1,651/2,290)	0.5 (11/2,290)

Abbreviation: EDSS, Expanded Disability Status Scale.

[a] Individual-level predictions are the fixed effects plus the individual-level residuals.

during follow-up, possibly due to the longer length of follow-up. There were a slightly greater number of relapses in the 2 years before ABN eligibility and a moderately higher EDSS at baseline in the UoWMS compared with the BCMS cohort. The BCMS cohort had a higher average time between observations.

3.1. Choice of time axis

For the UoWMS cohort, Table 2 compares the linear fixed effects, linear random intercept, and linear random slope and intercept models with age and time since onset as the time axes. The models with time since onset as the time axis had a substantially lower AIC, a lower RMSE, higher proportions of observations within 0.5 EDSS (apart from the fixed effects only model), and a similar proportion of observations out by two or more EDSS. The random intercept and slope models had better fit by all the criteria than the random intercept and fixed effects only models. When comparing the best-fitting models with degree 2 fractional polynomials for the individual trajectories (i.e., for fixed effects and individual-level random effects) for age vs. time since onset (Supplementary Table 1), the latter consistently had lower AICs, implying that time since onset should be chosen as the time axis.

3.2. Choice of powers of time

There was strong evidence of an improvement in model fit when comparing degree 2 fractional polynomials to degree 1 [$P < 0.001$, degrees of freedom (d.f.) = 5]. Supplementary Table 2 shows that models including linear and log time, or square root and log time, are consistently the models with the two smallest AICs. The exception was for the data set restricted to 0–15 years since onset; however, the AIC in this case was close to that of the best model (a difference of only 5). Comparing the RMSE of these two models, Table 2, however, shows that the model with linear and log time tends to fit the observed data better. Thus, the final simple model included time since onset and log of time since onset for the individual trajectories (i.e., for the fixed and individual-level random effects).

3.3. Observation-level variation

The observation-level residuals from the best-fitting simple model appear to decrease over time (Fig. 2).

Fractional polynomials of degree 2 for the observation-level random effects tended not to converge, so we only considered fractional polynomials of degree 1. Adding a linear time term to the observation-level random effects showed a clear improvement in the model fit ($P < 0.001$, d.f. = 3) compared with the model where observation-level random effects had constant variance. The best-fitting model included the square root of time in the

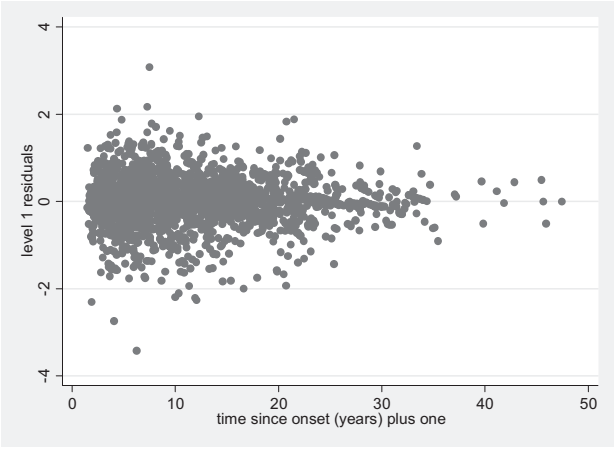


Fig. 2. Observation-level (level 1) residuals plotted over time since onset for the UoWMS best-fitting simple model with linear and log time since onset (n = 2,290).

observation-level random effects. The difference between the models with linear time and square root of time as observation-level random effects was small (difference in AIC = 1.1), so the model with linear time in the observation-level random effects was included for ease of interpretation.

We constrained the observation-level time variance term to be equal to zero because its 95% confidence interval (CI) included zero. Hence, we only allowed the covariance term between the constant and observation-level time term to be freely estimated, which amounts to the assumption that observation-level variance decreases linearly over time. This constraint had minimal impact on the model fit (Table 2). Hence, the complex model is of the form

$$\begin{aligned} y_{ij} &= \beta_0 + u_{0i} + e_{1ij} + \left(\beta_1 + u_{1i} + e_{2ij}\right) t_{ij} \\ &\quad + \left(\beta_2 + u_{2i}\right) \cdot \log t_{ij}, \\ &\text{where } \left\{e_{lij}\right\} \sim N_2(0, D_e),\ \left\{u_{li}\right\} \sim N_3(0, D_u). \end{aligned} \tag{2}$$

As described in the methods, we have added one to time since onset, to ensure strict positivity of log time. Unstructured covariance matrices, D_e and D_u, were used for the individual and observation-level random effects with a slight modification to the observation-level variance as discussed above, see Web Appendix C at www.jclinepi.com.

3.4. Autocorrelation

Supplementary Table 3 shows that in our model, we have some autocorrelation, with a correlation coefficient between consecutive observation-level residuals of 0.17 with some evidence that this increases with an individual's number of observations presumably due to a greater chance of having an observation close together. Using quarter year intervals (see methods) reduced the UoWMS data set from 2,290 to 1,876 observations with a maximum number of observations for a single individual being reduced from 72 to 26. Supplementary Table 3 shows that the correlations between lagged residuals were lower in this reduced data set.

3.5. Relapses

There was little difference in the fixed-effect estimates and individual-level random effects between the three models when EDSS scores were removed at 1, 3, or 6 months postrelapse, with all the 95% CIs overlapping (based on Equation (2), see Supplementary Table 4). However, when we consider the observation-level random effects, there is some evidence that the variance of the constant term is lower in the 6 month model compared with the 1 month model. For our final model, we choose the 3 month model, which seemed to have similar variance in the constant term when compared with the 6 month model.

3.6. Assessing model assumptions and fit

The QQ plots from the complex model (Equation (2)) with observations up to 3 months postrelapse removed and accounting for autocorrelation are shown in Supplementary Figure 1 and are close to normal.

Supplementary Figure 2 shows the observed vs. predicted values from the complex model adjusted for autocorrelation and relapses, which are all relatively close to the reference line of perfect predictions. The difference between observed and predicted values has an average of −0.003, a 95% central range (2.5th to 97.5th percentiles) of −1.13 to 0.97 and an RMSE of 0.48. Supplementary Figure 3 shows the observations and fitted patient-specific lines for six randomly chosen individuals with at least three observations.

3.7. Indication bias

When including a binary variable "ever used DMT" as a fixed effect, we found little evidence that those who started treatment have a different intercept ($P = 0.88$) or different progression ($P = 0.83$ for interaction with time and $P = 0.86$ for interaction with log time). This provides some evidence that "indication bias" was not a major issue in the UoWMS cohort.

3.8. Model development with the BCMS data

Developing the model on the BCMS data gave very similar results. Time since MS symptom onset was found to be a better time axis than age. Linear and log time for the individual trajectories gave very good fit to the data, although linear and square root time did give slightly better fit. Adding a linear time term to the observation-level random effects gave a model with better fit, and fractional polynomials of degree 2 showed little improvement.

In contrast to the UoWMS results, the model fitted to the data with all observations 1 month postrelapse removed gave similar results as those with all observations 3 and 6 months postonset of relapse removed (data not shown). Hence, our final BCMS model was based on the model with observations 1 month postrelapse removed.

3.9. Comparison of UoWMS and BCMS models and model validation

The UoWMS model had a higher fixed effect, higher individual-level random effect, and lower observation-level random effect for the constant term than the BCMS model (Table 3).

This indicates that the UoWMS patient population had a higher average EDSS at the presumed onset of disease, greater variation between individuals in EDSS at the onset, and slightly lower variation within individuals in EDSS at the onset (Table 1). The coefficients for the trajectory of EDSS over time and log time, and the other variances

Table 3. Parameter estimates, mean (95% CI), of the final UoWMS and BCMS models

Variable	BCMS (n = 6,447)	UoWMS (n = 1,589)
Fixed effects		
Intercept	1.05 (0.79, 1.31)	2.63 (2.00, 3.27)
Time since onset	0.22 (0.19, 0.26)	0.16 (0.10, 0.22)
Log time since onset	−0.13 (−0.39, 0.14)	−0.15 (−0.70, 0.40)
Individual-level (level 2) random effects		
Var(intercept)	2.80 (1.87, 3.73)	8.67 (5.05, 12.29)
Cov(intercept, time)	0.09 (−0.05, 0.24)	0.09 (−0.23, 0.40)
Var(time)	0.10 (0.08, 0.12)	0.08 (0.05, 0.12)
Cov(intercept, log time)	−2.73 (−3.82, −1.63)	−5.38 (−8.57, −2.19)
Cov(time, log time)	−0.65 (−0.81, −0.48)	−0.60 (−0.92, −0.28)
Var(log time)	6.14 (4.78, 7.49)	7.13 (4.01, 10.27)
Observation-level (level 1) random effects		
Var(intercept)	0.76 (0.70, 0.82)	0.40 (0.35, 0.45)
Cov(intercept, time)	−0.004 (−0.005, −0.002)	−0.003 (−0.005, −0.002)
Var(time)	Set equal to zero	Set equal to zero

Abbreviations: CI, confidence interval; UoWMS, the University of Wales MS; BCMS, British Columbia MS.

and covariances of the individual-level and observation-level random effects are remarkably similar between the two cohorts. The fixed effects would correspond to an average increase over 10 years from the onset of 1.9 EDSS and 1.3 EDSS points within the BCMS and UoWMS models, respectively.

We used the coefficients from the final BCMS model to predict EDSS in the UoWMS data set (with the EDSS scores removed 3 months postrelapse) conditional on the baseline EDSS, rounding the continuous prediction to the nearest true EDSS score. We observed a reasonable model fit with a mean difference between prediction and observed being −0.44 [standard deviation (SD): 1.36] and an RMSE of 1.46; 49.2% of predictions were within 0.5 units of the observed EDSS, whereas 22.5% of predictions were out by 2 or more EDSS units. These conditional predictions from the BCMS model are shown along with the observed EDSS in the UoWMS cohort, averaged within yearly bands, in Fig. 3A and Supplementary Table 5.

We also used the coefficients from the final UoWMS model to predict EDSS in the BCMS data set conditional on the baseline EDSS. The mean difference between prediction and observed was −0.61 (SD: 1.83), with some evidence of underprediction over time and RMSE of 1.93. Only 35.9% of predictions were within 0.5 units of the observed EDSS, whereas 35.8% of predictions were out by 2 or more EDSS. These conditional predictions from the UoWMS model are shown along with the observed EDSS in the BCMS cohort, averaged within yearly bands, in Fig. 3B and Supplementary Table 6.

4. Discussion

We used two large independent cohorts of MS patients from Canada and the United Kingdom to build a complex multilevel model for EDSS trajectory. Using the same model building strategy resulted in the models for both cohorts having the same time axis, powers of time that were consistent with each other and the same parameterization of observation-level variation. The average levels of disability in the two cohorts were different, but the average patterns of change were similar. This provided evidence that our model is transportable to other populations. When two cohorts are not available, it may be necessary to use other validation techniques such as randomly splitting the data set or using bootstrap or jackknife techniques.

Two key choices made here related to the functional form of the time axis. When both ‘‘time since onset’’ and ‘‘age’’ were considered, the best-fitting time axis was the former. The pattern of EDSS progression was not linear and fractional polynomials allowed us to identify the best-fitting trajectory. This approach has been developed and used for identification of nonlinear relationships in nonrepeated measures regression [40,41]. However, their use in multilevel models has been less widespread [4,42,43]. This may be because of increasing complexity of the models and also because of the additional need to parameterize the individual-level and observation-level random effects.

We found that observation-level variation (comprising measurement error and short term fluctuations in the EDSS) reduced as time progressed [i.e., as disability (EDSS) increased]. Previous work has shown that measurement error is lower for higher EDSS values [33,34] and that short-term improvements in disability are more likely earlier in the disease course [29]. Identification of complex observation-level variation highlights the importance of model checking when fitting multilevel longitudinal models.

Our ad hoc approach to adjust for the autocorrelation present was successful in reducing it further. A more robust approach might be to adjust the interval for merging multiple observations into one and then estimate the lagged residual correlation for each different interval. However, this would need to be balanced by the effect of concatenating too many observations and reducing any real fluctuations in EDSS due to using averaged data points. There is some evidence [44] that moderate misspecification of the observation-level variation has little impact on the fixed-effects estimates. The

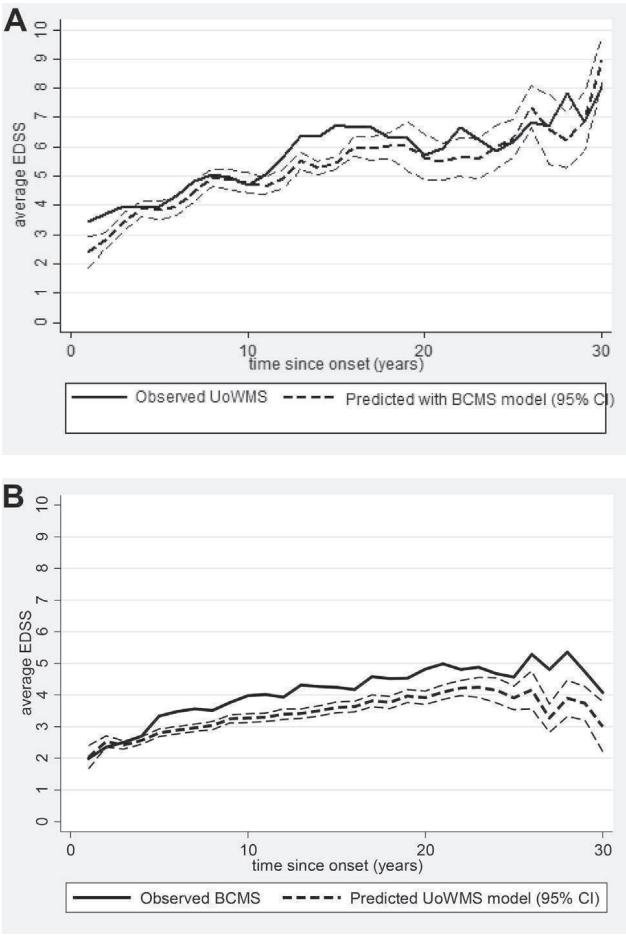


Fig. 3. (A) and (B) The upper graph shows the observed EDSS within the UoWMS cohort and the conditional predictions using the BCMS model. The lower graph shows the observed EDSS within the BCMS cohort and the conditional predictions using the UoWMS model. The plotted data, over a 30-year period, are the annual means at each year since time of the onset where data was grouped into yearly bins that is 0–0.5, 0.5–1.5, and so forth. EDSS, Expanded Disability Status Scale; UoWMS, the University of Wales MS; BCMS, British Columbia MS; CI, confidence interval.

moderate amount of autocorrelation present here had little influence on our estimates of average EDSS progression (the fixed-effect estimates were very similar in the models with and without concatenated observations).

Removal of observation (EDSS) scores within 1 month postrelapse appeared adequate for the BCMS cohort, but in the UoWMS cohort, the optimal window was 3 months postrelapse. This is consistent with the general approach to data collection within the BCMS database (EDSS scores were intended to be collected outside the influence of an acute relapse [8]), as well as with previous findings from the BCMS study [29]. Others have also shown that most improvement in physical disability occurs within 2 months postrelapse [36]. This underscores the importance of exploring local effects within cohorts.

We treated the ordinal EDSS score as a continuous measure, which facilitates the interpretation of the model parameters but does not imply that the meaning of a point change in EDSS is equivalent in terms of disability progression across the range of scores. All the models showed good fit to the observed data with normally distributed residuals. Researchers seeking to model repeated measures of ordinal scores should assess the normality of the residuals as a key model check, but our results show that this assumption may be satisfied even if the outcome measure itself is ordinal and/or not normally distributed.

Modeling the trajectory of EDSS against the time since MS symptom onset showed a good fit in both cohorts. Cross-validating our models by predicting the future EDSS trajectory conditional on the first observation in one data set using the model derived on the other data set showed reasonably good fit, with about half of all predicted EDSS scores within 0.5 EDSS of the observed when validating the BCMS model in the UoWMS data. However, the UoWMS model performed less well in the BCMS data, with evidence of underprediction at almost all time points. This is not surprising given the relative sizes of the two data set. The higher within-individual variation within the BCMS data set means that predictions within the BCMS data are likely to show greater variation than predictions made on the UoWMS data. Our ability to examine model fit in two different populations is important in validating any prognostic model. Using a repeated measures model to predict individual trajectories based on one or more observations has been done before [4,15,45], but further research needs to examine whether additional covariates can explain why the trajectory of some subjects was not well predicted in our current model. Bayesian methods could also be used, such that the previous model estimates form priors for parameters to be estimated using a small number of observations from an individual [46].

Truncating data once a DMT was started (UoWMS) or to 1995 (BCMS data) resulted in an average follow-up time for individuals of 3 and 5.8 years, respectively. We had limited data with 10 years of follow-up or greater (16% and 4% of individuals, respectively). However, because individuals enter the study at different times since onset, we are able to model this without extrapolation. Examining Fig. 3 shows little evidence that the fit to our data is worse at the upper end of 20–30 years since time of the onset.

Although the models are generally similar for most parameters, it seems there is higher between-individual variation in the UoWMS model and higher within-individual variation in the BCMS model. One reason for this could be an inherent difference between individuals from Canada and Wales or a difference in when and how individuals accessed the health systems in each area. Also, the BCMS cohort was mainly seen by the same four core neurologists, which would reduce any between-rater variation.

The corresponding Markov model that was fitted to the BCMS cohort [13] showed about a 2.2 EDSS increase over a 10-year period since ABN eligibility. Looking at our fixed effects over the same period (approx. 8–18 years since

onset) would correspond to a similar average increase of 2.1 and 1.5 EDSS within the BCMS and UoWMS models, respectively. Using the fixed effects, we would expect the average time from the onset to reaching an EDSS of 6 to be 23.1 and 23.4 years using the UoWMS and BCMS models, respectively. Scalfari [7] reported a mean time from the onset of 21.2 years (95% CI: 19.8, 22.6) to reaching EDSS 6, and Tremlett [8], who looked at six studies, reported the median times ranging from 15 to 32 years, consistent with our estimates. Comparing our study to one previous multilevel model for MS [18] is difficult given their transformation of EDSS into weighted change of EDSS. However, they did include a quadratic term whose estimate was positive, which creates a similar shape to a negative log term with rate of change increasing over time.

Our results have implications for the design and analysis of MS intervention trials, leading to more sensitive assessment of treatment effects over time than cross-sectional analyses of EDSS at fixed time points. Average EDSS could be similar in an intervention and control arm after a fixed follow-up period, but if the trajectory was different, the intervention might still be considered an effective treatment in the short and medium terms.

The development of this multilevel model is an important methodological achievement that will enable prediction of the expected long-term disease trajectory in other populations. This is especially necessary for extrapolating the findings from randomized controlled trials (which are often short term, lasting no more than 2−3 years) or for interpreting findings from DMT exposed cohorts of patients with no specific unexposed control cohort. In particular, these models facilitate natural history predictions within the UK MS risk sharing scheme to determine the efficacy of DMTs. It is only by modeling longer term follow-up of such studies that one can advise patients about the potential therapeutic benefits that accrue in the long term.

Acknowledgments

The authors' appreciation is extended to the BC MS Clinic neurologists who contributed to the study through patient examination and data collection (current members listed here by primary clinic): UBC MS Clinic: A. Traboulsee, MD, FRCPC (UBC Hospital MS Clinic Director and Head of the UBC MS Programs); A-L. Sayao, MD, FRCPC; V. Devonshire, MD, FRCPC; S. Hashimoto, MD, FRCPC (UBC and Victoria MS Clinics); J. Hooge, MD, FRCPC (UBC and Prince George MS Clinic); L. Kastrukoff, MD, FRCPC (UBC and Prince George MS Clinic); J.O., MD, FRCPC. Kelowna MS Clinic: D. Adams, MD, FRCPC; D. Craig, MD, FRCPC; S. Meckling, MD, FRCPC. Prince George MS Clinic: L. Daly, MD, FRCPC. Victoria MS Clinic: O. Hrebicek, MD, FRCPC; D. Parton, MD, FRCPC; K Atwell-Pope, MD, FRCPC. The views expressed in this article do not necessarily reflect the views of each individual acknowledged.

The authors thank Professor Mark Gilthorpe, Professor Andrew Pickles and Professor Alasdair Coles for commenting on the development of these models.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.05.003.

References

[1] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381.

[2] Liu H, Miller LG, Hays RD, Golin CE, Wu T, Wenger NS, et al. Repeated measures longitudinal analyses of HIV virologic response as a function of percent adherence, dose timing, genotypic sensitivity, and other factors. J Acquir Immune Defic Syndr 2006;41:315−22.

[3] Proust-Lima C, Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics 2009;10:535−49.

[4] Tilling K, Sterne JA, Wolfe CD. Multilevel growth curve models with covariate effects: application to recovery after stroke. Stat Med 2001;20:685−704.

[5] Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 1996;46:907−11.

[6] Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983;33:1444−52.

[7] Scalfari A, Neuhaus A, Daumer M, Ebers GC, Muraro PA. Age and disability accumulation in multiple sclerosis. Neurology 2011;77:1246−52.

[8] Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. Neurology 2010;74:2004−15.

[9] Vukusic S, Confavreux C. Natural history of multiple sclerosis: risk factors and prognostic indicators. Curr Opin Neurol 2007;20:269−74.

[10] Achiron A. Predicting the course of relapsing-remitting MS using longitudinal disability curves. J Neurol 2004;251(Suppl 5):v65−8.

[11] Achiron A, Barak Y, Rotstein Z. Longitudinal disability curves for predicting the course of relapsing-remitting multiple sclerosis. Mult Scler 2003;9:486−91.

[12] Gauthier SA, Mandel M, Guttmann CR, Glanz BI, Khoury SJ, Betensky RA, et al. Predicting short-term disability in multiple sclerosis. Neurology 2007;68:2059−65.

[13] Palace J, Bregenzer T, Tremlett H, Oger J, Zhu F, Boggild M, et al. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. BMJ open 2014;4:e004073.

[14] Goldstein H. Multilevel statistical models. Wiley series in probability and statistics. 4th ed. Chichester, West Sussex: Wiley; 2011:358. :xxi.

[15] Tilling K, Sterne JA, Rudd AG, Glass TA, Wityk RJ, Wolfe CD. A new method for predicting recovery after stroke. Stroke 2001;32:2867−73.

[16] Bosch JL, Tilling K, Bohnen AM, Donovan JL. Establishing normal reference ranges for PSA change with age in a population-based study: the Krimpen study. Prostate 2006;66:335−43.

[17] Tilling K, Garmo H, Metcalfe C, Holmberg L, Hamdy FC, Neal DE, et al. Development of a new method for monitoring prostate-specific antigen changes in men with localised prostate cancer: a comparison of observational cohorts. Eur Urol 2010;57:446−52.

[18] Di Serio C, Lamina C. Investigating determinants of multiple sclerosis in longitunal studies: a Bayesian approach. J Probab Stat 2009;2009.
[19] Boggild M, Palace J, Barton P, Ben-Shlomo Y, Bregenzer T, Dobson C, et al. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. BMJ 2009; 339:b4677.
[20] Swingler RJ, Compston DA. The prevalence of multiple sclerosis in south east Wales. J Neurol Neurosurg Psychiatry 1988;51:1520−4.
[21] Hennessey A, Robertson NP, Swingler R, Compston DA. Urinary, faecal and sexual dysfunction in patients with multiple sclerosis. J Neurol 1999;246:1027−32.
[22] Hennessy A, Swingler RJ, Compston DA. The incidence and mortality of multiple sclerosis in south east Wales. J Neurol Neurosurg Psychiatry 1989;52:1085−9.
[23] Hirst C, Ingram G, Pickersgill T, Swingler R, Compston DA, Robertson NP. Increasing prevalence and incidence of multiple sclerosis in South East Wales. J Neurol Neurosurg Psychiatry 2009;80: 386−91.
[24] Tremlett H, Paty D, Devonshire V. Disability progression in multiple sclerosis is slower than previously reported. Neurology 2006;66:172−7.
[25] Evans C, Kingwell E, Zhu F, Oger J, Zhao Y, Tremlett H. Hospital admissions and MS: temporal trends and patient characteristics. Am J Manag Care 2012;18:735−42.
[26] Koch M, Kingwell E, Rieckmann P, Tremlett H. The natural history of secondary progressive multiple sclerosis. J Neurol Neurosurg Psychiatry 2010;81:1039−43.
[27] Shirani A, Zhao Y, Karim ME, Evans C, Kingwell E, van der Kop ML, et al. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. JAMA 2012;308:247−56.
[28] Tremlett H, Zhao Y, Joseph J, Devonshire V, Neurologists UC. Relapses in multiple sclerosis are age- and time-dependent. J Neurol Neurosurg Psychiatry 2008;79:1368−74.
[29] Tremlett H, Zhu F, Petkau J, Oger J, Zhao Y. Natural, innate improvements in multiple sclerosis disability. J Mult Scler 2012;18(10): 1412−21.
[30] Shirani A, Zhao Y, Kingwell E, Rieckmann P, Tremlett H. Temporal trends of disability progression in multiple sclerosis: findings from British Columbia, Canada (1975-2009). Mult Scler 2012;18:442−50.
[31] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates—parsimonious parametric modeling. Appl Statistics-Journal R Stat Soc Ser C 1994;43(3):429−67.
[32] Howe LD, Tilling K, Matijasevich A, Petherick ES, Santos AC, Fairley L, et al. Linear spline multilevel models for summarising childhood growth trajectories: a guide to their application using examples from five birth cohorts. Stat Methods Med Res 2013. (in press).
[33] Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino P, Scherokman B, et al. Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke expanded disability status scale (EDSS). Multiple sclerosis collaborative research group. Neurology 1992;42: 859−63.
[34] Hughes S, Spelman T, Trojano M, Lugaresi A, Izquierdo G, Grand'maison F, et al. The Kurtzke EDSS rank stability increases 4 years after the onset of multiple sclerosis: results from the MSBase Registry. J Neurol Neurosurg Psychiatry 2012;83:305−10.
[35] Rasbash J, Charlton C, Jones K, Pillinger R. Manual supplement to MLwiN v2.26. UK: Centre for Multilevel Modelling, University of Bristol; 2012.
[36] Hirst CL, Ingram G, Pickersgill TP, Robertson NP. Temporal evolution of remission following multiple sclerosis relapse and predictors of outcome. J Mult Scler 2012;18(8):1152−8.
[37] StataCorp. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP; 2009.
[38] Leckie G, Charlton C. runmlwin: Stata module for fitting multilevel models in the MLwiN software package. UK: Centre for Multilevel Modelling, University of Bristol; 2011.
[39] Pan HQ, Goldstein H. Multi-level models for longitudinal growth norms. Stat Med 1997;16:2665−78.
[40] Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999;28:964−74.
[41] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc Ser A (Statistics Society) 1999;162(1): 71−94.
[42] Long J, Ryoo J. Using fractional polynomials to model non-linear trends in longitudinal data. Br J Math Stat Psychol 2010;63:177−203.
[43] Wen X, Kleinman K, Gillman MW, Rifas-Shiman SL, Taveras EM. Childhood body mass index trajectories: modeling, characterizing, pairwise correlations and socio-demographic predictors of trajectory characteristics. BMC Med Res Methodol 2012;12:38.
[44] Taylor JM, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? Stat Med 1998;17:2381−94.
[45] Palmer TM, Macdonald-Wallis CM, Lawlor DA, Tilling K. Estimating adjusted associations between random effects from multilevel models: The reffadjust package. Stata Journal 2014;14(1):119−40.
[46] Taylor JM, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, et al. Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics 2013;69:206−13.

Supplemental Appendix

A. Fitting fractional polynomials to find function of time

A fractional polynomial of degree *m* contains *m* different powers of the time variable. The power zero denotes the logarithmic transformation and a combination of two identical powers is interpreted as the product of the power and the logarithmic transformation. Hence a fractional polynomial in *t* of degree 2 with powers = (1, 1) would include *t* and $t \cdot \log t$. In order to fit fractional polynomials, the time variable must be strictly positive (as the logarithm of zero or a negative number is undefined). We therefore added one year to time since MS onset and used 17 as the origin for age.

Models with different degrees of fractional polynomials of time were compared using the deviance [1]. Fractional polynomials of degree *m* and (*m*-1) are usually compared with 2 degrees of freedom (one for the coefficient and one for the power) and we used a p-value of 0.05 for our significance level, [1]. In our case we must also include extra degrees of freedom for estimation of the variance and covariance parameters within the random effects. The powers of time we considered for inclusion in the fractional polynomials were -2, -1, -0.5, 0, 0.5, 1, 2 and 3. In our model, for ease of interpretation, the same powers of time were used for the fixed effects and the individual-level random effects. Hence the following multilevel model was considered:

$$y_{ij} = \beta_0 + u_{0i} + \sum_{k=1}^{m} (\beta_k + u_{ki}) t_{ij}^{r_k} + e_{ij} \quad (1)$$

where *m* is the degree of fractional polynomial. y_{ij}, β_0 and u_{0i} have the same meaning as within equation (1) from section 3 in the main manuscript. The u_{ki} for k=0,1,...,*m*, are

normally distributed with mean zero and unstructured covariance matrix D_u. The r_k are the powers of the fractional polynomials for the fixed effects and individual-level random effects.

B. Fitting fractional polynomials to find observation level-variation

Powers of time used for the observation level random effects were allowed to differ from those used for the fixed and individual level random effects, i.e. we considered the following model:

$$y_{ij} = \beta_0 + u_{0i} + \sum_{k=1}^{m} (\beta_k + u_{ki}) t_{ij}^{r_k} + \sum_{l=1}^{p} e_{lij} t_{ij}^{s_l} + e_{ij} \quad (2)$$

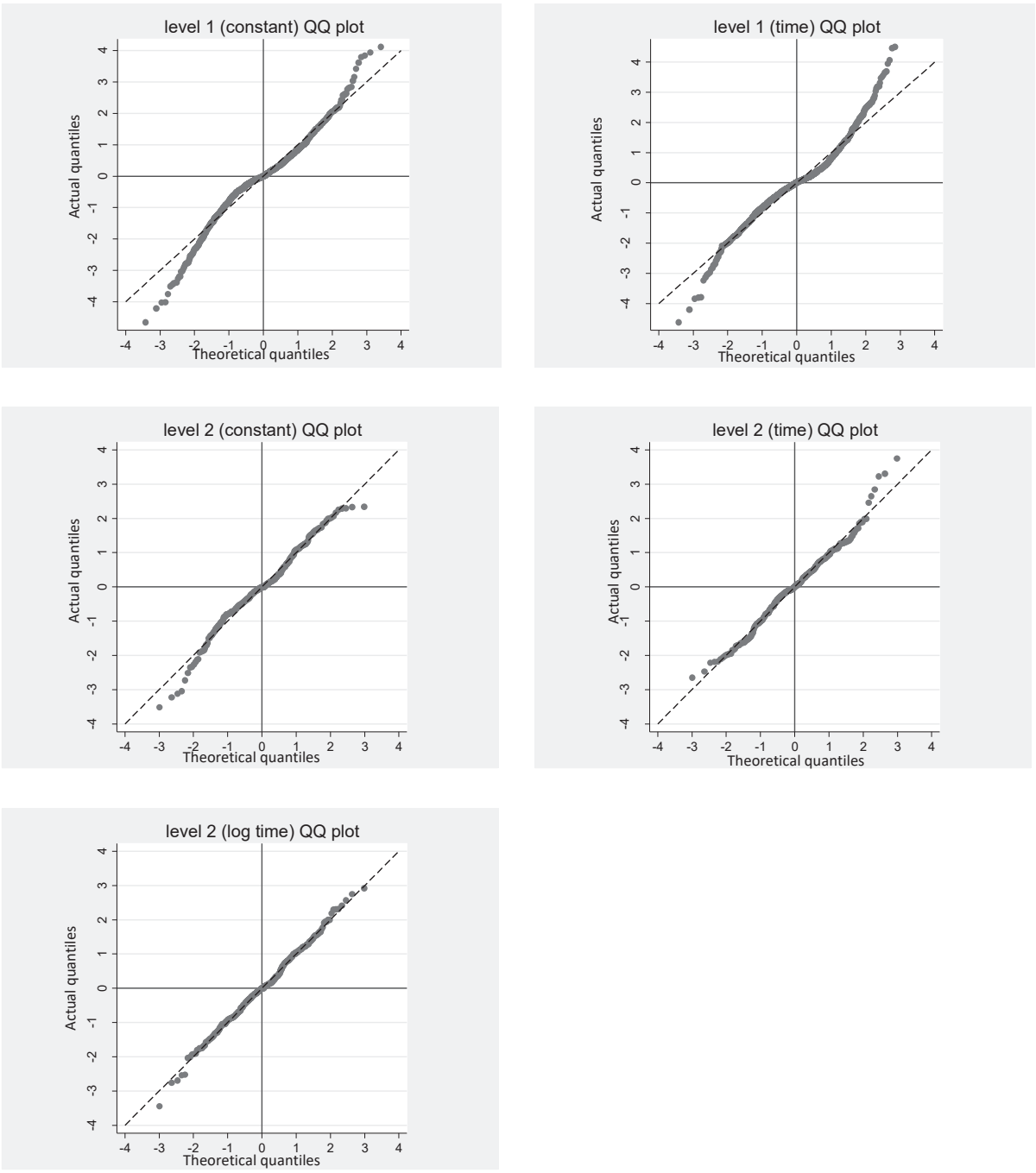
where the e_{lij}, e_{ij} are normally distributed with unstructured covariance matrix D_e. The s_l are the powers of the fractional polynomials for the observation-level random effects.

C. Covariance matrices for the final model

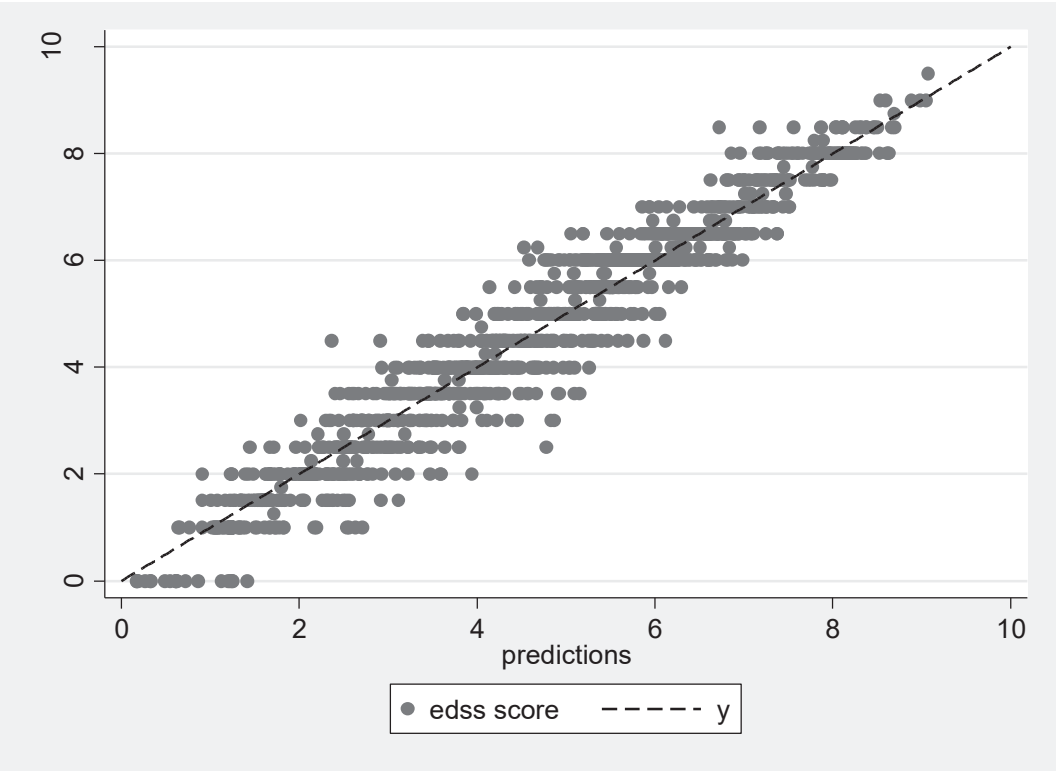
$$D_e = \begin{bmatrix} var(e_{1ij}) & cov(e_{1ij}, e_{2ij}) \\ cov(e_{1ij}, e_{2ij}) & 0 \end{bmatrix} \text{ and}$$

$$D_u = \begin{bmatrix} var(u_{0i}) & cov(u_{0i}, u_{1i}) & cov(u_{0i}, u_{2i}) \\ cov(u_{0i}, u_{1i}) & var(u_{1i}) & cov(u_{1i}, u_{2i}) \\ cov(u_{0i}, u_{2i}) & cov(u_{1i}, u_{2i}) & var(u_{2i}) \end{bmatrix} \quad (3)$$

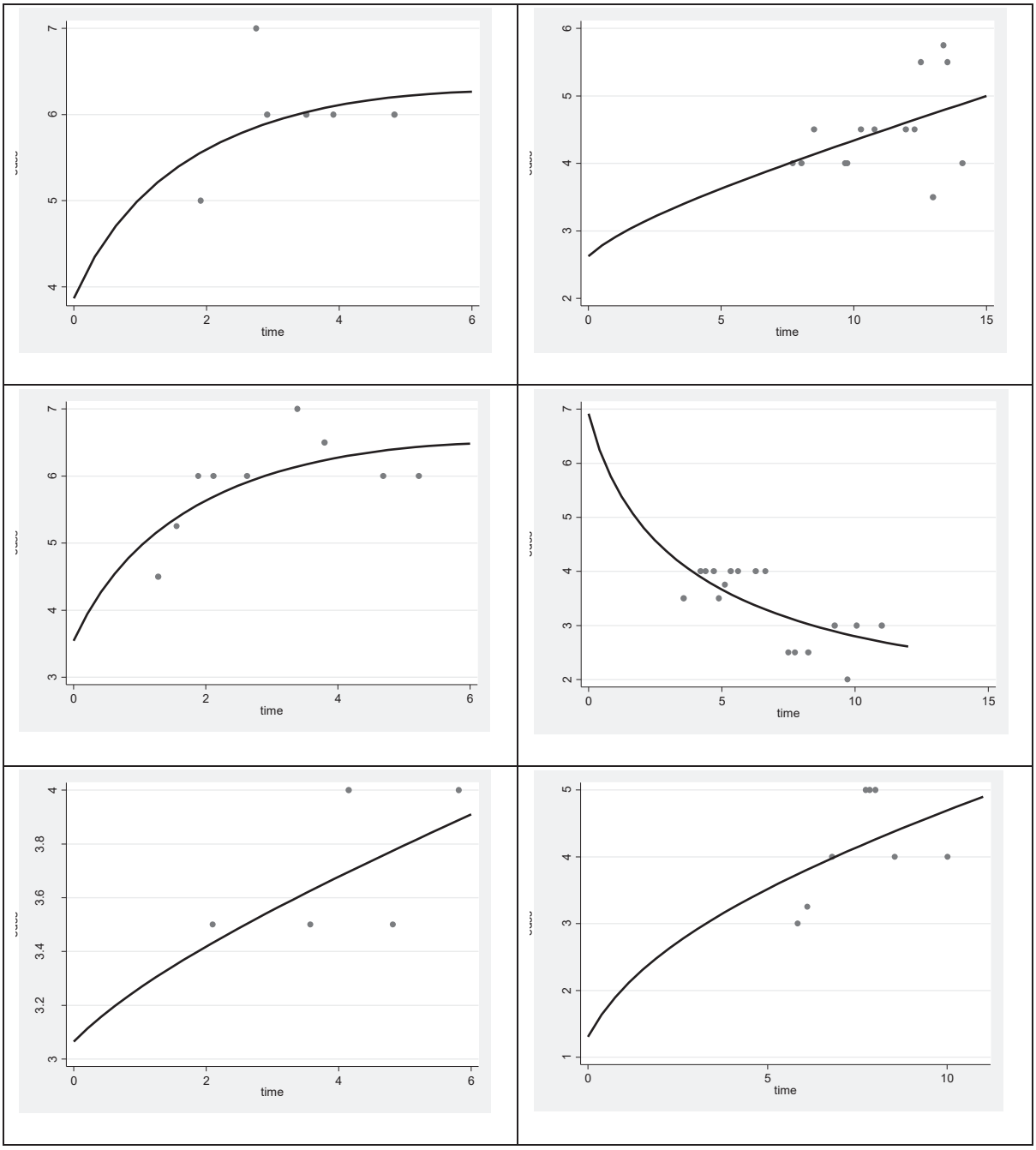
1. Royston, P. and D.G. Altman, *Regression Using Fractional Polynomials of Continuous Covariates - Parsimonious Parametric Modeling.* Applied Statistics-Journal of the Royal Statistical Society Series C, 1994. **43**(3): p. 429



Supplementary Figure 1. The Q-Q plots for all the residuals within the UoWMS complex model accounting for autocorrelation and relapse with the functional form as described within equation (2). Level 1 are observation-level random effects and level 2 individual-level random effects. (n=1589)



Supplementary Figure 2. Observed vs. predicted values the UoWMS model with the functional form as described within equation (2). (n=1589)



Supplementary Figure 3. EDSS observations plotted against time since onset along with the fitted patient specific lines from the UoWMS complex model accounting for autocorrelation and relapse for six randomly chosen individuals.

Supplementary Table 1. Using different time axes the five best models for the UoWMS ABN eligible cohort, their fixed and individual-level random effects of time, and within the brackets their AIC.

Powers of time for fixed and individual level random effects with age as time axis (AIC) (n=2290)		**Powers of time for fixed and individual level random effects with time since onset as time axis (AIC) (n=2290)**	
$t^2, \frac{1}{t}$	(6177.79)	$\sqrt{t}, \log t$	(6063.25)
$t^3, \frac{1}{t^2}$	(6238.95)	$t, \log t$	(6066.72)
$\sqrt{t}\ , (\sqrt{t} \cdot \log t)$	(6252.38)	$t^2, \sqrt{t}$	(6073.26)
$\log t\ , (\log t \cdot \log t)$	(6254.55)	t, t^2	(6073.89)
$t, \frac{1}{\sqrt{t}}$	(6264.65)	$t, \frac{1}{\sqrt{t}}$	(6076.75)

Supplementary Table 2. Powers of time since onset for fixed and individual-level random effects (AIC) for UoWMS data subsets including observations made within different time frames

Entire dataset (n=2290)		**Time from 0-30 years since onset (n=2244)**		**Time from 0-15 years since onset (n=1713)**	
$\sqrt{t}, \log t$	(6063.25)	$\sqrt{t}, \log t$	(5954.79)	$\log t, (\log t)^2$	(4695.39)
$t, \log t$	(6066.72)	$t, \log t$	(5959.78)	$\sqrt{t}, \log t$	(4696.63)
$t^2, \sqrt{t}$	(6073.26)	$t^2, \sqrt{t}$	(5963.89)	$t, \frac{1}{\sqrt{t}}$	(4697.93)
t, t^2	(6073.89)	t, t^2	(5964.08)	$\sqrt{t}, \frac{1}{t}$	(4698.92)
$t, \frac{1}{\sqrt{t}}$	(6076.75)	$t^2, \log t$	(5973.41)	$t, \log t$	(4700.42)

Supplementary Table 3. Correlation coefficients for "differences" between observations and patient specific lines with a lag of one when considering individuals with different numbers of observations for the UoWMS complex model. Individuals are grouped in this table by the number of observations they had before adjusting for autocorrelation so we compare the same individuals within each column. However since some individuals have only one observation (correlation cannot be estimated) the numbers of individuals within each row and column are given.

Number of observations per person included	**Lagged "difference" correlation for the complete dataset (number of individuals : %)**	**Lagged "difference" correlation for the dataset where observations within quarter year intervals are merged (number of individuals : %)**
Entire dataset	0.169 (303/404 : 75%)	-0.011 (297/404 : 73.5%)
3 or more	0.197 (233/404 : 57.7%)	0.021 (232/404 : 57.4%)
4 or more	0.222 (187/404: 46.3 %)	0.049 (186/404: 46.0 %)
5 or more	0.128 (157/404: 38.9%)	0.077 (157/404: 38.9%)
6 or more	0.162 (130/404: 32.2%)	0.105 (130/404: 32.2%)

7 or more	0.182 (111/404: 27.5%)	0.134 (111/404: 27.5%)
8 or more	0.206 (96/404: 23.8%)	0.143 (96/404: 23.8%)
9 or more	0.224 (80/404: 19.8%)	0.158 (80/404: 19.8%)
10 or more	0.248 (68/404: 16.9%)	0.183 (68/404: 16.9%)

Supplementary Table 4. Parameter estimates, mean (95% CI), for models based on datasets with observations removed at different durations post onset of a relapse for the UoWMS complex model adjusted for autocorrelation

Variable	1 month post relapse (n=1876)	3 months post relapse (chosen model) (n=1589)	6 months post relapse (n=1303)
Fixed Effects			
Intercept	2.63 (2.09, 3.17)	2.63 (2.00, 3.27)	2.53 (1.85, 3.21)
Time since onset	0.16 (0.10, 0.21)	0.16 (0.10, 0.22)	0.14 (0.08, 0.20)
Log time since onset	-0.15 (-0.63, 0.34)	-0.15 (-0.70, 0.40)	0.03 (-0.51, 0.57)
Individual Level (level 2) random effects			
Var(intercept)	6.81 (3.97, 9.65)	8.67 (5.05, 12.29)	7.11 (3.48, 10.75)
Cov(intercept, time)	0.04 (-0.26, 0.34)	0.09 (-0.23, 0.40)	-0.17 (-0.38, 0.03)
Var(time)	0.07 (0.04, 0.11)	0.08 (0.05, 0.12)	0.05 (0.03, 0.08)
Cov(intercept, log time)	-4.05 (-6.71, -1.39)	-5.38 (-8.57, -2.19)	-3.04 (-5.69, -0.39)
Cov(time, log time)	-0.51 (-0.81, -0.22)	-0.60 (-0.92, -0.28)	-0.31 (-0.53, -0.09)
Var(log time)	5.86 (3.14, 8.58)	7.13 (4.01, 10.27)	4.41 (2.04, 6.79)
Observation level (Level 1) random effects			
Var(intercept)	0.51 (0.46, 0.56)	0.40 (0.35, 0.45)	0.38 (0.33, 0.43)
Cov(intercept, time)	-0.004 (-0.006, -0.003)	-0.003 (-0.005, -0.002)	-0.003 (-0.005, -0.002)
Var(time)	**Set equal to zero**	**Set equal to zero**	**Set equal to zero**

Supplementary table 5. A table showing the observed EDSS data over time since onset for the UoWMS dataset as well as the predicted EDSS and the EDSS difference (predicted-observed) from the final UoWMS and BCMS models

Time since onset years	UoWMS N obs (N individuals)	UoWMS observed; mean(sd)	BCMS predicted; mean(sd)	BCMS pred-obs; mean(sd)	UoWMS predicted; mean (sd)	UoWMS pred-obs; mean (sd)
0.5-1.5	13 (10)	3.42 (1.51)	2.38 (0.98)	-1.04 (0.69)	3.04 (1.35)	-0.38 (0.55)
1.5-2.5	59 (35)	3.72 (1.93)	2.81 (1.10)	-0.92 (1.12)	3.33 (1.33)	-0.39 (0.96)
2.5-3.5	93 (47)	3.99 (2.01)	3.43 (1.49)	-0.56 (1.15)	3.75 (1.57)	-0.24 (1.10)
3.5-4.5	101 (51)	3.93 (2.03)	3.89 (1.43)	-0.04 (1.30)	3.91 (1.34)	-0.02 (1.30)
4.5-5.5	86 (53)	3.99 (1.87)	3.83 (1.56)	-0.16 (1.44)	3.76 (1.39)	-0.23 (1.35)
5.5-6.5	98 (53)	4.32 (1.86)	3.98 (1.64)	-0.34 (1.65)	3.88 (1.47)	-0.43 (1.50)
6.5-7.5	105 (50)	4.82 (1.54)	4.46 (1.75)	-0.36 (1.37)	4.35 (1.69)	-0.47 (1.30)
7.5-8.5	95 (47)	5.03 (1.56)	4.96 (1.48)	-0.07 (1.25)	4.74 (1.34)	-0.29 (1.18)
8.5-9.5	70 (49)	4.98 (1.72)	4.86 (1.48)	-0.11 (1.31)	4.66 (1.25)	-0.32 (1.19)
9.5-10.5	65 (42)	4.68 (1.78)	4.78 (1.43)	0.11 (1.15)	4.64 (1.37)	-0.04 (1.15)
10.5-11.5	78 (43)	5.04 (1.77)	4.66 (1.31)	-0.38 (1.28)	4.58 (1.28)	-0.46 (1.25)
11.5-12.5	64 (27)	5.63 (1.80)	4.91 (1.36)	-0.73 (1.35)	4.66 (1.16)	-0.98 (1.35)
12.5-13.5	64 (23)	6.38 (1.50)	5.52 (1.21)	-0.85 (1.28)	5.17 (1.11)	-1.20 (1.41)
13.5-14.5	77 (28)	6.36 (1.47)	5.27 (1.07)	-1.08 (1.38)	4.94 (1.03)	-1.42 (1.52)
14.5-15.5	92 (31)	6.77 (1.49)	5.45 (1.01)	-1.32 (1.38)	5.13 (0.90)	-1.65 (1.40)
15.5-16.5	69 (23)	6.66 (1.65)	6.01 (1.41)	-0.65 (1.22)	5.66 (1.29)	-1.00 (1.25)
16.5-17.5	39 (22)	6.65 (1.37)	5.94 (1.28)	-0.72 (0.97)	5.72 (1.20)	-0.94 (1.08)
17.5-18.5	46 (20)	6.30 (1.48)	6.03 (1.48)	-0.27 (0.97)	5.82 (1.38)	-0.49 (0.91)
18.5-19.5	26 (15)	6.31 (1.70)	6.02 (2.15)	-0.29 (1.44)	5.75 (2.08)	-0.56 (1.48)
19.5-20.5	23 (17)	5.74 (1.44)	5.65 (1.91)	-0.09 (1.29)	5.35 (1.78)	-0.39 (1.26)
20.5-21.5	39 (17)	5.94 (1.49)	5.49 (2.03)	-0.45 (1.26)	5.19 (1.82)	-0.74 (1.16)
21.5-22.5	34 (18)	6.68 (1.54)	5.66 (1.97)	-1.01 (1.47)	5.31 (1.87)	-1.37 (1.48)
22.5-23.5	17 (9)	6.26 (0.69)	5.59 (1.44)	-0.68 (1.12)	5.62 (1.45)	-0.65 (1.22)
23.5-24.5	18 (10)	5.83 (1.60)	5.97 (1.61)	0.14 (0.97)	5.58 (1.49)	-0.25 (0.97)
24.5-25.5	14 (9)	6.18 (1.37)	6.29 (1.27)	0.11 (1.00)	5.93 (1.07)	-0.25 (1.03)
25.5-26.5	12 (5)	6.83 (1.15)	7.38 (1.28)	0.54 (0.50)	6.96 (1.01)	0.13 (0.38)
26.5-27.5	18 (7)	6.69 (2.09)	6.61 (2.58)	-0.08 (1.10)	6.25 (2.17)	-0.44 (0.76)
27.5-28.5	13 (3)	7.81 (1.16)	6.23 (1.73)	-1.58 (1.46)	6.31 (1.11)	-1.50 (0.79)
28.5-29.5	12 (7)	6.83 (1.75)	6.88 (1.82)	0.04 (1.16)	6.46 (1.62)	-0.38 (0.91)
29.5-30.5	10 (3)	8.05 (0.50)	8.95 (1.23)	0.90 (0.81)	8.05 (0.72)	0.00 (0.75)

Supplementary table 6. A table showing the observed EDSS data over time since onset for the BCMS dataset as well as the predicted EDSS and the EDSS difference (predicted-observed) from the final UoWMS and BCMS models

Time since onset years	BCMS N obs (N individuals)	BCMS observed; mean(sd)	BCMS predicted; mean(sd)	BCMS pred-obs; mean(sd)	UoWMS predicted; mean (sd)	UoWMS pred-obs; mean (sd)
0.5-1.5	44 (37)	1.99 (1.46)	1.75 (0.74)	-0.24 (1.03)	2.03 (1.24)	0.05 (0.99)
1.5-2.5	178 (126)	2.36 (1.81)	2.07 (0.95)	-0.29 (1.26)	2.53 (1.20)	0.17 (1.17)
2.5-3.5	233 (152)	2.51 (1.80)	2.04 (0.97)	-0.47 (1.37)	2.42 (0.99)	-0.09 (1.38)
3.5-4.5	311 (185)	2.71 (1.87)	2.24 (1.05)	-0.46 (1.48)	2.57 (1.06)	-0.14 (1.53)
4.5-5.5	329 (200)	3.34 (2.15)	2.53 (1.21)	-0.81 (1.74)	2.81 (1.13)	-0.53 (1.81)
5.5-6.5	361 (227)	3.47 (2.38)	2.63 (1.18)	-0.84 (1.87)	2.89 (1.15)	-0.58 (1.95)
6.5-7.5	381 (241)	3.56 (2.35)	2.78 (1.17)	-0.78 (1.86)	2.97 (1.16)	-0.59 (1.97)
7.5-8.5	385 (258)	3.51 (2.36)	2.89 (1.30)	-0.62 (1.81)	3.04 (1.27)	-0.48 (1.94)
8.5-9.5	391 (263)	3.77 (2.37)	3.12 (1.37)	-0.64 (1.84)	3.25 (1.31)	-0.52 (1.89)
9.5-10.5	382 (233)	3.98 (2.23)	3.20 (1.40)	-0.78 (1.72)	3.27 (1.35)	-0.71 (1.83)
10.5-11.5	379 (230)	4.01 (2.32)	3.25 (1.35)	-0.77 (1.91)	3.30 (1.32)	-0.72 (2.00)
11.5-12.5	307 (211)	3.94 (2.46)	3.40 (1.50)	-0.54 (1.85)	3.39 (1.49)	-0.55 (1.90)
12.5-13.5	295 (192)	4.32 (2.25)	3.46 (1.34)	-0.86 (1.95)	3.42 (1.31)	-0.90 (1.99)
13.5-14.5	300 (194)	4.27 (2.32)	3.59 (1.48)	-0.68 (1.89)	3.50 (1.41)	-0.77 (1.93)
14.5-15.5	259 (156)	4.24 (2.32)	3.71 (1.46)	-0.53 (1.85)	3.61 (1.38)	-0.63 (1.86)
15.5-16.5	232 (135)	4.18 (2.05)	3.77 (1.42)	-0.41 (1.72)	3.64 (1.30)	-0.54 (1.74)
16.5-17.5	213 (117)	4.58 (1.98)	3.94 (1.49)	-0.64 (1.83)	3.82 (1.38)	-0.76 (1.81)
17.5-18.5	196 (114)	4.52 (2.07)	3.91 (1.50)	-0.60 (1.84)	3.77 (1.39)	-0.75 (1.78)
18.5-19.5	178 (107)	4.53 (2.02)	4.12 (1.56)	-0.41 (1.76)	3.97 (1.37)	-0.56 (1.72)
19.5-20.5	165 (103)	4.82 (2.03)	4.07 (1.52)	-0.75 (1.79)	3.92 (1.39)	-0.90 (1.77)
20.5-21.5	153 (89)	4.99 (1.73)	4.29 (1.57)	-0.70 (1.67)	4.09 (1.46)	-0.89 (1.56)
21.5-22.5	129 (76)	4.81 (1.80)	4.42 (1.53)	-0.39 (1.66)	4.22 (1.38)	-0.59 (1.61)
22.5-23.5	89 (67)	4.88 (1.72)	4.49 (1.61)	-0.39 (1.66)	4.24 (1.51)	-0.63 (1.58)
23.5-24.5	79 (62)	4.67 (1.96)	4.43 (1.91)	-0.24 (1.51)	4.15 (1.81)	-0.52 (1.45)
24.5-25.5	79 (49)	4.57 (1.96)	4.13 (1.83)	-0.44 (1.76)	3.91 (1.70)	-0.66 (1.71)
25.5-26.5	51 (39)	5.28 (2.17)	4.48 (2.32)	-0.80 (1.91)	4.17 (2.18)	-1.12 (1.83)
26.5-27.5	57 (35)	4.81 (1.87)	3.47 (1.82)	-1.33 (2.11)	3.27 (1.72)	-1.54 (2.02)
27.5-28.5	39 (25)	5.36 (1.93)	4.23 (2.02)	-1.13 (1.67)	3.90 (1.83)	-1.46 (1.57)
28.5-29.5	21 (14)	4.74 (1.81)	4.02 (1.30)	-0.71 (1.66)	3.74 (1.24)	-1.00 (1.80)
29.5-30.5	19 (12)	4.08 (2.05)	3.21 (1.66)	-0.87 (2.01)	3.00 (1.76)	-1.08 (2.09)

Supplementary table 7. A table showing the notation used in the main paper and the web-only appendix equations.

Notation	**Meaning**
y_{ij}	EDSS score for the *i*th individual at the *j*th time point
t_{ij}	Time of observation y_{ij}
β_0	Average intercept (constant fixed effect)
β_k	Fixed effect for the *k*th power of time
u_{0i}	Individual-level random effect of the constant term for *i*th individual
u_{ki}	Individual-level random effect of the *k*th power of time for *i*th individual
e_{ij}	Constant observation-level random effect for observation y_{ij}
e_{lij}	Observation-level random effect of *l*th power of time for observation y_{ij}
D_u	Unstructured covariance matrix unless otherwise stated.
$cov(u_{0i},u_{1i})$	Covariance between individual-level random effect u_{oi} individual-level random effect u_{1i}
$Var(u_{0i})$	Variance of individual-level random effect u_{0i}

Effectiveness and cost-effectiveness of interferon beta and glatiramer acetate in the UK Multiple Sclerosis Risk Sharing Scheme at 6 years: a clinical cohort study with natural history comparator

Jacqueline Palace, Martin Duddy, Thomas Bregenzer, Michael Lawton, Feng Zhu, Mike Boggild, Benjamin Piske, Neil P Robertson, Joel Oger, Helen Tremlett, Kate Tilling, Yoav Ben-Shlomo, Charles Dobson

Summary

Background In 2002, the UK's National Institute for Clinical Excellence (NICE) concluded that interferon beta and glatiramer acetate would be cost effective as disease-modifying therapies (DMTs) for multiple sclerosis only if the short-term disability benefits reported in clinical trials were maintained. The UK Multiple Sclerosis Risk Sharing Scheme (RSS) was established to assess whether disability progression was consistent with a cost-effectiveness target of £36 000 per quality-adjusted life-year projected over 20 years. We aimed to evaluate the long-term effectiveness and cost-effectiveness of these DMTs by comparing a cohort of patients with relapsing-remitting multiple sclerosis enrolled in the UK RSS with a natural history cohort from British Columbia, Canada.

Methods In our clinical cohort we included patients starting a DMT who were enrolled in the UK RSS who had relapsing multiple sclerosis at baseline and had at least one further clinical assessment. In our control cohort we included patients in the British Columbia multiple sclerosis database (BCMS; data collection 1980–96) who met the same eligibility criteria as for the RSS cohort. We compared disability progression at 6 years for RSS patients with untreated progression modelled from BCMS patients using continuous Markov and multilevel models. The primary outcomes were the progression ratio (treated *vs* untreated) measured both in Expanded Disability Status Scale (EDSS) score and utility. A ratio of less than 100% for EDSS implied slower than expected progression on treatment compared with off treatment; a utility ratio of 62% or less implied that the DMTs were cost effective.

Findings 5610 patients starting a DMT were enrolled in the UK RSS between Jan 14, 2002, and July 13, 2005 (72 sites), of whom 4137 were included in our clinical cohort. We included 898 BCMS patients in the control cohort who met the RSS inclusion criteria and had at least one EDSS score after baseline. RSS patients had a mean follow-up of 5·1 years (SD 1·4). Both models showed slower EDSS progression than predicted for untreated controls (Markov model, 75·8% [95% CI 71·4–80·2]; multilevel model, 60·0% [56·6–63·4]). Utility ratios were consistent with cost-effectiveness (Markov model, 58·5% [95% CI 54·2–62·8]; multilevel model, 57·1% [53·0–61·2]).

Interpretation Findings from this large observational study of treatment with interferon beta or glatiramer acetate provide evidence that their effects on disability in patients with relapsing-remitting multiple sclerosis are maintained and cost effective over 6 years. Similar modelling approaches could be applied to other chronic diseases for which long-term controlled trials are not feasible.

Funding Health Departments of England, Wales, Scotland, and Northern Ireland, Biogen Idec, Merck Serono, Bayer Schering Pharmaceuticals, Teva Pharmaceuticals Industries, UK National Institute of Health Research's Health Technology Assessment Programme.

Lancet Neurol 2015; 14: 497–505

Published Online
April 2, 2015
http://dx.doi.org/10.1016/S1474-4422(15)00018-6

See Comment page 460

Department of Clinical Neurology, Oxford University Hospitals Trust, Oxford, UK (J Palace DM); **Department of Neurology, Newcastle upon Tyne Hospitals, Newcastle upon Tyne, UK** (M Duddy MD); **Department of Biostatistics, PAREXEL International, Berlin, Germany** (T Bregenzer PhD, B Piske MSc); **School of Social and Community Medicine, University of Bristol, Bristol, UK** (M Lawton MPhil, Prof K Tilling PhD, Prof Y Ben-Shlomo PhD); **Department of Medicine (Neurology), University of British Columbia, Vancouver, BC, Canada** (F Zhu MSc, J Oger DM, H Tremlett PhD); **Department of Neurology, The Townsville Hospital, Townsville, QLD, Australia** (M Boggild MD); **Institute of Psychological Medicine and Clinical Neuroscience, Cardiff University, University Hospital of Wales, Heath Park, Cardiff, UK** (Prof N P Robertson MD); **and Department of Health, Leeds, UK** (C Dobson PhD)

Correspondence to:
Dr Jacqueline Palace, Department of Clinical Neurology, Oxford University Hospitals Trust, Oxford OX3 9DU, UK
jacqueline.palace@ndcn.ox.ac.uk

Introduction

Multiple sclerosis presents substantial health economic challenges. The disease course extends over decades and the bulk of direct and indirect health costs related to disability are not apparent until many years after diagnosis. The 1990s saw the first wave of published randomised controlled trials[1–4] of disease-modifying therapies (DMTs) in relapsing-remitting multiple sclerosis, with the subsequent use in clinical practice of interferon beta-1b (Betaferon or Betaseron, Bayer Schering Pharmaceuticals, Berlin, Germany), two formulations of interferon beta-1a (Avonex, Biogen Idec, Cambridge, MA, USA; and Rebif, Merck Serono, Darmstadt, Germany), and glatiramer acetate (Copaxone, Teva Pharmaceuticals Industries, Petah Tikva, Israel) for relapsing-remitting multiple sclerosis. Betaferon and Rebif were also later licensed for the treatment of relapsing secondary progressive multiple sclerosis.[5–7] Findings from the randomised controlled trials in relapsing-remitting multiple sclerosis showed a robust effect of all these drugs in reduction of relapse rates (about 30%) and MRI brain lesion activity. However, the effects on acquisition of disability were less certain, which has led to extensive debate[8–10] about whether the disability effects reported in

Research in context

Evidence before this study

The efficacy of the first-line disease-modifying therapies (DMTs) for multiple sclerosis—interferon beta and glatiramer acetate—has been demonstrated in short-term randomised controlled trials from 1993, although these benefits were not shown to be cost effective. Whether these DMTs have longer-term benefits for disability remains unclear in the absence of long-term randomised controlled trials. Several postmarketing studies have been done to assess the longer-term effectiveness of interferon beta or glatiramer acetate, with mixed findings. Methodological concerns have been raised about immortal time bias and the most appropriate control group; one study had differing results when it used a historical instead of contemporary comparison population. Up to 2014, no large-scale long-term prospective clinical cohort study with regular assessments has compared the rate of disability progression in a treated cohort of patients with that predicted from a natural-history untreated comparator with the aim of measuring effectiveness and cost-effectiveness of these treatments in multiple sclerosis over the long term.

Added value of this study

This study's findings show that interferon beta and glatiramer acetate result in a 24–40% reduction in disability progression in relapsing-remitting multiple sclerosis and are cost effective (using UK prices and a target of £36 000 per quality-adjusted life-year) over a 6 year follow-up period when modelled over a 20 year trajectory.

Implications of all the available evidence

Our findings suggest that these drugs represent value for money for the treatment of relapsing-remitting multiple sclerosis, and support the continued commissioning of these DMTs by health-care providers, albeit on the basis of observational data. The results are also useful in providing an evidence base to assess new DMTs by weighing their additional benefits and costs relative to these first-generation treatments. The final 10 year analysis will confirm whether these benefits are maintained. Our approach could be applicable to therapies for other chronic diseases for which long-term randomised controlled trials are unavailable or not deemed suitable.

randomised controlled trials are sustained or simply relapse related, and thus whether these short-term effects can be reasonably extrapolated over much longer periods.

These DMTs were appraised by the UK National Institute for Clinical Excellence (NICE; now known as the National Institute for Health and Care Excellence),[11] which concluded in 2002 that these drugs could not be classified as cost-effective unless the estimated effects on disability reported in the short term persisted at much the same magnitude over a period of at least 20 years. In February, 2002, the UK's Risk Sharing Scheme (RSS) for multiple sclerosis DMTs was launched,[12] representing a managed-entry agreement between the UK health departments and the manufacturers, with the MS Trust as the data custodian. Drug costs were reduced when necessary to achieve a cost-effectiveness of £36 000 (at current exchange rates roughly €50 000 or US$54 000) per quality-adjusted life-year (QALY) projected over 20 years. Projections were made with a discrete Markov model originally developed for NICE's appraisal,[13] using natural history data from London, ON, Canada, and hazard ratios (HRs) for attenuated disability progression derived from the individual randomised controlled trials of each drug. Actual disability progression was monitored in a cohort of 5600 patients, with the intention of making price adjustments to maintain cost-effectiveness should their progress differ substantially from the projections. The first interim analysis at 2 years[14] highlighted various methodological issues and provided ambiguous evidence, making interpretation difficult.

On the basis of lessons from the 2 year analysis, we identified a more suitable natural history cohort and adopted a more flexible continuous Markov model.[15] Additionally, we ran a parallel analysis using multilevel models. Here, we present RSS results for disability progression and cost-effectiveness at 6 years, to test two hypotheses: that patients on DMTs would have slower disability progression than untreated controls modelled from a natural history dataset; and that the treatment effect would be consistent with that predicted from the shorter-term trials so that the drugs remain cost effective when projected over 20 years according to NICE criteria.

For the **MS Trust** see http://www.mstrust.org.uk

Methods

Study design and participants

We included patients enrolled in the UK RSS in our clinical cohort. Patients had to fulfil the Association of British Neurologists (ABN) 2001 guidelines[12] for prescribing multiple sclerosis drugs: for relapsing-remitting multiple sclerosis (all drugs), age 18 years or older, two clinically significant relapses in the previous 2 years, and Expanded Disability Status Scale (EDSS) score 5·5 or lower; and for secondary progressive multiple sclerosis (interferon beta only), ambulant and with relapses as the main driver of advancing disability. In consultation with their physicians, patients could commence one of three formulations of interferon beta or glatiramer acetate. Patients were not randomly assigned, and the use of individual drugs represents patient and physician preferences at the time of prescribing. Patients were scheduled for annual review and assessment of EDSS score (except when the patient was in relapse),[16] irrespective of whether they stopped or switched treatment.[14] We excluded patients with secondary progressive multiple sclerosis at baseline for the primary analyses, but we did not censor patients if they progressed to secondary progres-

sive multiple sclerosis because we wanted to capture information about long-term disability progression.

We used the British Columbia multiple sclerosis database (BCMS)[17,18] as our control cohort. Established in the 1980s as a population-based database capturing about 80% of the multiple sclerosis population in British Columbia, Canada,[18] EDSS scores are recorded after a face-to-face consultation with a multiple sclerosis neurologist, with the same four core neurologists examining more than 85% of the patients during the period of our study. We used the same eligibility criteria as for the RSS to identify a subset of patients from the BCMS for the natural history cohort

All patient data for the BCMS were censored at the end of 1995 (after which DMTs became generally available in British Columbia) irrespective of whether patients commenced treatment, thereby avoiding confounding by indication. In our previously reported 2 year analysis we used the London, Ontario, dataset.[14] The main reasons for changing to the BCMS dataset were that it contained complete, contemporaneously recorded longitudinal EDSS scores (unlike the London, Ontario, dataset in which observations were retrospectively smoothed to ensure that EDSS values could only remain the same or increase), was a larger cohort (thus reducing sampling variability), covered a more recent time period (so would be less confounded by any potential secular changes), and allowed us direct access to the data to enable RSS analysts to validate different models.[15]

Ethical approval for the RSS was given by the South East Medical Research Ethics Committee (MREC 2/01/78) and all patients gave written consent. BCMS patients gave consent to be enrolled in the BCMS database and the University of British Columbia's Clinical Research Ethics Board approved the study.

Outcomes

Our primary outcome was accumulation of disability, measured both as EDSS progression and as loss of utility. Utility is a measure of society's perception of the quality of life of a patient in a given state of health. A utility of 1 represents perfect health; a utility of 0·5 implies that on average people would regard 12 months of life in that health state as equally preferable to 6 months of life in perfect health. We derived this measure from EDSS scores using previous data that reported EQ5D scores for different EDSS states (appendix; Nicola Russell, MS Trust, personal communication).[19] The EDSS progression ratio is the ratio of observed change in mean EDSS from baseline to the change predicted by the natural history model. We used this ratio to assess whether treatment had any effect on the rate of EDSS progression. A ratio of less than 100% means that EDSS progression on therapy is better than expected progression without therapy. Similarly we defined the utility progression ratio as the ratio of observed change in mean patient utility to the change predicted by the natural history model. We used this ratio to test whether the effect of the DMTs on utility progression was in line with that expected on the basis of findings from short-term randomised controlled trials; a utility progression ratio of 62% or lower means that utility progression for the drugs in aggregate is in line with, or slower than, the NICE target. The utility progression ratio can also be converted into a deviation score, which measures the difference between the required cost-effective benefit from treatment (derived from the randomised controlled trial HRs) and the observed benefit. Additionally, we compared the mean differences in EDSS with and without treatment at year 2 to compare our results with the findings from the shorter-term trials.

Statistical analysis

We used two independent modelling approaches (a continuous-time Markov model and a multilevel model) each to calculate expected disease progression, both on and off treatment, using data for both years 4 and 6; this report focuses on the year 6 results because of their longer follow-up period. Additional technical details are available in the appendix and have been previously reported.[15]

Markov modelling uses a set of transition probabilities (of moving between disease states, derived from the natural history data) which are applied to the RSS baseline EDSS distribution to predict mean EDSS and utility at year 6. Confidence intervals are derived by bootstrapping. We developed a continuous-time Markov model[20] from the BCMS data in place of the discrete Markov model used for the 2 year analysis, meaning that out-of-window EDSS assessments could be included. To permit a workable model, we rounded EDSS levels down to the nearest integer (eg, EDSS levels of 5 and 5·5 were grouped as a single state). We tested a set of candidate covariates in the development phase using the BCMS data alone. Age at onset, dichotomised at the median, resulted in the most parsimonious model. Internal validation within the BCMS was better than that with the original discrete Markov model.[15]

We derived a multilevel model from the BCMS cohort, treating EDSS as a continuous variable, allowing for variability between and within patients. The parameters of this multilevel model were then applied to the RSS baseline EDSS distribution to predict EDSS at year 6 for each individual. For the primary analysis with the multilevel model, we predicted the exact (half-integral) EDSS score for each patient, although for comparison with the Markov model we also present results after rounding down to the nearest integer EDSS. For consistency with the Markov model, we included binary age at onset as a covariate. Validation of the model was undertaken on the BCMS cohort (internal validation) and also with use of a natural history dataset from the University Hospital of Wales, UK (external validation).[21]

See Online for appendix

We used bootstrapping from the RSS cohort (with replacement) to derive 95% CIs. To allow for the

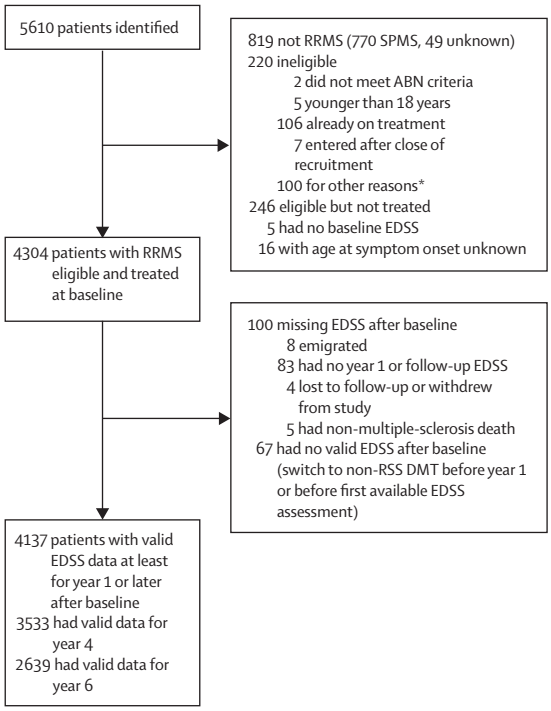


Figure: **Study profile for the UK Risk Sharing Scheme**
Data was taken from the year 8 locked dataset; follow-up data after year 6 were used only in secondary analyses to impute missing year 6 values. RSS=Risk Sharing Scheme. RRMS=relapsing-remitting multiple sclerosis. SPMS=secondary progressive multiple sclerosis. EDSS=Expanded Disability Status Scale. DMT=disease-modifying therapy. ABN=Association of British Neurologists. *Other eligibility criteria not met or "not eligible" ticked on withdrawal form.

	BCMS (n=898)	RSS, all patients* (n=4304)	RSS, primary analysis cohort (n=4137)
Sex			
Men	232 (26%)	1071 (25%)	1013 (24%)
Women	666 (74%)	3233 (75%)	3124 (76%)
Age at onset (years)			
Mean	29·2 (8·7)	30·5 (8·4)	30·5 (8·4)
Median	28 (23–35)	30 (24–36)	30 (24–36)
EDSS at baseline			
Mean	2·44 (1·70)	3·08 (1·52)	3·06 (1·52)
Median†	2 (1–3·5)	3 (2–4)	3 (2–4)
Time since symptom onset (years)			
Mean	7·9 (6·9)	7·7 (6·6)	7·7 (6·6)
Median	5·9 (2·5–11·5)	5·7 (2·6–11·0)	5·7 (2·6–11·1)
Number of confirmed relapses in past 2 years			
Mean	2·9 (1·2)	3·0 (1·3)	3·0 (1·3)
Median	2 (2–3)	3 (2–3)	3 (2–3)

Data are n (%), mean (SD), or median (IQR). BCMS=British Columbia multiple sclerosis database. RSS=Risk Sharing Scheme. *All eligible and treated patients with relapsing-remitting multiple sclerosis at baseline. †EDSS scores as half-integers.

Table 1: **Baseline characteristics of patients in the British Columbia multiple sclerosis database and Risk Sharing Scheme cohorts**

uncertainty in the BCMS model parameters, we sampled from the parameter distribution of our natural history model 500 times, predicted EDSS at 6 years for each individual, and used bootstrapping to estimate the variance of the predicted EDSS on and off treatment. The mean and variance of these 500 predictions were then combined with use of Rubin's rules.[22]

We used all EDSS points from patients who were on a scheme drug or who had ceased treatment during the study for the primary analysis. In modelling the group's predicted progression on treatment, the HR for drug effect was applied only to the time an individual was on the active therapy, and we assumed that patients would revert to the natural history progression rate after all treatment had been discontinued. If a patient switched to a non-scheme drug, they were censored at that point.

We used a wide range of prespecified supplementary analyses (appendix) to assess the sensitivity of conclusions to assumptions about stopping therapy, loss to follow-up, and inclusion of patients with secondary progressive multiple sclerosis at baseline. We did similar analyses using data to year 4 to identify any evidence of a trend. We undertook the equivalent of an intention-to-treat analysis, as might be done in a randomised controlled trial, in which we assumed that all patients continued on treatment throughout the 6 years of follow-up. Finally, we assessed whether the results were sensitive to the choice of natural history comparator by comparing EDSS progression with the transition matrices derived from the London, Ontario, and BCMS datasets.

Role of the funding source

The administration costs of the RSS are split equally among the funding parties (the Health Departments for England, Wales, Scotland, and Northern Ireland and the four manufacturers of the DMTs [Biogen Idec, Merck Serono, Bayer Schering Pharmaceuticals, and Teva Pharmaceuticals Industries]). The pharmaceutical representatives were observers at the independent scientific advisory group meetings and had no role in the study design, in collection, analysis, or interpretation of data, or in writing of the report. The statistical analysis plan was developed by the scientific advisory group and the analyses were undertaken by authors who had access to patient-level data (BCMS data, FZ, HT; RSS data, TB, ML, BP, KT, and CD). All members of the scientific advisory group and the other authors of this paper had access to all the results and were free to request further analyses. The corresponding author had final responsibility for the decision to submit for publication.

Results

Between Jan 14, 2002, and July 13, 2005, 5610 patients from 72 UK sites were enrolled into the RSS, most of whom were initiated onto DMTs (about 80% of all patients starting DMTs in the UK in this period).[14] Of

these individuals, 4304 were eligible, treated patients with relapsing-remitting multiple sclerosis (figure); 167 had no subsequent valid EDSS follow-up data, leaving 4137 patients for the primary analysis. 2639 patients had valid data collected at the year 6 visit (64%) and a further 894 patients had valid data for year 4 or year 5 (3533 [85%] had data for at least 4 years). The mean follow-up was 5·1 years (SD 1·4) and median follow-up was 6 years (IQR 5–6), equivalent to 21 260 person-years at risk. Additionally, 582 patients with missing year 6 data had data at year 7 or year 8 available at the time of analysis for use in imputation, thereby reducing the effect of participants missing at year 6 (appendix). 978 patients in the BCMS cohort met the inclusion criteria of our study, of whom 898 had at least one EDSS score after baseline; BCMS patients had a mean follow-up of 6·4 years (SD 3·5).

Table 1 compares the baseline characteristics of the BCMS and RSS cohorts. The sex distribution, time since symptom onset, and number of relapses in the 2 years before entry in the BCMS cohort were similar to those in the RSS cohort. Participants in the RSS cohort were older at symptom onset and had a higher EDSS at baseline. The modelling approaches were designed to adjust for these differences.

Table 2 shows the main outcomes for year 6. For both outcomes and with use of both methods, patients in the RSS progressed more slowly than predicted from natural history models, because all ratios were less than 100%. The Markov model resulted in a 75·8% EDSS progression ratio (95% CI 71·4–80·2) and the multilevel model resulted in a 60·0% ratio (95% CI 56·6–63·4), equivalent to a 24·2% and 40·0% relative reduction in EDSS progression, respectively. Adding in the additional uncertainty for the BCMS parameters resulted in slightly wider 95% CIs (53·8–67·5) for the multilevel model. For cost-effectiveness we noted that the central estimate of the utility progression ratio was better than the expected 62% target (Markov 58·5%, 95% CI 54·2–62·8; multilevel model 57·1%, 95% CI 53·0–61·2). In absolute terms, patients in the RSS had a mean EDSS score that was 0·28 (95% CI 0·23–0·34; Markov model) or 0·59 (95% CI 0·54–0·64; multilevel model) units less than would have been predicted off therapy.

Table 3 shows the most important supplementary analyses; a wider range of results are available in the appendix. Generally, use of intention-to-treat and imputation analyses made little difference, except when we assumed a missing-not-at-random pattern and added 0·5 of an EDSS point on top of the imputation prediction, resulting in weaker evidence of benefit. Similarly, inclusion of patients with secondary progressive multiple sclerosis at baseline also decreased the treatment effect, although in all analyses the observed progression remained slower than the natural history predictions.

The appendix shows characteristics of patients in the BCMS cohort according to the relative frequency of EDSS scores available and length of follow-up. In both the RSS and BCMS, patients with a worse prognosis tended to contribute fewer EDSS scores at longer periods

	Actual progression (95% CI)	Predicted progression (95% CI)		Absolute treatment effect (95% CI)		Relative rate of disease progression (95% CI)		
		Natural history (BCMS data)	Treated patients (UK RSS data)	Actual (predicted progression [natural history] minus actual progression)	Predicted (predicted progression [natural history] minus predicted progression [treated patients])	Actual progression divided by predicted progression (natural history)	Difference (actual absolute treatment effect minus predicted absolute treatment effect)	Deviation measure*
Utility								
Markov	0·057 (0·053 to 0·061)	0·098 (0·097 to 0·099)	0·061 (0·060 to 0·062)	0·041 (0·037 to 0·045)	0·037 (0·037 to 0·038)	58·5% (54·2 to 62·8)	0·004 (0·000 to 0·008)	–10% (–21 to 1)
MLM	0·057 (0·053 to 0·061)	0·100 (0·098 to 0·103)	0·063 (0·061 to 0·064)	0·043 (0·039 to 0·047)	0·038 (0·036 to 0·040)	57·1% (53·0 to 61·2)	0·005 (0·001 to 0·009)	–14% (–25 to –2)
Rounded EDSS†								
Markov	0·888 (0·836 to 0·940)	1·172 (1·157 to 1·188)	0·762 (0·748 to 0·776)	0·284 (0·232 to 0·336)	0·410 (0·405 to 0·415)	75·8% (71·4 to 80·2)	–0·126 (–0·178 to –0·075)	31% (18 to 43)
MLM	0·888 (0·837 to 0·940)	1·448 (1·430 to 1·466)	0·861 (0·846 to 0·875)	0·560 (0·507 to 0·613)	0·588 (0·573 to 0·603)	61·3% (58·0 to 64·7)	–0·028 (–0·079 to 0·024)	5% (–3 to 12)
EDSS								
MLM	0·881 (0·830 to 0·932)	1·468 (1·454 to 1·482)	0·894 (0·883 to 0·906)	0·587 (0·535 to 0·639)	0·573 (0·565 to 0·582)	60·0% (56·6 to 63·4)	0·013 (–0·038 to 0·064)	–2% (–10 to 5)

Data for the primary analysis cohort (n=4137), with a mean follow-up of 5·1 years (SD 1·4). Utility is a ratio varying from 0 to 1; EDSS is an ordinal scale varying from 0 to 10. BCMS=British Columbia multiple sclerosis database. RSS=Risk Sharing Scheme. EDSS=Expanded Disability Status Scale. MLM=multilevel model. *Deviation measure refers to actual treatment effect minus predicted treatment effect, divided by predicted treatment effect. †For the Markov model, half-integer EDSS states are combined with the next lower integer EDSS state. For the MLM, predicted EDSS values are rounded to the nearest point on the EDSS scale and then rounded down to the next lower integer EDSS value.

Table 2: **Accumulation of disability from the primary analysis**

	Number of patients	Mean follow-up (SD)	Absolute treatment effect (EDSS)			Deviation scores (utility)		
			Markov model	Multilevel model		Markov model	Multilevel model	
				Mean	95% CI		Mean	95% CI
Primary analysis	4137	5·1 (1·4)	0·284	0·587	0·535 to 0·639	–10%	–14%	–25 to –2
Intention-to-treat analysis*	4137	5·4 (1·2)	0·273	0·590	0·537 to 0·642	..	..	..
Imputation—Markov model								
Last value carried forward	4209	6·0 (0·2)	0·378	..	..	–32%	..	..
Linear interpolation-extrapolation	4209	6·0 (0·3)	0·226	..	..	34%	..	..
Imputation—multilevel model†								
Single imputation	4137	6·0 (0·2)	..	0·644	0·593 to 0·694	..	–7%	–17 to 4
Multiple imputation	4137	6·0 (0·2)	..	0·643	0·590 to 0·697	..	–2%	–13 to 10
Single imputation plus 0·5 EDSS (points for each imputed value)	4137	6·0 (0·2)	..	0·464	0·411 to 0·516	..	38%	26 to 49
Including patients with SPMS at baseline	4780	5·1 (1·5)	0·173	0·521	0·473 to 0·568	3%	11%	2 to 20

EDSS=Expanded Disability Status Scale. SPMS=secondary progressive multiple sclerosis. *Because the intention-to-treat analysis is essentially a comparison of actual disease progression with predicted progression off treatment (rather than predicted progression on treatment), the deviation measure cannot meaningfully be calculated for these variants. †In this analysis we assumed that patients went off treatment when they were lost to follow-up.

Table 3: **Sensitivity analyses**

of follow-up (appendix), suggesting that the less frequent availability of EDSS scores for these patients might not bias the relative comparison. By contrast, we did note differences in the quantity of data available at shorter periods of follow-up; in the BCMS dataset (but not in the RSS), patients with more severe disease tended to contribute more frequent data than did those with less severe disease. This finding could inflate the apparent treatment effect, but our calculations suggested that any such effect was small (appendix).

The results based on the year 4 data (appendix) showed slightly more benefit of treatment than did the year 6 analyses. At 2 years of follow-up we noted an absolute mean reduction of 0·22 units (95% CI 0·19–0·25) in EDSS change in the RSS by comparison with the natural history prediction. This estimate could be biased toward overly positive results because 574 (14%) of 4030 patients with data for years 1 or 2 lacked year 2 data and patients with missing data might be more likely to have worse EDSS scores.

Finally, when we applied transition probabilities from both the BCMS and the London, Ontario, datasets to the baseline EDSS distribution from the RSS, we predicted slightly more rapid progression over 10 years using the data from London, Ontario (appendix), suggesting that the switch in databases reduced the estimated treatment effect.

Discussion

We believe our results provide the best available observational data about whether DMTs change the natural history of multiple sclerosis and are cost effective, at least within the setting of the UK RSS for multiple sclerosis. The RSS provides the first example of a managed-entry agreement that incorporates long-term monitoring of outcomes to allow analysis of cost-effectiveness. Patients with active relapsing-remitting multiple sclerosis meeting the 2001 ABN prescribing guidelines, treated with either interferon beta or glatiramer acetate, seemed to progress more slowly over the first 6 years of therapy than would be predicted from a natural history cohort. If this benefit were sustained over 20 years, it would meet a cost-effectiveness target of £36000 per QALY, although cost-effectiveness would not necessarily be achieved for countries where drug costs are substantially higher.[23] We cannot comment on the benefits of specific DMTs, because our analyses included DMTs only in aggregate.

The prespecified primary analyses showed similar results across two independent analytical approaches. Generally, the multilevel model showed that the observed progression was consistent with the predicted treatment effect for utility and EDSS progression. The Markov model showed slightly worse progression than did the multilevel model, and the 95% CIs for the utility measure just overlapped the target for cost-effectiveness, although actual EDSS progression was still better (24·2%) than predicted from natural history. This finding might result from differences in how the models dealt with measurement error, and the slower progression under the natural history model predicted by the Markov model than predicted by the multilevel model. The greater similarity between the two models for utility than for EDSS probably results from the fact that different EDSS scores can have the same utility value.

An extensive range of supplementary analyses showed consistently positive, although variable, treatment effects, including when missing data were imputed under various assumptions. The addition of patients with relapsing secondary progressive multiple sclerosis at baseline weakened the observed benefit. We did not examine conversion of relapsing-remitting multiple sclerosis to secondary progressive multiple sclerosis as an outcome, because clinicians in the RSS tended to delay the diagnosis of secondary progressive multiple sclerosis (because it could result in treatment withdrawal) whereas clinicians in British Columbia would not have (because treatment was not available). This difference would have exaggerated the treatment effect.

Findings from our 2 year analysis[14] showed that the outcome was extremely sensitive to the use of a no-improvement smoothing rule to the RSS dataset so that it could be compared with the London, Ontario, dataset. The use of the BCMS database removed the need for smoothing rules to be applied to the RSS dataset. Modelling the two sets of transition probabilities to the baseline EDSS of the RSS cohort over 10 years produced similar mean EDSS trajectories.

We cannot directly compare our models with the published randomised controlled trials, partly because of differences in measured outcomes (eg, many trials used point increases in the EDSS confirmed at 3 or 6 months). Reassuringly, however, the difference in the mean EDSS change after 2 years between the treated RSS patients and predicted progression off treatment was 0·22, consistent with a mean EDSS difference of 0·25 (95% CI 0·05–0·46) reported in a meta-analysis of studies of interferon beta in relapsing-remitting multiple sclerosis.[8]

Several post-marketing studies have been done, aimed at assessment of the longer-term effectiveness of beta interferon or glatiramer acetate, with mixed findings.[24–28] Positive treatment effects were reported by investigators of an Italian study,[26] but a reanalysis[29] suggested that the benefit could have been due to immortal time bias. Immortal time bias arises because patients treated with a drug start their follow-up at first exposure, whereas those not treated start their follow-up earlier, when entering the study; because outcomes (eg, disability milestones) cannot occur in this gap for the treated individuals, they experience immortal time.[29] A previous analysis of the BCMS database used historical and contemporary control groups to determine the association between DMT exposure and disability progressions,[27,28] with use of time to EDSS 6 as the outcome. The HR when comparing their contemporary treated patients with a historical control group (0·77, 95% CI 0·58–1·02) was consistent with our results; we noted the EDSS progression ratio for most of our analyses to be about 70%. However, the results calculated with use of a contemporary control group did not show evidence of any benefit for DMTs, which the investigators attributed to possible residual indication bias despite using sophisticated analytical methods.[27,28]

Our results, like those from other studies, suggest that these treatments only slow down and do not completely halt the progress of disability. However, our findings would support the notion that treatment of the inflammatory processes of multiple sclerosis in the relapsing-remitting stage leads to a downstream effect on neurodegeneration and reduced long-term disability.

Do these results represent a treatment effect? There are several potential non-causal explanations. First are geographical differences. We would have noted an apparent treatment effect if the natural history of Canadian patients with multiple sclerosis were more aggressive than that of our UK patients. Although British Columbia has an ethnically diverse population, most patients were of European ancestry during the study period.[30] We validated the BCMS multilevel model in an independent cohort of untreated patients and found that the model fitted the Welsh data well (appendix).

Second are temporal differences. The BCMS data come from an earlier period than the RSS, and relapse rates in the placebo groups of clinical trials in multiple sclerosis have been reducing since the 1980s,[31] suggesting that the natural history of multiple sclerosis might be improving. However, recent randomised controlled trials might be enrolling milder cases of multiple sclerosis as patients with more severe disease are treated in clinical practice with available drugs (and could be unwilling to receive a placebo). There have been other changes such as the use of early rescue protocols and stricter definitions for relapses.[31] Population-based cohorts are less likely to be biased by selection and neither the BCMS dataset[32] nor recent data from south Wales (Robertson NP, unpublished) have shown evidence for milder prognosis in more recent cohorts.

The third explanation is confounding by indication. A well recognised problem in pharmacoepidemiology, variables such as disease severity and comorbidity confound treatment decisions and clinical outcomes. This issue did not arise in our analysis because all the BCMS patients were untreated and the sample was selected on the basis of explicit inclusion criteria that applied to both cohorts. Both models took account of differences in the baseline EDSS distribution and age at entry; other covariates added little to the prediction of progression (appendix).[15]

The fourth non-causal explanation is selection bias. The RSS cohort probably would have excluded some patients with rapid disease progression who would have been eligible for treatment according to the ABN criteria in the past but were no longer eligible in 2002. This bias would favour a treatment effect. Working in the opposite direction, patients with mild disease who were no longer actively relapsing at the time of recruitment would have been excluded. Similarly, the presence of prevalent as well as incident ABN-eligible patients would have led to

patients with later disease being treated, and DMTs might be less effective later on in the disease course. Reassuringly, time from symptom onset to cohort entry was almost identical for both datasets (table 1).

The fifth potential explanation is loss-to-follow-up bias. The pattern of loss to follow-up could differ between the cohorts. The available evidence (appendix) suggests that, although there are some differences as well as some similarities in the patterns of follow-up between the two datasets, the differences are unlikely to account for more than a small proportion of the apparent treatment effect.

The strengths of our study are the large number of patients followed up prospectively with regular EDSS assessments, an acceptable follow-up rate, and a high-quality, historical, natural history comparator. To some researchers, no observational study—no matter how rigorous—will ever provide as convincing evidence for treatment effects as a large, well conducted randomised controlled trial. However, long-term randomised controlled trials are difficult to implement after drugs have been shown to have short-term benefits, and cannot generally assess the effectiveness of treatments in routine clinical practice. Careful assessment of observational data can then provide important supportive evidence about long-term benefits.[33,34] NICE allows only direct costs to be incorporated into its cost-effectiveness analysis, so cost-effectiveness might be even greater if indirect costs such as loss of work for the patient or their carers is included.

This is an exciting time in multiple sclerosis therapy, with new drugs becoming available. Ensuring the cost-effectiveness of increasingly expensive drugs is becoming imperative. Managed-entry and risk-sharing agreements between commissioners and manufacturers are increasingly used to deliver value for money of new and expensive therapies. The application of prognostic models supports the use of this type of scheme for other chronic diseases for which long-term trials are not thought appropriate. This 6 year analysis supports a predicted long-term effect of multiple sclerosis DMTs in patients with relapsing-onset disease, consistent with their UK cost-effectiveness at an aggregate level. The final 10 year analysis will confirm whether these benefits are maintained.

Contributors

JP, MB, and CD designed the study. JP, MD, MB, NPR, JO, and HT collected the data. TB, ML, FZ, BP, KT, YB-S, and CD analysed the data. JP, MD, TB, ML, FZ, MB, BP, NPR, JO, HT, KT, YB-S, and CD interpreted the data. JP, MD, TB, ML, FZ, MB, BP, NPR, JO, HT, KT, YB-S, and CD wrote the report. KT designed the multilevel model approach.

UK Risk Sharing Scheme investigators

72 centres have participated in the collection of data for the Risk Sharing Scheme from across the UK. Data have been collected over a 10 year period and thus changes have occurred within the clinical teams at the centres. The principal investigators are: Amir Ali-Din, Sandeep Ankolekar, David Barnes, Bob Brenner, Abhijit Chaudhuri, Sam Chong, Peter Cleland, Cris Constantinescu, Shane Delamont, David Dick, Martin Duddy, Edward Fathers, David Footitt, Helen Ford, David Francis, Andrew Gale, Andrew Graham, Nicholas Gutowski, James Harley, Clive Hawkins, Stanley Hawkins, Andrew Heald, Abdulhamied Hewazy, Charles Hillier, Matthew Jackson, Brian Kendall, Angus Kennedy, Paul Lyons, Andreas Malaspina, Omar Malik, Roswell Martin, Paul Mattison, Gordon Mazibrada, Brendan McLean, Shafqat Ullah Memon, David Miller, Roger Mills, Christian Neumann, Piers Newman, Esmaeil Nikfekr, Colin O'Leary, Jonathon O'Riordan, Michael Osei-Bonsu, Jacqueline Palace, Owen Pearson, Wojciech Pietkiewicz, Sian Price, Jennifer Quirk, Ian Redmond, Neil P Robertson, Gerard Saldanha, Lara Sanvito, Neil Scolding, Mohammad Sharief, Basil Sharrack, Abdullah Shehu, Eli Silber, Uwe Spelmeyer, Stephen Sturman, Paul Talbot, John Thorpe, Alan Turner, Ben Turner, Lal Vaithianathar, Rodney Walker, Andrew Weir, Belinda Weller, Victoria Williams, Tilo Wolf, Damien Wren, Carolyn Young, John Zajicek, Ioannis Zoukos, Tadas Zuromskirs.

British Columbia multiple sclerosis (BCMS) neurologists

The BCMS database was created by Donald Paty (deceased) and Donald Studney, and has been maintained over the years by grants from various sources, including the MS Society of Canada, the MS/MRI Research Group, the US National MS Society, and the Canadian Institutes of Health Research. The current BCMS neurologists are: A Traboulsee, A-L Sayao, V Devonshire, S Hashimoto, J Hooge, L Kastrukoff, J Oger, D Adams, D Craig, S Meckling, L Daly, O Hrebicek, D Parton, K Attwell-Pope.

Declaration of interests

JP serves on the scientific advisory board for the Charcot Foundation, and has performed advisory work for Biogen Idec, Merck Serono Ltd, Bayer, Novartis UK Ltd, Teva UK Ltd, Ono Pharmaceutical Co Ltd, Primary i-research, Chugai Pharma Europe, and CI Consulting. She receives research support from the MS Society, QIDIS, Merck Serono Ltd, Novartis and Bayer, plus conference expenses from Novartis and Merck Serono Ltd. Her NHS trust has received funding for her RSS clinical lead role. MD has received speaker honoraria, consulting fees, and travel grants from, Bayer, Biogen Idec, Novartis UK Ltd, Merck Serono Ltd, and Teva UK Ltd over the past 5 years. His NHS trust has received funding for his RSS clinical lead role. TB is an employee of PAREXEL International (Department of Biostatistics) and in this role has worked for numerous pharmaceutical companies, including those participating in the UK MS Risk Sharing Scheme. ML received funds from the Health Technology Assessment Programme to develop the multilevel model with additional analysis time funded by RSS funders. MB sits on advisory boards for Bio CSL, Genzyme, and Biogen Idec. He has received sponsorship to attend international meetings from Novartis and BioCSL. His department has received funding to develop services from Biogen Idec, Genzyme, and Novartis. The Walton Centre, Liverpool, previously received funding for his RSS clinical lead role. BP is an employee of PAREXEL International (Department of Biostatistics) and in this role has worked for numerous pharmaceutical companies, including those participating in the UK MS Risk Sharing Scheme. NPR has received support for attendance at scientific meetings and honoraria for advisory work from Biogen Idec, Merck Serono, Novartis, Sanofi, and Bayer, and has received unrestricted research grants from Genzyme. JO has received, in the past 5 years, speaker honoraria, consulting fees, travel grants, research grants, or educational grants of less than CAD$5000 each from, Aspreva, Aventis, Bayer, Biogen Idec, BioMS, Berlex, Bristol-Myers Squibb, Genentech, GlaxoSmithKline, Novartis, Merck Serono, Schering, Talecris, and Teva Neurosciences, and has received fees for services from Bayer, Novartis, and Biogen Idec to serve on advisory committees. HT is funded by the Multiple Sclerosis Society of Canada (Don Paty Career Development Award), and is a Michael Smith Foundation for Health Research Scholar and the Canada Research Chair for Neuroepidemiology and Multiple Sclerosis. She has received research support from the National Multiple Sclerosis Society, the Canadian Institutes of Health Research, and the UK MS Trust; speaker honoraria or travel expenses to attend conferences from the Consortium of MS Centres (2013), the National MS Society (2012–14), Bayer Pharmaceuticals (2010), Teva Pharmaceuticals (2011), ECTRIMS (2011–13), UK MS Trust (2011), the Chesapeake Health Education Program, US Veterans Affairs (2012), Novartis Canada (2012), Biogen Idec (2014), and the American Academy of Neurology (2013–14). All speaker honoraria were donated as an unrestricted grant for use by her research group. KT has received funds from the Health Technology

Assessment Programme to develop the multilevel model with additional analysis time funded by the RSS funders. YB-S's department received funds from the Health Technology Assessment Programme to develop the multilevel model with additional analysis time funded by the RSS funders. He has a relative who is on a disease-modifying therapy for multiple sclerosis. CD is an employee of the Department of Health of England, which is one of the parties to the Risk Sharing Scheme. He declares no financial interests. FZ declares no competing interests.

Acknowledgments

The multilevel modelling project was funded by the NIHR Health Technology Assessment programme (HTA project 10/55/01) and will be published in full in the Health Technology Assessment journal series. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Department of Health. The MS Trust has had an administrative coordinating role in the Risk Sharing Scheme and PAREXEL has done the data collection from the sites and undertaken the analytical work. The scientific advisory group is chaired by Richard Lilford and consists of Pelham Barton, Richard Gray, and Yoav Ben-Shlomo. The views expressed in this paper do not necessarily reflect the views of each individual acknowledged.

References

1 The IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* 1993; **43:** 655–61.

2 Jacobs LD, Cookfair DL, Rudick RA, et al. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG). *Ann Neurol* 1996; **39:** 285–94.

3 PRISMS (Prevention of Relapses and Disability by Interferon β-1a Subcutaneously in Multiple Sclerosis) Study Group. Randomised double-blind placebo-controlled study of interferon β-1a in relapsing/remitting multiple sclerosis. *Lancet* 1998; **352:** 1498–504.

4 Johnson KP, Brooks BR, Cohen JA, et al. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: results of a phase III multicenter, double-blind placebo-controlled trial. The Copolymer 1 Multiple Sclerosis Study Group. *Neurology* 1995; **45:** 1268–76.

5 European Study Group on interferon β-1b in secondary progressive MS. Placebo-controlled multicentre randomised trial of interferon β-1b in treatment of secondary progressive multiple sclerosis. *Lancet* 1998; **352:** 1491–97.

6 Secondary Progressive Efficacy Clinical Trial of Recombinant Interferon-Beta-1a in MS (SPECTRIMS) Study Group. Randomized controlled trial of interferon-beta-1a in secondary progressive MS: Clinical results. *Neurology* 2001; **56:** 1496–504.

7 North American Study Group on Interferon β-1b in Secondary Progressive MS. Interferon beta-1b in secondary progressive MS: results from a 3-year controlled study. *Neurology* 2004; **63:** 1788–95.

8 Filippini G, Munari L, Incorvaia B, et al. Interferons in relapsing remitting multiple sclerosis: a systematic review. *Lancet* 2003; **361:** 545–52.

9 Clegg A, Bryant J. Immunomodulatory drugs for multiple sclerosis: a systematic review of clinical and cost effectiveness. *Expert Opin Pharmacother* 2001; **2:** 623–39.

10 La Mantia L, Munari LM, Lovati R. Glatiramer acetate for multiple sclerosis. *Cochrane Database Syst Rev* 2010; **5:** CD004678.

11 National Institute for Clinical Excellence. Beta interferon and glatiramer acetate for the treatment of multiple sclerosis. NICE technology appraisal guidance no. 32. London: National Institute for Clinical Excellence, 2002.

12 Department of Health. Cost effective provision of disease modifying therapies for people with multiple sclerosis. Health Service Circular (2002/004). London: Stationery Office, 2002.

13 Chilcott J, McCabe C, Tappenden P, Cooper NJ, Abrams K, Claxton K. Modelling the cost-effectiveness of interferon beta and glatiramer acetate in the management of multiple sclerosis. *BMJ* 2003; **326:** 522–26.

14 Boggild M, Palace J, Barton P, et al. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. *BMJ* 2009; **339:** b4677.

15 Palace J, Bregenzer T, Tremlett H, et al. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. *BMJ Open* 2014; **4:** e004073.

16 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983; **33:** 1444–52.

17 Tremlett H, Paty D, Devonshire V. Disability progression in multiple sclerosis is slower than previously reported. *Neurology* 2006; **66:** 172–77.

18 Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. *Neurology* 2010; **74:** 2004–15.

19 Orme M, Kerrigan J, Tyas D, Russell N, Nixon R. The effect of disease, functional status, and relapses on the utility of people with multiple sclerosis in the UK. *Value Health* 2007; **10:** 54–60.

20 Jackson CH, Sharples LS, Thompson SG, et al. Multistate Markov models for disease progression with classification error. *J R Stat Soc* 2003; **52:** 193–209.

21 Harding KE, Wardle M, Moore P, et al. Modelling the natural history of primary progressive multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2015; **86:** 13–19.

22 Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: J Wiley & Sons, 1987.

23 Noyes K, Bajorska A, Chappel A, et al. Cost-effectiveness of disease-modifying therapy for multiple sclerosis: a population-based study. *Neurology* 2011; **77:** 355–63.

24 Brown MG, Kirby S, Skedgel C, et al. How effective are disease-modifying drugs in delaying progression in relapsing-onset MS? *Neurology* 2007; **69:** 1498–507.

25 Ebers GC, Traboulsee A, Li D, et al, for the Investigators of the 16-Year Long-Term Follow-up Study. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal IFNB-1b trial. *J Neurol Neurosurg Psychiatry* 2010; **81:** 907–12.

26 Trojano M, Pellegrini F, Fuiani A, et al. New natural history of interferon-beta-treated relapsing multiple sclerosis. *Ann Neurol* 2007; **61:** 300–06.

27 Shirani A, Zhao Y, Karim ME, et al. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. *JAMA* 2012; **308:** 247–56.

28 Karim ME, Gustafson P, Petkau J, et al. Marginal structural Cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *Am J Epidemiol* 2014; **180:** 160–71.

29 Renoux C, Suissa S. Immortal time bias in the study of effectiveness of interferon-beta in multiple sclerosis. *Ann Neurol* 2008; **64:** 109–10.

30 Lourenco P, Shirani A, Saeedi J, Oger J, Schreiber WE, Tremlett H, for the UBC MS Clinic Neurologists. Oligoclonal bands and cerebrospinal fluid markers in multiple sclerosis: associations with disease course and progression. *Mult Scler* 2013; **19:** 577–84.

31 Inusah S, Sormani MP, Cofield SS. Assessing changes in relapse rates in multiple sclerosis. *Mult Scler* 2010; **16:** 1414–21.

32 Shirani A, Zhao Y, Kingwell E, Rieckmann P, Tremlett H. Temporal trends of disability progression in multiple sclerosis: findings from British Columbia, Canada (1975–2009). *Mult Scler* 2012; **18:** 442–50.

33 Smeeth L, Douglas I, Hall AJ, Hubbard R, Evans S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. *Br J Clin Pharmacol* 2009; **67:** 99–109.

34 Hodgson R, Bushe C, Hunter R. Measurement of long-term outcomes in observational and randomised controlled trials. *Br J Psychiatry* 2007; **50** (suppl)**:** s78–84.

For the **HTA Programme** see http://www.nets.nihr.ac.uk/projects/hta/105501

THE LANCET Neurology

Supplementary webappendix

This webappendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Palace J, Duddy M, Bregenzer T, et al. Effectiveness and cost-effectiveness of interferon beta and glatiramer acetate in the UK Multiple Sclerosis Risk Sharing Scheme at 6 years: a clinical cohort study with natural history comparator. *Lancet Neurol* 2015; published online April 2. http://dx.doi.org/10.1016/S1474-4422(15)00018-6.

Appendices to "Assessing the long-term effectiveness and cost-effectiveness of interferon-beta and glatiramer acetate in the UK multiple sclerosis risk sharing scheme at six years: a clinical cohort study with natural history comparator"

Appendix 1: Background to the UK MS Risk Sharing Scheme

1.1 The MS disease modifying therapies (DMTs) are licensed for relapsing MS ie for early stage disease, whereas the major disability is incurred in the later progressive phase. However none of the DMTs have been shown to be effective in reducing progressive disability when instituted during the later stages of disease when relapse-free or independent free progressive disability occurs. Thus the main debate in extrapolating short-term effects of MS treatments to the longer-term is around whether the later progressive disability phase is attributable to downstream events of remote acute inflammation (which are affected by DMTs) or to ongoing compartmentalised chronic CNS inflammation and/or neurodegeneration (which appear not to be directly affected by interferon beta or glatiramer acetate).

1.2 In order for the DMTs to be cost-effective they must demonstrate that they do more than only prevent early relapses in RRMS. They must delay an individual reaching higher levels of disability particularly because this is where the major costs are incurred. For example, transient clinical relapses have only a modest impact on cost (in the UK , mean £1500/attack). Fixed disability related to MS has however a substantial cost, for example in 2001 the NICE model[1] estimated £5,678/yr for an individual with an EDSS of 6 (requiring a stick to walk 100m) and £17,327 at EDSS 7.0. The NICE models have only allowed direct costs to be taken into consideration, adding in indirect costs would make these values even greater.

1.3 The EDSS used to calculate MS disability is a 20 point scale[2] (0 and then 1 to 10 in 0.5 steps) incorporating clinical signs on examination and observed and reported impairments and disability. Population studies have ascribed a utility (quality of life) value to each integral point of the EDSS scale[3,4], where 1 is the maximum and represents a year of normal health, 0.5 would represent the general population's view that 2 years at this health state would be worth 1 in full health and 0 is equivalent to death. By converting different outcomes into a common measure of benefit ie quality adjusted life years (QALYs) it is possible to compare cost-effectiveness of different treatments even across different diseases.

1.4 The UK MS risk sharing scheme (RSS)[5] has two components. Firstly, the then current UK prices for the four main DMTs were reduced, where necessary, in order to achieve a cost effectiveness target of £36,000 per QALY, using the model of disability progression developed for NICE's 2001 technology appraisal and target treatment effects (relative rates of disability progression) derived from the pivotal RCTs. Secondly, the parties to the scheme agreed to track a prospectively observed MS cohort on DMTs against the trajectory required to offer cost effectiveness. The plan was to assess the data every two years, adjusting the price if necessary after each analysis to maintain the 20 year cost-effectiveness target should the observed results differ significantly from the required trajectory.

1.5 Considering the practicalities of running such a scheme, the decision was made to follow the cohort for the first 10 years of the twenty year model. Over 5,500 patients were enrolled in 2002/2003. EDSS scores were collected pragmatically, ie as part of normal clinical practice. Although the EDSS was not assessed under trial standard protocols, it was performed by MS neurologists who were experienced in this scoring with the same neurologist being encouraged to continue scoring an individual patient throughout if possible.

1.6 Since 2005, the parties to the scheme have been advised by a Scientific Advisory Group (SAG) chaired by Professor Richard Lilford. The group's main function is to advise on the analysis plan for the interim and final analyses and on the interpretation of the results. In addition, the group advise the MS Trust (as the custodian of the data) on applications from other researchers to access the data on the RSS cohort.

1.7 As published previously, the first interim analysis at two years[6] revealed shortcomings in the original modelling and highlighted difficulties in interpretation of the results because the Ontario dataset did not record any EDSS improvements due to smoothing of collected data. Applying a 'no improvement' rule to the treated RSS patients was felt to bias against a treatment effect and indeed

removing this rule changed the results from a negative treatment effect (disability progression for treated patients worse than for untreated patients) to a marked 'over target' effectiveness. This led the scientific advisory group to advise that the results were too unreliable to recommend price adjustment at the 2 year stage and to produce a new scientific analysis plan agreed by all parties.

1.8 The cost-effective target for the four drugs is presented in aggregate, and this has been calculated to be equivalent to a relative rate of disability progression (on treatment versus off treatment, with disability measured using the EDSS scale) of 62% (ie at least a 38% treatment effect). This cost-effectiveness target is set for the utility outcome. Despite the utility score being obtained from the EDSS, because these scores have a non-linear relationship the results on these outcomes will not always be the same. The ratios for individual DMTs were agreed between the companies concerned and the Department of Health and remain commercially confidential.

Appendix 2: Definition of the deviation measure

2.1 According to the fundamental principles of the UK risk sharing scheme (RSS), adjustments to the prices paid by the NHS to the companies for DMTs are calculated on the basis of any deviation between the actual outcomes for patients in the RSS cohort and the "target outcomes" predicted on the basis of the Markov model. In assessing these outcomes (benefits of treatment), a comparison is made between the disability progression of treated patients and the expected progression of a similar cohort of untreated patients, modelled using the same Markov model (a "virtual placebo group").

2.2 The basic calculation is set out in paragraphs 19 and 20 of the Health Circular[3]. The circular refers to the deviation as a 'shortfall', but the term 'deviation' is now preferred since it does not imply an expectation that the actual outcomes will be worse than their targets.

2.3 Denoting the deviation measure at year *t* by $S(t)$, this is defined as the difference between actual and expected benefit expressed as a percentage of expected benefit, or:

$$S(t) = \frac{B_e(t) - B_a(t)}{B_e(t)} \cdot 100\%, \qquad (1)$$

where

$B_e(t)$ is the *expected* benefit from treatment in terms of slowing disease progression for the cohort, and

$B_a(t)$ is the corresponding actual benefit.

These quantities are defined further as:

$$B_e(t) = D_e(t) - D_e^*(t), \quad B_a(t) = D_e(t) - D_a^*(t), \qquad (2)$$

where

$D_e^*(t)$ is the *expected* value of disease progression at year *t* with treatment,

$D_a^*(t)$ is the *actual* value of the disease progression at year *t* with treatment,

$D_e(t)$ is the *expected* value of the disease progression at year *t* without treatment.

2.4 Disease progression can be defined in either utility terms (the primary outcome measure) or in terms of EDSS progression (secondary outcome). In utility terms, disease progression at *t* years after recruitment to the scheme is defined as:

$$D(t) = \sum_i (P_i(t) - P_i(t_0))(1 - u_i) \quad (3)$$

where

$P_i(t)$ is the proportion of patients in EDSS i at year t (t_0: baseline),

u_i is the utility for patients in EDSS i.

The summation in equation (3) is over all possible EDSS scores from 0 to 10. To calculate the deviation measure In terms of EDSS progression, the EDSS score i is substituted for the term (1 - u_i) in the equation.

Appendix 3: Changes to the model after the year 2 analysis

3.1 As noted in Appendix 1, the problems with the year 2 analysis[6] led the scheme's Scientific Advisory Group (SAG) to advise that the results should not be used as the basis of price adjustment. In addition, the group advised that the parties to the scheme should seek an alternative natural history comparator.

3.2 For the purpose of the RSS the main issue with the London Ontario natural history dataset is that the data were retrospectively adjusted to avoid EDSS scores improving from one year to the next. In contrast, one-year improvements in measured EDSS scores in the RSS cohort are very frequent. It therefore becomes very difficult to ensure a like-for-like comparison between the "virtual placebo group" derived from the natural history data, and the treated patients in the RSS cohort. In the year 2 analysis, an attempt was made to adjust the RSS data to mimic the "no improvement rule" in the natural history data, but sensitivity analysis showed that the outcomes were highly sensitive to the way in which this rule was applied.

3.3 After reviewing a number of alternatives, SAG recommended that the scheme should use the British Columbia MS (BCMS) dataset[7] as the natural history comparator. This is not only the largest dataset of its kind, with a high proportion of patients followed up for 10 or more years, but also contains "raw" (unadjusted) EDSS data showing the same sort of year to year improvements as the RSS data.

3.4 The opportunity was taken to make a number of other improvements to the Markov model used for the calculations, including modelling disability improvement and allowing the disability trajectory for individual patients to depend on baseline coordinates. Details are given in a published paper[8].

3.5 Given these changes, it is not surprising that the year 4 and year 6 results which we present in this paper are so different from the primary analysis in the published year 2 results[6]. However, one of the sensitivity analyses reported at year 2 (involving minimal use of the "no improvement" algorithm) gave very similar results to the ones we present here. In addition, although we had not planned to repeat the year 2 analysis using the new dataset and model, a limited retrospective analysis suggests that had we done so the results for year 2 would be fully in line with those for year 4 and year 6.

Appendix 4: Relation between EDSS and utility

4.1 We examined three possible sources for the relation between EDSS and utility (quality of life): two surveys of patients with MS[3,4], using patient-determined measures of disability (the "MS Trust" and "Heron" datasets), and an unpublished paper[9] drawing on the clinician-determined EDSS scores in the RSS itself (the "Boggild dataset"). In each case, patients were asked to completed the EQ5D questionnaire, an instrument which categorises the patient's perceived state of health according to the dimensions of mobility, self-care, ability to take part in usual activities (eg work), pain/discomfort and anxiety/depression. The EQ5D scores can then be converted into utility, an overall measure of society's perception of the patient's quality of life, using standard tariffs[10]. A utility of one represents perfect health; a utility of 0.5 implies that on average members of the general population would be regard 12 months of life in that health state as equally preferable as 6 months of life in perfect health.

4.2 On advice from our Scientific Advisory Group, we decided to use a synthesis of the MS Trust and Heron datasets for the primary analysis, primarily because they contained more data for the higher EDSS scores. We also carried out a sensitivity analysis using a synthesis of all three datasets. The utility values we adopted are given in the table below:

Utility values used in the year 6 analysis

EDSS	Utility (primary analysis)	Utility (secondary analysis)
0	0.9248	0.8722
1 or 1.5	0.7614	0.7590
2 or 2.5	0.6741	0.6811
3 or 3.5	0.5643	0.5731
4 or 4.5	0.5643	0.5731
5 or 5.5	0.4906	0.5040
6 or 6.5	0.4453	0.4576
7 or 7.5	0.2686	0.2825
8 or 8.5	0.0076	0.0380
9 or 9.5	-0.2304	-0.2246

4.3 Details of the three datasets and of the methodology used for their synthesis can be found in a report by IMS Health[11].

Appendix 5: Further details on the Markov and Multi-Level Models

5.1 The parameters for the two models representing disease progression for untreated patients were estimated using a subset of the BCMS dataset selected according to two key criteria:

a. at some clinic visit before 31 December 1995 they would have met the 2001 criteria of the UK's Association of British Neurologists for eligibility for treatment with a DMT;

b. they had at least one further EDSS measurement before the cut-off date of 31 December 1995 (or before first treatment with a DMT if this was earlier).

This left a total of 898 patients for analyses, followed up for a median period of 6.4 years before the cut-off date. The MLM included an additional 80 patients with only one EDSS

Markov model

5.2 The Markov model defines its states in terms of the patient's EDSS score rounded down to the nearest integer. Thus patients at EDSS 0 are allocated to state 1, patients at EDSS 1 or 1.5 are allocated to state 2, and so on. The model assumes a constant probability of making a transition from state i to state j conditional on the vector of baseline covariates $\boldsymbol{x}$ for the individual patient. For the purposes described in this paper (calculation of the "deviation measure" and other outcomes) death from non-MS causes was not explicitly modelled and patients in the RSS who died before the final analysis year were treated as lost to follow up.

5.3 Estimation was by the method of Jackson[12] and, after assessing a number of possible combinations of baseline covariates, a relatively simple model with a single baseline covariate, age at onset, was chosen[8].

5.4 For treated patients, it is assumed that DMTs affect only the probability of forward transitions (transitions to a higher EDSS state) and not the probability of backward transitions (transitions to a lower EDSS state). The instantaneous hazard ratios applied to the forward transition probabilities, which are different for each DMT, are related to the "target hazard ratios" agreed between the UK Health Departments and the 4 companies at the outset of the scheme, but have to be adjusted for use in a model allowing backward as well as forward transitions. The basis of this adjustment is described in the report from the scheme's Scientific Advisory Group[13].

Repeated measures multi-level model

5.5 We modelled the EDSS scores of individuals with MS using multilevel models[14]. Our model had two levels; observations (level 1) within individuals (level 2). We used a model with a random intercept and two random powers of time since ABN eligibility: time and the log of time. We also allowed level-1 variation to change linearly with time, to take into account varying measurement error in EDSS scores at different levels of disability. Thus the basic model is of the form:

$$y_{ij} = \beta_0 + u_{0i} + e_{1ij} + (\beta_1 + u_{1i} + e_{2ij})t_{ij} + (\beta_1 + u_{2i}) \cdot \log t_{ij},$$

$$\text{where } \{e_{lij}\} \sim N_2(0, D_e), \{u_{li}\} \sim N_3(0, D_u),$$

$$D_e = \begin{bmatrix} var(e_{1ij}) & cov(e_{1ij}, e_{2ij}) \\ cov(e_{1ij}, e_{2ij}) & 0 \end{bmatrix} \text{ and} \qquad (2)$$

$$D_u = \begin{bmatrix} var(u_{0i}) & cov(u_{0i}, u_{1i}) & cov(u_{0i}, u_{2i}) \\ cov(u_{0i}, u_{1i}) & var(u_{1i}) & cov(u_{1i}, u_{2i}) \\ cov(u_{0i}, u_{2i}) & cov(u_{1i}, u_{2i}) & var(u_{2i}) \end{bmatrix}$$

where y_{ij} is the EDSS for individual *i* at occasion *j* and t_{ij} is the time since ABN eligibility (plus one year) for individual *i* at occasion *j*.

5.6 We then included the binary covariate of age at onset of MS (as in the Markov model), allowing this to be associated with intercept, time and log of time. We assessed the Normality of the residuals, and the fit of the model by comparing the actual and predicted EDSS values. All analyses were carried out using Stata software[15], and all multilevel models estimated using the runmlwin command[16]. We then used the coefficients from the BCMS model to predict the EDSS value for all patients in the RSS cohort, conditional on their first observed EDSS[17]. We assumed that the hazard ratio multiplied the BCMS rate of progression to give the on-treatment rate of EDSS progression.

5.7 We used the random effects matrices from the BCMS model to estimate the "natural history" EDSS for those in the RSS cohort (at every time at which they had an observed EDSS), conditional on their observed baseline EDSS. The same approach was used to calculate the "treated" EDSS for those in the RSS cohort (by multiplying the progression by the hazard ratio). When calculating the observed, natural history and treated progression, we used the observed EDSS as the comparator, for consistency with the Markov analysis. We assessed the sensitivity of our analysis to this assumption by also using the estimated EDSS as baseline as the comparator. This did not change the deviation scores (as the choice of comparator cancels out in the numerator and the denominator), but led to a slightly higher "actual" progression (1.003), natural history progression (1.589) and treated progression (1.016). The estimated relative rate of disease progression was 63.1% (95% CI 56.6% to 63.4%).

Appendix 6: Descriptive analyses of possible sources of bias in RSS dataset

6.1 We carried out a number of descriptive analyses of the RSS dataset in order to identify possible factors which might result in bias. No obvious differences were noted in the baseline characteristics and follow up outcomes of those patients who registered but were not eligible for the RSS analysis versus those that were included in the analysis, and the baseline characteristics were similar between those who had no EDSS data after baseline compared to those with. No apparent differences were seen in the outcomes from centres with the most complete available data versus those centres with the most missing data.

6.2 In two instances, where this preliminary work suggested that there might be a significant source of bias, appropriate sensitivity analyses were added to the pre-specified list. These related to (a) patients who did not start treatment within the 3 months of baseline specified in the Statistical Analysis Plan, who might be expected to have worse outcomes as a result of the delay in starting treatment; (b) patients who switched to a DMT not part of the scheme, who tend to have worse prognostic factors at baseline and worse outcomes than those in the primary analysis population (as expected, since these are generally treatments for more aggressive MS).

Appendix 7: Potential bias due to differential patterns of data collection between the BCMS and RSS datasets

7.1 We paid considerable attention to the possibility of bias resulting from different patterns of data collection in the RSS cohort as compared to the BCMS dataset used to estimate the parameters for the Markov and multi-level models.

7.2 In the RSS cohort, we have observed a tendency for patients with worse disease progression to fail to attend at subsequent annual reviews, probably because there is little incentive for a patient to attend a review if they have already decided to discontinue DMT treatment. This tendency was already noticeable in the year 2 analysis[6] and was confirmed by the descriptive analyses carried out as part of this year 6 analysis. We have attempted to quantify the potential impact of this differential loss to follow up through the various imputation methods described in the main paper.

7.3 The BCMS dataset was not collected as part of a specific observational study but through routine clinic visits. During the period in which the data used in this study was collected (1980-1995) effective disease modifying treatments were not available. Patterns of data collection could therefore be different from those in the RSS.

7.4 Web table 4 shows the results of a descriptive analysis of data from the BCMS dataset. It compares baseline parameters and mean EDSS progression over the first 5 years from baseline for two pairs of subsets of the total cohort:

a. patients contributing relatively frequent vs relatively infrequent data, where "frequency" was defined as the number of EDSS scores divided by the interval between first and last scores;

b. patients who had a recorded EDSS score either after or within 18 months of the cut-off date of 31 December 1995, vs patients without such a score (who could therefore be regarded as "lost to follow-up").

7.5 The first comparison shows that patients with relatively frequent EDSS scores tended to have worse prognostic factors at baseline and worse disease progression in the first five years. (British Columbia clinicians have confirmed that, in their experience, patients in the province were more likely to attend clinic if they had concerns over the progress of their disease.) Since the patients with faster disease progression are contributing more EDSS scores to the estimation process, there is at least a theoretical possibility that they could bias the estimates in the Markov model towards predicting higher rates of disease progression for untreated patients, and thus inflate the apparent treatment effect when compared with the actual rates of disease progression in the treated RSS cohort. (In the MLM, this bias may not be operating if the data are "missing at random", ie if the probability of data being missing depends only on parameters explicitly included in the model.)

7.6 In contrast, the second comparison suggests that patients with relatively fast disease progression are more likely to be "lost to follow-up", as they are in the RSS cohort, and will thus contribute less data at longer periods from baseline. This would be expected to bias the estimates towards predicting lower rates of disease progression and would tend to offset any bias resulting from the similar differential loss to follow-up in the RSS dataset.

7.7 To help quantify the possible impact of the first factor, we used the MLM to impute additional EDSS scores for any patients in the BCMS dataset with a gap of more than 2 years between successive values (this used an identical BCMS dataset as used for the continuous Markov model). This would be expected to increase the weighting for patients with relatively slowly progressing disease and thus to compensate for the expected bias. We then re-estimated the transition probabilities for the Markov model using the same method as for the primary analysis. The result of this calculation was to increase very slightly the mean rate of disease progression predicted for the untreated RSS cohort and thus to increase the estimated treatment effect (deviation score of -10.6% compared to -9.9% for the primary analysis).

7.8 Our tentative conclusion is that, if the differential patterns of attendance at clinics in the BCMS cohort compared to the RSS cohort are responsible for any bias in our estimates of the treatment effect, it is at most very small.

References to online appendices

1 Chilcott J, McCabe C, Tappenden P, Cooper NJ, Abrams K, Claxton K. Modelling the cost-effectiveness of interferon beta and glatiramer acetate in the management of multiple sclerosis. *BMJ* 2003; **326**: 522-6.

2 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983; **33**: 1444-52.

3 MS Trust, personal communication.

4 Orme M, Kerrigan J, Tyas D, Russell N, Nixon R. The effect of disease, functional status, and relapses on the utility of people with multiple sclerosis in the UK. *Value Health* 2007; **10**: 54-60.

5 Department of Health. Cost effective provision of disease modifying therapies for people with multiple sclerosis. London: Health Service Circular (2002/004) London: Stationery Office; 2002.

6 Boggild M, Palace J, Barton P, et al. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. *BMJ* 2009; **339**: b4677.

7 Tremlett H, Paty D, Devonshire V. Disability progression in multiple sclerosis is slower than previously reported. *Neurology* 2006; **66**: 172–7.

8 Palace J, Bregenzer T, Tremlett H, et al. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. *BMJ Open* 2014; **4**: e004073.

9 Boggild M, personal communication.

10 Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. University of York Centre for Health Economics, Discussion Paper 138 1995.

11 IMS Health. Utilities in Multiple Sclerosis patients (update July 2013): report for the MS Trust. London: IMS Health; 2013.

12 Jackson CH, Sharples LS, Thompson SG, et al. Multistate Markov models for disease progression with classification error. *J R Stat Soc* 2003; **52**: 193–209.

13 Scientific Advisory Group to the UK MS Risk Sharing Scheme. Analysis of the year 4 and year 6 data. London: Department of Health; 2015.

14 Goldstein H. Multilevel Statistical Models (second edition). London: Edward Arnold; 1995.

15 Stata Corporation. College Station, Texas; 2007.

16 Leckie, G. and Charlton, C. runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. *Journal of Statistical Software* 2013: **52:** 1-40.

17 Tilling K, Sterne JA, Wolfe CD. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Stat Med*. 2001; **20(5)**: 685-704.

TABLE W1: LIST OF SUPPLEMENTARY ANALYSES UNDERTAKEN AND RATIONALE

Pre-specified?	Problem to be addressed	Analysis undertaken	Results at
Sensitivity analyses			
Yes	Application of "intention to treat" principles	Include scores recorded after a switch to a non-scheme DMT. In the counterfactual, predict disease progression with treatment as if patients stayed on treatment throughout period of follow-up.	Table 3
Yes	Application of "intention to treat" principles and assessment of possible bias due to incomplete follow-up	As for previous, with imputation of missing values.	Web table 3
No	Possible bias due to exclusion of patients not starting treatment within specified 3-month window	Include such patients in analysis set	Web table 3
No	Possible bias due to censoring of EDSS scores recorded after a patient has switched to a non-scheme DMT	Include such scores, including those for patients who switch to a non-scheme DMT before the first annual review.	Web table 3
Yes	Possible bias due to incomplete follow-up in RSS cohort	Impute missing values in the Markov framework on two extreme assumptions: (a) last value carried forward, (b) linear interpolation/extrapolation using all available valid EDSS scores	Table 3
Yes	Possible bias due to incomplete follow-up in RSS cohort	Impute missing values in the MLM framework using the model itself. As a sensitivity analysis, add 0.5 EDSS points to each imputed value	Table 3
Yes	Possible bias due to exclusion of patients with SPMS at baseline from RSS analysis cohort	Include such patients in analysis set	Table 3
Yes	Test sensitivity of results to precise mapping from EDSS to utilities	Use alternative set of utilities	Web table 3
Yes	Main outcome measure only uses information from patient's last recorded EDSS - this applies an "area under the curve" approach	Integrates actual and expected benefit for each year from baseline	Web table 3
No	Assess most extreme combination of variant assumptions - Markov model	Include patients with SPMS at baseline in analysis set and impute missing values using linear interpolation/ extrapolation	Web table 3
No	Assess most extreme combination of variant assumptions - MLM	Include patients with SPMS at baseline in analysis set and impute missing values using MLM	Web table 3
No	Assess the bias due to using observed EDSS at baseline to assess progression, rather than estimated "true" EDSS at baseline	Use estimated "true" EDSS at baseline (conditional on first observed EDSS) to calculate progression	Appendix 5
Other supplementary analyses			
No	Assess possible implications of different patterns of availability of data in BCMS dataset compared to RSS dataset	Compare baseline parameters and average rates of EDSS progression for patients (a) contributing relatively frequent/infrequent data, (b) not lost/lost to follow up (defined as contributing/not contributing any data within last 18 months before cut-off)	Web table 4
No	Compare rates of disease progression in the London Ontario and British Columbia datasets	Comparison of projections using a Markov model with transition matrices estimated from the two datasets and starting with the EDSS distribution in the RSS baseline population	Web figure 1

TABLE W2: NUMBER OF YEARS OF FOLLOW-UP DATA IN THE PER PROTOCOL ANALYSIS COHORT

Patient group	Number	On scheme DMT at final year?		Mean number of years on scheme DMT	Data available for year 7 and/or year 8
Total cohort	4137	3561	86.1%	4.526	2585
Number with:					
Year 6 data	2639	2263	85.8%	5.324	2003
Year 5 but no year 6 data	602	508	84.4%	4.322	338
Year 4 but no year 5 or 6 data	292	239	81.8%	3.401	94
Year 3 but no year 4-6 data	229	197	86.0%	2.576	64
Year 2 but no year 3-6 data	186	168	90.3%	1.742	42
Year 1 but no year 2-6 data	189	186	98.4%	0.868	44
Mean years follow-up	5.139				

TABLE W3: SENSITIVITY ANALYSES

Variant	Number of patients	Mean years of follow up	Absolute treatment effect (EDSS)				Deviation scores (utility)			
			Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.	Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.
Primary analysis	4,137	5.139	0.284	0.587	0.535	0.639	-10%	-14%	-25%	-2%
Supplementary/sensitivity analyses										
Intention to treat (ITT) analysis	4,137	5.409	0.273	0.590	0.537	0.642	(a)	(a)	(a)	(a)
ITT analysis with MLM imputation[(b)]:										
Single mean imputation	4,137	5.983	n/a	0.639	0.588	0.690	n/a	(a)	(a)	(a)
Multiple imputation	4,137	5.983	n/a	0.639	0.586	0.692	n/a	(a)	(a)	(a)
Single mean imputation with additional 0.5 EDSS points for each imputed value	4,137	5.983	n/a	0.499	0.447	0.551	n/a	(a)	(a)	(a)
Including patients starting treatment more than 91 days after baseline	4,328	5.130	0.274	(c)	(c)	(c)	-7%	(c)	(c)	(c)
Including EDSS scores recorded after a patient has switched to a non-scheme DMT (including patients switching before year 1)	4,197	5.400	0.260	(c)	(c)	(c)	-1%	(c)	(c)	(c)
Imputation - Markov model										
- last value carried forward	4,209	5.992	0.378	n/a	n/a	n/a	-32%	n/a	n/a	n/a
- linear interpolation/extrapolation	4,209	5.961	0.226	n/a	n/a	n/a	34%	n/a	n/a	n/a
Imputation - multilevel model										
- "on treatment" assumption	4,137	5.984	n/a	0.664	0.614	0.715	n/a	-6%	-16%	4%
- "off treatment" assumption:										
single imputation	4,137	5.984	n/a	0.644	0.593	0.694	n/a	-7%	-17%	4%
multiple imputation		5.984		0.643	0.590	0.697	n/a	-2%	-13%	10%
single imputation + 0.5 EDSS points for each imputed value	4,137	5.984	n/a	0.464	0.411	0.516	n/a	38%	26%	49%
Including patients with SPMS at baseline	4,780	5.082	0.173	0.521	0.473	0.568	3%	11%	2%	20%
Alternative utility estimates (pooled combination of 3 datasets including the Boggild data)	4,137	5.139	(d)	(d)	(d)	(d)	-14%	-7%	-18%	5%
Cumulative disease burden	4,137	5.139	1.332	(e)	(e)	(e)	-30%	(e)	(e)	(e)

Variant	Number of patients	Mean years of follow up	---------- Absolute treatment effect (EDSS) ----------				---------- Deviation scores (utility) ----------			
			Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.	Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.
Including patients with SPMS at baseline with imputation of missing values (Markov, linear extrapolation)	4,859	5.954	0.095	n/a	n/a	n/a	57%	n/a	n/a	n/a
Including patients with SPMS at baseline with imputation of missing values (MLM model, "off treatment" assumption")	4,780	5.987	n/a	0.571	0.525	0.617	n/a	14%	5%	24%

Notes:

(a) Since the ITT analysis is essentially a comparison of actual disease progression with predicted progression off treatment (rather than predicted progression on treatment), the deviation measure cannot meaningfully be calculated for these variants.

(b) In this analysis, it is assumed that patients go "off treatment" once they are lost to follow-up.

(c) These variants were only carried out for the Markov model

(d) As for primary analysis

(e) Results not available on a fully comparable basis

TABLE W4: CHARACTERISTICS OF PATIENTS IN THE BCMS DATASET ACCORDING TO (A) THE FREQUENCY OF EDSS SCORES, (B) WHETHER LOST TO FOLLOW UP

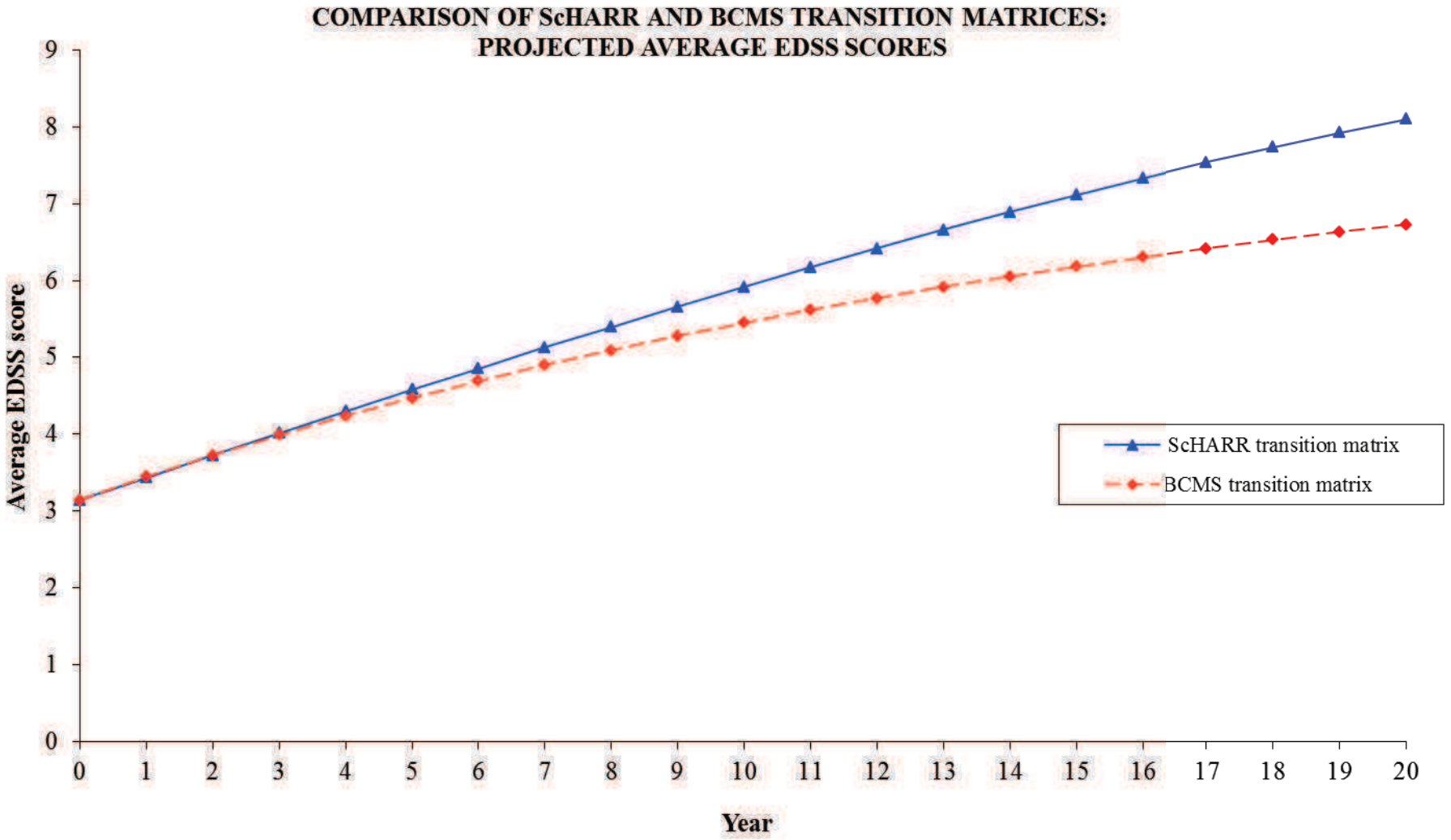
	Frequency of scores:		Whether "lost to follow up":	
	high frequency	low frequency	not lost	lost
Number	449	449	649	249
(%)	(50.0%)	(50.0%)	(72.3%)	(27.7%)
% female	71.7%	76.6%	75.5%	70.7%
Age at baseline (eligibility)	37.32	37.03	36.56	38.78
Age at onset	28.87	29.62	28.41	31.43
MSSS score at baseline	4.21	3.61	3.74	4.35
EDSS at baseline	2.67	2.20	2.35	2.66
Number with year 5 data	229	144	289	84
Mean EDSS increase (year 5 on baseline)	1.88	1.09	1.45	2.00

Notes:
Frequency = number of EDSS scores divided by interval between first and last scores
Lost to follow-up = no EDSS score after or within 18 months of cut-off date (31.12.1995)

TABLE W5: MAIN OUTCOMES FOR YEAR 4 COMPARED WITH YEAR 6

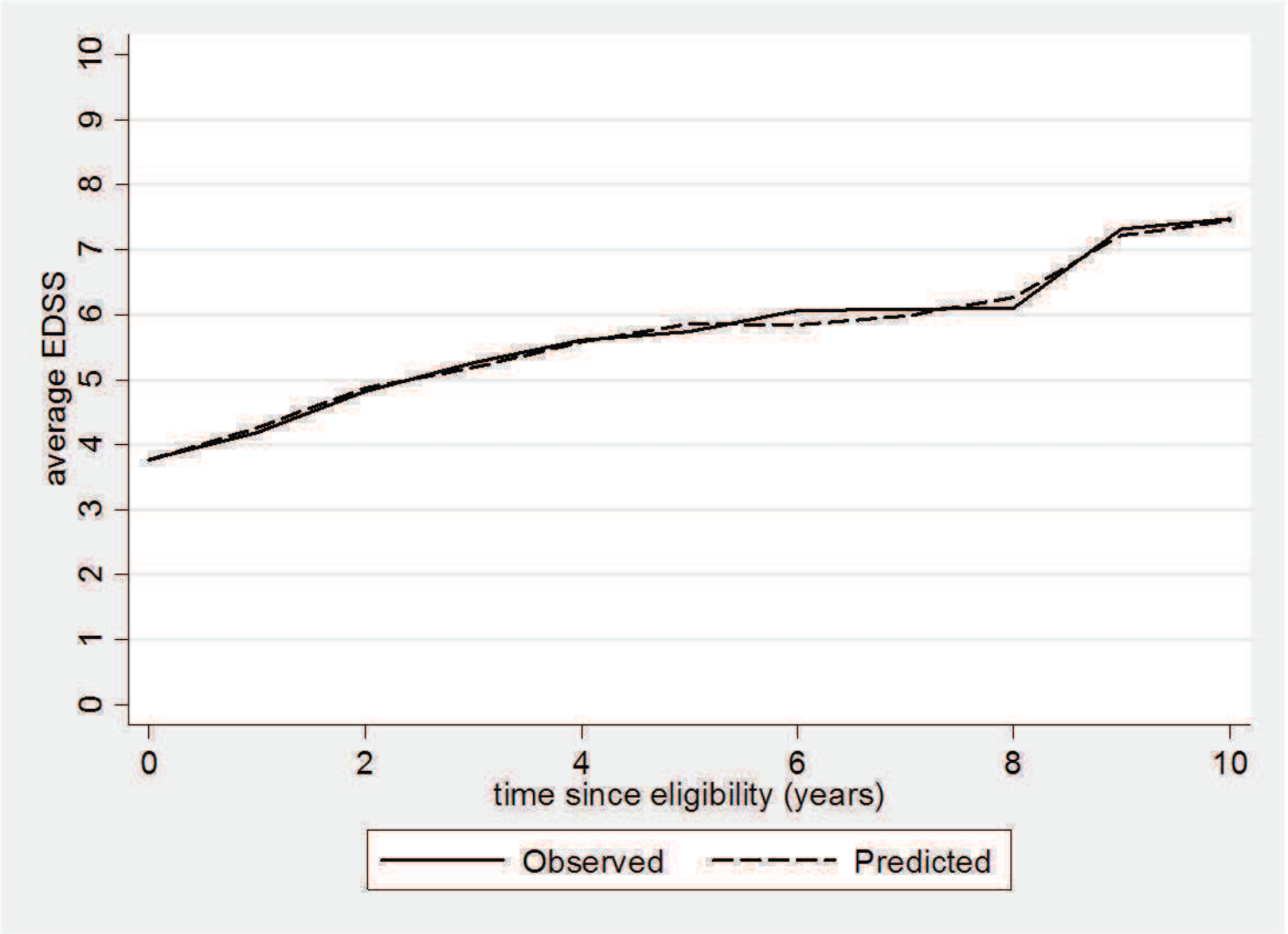
	Number of patients	Mean years of follow up	Absolute treatment effect (EDSS)				Deviation scores (utility)			
			Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.	Markov model	Multilevel model Mean	Multilevel model Lower C.I.	Multilevel model Upper C.I.
Year 4	4,110	3.532	0.268	0.514	0.466	0.561	-16%	-51%	-66%	-37%
Year 6	4,137	5.139	0.284	0.587	0.535	0.639	-10%	-14%	-25%	-2%

Web figure 1


COMPARISON OF ScHARR AND BCMS TRANSITION MATRICES:
PROJECTED AVERAGE EDSS SCORES
Average EDSS score
Year
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
ScHARR transition matrix
BCMS transition matrix

Web figure 2

EXTERNAL VALIDATION OF THE MULTILEVEL MODEL VERSUS DATA FROM THE UNIVERSITY OF WALES MULTIPLE SCLEROSIS DATASET



OPEN ACCESS

RESEARCH PAPER

Assessing the long-term effectiveness of interferon-beta and glatiramer acetate in multiple sclerosis: final 10-year results from the UK multiple sclerosis risk-sharing scheme

Jacqueline Palace,[1] Martin Duddy,[2] Michael Lawton,[3] Thomas Bregenzer,[4] Feng Zhu,[5] Mike Boggild,[6] Benjamin Piske,[4] Neil P Robertson,[7] Joel Oger,[5] Helen Tremlett,[5] Kate Tilling,[3] Yoav Ben-Shlomo,[3] Richard Lilford,[8] Charles Dobson[9]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jnnp-2018-318360).

For numbered affiliations see end of article.

Correspondence to
Dr Jacqueline Palace, Clinical Neurology, The Oxford University Hospitals Trust, Oxford OX3 9DU, UK; jacqueline.palace@ndcn.ox.ac.uk

YB-S and RL contributed equally.

Received 6 March 2018
Revised 14 June 2018
Accepted 7 July 2018

To cite: Palace J, Duddy M, Lawton M, *et al*. *J Neurol Neurosurg Psychiatry* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jnnp-2018-318360

ABSTRACT

Background Because multiple sclerosis (MS) is a chronic disease causing disability over decades, it is crucial to know if the short-term effects of disease-modifying therapies reported in randomised controlled trials reduce long-term disability. This 10-year prospective observational study of disability outcomes (Expanded Disability Status Scale (EDSS) and utility) was set up, in conjunction with a risk-sharing agreement between payers and producers, to investigate this issue.

Methods The outcomes of the UK treated patients were compared with a modelled untreated control based on the British Columbia MS data set to assess the long-term effectiveness of these treatments. Two complementary analysis models were used: a multilevel model (MLM) and a continuous Markov model.

Results 4862 patients with MS were eligible for the primary analysis (mean and median follow-up times 8.7 and 10 years). EDSS worsening was reduced by 28% (MLM), 7% (Markov) and 24% time-adjusted Markov in the total cohort, and by 31% (MLM) and 14% (Markov) for relapsing remitting patients. The utility worsening was reduced by 23%–24% in the total cohort and by 24%–31% in the RR patients depending on the model used. All sensitivity analyses showed a treatment effect. There was a 4-year (CI 2.7 to 5.3) delay to EDSS 6.0. An apparent waning of treatment effect with time was seen. Subgroup analyses suggested better treatment effects in those treated earlier and with lower EDSS scores.

Conclusions This study supports a beneficial effect on long-term disability with first-line MS disease-modifying treatments, which is clinically meaningful. However the waning effect noted requires further study.

INTRODUCTION

Multiple sclerosis (MS) is a major cause of serious physical disability in adults of working age. The majority of patients start with a relapsing remitting phase (RRMS), which then becomes secondary progressive (SPMS) with or without superimposed relapses at a median disease duration of 20 years.[1] It is during this latter phase that the majority of the disability is manifest.

The first randomised controlled trials (RCT) of MS disease-modifying treatments (DMTs) were reported in the 1990s.[2–5] These led to the licensing of three formulations of interferon beta (Betaferon/Betaseron, Rebif and Avonex) and of glatiramer acetate (Copaxone) for RRMS. The licences of Betaferon and Rebif were subsequently extended in Europe to include patients with relapsing SPMS.[6–8]

All four treatments were shown, over the period of 2–3 years of the trials, to be efficacious in reducing relapses (by approximately 30%) and in reducing MRI activity, with less robust evidence on disability. The most important outcome, that is, the long-term effect of treatment on disability, could not be addressed within this timescale, and the predictive value of short-term treatment outcomes on longer term disability remains unproven. Only longer term follow-up data can provide information on the sustainability of treatment effects and on delaying time to loss of independent ambulation.

Because of the uncertainty of the long-term benefit, the UK's National Institute for Health and Care Excellence (NICE) in 2002 concluded that it was unable to recommend the use of these DMTs within the UK National Health Service,[9] but recognised that these drugs could be cost-effective over the longer term if early treatment effects on disability persisted. Consequently a novel risk-sharing scheme (RSS) was set up between the UK Department of Health, the pharmaceutical companies, professional and patient groups to deliver these DMTs cost-effectively[10] (see online supplementary appendix 1 for details).

A key feature of the scheme was the monitoring of disability progression in a cohort of patients to test whether observed outcomes were in line with those required for cost-effectiveness. This observational study recruited over 5000 UK patients prescribed the interferon-βs and glatiramer acetate between 2002 and 2005 and followed them up over a 10-year period in order to measure the long-term effectiveness of the drugs when compared with a modelled natural history cohort. Two yearly analyses were performed and if observed outcomes deviated from target by more than an agreed margin, the price of the drugs would be adjusted to restore cost-effectiveness.

At 2 years the initial Ontario natural history data set was deemed not suitable to model the untreated

control group because the disability scores had been smoothed,[11] and the British Columbia MS database was selected and validated.[12 13] A second independent analysis model (a multilevel model (MLM)) was added[14] to corroborate results from the Markov model specified in the RSS.[15]

After 6 years of follow-up,[13] the drugs in aggregate showed a 40% reduction in the rate of deterioration in disability on the MLM in patients with relapsing remitting disease (24% on the Markov model); this translated into a reduction of 43% (multilevel) or 42% (Markov) in the rate at which utility worsened, which was on track for the cost-effective target.[10]

The scientific analysis plan for this 10-year analysis[16] was further revised in line with intention-to-treat principles, with additional analyses of (1) subgroups, (2) the influence of time on treatment effectiveness and (3) the effect of treatment on the time to loss of unaided ambulation. Here we report the final results of this study, focusing on the longer term effect of DMTs on disability progression.

METHODS

The detailed protocol can be found elsewhere[12–14] and the following is a summary.

Patient recruitment and follow up

Between May 2002 and July 2005, 72 sites across the UK recruited patients with RRMS or SPMS fulfilling the Association of British Neurologists (ABN) 2001 criteria for treatment (Expanded Disability Status Scale (EDSS) ≤6.5; ≥18 years old; two relapses in the last two calendar years).[17] Drug selection reflected clinical practice and was led by individual patient and physician choice within the licensed indications in the UK. Ten-year follow-up was planned with the annual EDSS[18] scores, whether treatment was continued or not. Telephone assessments were permitted for EDSS values over 6.[19]

For this final analysis the main outcome was an 'intention to treat' analysis which included all recruited patients, both RRMS and SPMS at baseline, and all follow-up scores including those from patients switching to non-scheme DMTs. However, subgroup analysis also focused on the RRMS at baseline patients, since this was the cohort for which the price adjustment scheme was set up.

Natural history comparator

The British Columbia MS data set is a population-based database, established in 1980, with the EDSS scored by the MS neurologist at outpatient visits.[1] Patients (n=978) were identified who, between 1980 and 1995, fulfilled similar eligibility criteria to the RSS cohort. The 'baseline' for these patients was taken as the first clinic visit at which they fulfilled the criteria. Patient information was included only up to the end of 1995, after which DMTs were widely available in Canada.

Outcomes

For scientific purposes, our primary interest was to compare the rate of disability worsening (ie, disease progression) of patients treated with DMTs, as observed in the RSS, with that in an untreated modelled comparator control group ('comparison against control'). Disease progression was modelled in the patients in the British Columbia data set adjusting for differences in prognostic factors (EDSS at baseline, age at onset) between the two data sets. The technical details are explained below. Disease progression was measured in terms of the accumulation of disability (as measured by EDSS) and worsening of quality of life (expressed as utility, derived from the EDSS score—see online supplementary appendix 2), for eligible and treated patients with at least one EDSS score after baseline. Outcomes were expressed both as absolute differences between progression observed in the RSS treated group and the comparator control group, and the relative difference in progression (expressed as a percentage).

The secondary outcomes included the area under the curve (a cumulative measure of reduction in disability using the absolute value of area under EDSS–time curve calculated using the trapezoid rule) and the delay in the median time required for sustained progression to the clinically relevant milestone of EDSS 6.0 (needing a stick to walk 100 m), confirmed at least 6 months with no subsequent lower scores.[20]

Supplementary analyses

A wide range of prespecified sensitivity analyses (online supplementary table 1 and 2) were conducted to assess assumptions related to modelling and loss to follow-up. Prespecified subgroup analyses allowed us to examine treatment effect by sex, EDSS and disease course (RRMS/SPMS) at baseline, disease duration and recruitment date. (Patients recruited later in the scheme may be more typical of normal clinical practice than those recruited earlier, many of whom had been waiting for treatment and had more advanced disease.) We also used a variety of methods to explore the apparent changes in the treatment effect over time (online supplementary appendix 3).

Statistical design

In order to calculate the expected progression for untreated patients, two independent models were developed, a Markov model and an MLM—for further details see online supplementary appendix 4. Both models used data from the natural history comparator data set and these were validated in previous studies[12 14] as being reliable methods for predicting the untreated outcome. The essential role of the modelling is to adjust for the (relatively slight) differences in baseline characteristics between the natural history and RSS data sets.

Markov model

The Markov model uses the annual probabilities of moving between EDSS scores (or remaining on the same score) from the untreated values derived from the British Columbia data set. These probabilities are then applied to the baseline EDSS scores of the RSS cohort and to subsequent modelled EDSS states over 10 years to give the expected EDSS progression for untreated patients. The difference between the expected untreated and observed RSS treated mean EDSS score represents the treatment effect.

Estimation was by the Jackson continuous time method[21] using age at onset dichotomised at the median as a predetermined covariate.[12] Because the assumption in the Markov model is that the transition probabilities from one specific EDSS score do not vary over time (and this assumption is unlikely to be precise), we used a 'time-varying' model, with separate transition matrices estimated for the first 2 years after baseline and for the rest of the follow-up period, as a sensitivity analysis. CIs on projections using the Markov model are derived by bootstrapping.

Multilevel model

The MLM[22 23] uses the mean trajectory for the whole population, the variation of individual trajectories about this mean and the fluctuation of individual EDSS scores about the trend for each individual. The untreated values were again derived

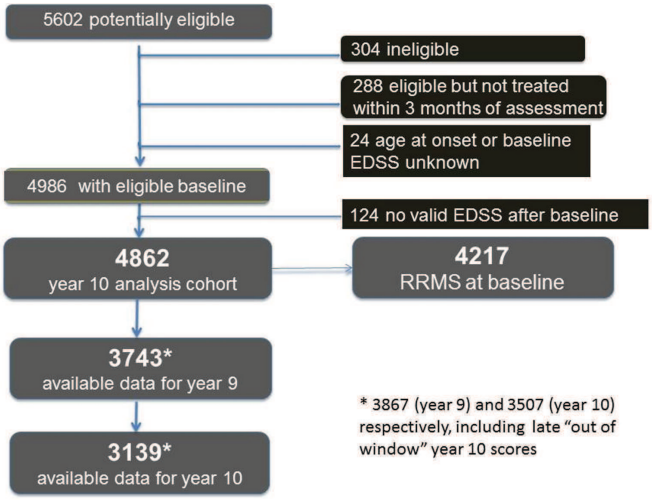


Figure 1 Consolidated Standards of Reporting Trials diagram for all patients in the UK MS risk-sharing scheme. EDSS, Expanded Disability Status Scale; MS, multiple sclerosis; RRMS, relapsing remitting MS.

from the British Columbia data set. The model is applied to the baseline data for patients in the RSS data set to predict the EDSS progression without treatment for each individual. The projections are then combined across all individuals to produce a predicted untreated mean EDSS progression for the whole RSS cohort. This is then compared with the observed data in the RSS (the 'comparison against control') to estimate the absolute treatment effect.

For consistency with the Markov model,[12] age at onset (binary variable) was included as the only covariate after previous covariate modelling found this to be sufficient. Validation of the model[14 24] was undertaken on the British Columbia cohort (internal validation) and also using a natural history data set from the University Hospital of Wales (external validation). Bootstrapping was used to derive 95% CIs.

Time to sustained EDSS 6.0

To estimate the time to sustained EDSS 6.0, a parametric (Weibull) model was fitted to the British Columbia Multiple Sclerosis (BCMS) data, with gender and baseline EDSS as covariates. A similar model was fitted to the RSS data. To adjust for differences in baseline distributions, the strata were combined using as weights the proportion of patients at baseline in the RSS data set (see online supplementary appendix 5 for details).

Governance of the study

An independent Scientific Advisory Group consisting of Pelham Barton, Yoav Ben-Shlomo, Richard Gray and chaired by Richard Lilford developed and approved the statistical analysis plan. Representation from the authors, NICE, the MS Trust and the Department of Health attended the Scientific Advisory Group meetings. The representatives from each company were observers at the Scientific Advisory Group meetings but had no role in the data collection, analysis or preparation of the manuscript.

RESULTS

Patient disposition and characteristics

Out of 5602 patients registered for recruitment, 4986 were consented and eligible for treatment, began treatment, and had baseline assessments. Of these 4986, 4862 (97.5%) had at least one follow-up EDSS score and were included in the primary analysis. Of these, 17% switched to non-scheme drugs during follow-up, with the majority using natalizumab or fingolimod, and 18% discontinued treatment during the study. The Consolidated Standards of Reporting Trials diagram is shown in figure 1 (details in online supplementary table 2). The median follow-up was 10 years, the mean follow-up was 8.7 years and 77% of the primary analysis group had a minimum of 9 years of follow-up data. An additional 2.6% had late year 10 data collected, used in the sensitivity analysis.

The baseline characteristics for the analysis cohort were similar to those in the British Columbia data set (online supplementary table 3), although RSS patients had on average slightly higher EDSS and age at onset and thus slightly longer disease duration on first assessment; however, the models adjust for baseline EDSS scores and age of onset.

Primary outcomes

The primary outcomes are shown in table 1. The absolute 10-year treatment effect was a reduction of 0.61 using the MLM (95% CIs 0.55 to 0.66) and of 0.12 in mean EDSS using the Markov model (95% CIs 0.07 to 0.17). The corresponding relative reduction in progression was 28% (95% CIs 26% to 31%) for the MLM and 7% (95% CIs 4% to 10%) for the Markov model. However, using the time-varying Markov model, the effect size closely mirrors the MLM results (see below). In utility terms, the reduction in progression was 23% (95% CIs 20% to 26%) on the MLM and 24% (95% CIs 21% to 27%) on the Markov model. Figure 2A,B shows how the observed disability progression deviates over the 10-year period from that expected for untreated patients (the comparator control group) for the MLM.

Secondary outcomes

The cumulative benefit accrued over the first 10 years of treatment for the whole cohort (the area under the curve for the full period) was 1.3 EDSS years (95% CIs 0.9 to 1.6—estimates available only for the Markov model). The estimated median time for the whole cohort to reach sustained EDSS 6.0 using the Weibull model was 12.5 years (95% CIs 11.8 to 13.3); this compares with 8.4 years (95% CIs 7.4 to 9.6) for untreated patients (figure 3), indicating that treatment is associated with a delay of 4.0 years (95% CIs 2.7 to 5.3) in reaching this relevant disability endpoint.

Sensitivity analyses

The results of the sensitivity analyses are shown in table 2 and online supplementary table 4. As expected, excluding EDSS scores after patients switched to alternative treatments—effectively censoring later data from patients on a worse trajectory—makes the results appear more favourable to the DMTs. Imputing missing values in the RSS data set makes relatively little difference to the outcomes, except in the deliberately extreme case (greater than a standard worse case scenario) in which we added 0.5 EDSS points to each imputed value (on average, doubling the imputed progression since the last available score), where the magnitude of the treatment effect on EDSS was reduced by roughly a quarter. In contrast, a variant in which we 'enriched' the British Columbia data set by imputing additional values for patients with relatively sparse follow-up and re-estimated the transition probabilities resulted in a substantially greater treatment effect. Finally, using variant of the Markov model with time-varying transition probabilities resulted in a greater treatment effect, with a much closer agreement for the EDSS results

Table 1 Outcomes of the primary analysis: primary analysis cohort, including patients with SPMS at baseline (n=4862; average follow-up 8.7 years)

Outcome measure	Model	Actual progression (95% CI)	Predicted progression (natural history) (95% CI)	Absolute treatment effect (predicted natural history progression less actual) (95% CI)	Relative rate of disease progression (actual divided by predicted natural history) (95% CI)
		(1)	(2)	(3)=(2–1)	(4)=(1)/(2)
Rounded EDSS*	Markov	1.53 (1.48 to 1.58)	1.65 (1.63 to 1.67)	0.12 (0.07 to 0.17)	93% (90% to 96%)
EDSS*	MLM	1.53 (1.47 to 1.58)	2.13 (2.11 to 2.15)	0.61 (0.55 to 0.66)	72% (69% to 74%)
Utility	Markov	0.122 (0.117 to 0.127)	0.161 (0.159 to 0.163)	0.039 (0.034 to 0.044)	76% (73% to 79%)
	MLM	0.122 (0.117 to 0.127)	0.159 (0.156 to 0.162)	0.037 (0.031 to 0.042)	77% (74% to 80%)

*For the Markov model, half-integer EDSS states are combined with the next lower integer EDSS state. Calculations with the MLM using rounded EDSS values gave almost identical results to those using the full EDSS scale.
EDSS, Expanded Disability Status Scale; MLM, multilevel model; SPMS, secondaryprogressive multiple sclerosis.

to those from the MLM. Online supplementary appendices 6 and 7 provide further details.

Subgroup analyses

Estimated treatment effects were rather larger for the subset of patients with RRMS at baseline (table 3, figure 2C,D; MLM), with 31% and 14% reduction in EDSS worsening and 24% and 31% reduction in utility worsening (MLM and Markov models, respectively).

Other prespecified subgroup analyses using the Markov model (tables 4 and 5, and online supplementary table 5) suggest that greater treatment effects are associated with a lower EDSS score at baseline, patients with shorter time from disease onset, women rather than men, patients on treatment throughout rather than those who come off treatment, and patients recruited later in the scheme. The MLM however did not find evidence of a reduction in treatment effect with increasing baseline EDSS.

Changes over time

Table 6 shows how the main outcomes using the MLM model change when the primary analysis is repeated with years 2, 4, 6, 8 and 10 as the end year. The results suggest that, although the treatment effect continues throughout the follow-up period, it becomes progressively smaller. We explored this through a number of prespecified supplementary analyses using both models, including the version of the Markov model with time-variant transition probabilities. The results (online supplementary appendix 8) confirm this trend of a large initial effect—in the first year or two after treatment initiation—followed by a smaller continuing effect. A sensitivity analysis using year 1 as

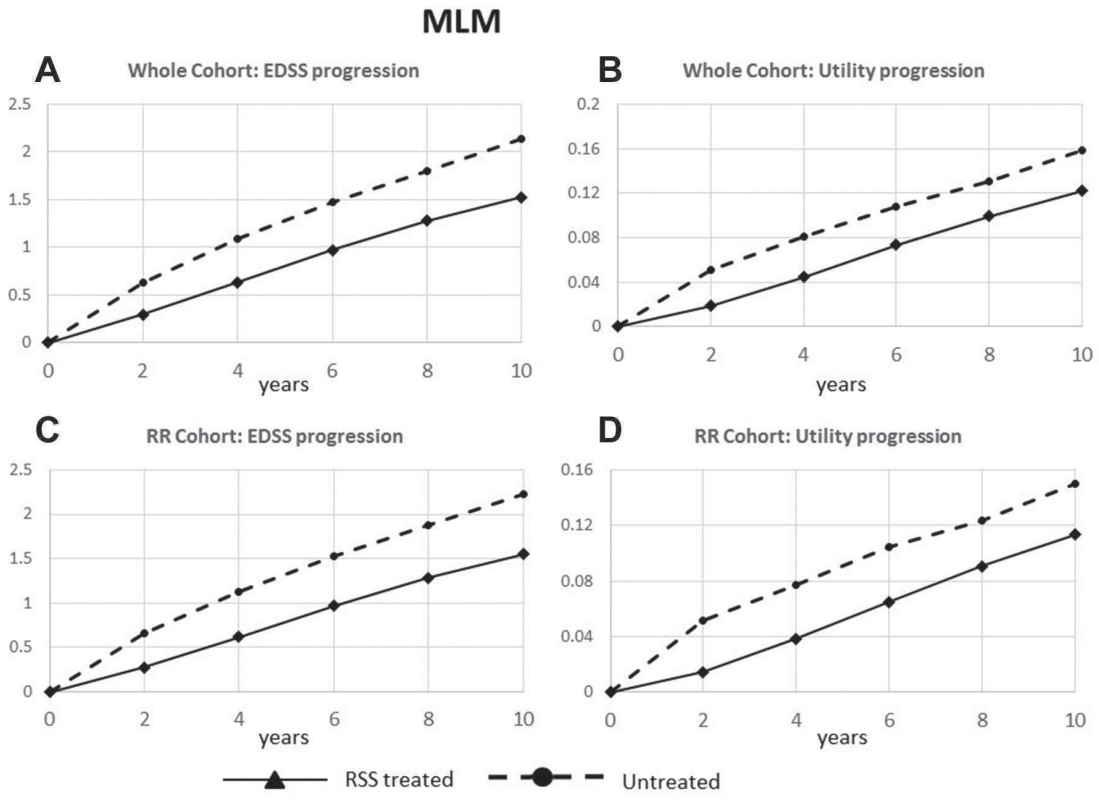


Figure 2 Comparison against control results using the MLM model: disability outcomes over 10 years in RSS patients and in the untreated comparator control group, for the whole cohort (A) EDSS progression and (B) utility progression, and for the patients with RRMS at baseline (C) EDSS progression and (D) utility progression. EDSS, Expanded Disability Status Scale; MLM, multilevel model; RRMS, relapsing remitting multiple sclerosis; RSS, risk-sharing scheme.

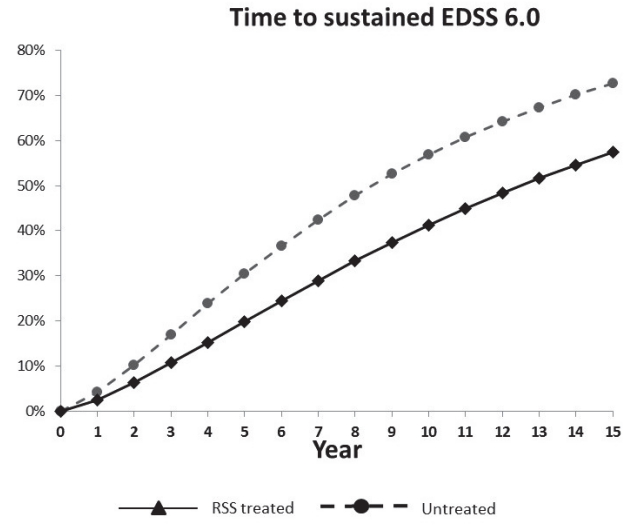


Figure 3 Time to sustained EDSS 6.0 in RSS treated patients and in the untreated comparator control group. EDSS, Expanded Disability Status Scale; RSS, risk-sharing scheme.

the baseline (table 2 and online supplementary table 4) also showed a positive but smaller treatment effect.

DISCUSSION

Main findings

Our findings show a clinically significant treatment effect maintained at 10 years, on both EDSS and quality of life outcomes, with a meaningful delay to sustained loss of unaided ambulation of around 4 years in an already impaired cohort at treatment onset (median EDSS 3.5).

Our analysis also highlighted a gradual attenuation of the treatment effect over time, or perhaps a large initial effect followed by a more modest continuing effect. The treatment effects appeared larger for patients with RRMS, those who start treatment earlier in their disease course and for women. The evidence for the effect of baseline EDSS on the treatment effect is less clear.

Methodological issues

This unique study addresses an important issue related to the long-term impact of the first-generation DMTs, which was not possible to address by a conventional RCT. By establishing a

Table 2 Sensitivity analyses (EDSS basis)

		Mean years		Absolute treatment effect (EDSS)		Relative treatment effect (EDSS)	
Variant	Patients (n)	Follow-up	On treatment	Markov model	MLM	Markov model	MLM
Primary analysis	4862	8.7	7.0	0.12 (0.07, 0.17)	0.61 (0.55, 0.66)	93% (90%, 96%)	72% (69%, 74%)
RRMS-only subgroup	4217	8.9	7.3	0.25 (0.20, 0.31)	0.68 (0.62, 0.74)	86% (83%, 89%)	69% (67%, 72%)
Supplementary/sensitivity analyses (including patients with SPMS at baseline except where noted)							
Excluding EDSS scores recorded after a patient has switched to a non-scheme DMT*							
Primary analysis cohort	4799	8.0	6.4	0.13 (0.08, 0.19)	0.59 (0.53, 0.64)	91% (88%, 95%)	70% (68%, 73%)
RRMS-only subgroup†	4157	8.1	6.7	0.26 (0.21, 0.32)	0.66 (0.60, 0.72)	84% (81%, 88%)	68% (65%, 71%)
Excluding EDSS scores recorded after a patient has switched to any other DMT*							
Primary analysis cohort	4475	7.0	5.7	0.16 (0.11, 0.22)	0.57 (0.52, 0.62)	88% (84%, 92%)	68% (65%, 70%)
RRMS-only subgroup	3871	7.0	5.9	0.29 (0.23, 0.35)	0.64 (0.59, 0.70)	80% (76%, 84%)	65% (62%, 68%)
Imputation—Markov model, using imputed values derived from the MLM							
'on treatment' assumption	4862	9.9	7.7	0.12 (0.07, 0.17)	NA	93% (91%, 96%)	NA
'off treatment' assumption	4862	9.9	7.0	0.10 (0.04, 0.15)	NA	95% (92%, 98%)	NA
Imputation—multilevel model							
'on treatment' assumption	4862	9.9	9.9	NA	0.67 (0.62, 0.72)	NA	72% (70%, 74%)
'off treatment' assumption							
Single imputation	4862	9.9	9.9	NA	0.65 (0.59, 0.70)	NA	73% (70%, 75%)
Multiple imputation	4862	9.9	9.9	NA	0.64 (0.59, 0.70)	NA	73% (71%, 75%)
Single imputation +0.5 EDSS points for each imputed value	4862	9.9	9.9	NA	0.48 (0.43, 0.53)	NA	80% (78%, 82%)
Supplement RSS data with imputed values derived from out-of-window year 10 scores‡	4862	8.9	7.0	0.12 (0.07, 0.18)	‡	93% (90%, 96%)	‡
Use transition matrices from BCMS data set supplemented by imputing additional data for patients with sparse follow-up‡							
Including patients with SPMS at baseline	4862	8.7	7.0	0.28 (0.23, 0.34)	‡	84% (81%, 87%)	‡
Excluding patients with SPMS at baseline	4217	8.9	7.3	0.43 (0.37, 0.49)	‡	78% (75%, 81%)	‡
Alternative Markov model with time-varying natural history transition matrices	4862	8.7	7.0	0.48 (0.42, 0.53)	NA	76% (74%, 79%)	NA
Year 1 baseline (see online supplementary appendix 8)	4360	7.9	6.2	0.07 (0.02, 0.12)	0.43 (0.38, 0.48)	95% (92%, 99%)	77% (74%, 79%)

*Patients switching before year 1 are also excluded from the analysis.
†This was the primary analysis for the interim year 6 analysis.
‡These variants were only carried out for the Markov model.
DMT, disease-modifying treatment; EDSS, Expanded Disability Status Scale; MLM, multilevel model; NA, not applicable; RRMS, relapsing remitting multiplesclerosis; RSS, risk sharing scheme; SPMS, secondary progressive multiple sclerosis; BCMS, British Columbia Multiple Sclerosis.

Table 3 Main outcomes, patients with RRMS at baseline: subgroup of patients with RRMS at baseline (n=4217; average follow-up 8.9 years)

Outcome measure	Model	Actual progression (95% CI)	Predicted progression (natural history)(95% CI)	Absolute treatment effect (predicted natural history progression less actual) (95% CI)	Relative rate of disease progression (actual divided by predicted natural history)(95% CI)
		(1)	(2)	(3)=(2–1)	(4)=(1)/(2)
Rounded EDSS*	Markov	1.55 (1.49 to 1.61)	1.80 (1.78 to 1.83)	0.25 (0.19 to 0.31)	86% (83% to 89%)
EDSS*	MLM	1.55 (1.49 to 1.61)	2.23 (2.21 to 2.25)	0.68 (0.62 to 0.74)	69% (67% to 72%)
Utility	Markov	0.113 (0.108 to 0.119)	0.164 (0.163 to 0.166)	0.051 (0.046 to 0.056)	69% (66% to 72%)
	MLM	0.113 (0.108 to 0.118)	0.150 (0.148 to 0.152)	0.037 (0.031 to 0.042)	76% (72% to 79%)

*For the Markov model, half-integer EDSS states are combined with the next lower integer EDSS state. Calculations with the MLM using rounded EDSS values gave almost identical results to those using the full EDSS scale.
EDSS, Expanded Disability Status Scale; MLM, multilevel model; RRMS, relapsing remitting multiple sclerosis.

novel RSS between the manufacturers and payers, we were able to collect high-quality long-term data on disease progression on a treated cohort and use historical data on untreated patients to construct a comparison control group.

Given the inevitable limitations of observational data, our study has important strengths. The RSS cohort is one of the largest cohorts ever used in a study of this type, with over 42 000 patient years of data, with 9 years or more of data for nearly 80% of patients—a dropout rate over a decade comparable with that seen in recent 2-year treatment trials.[25 26] Moreover these results reflect the 'real world' performance of the DMTs and include the effects of patients discontinuing treatment, and may be more generalisable than estimates from highly selected short-term RCT populations. Our natural history comparator, from the British Columbia data set, is also the largest and most complete data set of untreated patients available for MS research.[1] We used two completely independent modelling techniques, with complementary strengths and weaknesses: the multilevel repeated measures model can more easily take account of variations in the rate of disease progression between individual disease trajectories, whereas the Markov model predicted more accurately the number of patients at high EDSS levels which largely determine changes in mean utility (see (online supplementary appendix 4). A third model, used for the survival analysis, confirmed the finding of a clinically significant treatment effect on time to loss of unaided ambulation.

We have implemented 'intention to treat' principles, adapting the methods of our 6-year study[13] to include all eligible and consented patients with MS prescribed these DMTs, including those with SPMS as well as RRMS at baseline, and data from patients who switched to different DMTs—including non-scheme DMTs—during follow-up. We believe therefore that we have mitigated an important source of potential bias. Naturally, the possibility of bias from incomplete follow-up remains, but we have tested this by extensive sensitivity analyses with imputation of missing values. In addition, the patterns of follow-up seem broadly similar between the RSS and the British Columbia cohorts,[13] so any residual biases are likely to offset one another (see online supplementary appendix 7). If anything, our results may be biased against a treatment effect because we used telephone questionnaires in the UK cohort to capture data from patients with high EDSS scores who were unable to attend clinic, whereas there was no equivalent for the BCMS cohort. Our models are designed to adjust for the (relatively small) difference in baseline characteristics between the British Columbia and RSS data sets, but if this adjustment is imperfect this would be expected to bias the results *against* the treatment effect since RSS patients had on average marginally worse prognostic factors than British Columbia patients.

There are several other potential biases that must be considered before assuming a causal explanation for the DMTs that may have resulted in differences in the natural history of the two cohorts unrelated to treatment.

Geographical differences

Could our UK patients have a milder form of MS than the natural history British Columbia patients? Although British Columbia is more ethnically diverse, MS still predominates in Europeans there.[27] More direct evidence demonstrated similar untreated trajectories by using our Canadian data set to reproduce accurately 10-year outcomes in a Welsh untreated MS cohort.[24]

Temporal differences

Could untreated patients with MS have a slower disease progression than they used to? This concept has been supported by the observation that the relapse rates in the placebo arms of RCTs have been reducing. This is likely to be due to other reasons however. Now DMTs are widely available, ethical considerations make it likely that patients with more active disease are prescribed MS drugs rather than risk being allocated to a placebo arm. Additionally the later trials introduced early rescue therapy options and stricter relapse definitions, which would also reduce relapse rates recorded.[28] Population-based cohorts are less likely to be biased by selection, and there is no evidence for a better prognosis in recent MS cohorts from British Columbia[29] nor Wales (Neil Robertson, personal communication).

Confounding by indication

Could treatment decisions and clinical outcomes be confounded by variables such as disease severity and comorbidity, as seen in contemporary treated and untreated comparisons? This is not relevant in this study because there was no treated versus untreated comparison from within a single cohort. The BCMS patients were selected using the same inclusion criteria applied to the UK patients. Additionally both analysis methods adjusted for baseline EDSS and age at entry, and adding other covariates made little difference to the outcomes.[12] Previous results from the British Columbia data[20 30] demonstrated that indication bias may be important when comparing with contemporary untreated patients,[20] because untreated patients differed, having a milder disease even though fulfilling the same baseline criteria for treatment. When the authors compared treated patients with historical controls as in our study, they showed a trend towards a treatment benefit on time to reach EDSS 6.0 (relative risk 0.77, CI 0.58 to 1.02), which is comparable with our results considering they had a much smaller treated cohort (n=868 vs 4986) with shorter follow-up time (mean 5.2 and median 5.1 years vs

Table 4 Subgroup analyses by type of MS, date of baseline assessment and baseline EDSS (both models): EDSS outcomes

		Mean years		Mean	Markov model		Multilevel model		Multilevel model with imputation	
Population	n	At risk	On treatment	Baseline EDSS	Absolute treatment effect	Relative rate of progression	Absolute treatment effect	Relative rate of progression	Absolute treatment effect	Relative rate of progression
Primary analysis population	4862	8.7	7.0	3.18	0.12 (0.07, 0.17)	93% (90%, 96%)	0.61 (0.55, 0.66)	72% (69%, 74%)	0.65 (0.59, 0.70)	73% (70%, 75%)
Baseline EDSS ≤3.5	2940	9.0	7.4	2.02	0.35 (0.28, 0.42)	84% (81%, 87%)	0.63 (0.56, 0.71)	74% (71%, 77%)	0.67 (0.59, 0.74)	75% (73%, 78%)
Baseline EDSS 4–5.5	1275	8.6	6.7	4.42	−0.11 (−0.21, −0.02)	110% (102%, 119%)	0.59 (0.49, 0.68)	68% (63%, 73%)	0.64 (0.54, 0.73)	69% (65%, 73%)
Baseline EDSS ≥6	647	8.0	5.5	6	−0.47 (−0.57, −0.36)	233% (202%, 265%)	0.52 (0.41, 0.62)	63% (56%, 70%)	0.59 (0.48, 0.69)	65% (59%, 71%)
RRMS at baseline	4217	8.9	7.3	2.86	0.25 (0.20, 0.31)	86% (83%, 89%)	0.68 (0.62, 0.74)	69% (67%, 72%)	0.71 (0.66, 0.78)	71% (69%, 73%)
SPMS at baseline	645	7.8	5.1	5.29	−0.76 (−0.86, −0.66)	218% (202%, 234%)	0.10 (0.00, 0.20)	93% (86%, 101%)	0.19 (0.09, 0.29)	90% (85%, 95%)
RRMS at baseline and:										
Baseline EDSS ≤3.5	2882	9.0	7.5	2.00	0.39 (0.32, 0.46)	82% (79%, 86%)	0.67 (0.59, 0.74)	73% (69%, 76%)	0.70 (0.63, 0.77)	74% (71%, 77%)
Baseline EDSS 4–5.5	1093	8.7	7.0	4.40	−0.01 (−0.11, 0.10)	101% (91%, 110%)	0.68 (0.58, 0.79)	63% (58%, 68%)	0.72 (0.62, 0.83)	65% (60%, 70%)
Baseline EDSS ≥6	242	8.4	6.2	6	−0.16 (−0.35, 0.04)	143% (90%, 196%)	0.84 (0.64, 1.03)	44% (32%, 56%)	0.91 (0.72, 1.10)	47% (37%, 58%)
SPMS at baseline and:										
Baseline EDSS ≤3.5	58	8.2	5.9	2.69	−1.48 (−1.87, −1.09)	181% (158%, 203%)	−1.10 (−1.51, −0.70)	151% (133%, 170%)	−1.05 (−1.46, −0.64)	142% (125%, 159%)
Baseline EDSS 4–5.5	182	7.8	5.1	4.55	−0.77 (−0.95, −0.59)	182% (161%, 202%)	−0.01 (−0.20, 0.19)	101% (89%, 112%)	0.13 (−0.06, 0.32)	94% (85%, 102%)
Baseline EDSS ≥6	405	7.8	5.1	6	−0.65 (−0.77, −0.53)	292% (255%, 329%)	0.32 (0.21, 0.44)	76% (68%, 84%)	0.39 (0.28, 0.50)	76% (69%, 83%)
Date of baseline assessment										
On or before 28 August 2003	2351	8.7	6.9	3.32	0.02 (−0.06, 0.10)	99% (94%, 103%)	0.53 (0.46, 0.61)	75% (71%, 78%)	0.58 (0.51, 0.66)	75% (72%, 79%)
After 28 August 2003	2511	8.8	7.0	3.05	0.21 (0.14, 0.29)	88% (83%, 92%)	0.67 (0.60, 0.75)	69% (66%, 72%)	0.71 (0.64, 0.78)	71% (67%, 74%)
Baseline at or before 31 August 2003 and:										
Baseline EDSS ≤3.5	1352	8.9	7.5	2.07	0.27 (0.16, 0.38)	87% (82%, 92%)	0.57 (0.46, 0.68)	77% (72%, 81%)	0.61 (0.50, 0.72)	77% (73%, 82%)
Baseline EDSS 4–5.5	631	8.6	6.6	4.42	−0.21 (−0.34, −0.08)	119% (107%, 131%)	0.49 (0.36, 0.62)	73% (66%, 80%)	0.54 (0.41, 0.66)	74% (68%, 80%)
Baseline EDSS ≥6	368	7.9	5.4	6	−0.50 (−0.64, −0.36)	244% (202%, 287%)	0.48 (0.34, 0.62)	65% (56%, 74%)	0.55 (0.41, 0.69)	67% (59%, 75%)
Baseline after 31 August 2003 and										
Baseline EDSS ≤3.5	1588	9.0	7.4	1.97	0.42 (0.32, 0.52)	81% (77%, 85%)	0.69 (0.59, 0.79)	72% (68%, 76%)	0.71 (0.61, 0.81)	73% (69%, 78%)
Baseline EDSS 4–5.5	644	8.5	6.8	4.43	−0.02 (−0.16, 0.11)	102% (90%, 114%)	0.68 (0.54, 0.82)	62% (55%, 69%)	0.74 (0.60, 0.87)	64% (58%, 71%)
Baseline EDSS ≥6	279	8.2	5.6	6	−0.43 (−0.58, −0.27)	219% (175%, 264%)	0.57 (0.41, 0.72)	61% (50%, 71%)	0.63 (0.48, 0.78)	63% (54%, 71%)

EDSS, Expanded Disability Status Scale; MS, multiple sclerosis; RRMS, relapsing remitting multiple sclerosis; SPMS, secondary progressive multiple sclerosis.

Table 5 Other subgroup analyses (Markov model only)

Population	n	Mean years		Mean	EDSS		Utility	
		At risk	On treatment	Baseline EDSS	Absolute treatment effect	Relative rate of progression	Absolute treatment effect	Relative rate of progression
Primary analysis population	4862	8.7	7.0	3.18	0.12 (0.07, 0.17)	93% (90%, 96%)	0.039 (0.034, 0.044)	76% (73%, 79%)
One-way analyses								
On treatment throughout	2855	8.7	8.7	3.02	0.46 (0.39, 0.53)	73% (69%, 77%)	0.073 (0.067, 0.079)	55% (51%, 58%)
Ever off-treatment	2007	8.8	4.6	3.40	−0.36 (−0.44, −0.28)	123% (118%, 128%)	−0.010 (−0.018, −0.001)	106% (101%, 111%)
By gender								
Male	1224	8.6	6.8	3.19	−0.12 (−0.23, 0.00)	107% (100%, 114%)	0.013 (0.002, 0.024)	92% (85%, 99%)
Female	3638	8.8	7.0	3.17	0.20 (0.14, 0.26)	88% (84%, 92%)	0.048 (0.042, 0.053)	71% (67%, 74%)
Baseline at or before 28 February 2003	997	8.7	7.0	3.35	0.03 (−0.09, 0.14)	98% (91%, 106%)	0.035 (0.024, 0.046)	78% (71%, 85%)
Baseline after 28 February 2003	3865	8.8	7.0	3.13	0.14 (0.09, 0.20)	91% (88%, 95%)	0.040 (0.034, 0.045)	75% (72%, 79%)
Baseline at or before 31 August 2003	2351	8.7	6.9	3.32	0.02 (−0.06, 0.10)	99% (94%, 103%)	0.033 (0.025, 0.040)	79% (75%, 84%)
Baseline after 31 August 2003	2511	8.8	7.0	3.05	0.21 (0.14, 0.29)	88% (83%, 92%)	0.044 (0.037, 0.052)	73% (69%, 77%)
Disease duration at baseline								
≤3 years	1251	8.8	7.3	2.46	0.45 (0.33, 0.57)	77% (71%, 83%)	0.060 (0.050, 0.070)	65% (59%, 71%)
>3 years	3611	8.7	6.9	3.43	0.00 (−0.05, 0.06)	100% (96%, 103%)	0.032 (0.026, 0.037)	80% (76%, 84%)
≤6 years	2295	8.8	7.2	2.67	0.32 (0.24, 0.40)	83% (78%, 87%)	0.051 (0.043, 0.058)	70% (65%, 74%)
>6 years	2567	8.7	6.8	3.64	−0.06 (−0.12, 0.01)	104% (99%, 108%)	0.028 (0.022, 0.035)	82% (78%, 86%)
Two-way analyses								
Disease duration ≤3 years and:								
Baseline EDSS ≤3.5	981	8.9	7.5	1.80	0.60 (0.47, 0.73)	73% (67%, 79%)	0.068 (0.058, 0.079)	61% (55%, 67%)
Baseline EDSS 4–5.5	199	8.4	6.7	4.41	0.00 (−0.29, 0.28)	100% (74%, 127%)	0.039 (0.009, 0.068)	74% (56%, 94%)
Baseline ≥EDSS 6	71	8.0	5.7	6	−0.35 (−0.76, 0.05)	209% (83%, 338%)	0.001 (−0.050, 0.051)	100% (65%, 135%)
Disease duration >3 years and:								
Baseline EDSS ≤3.5	1959	9.0	7.4	2.12	0.23 (0.14, 0.31)	90% (85%, 94%)	0.044 (0.036, 0.051)	73% (69%, 78%)
Baseline EDSS 4–5.5	1076	8.6	6.7	4.43	−0.14 (−0.23, −0.04)	112% (104%, 121%)	0.038 (0.027, 0.048)	76% (69%, 82%)
Baseline ≥EDSS 6	576	8.0	5.4	6	−0.48 (−0.59, −0.37)	236% (204%, 268%)	−0.021 (−0.038, −0.004)	114% (103%, 125%)

EDSS, Expanded Disability Status Scale.

8.7 and 10 in our study). The use of marginal structured Cox modelling (which included the contemporary untreated patients) did not change interpretation of findings.[30]

Selection bias

UK patients eligible for DMTs prior to 2002 but who progressed rapidly so they were no longer eligible would have been excluded. This bias would favour a treatment effect. Working against a treatment effect is the exclusion of those who were previously eligible but were no longer active enough to be included. A bias against a treatment effect might have occurred due to the presence of prevalent as well as incident ABN eligible patients, which would have included patients with later disease being treated, and it is possible that DMTs may be less effective later on in the disease course. Indeed our results suggested a greater treatment effect when excluding patients recruited in the first 18 months (once the backlog of prevalent patients had been entered). It is also reassuring that disease duration at study entry was almost identical for both data sets (table 3).

Other methodological issues worth considering are the use of the EDSS as the main outcome measure and the requirements for the EDSS scoring. Although the EDSS is not sensitive to change over the shortterm and focuses on mobility, it is the most widely used and accepted measure of disability in MS. Using this measure also allows us to convert the outcome to quality of life measures for cost-effectiveness calculations and to

Table 6 Variation of the treatment effect with year of analysis (MLM)

Analysis	Relative progression rate (%) at year:				
	2	4	6	8	10
Primary analysis cohort					
EDSS basis (95% CI)	47% (41% to 52%)	58% (54% to 62%)	66% (63 to 69%)	71% (68 to 74%)	72% (69 to 74%)
Utility basis (95% CI)	36% (31% to 41%)	55% (51% to 60%)	68% (64 to 72%)	76% (73 to 80%)	77% (743 to 80%)
RRMS-only subgroup					
EDSS basis (95% CI)	42% (36% to 48%)	55% (51% to 59%)	63% (60% to 66%)	69% (66% to 72%)	69% (67% to 72%)
Utility basis (95% CI)	28% (22% to 34%)	50% (45% to 55%)	62% (58% to 66%)	74% (70% to 78%)	76% (72% to 79%)

EDSS, Expanded Disability Status Scale; MLM, multilevel model; RRMS, relapsing remitting multiple sclerosis.

assess whether our results are plausible by comparing the results with other studies. The clinicians in our study did not undergo formal EDSS training as is required in clinical trials, rather the EDSS was collected within the routine clinical setting. Sites were instructed to use the same clinician to perform the EDSS over time where possible, and it was performed mainly by neurologists with expertise in MS. This was similar to the way the British Columbia natural history data set EDSS measurements were performed and thus comparable. The interobserver variability in our study was previously reported[31] to have a kappa value of 0.59, 0.71 and 0.85: for full agreement, within 0.5 and within 1.0 EDSS scores, respectively, and it is expected that the intraobserver variability is even better.

On balance, therefore, we consider that there are sound reasons for believing that these results represent a true treatment effect. These reasons include the good agreement between our estimates of the absolute treatment effect at year 2 of a mean EDSS difference of 0.22[13] and those derived from a meta-analysis of RCTs which gave a mean EDSS difference of 0.25[32]; the generally good agreement between two very different modelling approaches (particularly when the Markov models allows time to be included in the model); the adjustments included in each model to allow for the (relatively small) differences in baseline characteristics between the British Columbia and RSS data sets; and the use of a wide range of sensitivity analyses to test for possible residual sources of bias.

Implications for the treatment of MS

Assessing the longer term effectiveness of treatment is vital in considering the cost-effectiveness of expensive treatments for chronic conditions, but RCTs have not and cannot address this in MS where disability is acquired over many years. We have performed the only prospective long-term study of the treatment effect of MS DMTs, and have managed to recruit and retain a very large cohort (with higher levels than many short-term RCTs). Our results show that the benefits seen in the short term for these drugs are maintained over a 10-year period, although the treatment effect appears to decrease over time and may not justify the prices for these drugs in some healthcare systems particularly where the drug costs are higher.[33] This apparent waning effect would be consistent with the recent meta-analysis[34]showing an inverse age-dependent association with efficacy. Because age, disease duration and EDSS scores are all inter-related and the latter two were associated with the treatment effect size in our study, it is not clear what is the driving factor. Although newer treatments are now available and are often preferred because of greater efficacy or ease of administration, their cost-effectiveness has in turn been assessed by NICE using incremental comparison with these first-generation drugs and more recently in direct head-to-head studies now that placebo studies are regarded as unethical. Thus the results of the MS RSS will have important consequences for assessing the cost-effectiveness of subsequent currently licensed and future drugs for MS.

Our finding that the treatment effect of these drugs is attenuated over time could have important implications for clinical practice, both for MS and perhaps for other longer term conditions. It might now be ethical to conduct an RCT to determine whether dicontinuation of treatment—after a given period on treatment, or after reaching a given level of disability—offers any disadvantage over continuing treatment. An RCT is currently under way to assess the effects of withdrawing treatment in older patients with MS who have had no relapses or new brain lesions for at least 5 years.[35] Future trials of treatments for MS could be designed so that one arm receives treatment for an initial period and then discontinues treatment. On the basis of our findings, it might be more cost-effective—and ultimately in the interests both of patients with MS and of patients with other conditions—to intervene earlier in the disease course, rather than to persist with treatment for long periods.

Author affiliations
[1]Clinical Neurology, The Oxford University Hospitals Trust, Oxford, UK
[2]Department of Neurology, The Newcastle upon Tyne Hospitals Trust, Newcastle upon Tyne, UK
[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[4]Biostatistics, PAREXEL International, Berlin, Germany
[5]Department of Medicine (Neurology), University of British Columbia, Vancouver, British Columbia, Canada
[6]The Townsville Hospital, Townsville, Queensland, Australia
[7]Institute of Psychological Medicine and Clinical Neuroscience, Cardiff University, University Hospital of Wales, Cardiff, UK
[8]Warwick Medical School, University of Warwick, Warwick, UK
[9]Department of Health, Leeds, UK

Acknowledgements We acknowledge the input of Pelham Barton and Richard Gray, who were members of the SAG, and Nicola Russell and Tracy Nicholson, who represented the MS Trust throughout this project. The multilevel modelling project was funded by the NIHR Health Technology Assessment programme (HTA project 10/55/01) and has been published in full in the Health Technology Assessment journal series.[23] Visit the HTA programme website for more details (http://www.nets.nihr.ac.uk/projects/hta/105501). 72 centres have participated in the collection of data for the risk-sharing scheme from across the UK. Data have been collected over a 10-year period, and thus changes have been seen within the clinical teams at the centres. The authors are grateful to all the clinicians, nurses and administrators for the work undertaken at the centres, which made the RSS possible. A list of the centres can be found at http://www.mstrust.org.uk/research/risksharingscheme. The MS Trust has performed an administrative coordinating role in the risk-sharing scheme, and PAREXEL has performed the data collection from the sites and undertaken the analytical work. The British Columbia MS database has been maintained over the years by grants from various sources, including the MS Society of Canada, the MS/MRI Research Group, the US National MS Society and the Canadian Institutes of Health Research. We gratefully acknowledge the BC MS Clinic neurologists who contributed to the study through patient examination and data collection (current members listed on https://www.mstrust.org.uk/risk-sharing-scheme-BC-clinic-neurologists). RL and YB-S are supported by the NIHR Collaboration for Leadership in Applied Health Research and Care, West Midlands and West, respectively.

Contributors JP, MD and MB were the clinical leads for the study involved in the day-to-day running of the study. JP and MD were responsible for writing the paper and advising on the interpretation of the results. MB was responsible for revising the paper. ML, KT and YB-S performed the MLM, revised the paper and interpreted the results; in addition YB-S was on the scientific advisory committee. BP and TB were responsible for the Markov model analysis, revising the paper and advising on interpretation of the results. NR supplied the Welsh natural history data set and was involved in the validation of the MLM and revising the paper. HT, FZ and JO oversaw the use of their BC natural data set, revised the paper and interpreted the results. RL was the chair of the scientific advisory board, advised on the analysis and interpretation of the results, and revised the paper. CD was responsible for the oversight of the study, the analysis and interpretation of both models, and writing and revising the paper.

Funding The administration costs of the risk-sharing scheme were split equally five ways among the UK health departments and the four manufacturers of the DMTs (Biogen Idec, Merck Serono, Bayer, Teva UK). Additional funding from the UK Health Technology Assessment programme was obtained for the development of the MLM model (HTA Project: 10/55/01)The views and opinions expressed therein are those of the authors.

Competing interests None declared.

Patient consent Obtained.

Ethics approval Ethical approval for the RSS was given by the South East Medical Research Ethics Committee (MREC 2/01/78). The University of British Columbia's Clinical Research Ethics Board approved the study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data storage governance is under review to assess who will store the data on behalf of the MS Trust, the four pharma companies and the Department of Health. Also the oversight committee to give permission for access to data is being discussed currently.

REFERENCES

1. Tremlett H, Zhao Y, Rieckmann P, *et al*. New perspectives in the natural history of multiple sclerosis. *Neurology* 2010;74:2004–15.
2. The IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. The IFNB Multiple Sclerosis Study Group. *Neurology* 1993;43:655–61.
3. Johnson KP, Brooks BR, Cohen JA, *et al*. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: results of a phase III multicenter, double-blind placebo-controlled trial. The Copolymer 1 Multiple Sclerosis Study Group. *Neurology* 1995;45:1268–76.
4. Jacobs LD, Cookfair DL, Rudick RA, *et al*. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG). *Ann Neurol* 1996;39:285–94.
5. Randomised double-blind placebo-controlled study of interferon beta-1a in relapsing/remitting multiple sclerosis. PRISMS (Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis) Study Group. *Lancet* 1998;352:1498–504.
6. Placebo-controlled multicentre randomised trial of interferon beta-1b in treatment of secondary progressive multiple sclerosis. European Study Group on interferon beta-1b in secondary progressive MS. *Lancet* 1998;352:1491–7.
7. Secondary Progressive Efficacy Clinical Trial of Recombinant Interferon-Beta-1a in MS (SPECTRIMS) Study Group. Randomized controlled trial of interferon- beta-1a in secondary progressive MS: Clinical results. *Neurology* 2001;56:1496–504.
8. Panitch H, Miller A, Paty D, *et al*. Interferon beta-1b in secondary progressive MS: results from a 3-year controlled study. *Neurology* 2004;63:1788–95.
9. National Institute for Clinical Excellence. *Beta interferon and glatiramer acetate for the treatment of multiple sclerosis. NICE Technology Appraisal Guidance No. 32*. London: NICE, 2002.
10. Department of Health. *Cost effective provision of disease modifying therapies for people with multiple sclerosis*. London: Health Service Circular (2002/004), Stationery Office, 2002.
11. Boggild M, Palace J, Barton P, *et al*. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. *BMJ* 2009;339:b4677.
12. Palace J, Bregenzer T, Tremlett H, *et al*. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. *BMJ Open* 2014;4:e004073.
13. Palace J, Duddy M, Bregenzer T, *et al*. Effectiveness and cost-effectiveness of interferon beta and glatiramer acetate in the UK Multiple Sclerosis Risk Sharing Scheme at 6 years: a clinical cohort study with natural history comparator. *Lancet Neurol* 2015;14:497–505.
14. Lawton M, Tilling K, Robertson N, *et al*. A longitudinal model for disease progression was developed and applied to multiple sclerosis. *J Clin Epidemiol* 2015;68:1355–65.
15. Chilcott J, Miller D, McCabe C. Modelling the cost effectiveness of interferon beta and glatiramer acetate in the management of multiple sclerosis. *BMJ* 2003;326:522.
16. Parexel/Department of Health. "Cost effective provision of disease modifying therapies for people with Multiple Sclerosis: Statistical Analysis Plan for year 10 data" (Department of Health, November 2015).
17. Association of British Neurologists "Guidelines for the use of beta interferons and glatiramer acetate in multiple sclerosis". January 2001.
18. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;33:1444–52.
19. Lechner-Scott J, Kappos L, Hofman M, *et al*. Can the Expanded Disability Status Scale be assessed by telephone? *Mult Scler* 2003;9:154–9.
20. Shirani A, Zhao Y, Karim ME, *et al*. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. *JAMA* 2012;308:247–56.
21. Jackson CH, Sharples LD, Thompson SG, *et al*. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D* 2003;52:193–209.
22. Goldstein H, Models MS. *Multilevel Statistical Models (second edition)*. London: Edward Arnold, 1995.
23. Tilling K, Sterne JA, Wolfe CD. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Stat Med* 2001;20:685–704.
24. Tilling K, Lawton M, Robertson N, *et al*. Modelling disease progression in relapsing-remitting onset multiple sclerosis using multilevel models applied to longitudinal data from two natural history cohorts and one treated cohort. *Health Technol Assess* 2016;20:1–48.
25. Gold R, Kappos L, Arnold DL, *et al*. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med* 2012;367:1098–107.
26. Fox RJ, Miller DH, Phillips JT, *et al*. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med* 2012;367:1087–97.
27. Lourenco P, Shirani A, Saeedi J, *et al*. Oligoclonal bands and cerebrospinal fluid markers in multiple sclerosis: associations with disease course and progression. *Mult Scler* 2013;19:577–84.
28. Inusah S, Sormani MP, Cofield SS, *et al*. Assessing changes in relapse rates in multiple sclerosis. *Mult Scler* 2010;16:1414–21.
29. Shirani A, Zhao Y, Kingwell E, *et al*. Temporal trends of disability progression in multiple sclerosis: findings from British Columbia, Canada (1975-2009). *Mult Scler* 2012;18:442–50.
30. Karim ME, Gustafson P, Petkau J, *et al*. Marginal structural Cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *Am J Epidemiol* 2014;180:160–71.
31. Pickin M, Cooper CL, Chater T, *et al*. The Multiple Sclerosis Risk Sharing Scheme Monitoring Study – early results and lessons for the future. *BMC Neurol* 2009;9:1.
32. Filippini G, Munari L, Incorvaia B, *et al*. Interferons in relapsing remitting multiple sclerosis: a systematic review. *The Lancet* 2003;361:545–52.
33. Hartung DM, Bourdette DN, Ahmed SM, *et al*. The cost of multiple sclerosis drugs in the US and the pharmaceutical industry: Too big to fail? *Neurology* 2015;84:2185–92.
34. Weideman AM, Tapia-Maltos MA, Johnson K, *et al*. Meta-analysis of the age-dependent efficacy of multiple sclerosis treatments. *Front Neurol* 2017;8:577.
35. Explore Our Portfolio of Funded Projects. http://www.pcori.org/research-results/2015/discontinuation-disease-modifying-therapies-dmts-multiple-sclerosis-mshttp://www.pcori.org/research-results/2015/discontinuation-disease-modifying-therapies-dmts-multiple-sclerosis-ms

Appendices to "Assessing the long-term effectiveness of interferon-beta and glatiramer acetate in multiple sclerosis: final ten year results from the UK multiple sclerosis risk sharing scheme"

Appendix 1: Background to the UK MS Risk Sharing Scheme

1.1 The MS disease modifying therapies (DMTs) are licensed for relapsing MS, ie for early stage disease, whereas the major disability is incurred in the later progressive phase. In order for the DMTs to be cost-effective they must demonstrate that they significantly delay progression to these later stages of the disease. When the UK's National Institute for Health and Clinical Excellence (NICE) first looked at the DMTs in 2002 it was satisfied that there was robust evidence for their short-term benefits, e.g. in reducing the frequency of relapses. It was however unable to conclude, on the basis of the evidence then available, that the benefits of treatment seen in early years could safely be extrapolated over the 20 or more years needed to achieve cost-effectiveness[1,2].

1.2 The UK MS risk sharing scheme (RSS)[3] has two components. Firstly, the then current UK prices for the four main DMTs were reduced, where necessary, in order to achieve a cost effectiveness target of £36,000 per QALY, using the model of disability progression developed for NICE's 2001 technology appraisal, a 20-year time horizon, and target treatment effects (relative rates of disability progression) derived from the pivotal RCTs. Secondly, the parties to the scheme agreed to track a prospectively observed MS cohort on DMTs against the trajectory required to offer cost effectiveness. The plan was to assess the data every two years, adjusting the price if necessary after each analysis to maintain the 20 year cost-effectiveness target should the observed results differ from the required trajectory by more than an agreed "margin of tolerance" (10% from year 4 onwards).

1.3 Considering the practicalities of running such a scheme, the decision was made to follow the cohort for the first 10 years of the twenty year model. Over 5,500 patients were enrolled in 2002/2003. EDSS scores were collected pragmatically, i.e. as part of normal clinical practice. Although the EDSS was not assessed under trial standard protocols, it was performed by MS neurologists who were experienced in this scoring with the same neurologist being encouraged to continue scoring an individual patient throughout if possible.

1.4 From 2005 to 2016 the parties to the scheme were been advised by a Scientific Advisory Group (SAG) chaired by Professor Richard Lilford. The group's main function was to advise on the analysis plan for the interim and final analyses and on the interpretation of the results. In addition, the group advised the MS Trust (as the custodian of the data) on applications from other researchers to access the data on the RSS cohort.

1.5 This paper focusses on the evidence generated by the scheme on the long-term effectiveness of the four DMTs in aggregate. Results on the extent to which the DMTs in aggregate achieved the targets set at the outset of the scheme will be presented elsewhere. The target outcomes for individual DMTs were agreed between the companies concerned and the Department of Health and remain commercially confidential.

Appendix 2: Relation between EDSS and utility

2.1 We examined three possible sources for the relation between EDSS and utility (quality of life): two surveys of patients with MS[4,5], using patient-determined measures of disability (the "MS Trust" and "Heron" datasets), and an unpublished paper[6] drawing on the clinician-determined EDSS scores in the RSS itself (the "Boggild dataset"). In each case, patients were asked to completed the EQ5D questionnaire, an instrument which categorises the patient's perceived state of health according to the dimensions of mobility, self-care, ability to take part in usual activities (e.g. work), pain/discomfort and anxiety/depression. The EQ5D scores can then be converted into utility, an overall measure of society's perception of the patient's quality of life, using standard tariffs[7]. A utility of one represents perfect health; a utility of 0.5 implies that on average members of the general population would regard 12 months of life in that health state as equally preferable as 6 months of life in perfect health.

2.2 On advice from our Scientific Advisory Group, we decided to use a synthesis of the MS Trust and Heron datasets for the primary analysis, primarily because they contained more data for the higher EDSS scores. The utility values we adopted are given in the table below:

Utility values used for the year 10 analysis

EDSS	Utility
0	0.9248
1 or 1.5	0.7614
2 or 2.5	0.6741
3 or 3.5	0.5643
4 or 4.5	0.5643
5 or 5.5	0.4906
6 or 6.5	0.4453
7 or 7.5	0.2686
8 or 8.5	0.0076
9 or 9.5	-0.2304

In our analysis of the year 6 data[8,9] we also carried out a sensitivity analysis using pooled data from all three datasets, but the results were very similar to those of the primary analysis and we did not repeat this sensitivity analysis at year 10.

2.3 Details of the three datasets and of the methodology used for their synthesis can be found in a report by IMS Health[10].

Appendix 3: Changes over time

3.1 Our analysis of the year 6 data suggested that the treatment effect may reduce with time from baseline (start of treatment). We therefore decided to explore this possible change over time in the treatment effect through a range of pre-specified analyses.

Multi-level model

3.2 A multi-level model was fitted to the year 10 RSS data of identical form to that already fitted to the BCMS data, and the two models were used to predict disability progression from the same baseline distribution (the RSS baseline)[a]. The resulting projections were compared both visually and by calculating the divergence for each year from baseline, which represents the cumulative treatment effect up to that point.

Markov model: time-varying hazard ratios

3.3 In the Markov framework, we used a rather different method (see Annex F of the statistical analysis plan[11] for details). The basic approach is to assume that the effect of treatment by a DMT can be modelled as a time-varying hazard ratio – specifically, a hazard ratio with different values for consecutive two-year periods. The hazard ratios are varied in order to minimise the deviation between mean disability progression in the observed RSS data and in the (on treatment) counterfactual. Two variant approaches were tried:

i a "non-parametric" approach, which did not impose any particular functional form on the hazard ratios; and

ii a "parametric" approach which assumed that the hazard ratios followed a simple parametric form (eg linear, step function or negative exponential).

Markov model – direct estimation of a model using RSS data

3.4 Finally, we sought to estimate a Markov model using RSS data in order to examine which particular transitions, in which time periods, were enhanced or reduced in the RSS (treated) patients compared with the BCMS (untreated) patients. Because of the large number of parameters and the risk of obtaining misleading results through random fluctuations, most of the work was done using a "constrained" estimation process in which elements of the instantaneous probability matrix in the RSS model were related to those in the BCMS model by equations of the form

$$q^{RSS}_{ijt} = q^{BCMS}_{ijt} \times r_{ijt} \qquad i \neq j \qquad (5)$$

where i is the EDSS state before the transition, j the state after the transition, t the time from baseline at the start of the transition, and the hazard ratio r_{ijt} is constrained to take one of a small number of possible values. We then used a maximum likelihood method to estimate the most parsimonious set of the hazard ratios r_{ijt} needed to obtain a reasonable fit to the data while exploring the variation of the hazard ratios with initial EDSS i and time t.

---

[a] Because the functional form chosen for the model includes both time and log time terms, it is difficult to compare the two models simply by inspection of the coefficients to test for a time-treatment interaction. We offer this device of projecting forward from a common baseline as an intuitively easier way of achieving the same end.

Appendix 4: Further details on the Markov and Multi-Level Models

6.1 The parameters for the two models representing disability progression for untreated patients were estimated using a subset of patients in the BCMS dataset[12] who, at some clinic visit before 31 December 1995, would have met the 2001 criteria of the UK's Association of British Neurologists for eligibility for treatment with a DMT. This gave a total of 978 patients for potential analysis, all of whom were used to estimate the MLM. For the Markov model, patients had to have had at least one further EDSS measurement before the cut-off date of 31 December 1995 (or before first treatment with a DMT if this was earlier). This left a total of 898 patients for analysis, followed up for a median period of 4.4 years before the cut-off date.

Markov model

4.2 The Markov model[13] defines its states in terms of the patient's EDSS score (half-integral scores are rounded down to the nearest integer). Thus patients at EDSS 0 are allocated to state 1, patients at EDSS 1 or 1.5 are allocated to state 2, and so on. The model assumes a constant probability of making a transition from state i to state j conditional on the vector of baseline covariates $\boldsymbol{x}$ for the individual patient. For the purposes described in this paper death from non-MS causes was not explicitly modelled and patients in the RSS who died before the final analysis year were treated as lost to follow up.

4.3 Estimation was by the continuous-time method of Jackson[14] which does not require the data to be collected at regular (eg annual) intervals, and allows the transition probabilities to depend on baseline and other covariates. After assessing a number of possible combinations of baseline covariates, a relatively simple model was chosen[13] with a single baseline covariate, age at onset as a binary variable split about the median value in the BCMS dataset. Since information on MS-related death was not available from the BCMS dataset, probabilities for MS-related death (transitions to "EDSS 10") were taken from the model developed for NICE's 2002 appraisal[2]. For one sensitivity analysis, we used a "time-varying" model with separate transition matrices estimated for the first two years after baseline and for the rest of the follow-up period (see Appendix 7).

4.4 The transition probabilities are then applied to the baseline EDSS scores of the RSS cohort and to subsequent modelled EDSS states over 10 years to give the expected EDSS progression for patients had they never received treatment. The difference between the observed and predicted mean EDSS score represents the "comparison against control". Confidence intervals on projections using the Markov model are derived by bootstrapping.

Repeated measures multi-level model

4.5 The repeated measures model was derived from the EDSS trajectories of individual patients in the British Columbia dataset. The basic approach is to estimate a mean trajectory for the whole cohort, the variation of individual trajectories about this mean, and the fluctuation of individual EDSS scores about the trend for each individual. The model is then applied to the baseline data for patients in the RSS dataset to predict the EDSS progression which would have been expected without treatment for each individual. The projections are then combined across all individuals to produce a predicted mean EDSS progression for the whole population.

4.6 The model was estimated by means of the method of multilevel models[15]. The EDSS score is regarded as a continuous variable although the observations can take only integral or half-integral values. Our model[16,17] had two levels; observations (level 1) within individuals (level 2). We used a model with a random intercept and two random powers of time since ABN eligibility: time and the log of time. We also allowed level-1 variation to change linearly with time, to take into account varying measurement error in EDSS scores at different levels of disability. Thus the basic model is of the form:

$$y_{ij} = \beta_0 + u_{0i} + e_{1ij} + (\beta_1 + u_{1i} + e_{2ij})\, t_{ij} + (\beta_2 + u_{2i}) \log t_{ij},$$

$$\text{where } \{e_{1ij}\} \sim N_2(0, D_e), \{u_{1i}\} \sim N_3(0, D_u),$$

$$D_e = \begin{bmatrix} var(e_{1ij}) & cov(e_{1ij}, e_{2ij}) \\ cov(e_{1ij}, e_{2ij}) & 0 \end{bmatrix} \text{ and}$$

$$D_u = \begin{bmatrix} var(u_{0i}) & cov(u_{0i}, u_{1i}) & cov(u_{0i}, u_{2i}) \\ cov(u_{0i}, u_{1i}) & var(u_{1i}) & cov(u_{1i}, u_{2i}) \\ cov(u_{0i}, u_{2i}) & cov(u_{1i}, u_{2i}) & var(u_{2i}) \end{bmatrix} \quad (4)$$

where y_{ij} is the EDSS for individual i at occasion j and t_{ij} is the time since ABN eligibility (plus one year) for individual i at occasion j, β_0 to β_2 are coefficients, and the e and u are random variables (residuals). As with the Markov model, non-MS death is not included in the model.

4.6 We then included the binary covariate of age at onset of MS (as in the Markov model), allowing this to be associated with intercept, time and log of time. We assessed the normality of the residuals, and the fit of the model by comparing the actual and predicted EDSS values. All analyses were carried out using Stata software[18], and all multilevel models estimated using the runmlwin command[19]. Bootstrapping was used to derive 95% confidence intervals.

4.7 We used the random effects matrices from the BCMS model to estimate the "natural history" EDSS for those in the RSS cohort (at every time at which they had an observed EDSS), conditional on their observed baseline EDSS[20]. When calculating the observed and natural history progression, we used the observed baseline EDSS as the comparator, for consistency with the Markov analysis. We assessed the sensitivity of our analysis to this assumption by also using the estimated EDSS as baseline as the comparator. This led to slightly higher estimates of "actual" and natural history progression but does not affect the estimated absolute treatment effect. See the web appendices to reference 8 for details.

Comparative strengths and weaknesses of the models

4.8 Validation work carried out as part of our year 6 analysis has been described elsewhere[13,16,17,]. The Markov model reproduces very accurately the disability progression n the BCMS dataset from which it was derived, both in terms of EDSS and utility. It also appeared to perform well in a "random split" test in which patients in the BCMS dataset were allocated randomly to one of two subsets, one of which was used to estimate the parameters for the Markov model and the other to test the predictions of the model. However, subsequent work with further random splits suggested that the model is not very sensitive in adjusting for changes in the baseline distribution of EDSS scores, ie the model tends to underpredict disability progression in patients with EDSS at baseline above the median in the BCMS dataset, and overpredict progression in patients with EDSS below the median.

4.9 We believe that this problem arises from the standard Markov assumption that the probability of a transition from one disease state to another is independent of time from baseline. In practice, most of the transitions used to estimate transition probabilities from (for example) EDSS 5 in the BCMS dataset occur some years after baseline; they may not be a good guide to the probability of a transition for a patient who is already at EDSS 5 at baseline. In principle, the use of baseline covariates (eg age at onset, generally considered a good predictor of speed of disability progression) should help to adjust for this difference but it would appear that the adjustment is insufficient. A possible way of overcoming this difficulty, by estimating and using different transition probabilities for different periods after baseline, is described in Appendix 7 below.

4.10 In contrast, multi-level models are inherently sensitive to the characteristics of individual patients such as baseline EDSS – in fact, in the version of the model used for this study there is a linear relation between the predicted EDSS for an individual patient at any given time and the deviation of the patient's baselined EDSS from the population mean in the BCMS dataset. As a result, the MLM performed well in an external validation, predicting disability progression in a dataset of patients from Cardiff, Wales on the basis of a model estimated from BCMS data[16,17].

4.11 Although in this study we believe that the MLM is likely to be more robust than the Markov model in modelling changes in the mean EDSS of the RSS population, the Markov model is superior in modelling the details of the EDSS distribution. Partly because of the unusual (non-linear) nature of the scale, the EDSS distribution of typical patient populations tends to be bi-modal, with one mode at around EDSS 2-3 and another at EDSS 4. The Markov model is quite successful in reproducing this bi-modal distribution, as the chart at figure A1 shows. (In the MLM, all residuals are assumed to be normally distributed so the predicted distribution of patients at any time is normally distributed about its predicted mean value.) Since changes in mean utility for populations of MS patients are highly sensitive to small changes in numbers at the extreme end of the distribution (EDSS 7 and upwards) this suggests that the results from the Markov model may be more robust than those of the MLM when considering results on the utility basis.

4.12 Finally, it may be worth pointing out that the survival analysis described in Appendix 5 uses, in effect, a third independent model, which adds weight to our view that we are seeing a real and clinically significant long-term treatment effect and not an artefact of the particular models used.

Appendix 5: Survival analysis (delay in median time to EDSS 6)

5.1 One new analysis in this year 10 analysis was to estimate the impact of treatment on the median time to reach the clinically significant milestone of EDSS 6 (needing a stick to walk 100 metres). The endpoint chosen, in line with other analyses carried out using BCMS data, was the first occurrence of an EDSS score of 6 or over, confirmed by at least one subsequent score at EDSS 6 or over, and with no later score at less than EDSS 6 ("confirmed and sustained progression to EDSS 6").

5.2 Our objective was to compare the median time required to reach this endpoint between the BCMS dataset (untreated patients) and the RSS dataset (treated patients), adjusting for

differences in the baseline distribution between the two datasets. Ethical constraints (lack of consent given for transfer of BCMS individual patient data to a third party) meant that we could not carry out a conventional analysis of the combined data and test directly for the significance of a co-efficient representing the effect of treatment. Instead we used an indirect method of comparison. We fitted parametric models of identical form to the two datasets (see table below), after first testing (in the BCMS data) that the hazard ratios for the baseline covariates in the parametric model were similar to those for a non-parametric Cox proportional hazards model. The model chosen was a Weibull model with gender and baseline EDSS (stratified as EDSS 0 to 1.5, EDSS 2 to 3.5, EDSS 4 to 5.5, and EDSS 6/6.5) as covariates. Adding age at onset as a further covariate did not significantly improve the model.

Fitted parameters for Weibull model (in log form, standard errors in brackets)

Parameter	BCMS dataset (n = 836)	RSS dataset (n = 4,310)
"Shape" parameter *k*	0.327 (0.055)	0.315 (0.023)
Components of "scale" parameter λ:		
Intercept (male, EDSS 0-1.5)	3.441 (0.167)	3.477 (0.090)
Female	0.076 (0.109)	0.227 (0.041)
EDSS 2 to 3.5	-1.011 (0.146)	-0.709 (0.076)
EDSS 4 to 5.5	-1.648 (0.180)	-1.434 (0.075)
EDSS 6 to 6.5	-1.723 (0.240)	-1.411 (0.109)

Cumulative proportion reaching endpoint by year $x = 1 - \exp(- (x/\lambda)^k)$

5.3 For each subpopulation *i* (each combination of gender and stratified EDSS) in the BCMS dataset we calculated the "survival curve" $P_i^{BCMS}(t)$, ie the proportion of patients who have not yet reached the endpoint at or before time *t*. We then calculated the weighted average survival curve

$$P^{BCMS}(t) \quad = \quad \Sigma_i \; w_i \; P_i^{BCMS}(t) \qquad (5)$$

where the weights w_i represent the proportion of patients in subpopulation *i* in the RSS dataset. This weighted average curve thus represents the survival curve we would expect for an untreated population with the same baseline distribution as the RSS population (ie it corresponds to the "comparator control group" described in the methods section of the main paper). The median time to EDSS for this untreated population can be readily found from equation (5) by seeking the time *t* for which the survival proportion is exactly 50%. The median time to EDSS 6 for the treated patients in the RSS population is found in an analogous way.

5.4 Calculating the confidence intervals on these estimates is not straightforward, because of the possibility of correlation between the sampling errors in the estimated survival proportions for different subpopulations. However, using the standard outputs from the statistical packages used we were able to estimate for each of the models the correlation between the various Weibull parameters (the correlation between the "shape" and "scale" parameters and between the various components of the "scale" parameter). We then used stochastic simulation (50,000 replications) to estimate the 95% confidence intervals on the

weighted average survival functions, and hence on the estimated median times. To calculate the 95% confidence intervals on the difference in median times (the delay in reaching EDSS 6) we relied on the fact that the BCMS and RSS datasets refer to entirely different populations and therefore there will be no correlation between the respective sampling errors.

Appendix 6: Further detail on the sensitivity analyses, including the results from the "time-variant" Markov model

6.1 The sensitivity analyses for this final year 10 analysis, like those we carried out at year 6[8], were intended to assess the possible impact of the most likely sources of bias, in particular resulting from missing values. For the year 6 analysis, we were guided by some additional descriptive analyses of the RSS and BCMS data which showed where bias was most likely to occur (see ref 8, in particular online appendices 6 and 7; see also appendix 7 below). A number of the analyses carried out at year 6 which showed only a very small impact on the outcomes were omitted in the year 10 analysis.

6.2 The main addition at year 10 was the inclusion of a sensitivity analysis using a "time-variant" Markov model. This model was originally developed for use as part of the "waning" analysis described in appendix 5 above; a further motivation was that the model used for our primary analysis replicates very accurately changes over time in mean EDSS and mean utility in the BCMS dataset as a whole, but tends to underestimate disability progression in patients with low EDSS at baseline and overestimate disability progression in patients with high EDSS. We thought it would be of interest to examine the impact of using this model on our primary analysis, and on the main subgroup analyses, although this had not been pre-specified in our analysis plan.

6.3 In the Jackson method for estimating Markov models[14] the basic unit of analysis is the "transition", where a transition is defined as any two consecutive measurements on the same patient. For our model, the "transition" contains information on the initial and final time, the initial and final EDSS measurements, and any relevant baseline covariates. It is therefore straightforward to separate those transitions starting 0-2 years, 2-4 years, 4-6 years … from baseline, and to estimate the matrix of transition probabilities separately for each interval. For the BCMS dataset, the number of available transitions decreases with time from baseline so we pragmatically chose to focus on a model with just two time intervals, 0-2 years after baseline and over 2 years after baseline. The respective transition matrices were estimated in exactly the same way as for the model used in our primary analysis.

6.4 Validation of this time-variant model showed that it was superior to the original model in replicating disability progression in patients starting at individual EDSS levels, and nearly as good in replicating mean disability progression in the BCMS dataset as a whole.

6.5 Applying the time-variant model to the primary RSS population, we found significantly larger estimates of treatment effects in terms both of EDSS (relative rate of disability progression 76% (CIs 74%,79%) vs 93% (90%,96%) for the original model) and utility (relative rate of utility progression 64% (CIs 61%,66%) vs 76% (73%,79%) for the original model).

6.6 Using the time-variant model in conjunction with the sub-group analysis by initial EDSS also resulted in changes in the absolute estimates, though not in the qualitative conclusion that the treatment effect is largest for patients starting at low EDSS:

Baseline EDSS	**Relative EDSS progression:**		**Relative utility progression:**	
	Time-variant model	Original model	Time-variant model	Original model
0 to 3.5	70% (68%, 73%)	84% (81%, 87%)	57% (54%, 60%)	69% (65%, 73%)
4 to 5.5	86% (80%, 93%)	110% (102%, 119%)	64% (59%, 69%)	75% (69%, 82%)
6 and 6.5	148% (129%, 168%)	233% (202%, 265%)	98% (89%, 108%)	112% (102%, 123%)

Appendix 7: Potential bias due to differential patterns of data collection between the BCMS and RSS datasets

7.1 We paid considerable attention to the possibility of bias resulting from different patterns of data collection in the RSS cohort as compared to the BCMS dataset used to estimate the parameters for the Markov and multi-level models.

7.2 In the RSS cohort, we have observed a tendency for patients with worse disease progression to fail to attend at subsequent annual reviews, probably because there is little incentive for a patient to attend a review if they have already decided to discontinue DMT treatment. This tendency was already noticeable in the year 2 analysis[21] and was confirmed by the descriptive analyses carried out as part of the year 6 analysis. We have attempted to quantify the potential impact of this differential loss to follow up through the various imputation methods described in the main paper.

7.3 The BCMS dataset was not collected as part of a specific observational study but through routine clinic visits. During the period in which the data used in this study was collected (1980-1995) effective disease modifying treatments were not available. Patterns of data collection could therefore be different from those in the RSS.

7.4 The Table below shows the results of a descriptive analysis of patterns of follow up from the BCMS dataset. It compares baseline parameters and mean EDSS progression over the first 5 years from baseline for two pairs of subsets of the total cohort:

a. patients contributing relatively frequent vs relatively infrequent data, where "frequency" was defined as the number of EDSS scores divided by the interval between first and last scores;

b. patients who had a recorded EDSS score either after or within 18 months of the cut-off date of 31 December 1995, vs patients without such a score (who could therefore be regarded as "lost to follow-up").

Population	Frequency of scores:		Whether "lost to follow up":	
	High frequency	**Low frequency**	**Not lost**	**Lost**
Number (%)	449 (50)	449 (50)	649 (72.3)	249 (27.7)
% female	72%	77%	76%	71%
Age at baseline (eligibility)	37.3	37.0	36.6	38.8
Age at onset	28.9	29.6	28.4	31.4
MSSS score at baseline	4.21	3.61	3.74	4.35
EDSS at baseline	2.67	2.20	2.35	2.66
Number with year 5 data	229	144	289	84
Average increase (year 5 on baseline)	1.88	1.09	1.45	2.00

Frequency = number of EDSS scores divided by interval between first and last scores

Lost to follow-up = no EDSS score after or within 18 months of cut-off date (31.12.1995

7.5 The first comparison shows that patients with relatively frequent EDSS scores tended to have worse prognostic factors at baseline and worse disease progression in the first five years. (British Columbia clinicians have confirmed that, in their experience, patients in the province were more likely to attend clinic if they had concerns over the progress of their disease.) Since the patients with faster disease progression are contributing more EDSS scores to the estimation process, there is at least a theoretical possibility that they could bias the estimates in the Markov model towards predicting higher rates of disease progression for untreated patients, and thus inflate the apparent treatment effect when compared with the actual rates of disease progression in the treated RSS cohort. (In the MLM, this bias may not be operating if the data are “missing at random”, ie if the probability of data being missing depends only on parameters explicitly included in the model.)

7.6 In contrast, the second comparison suggests that patients with relatively fast disease progression are more likely to be “lost to follow-up”, as they are in the RSS cohort, and will thus contribute less data at longer periods from baseline. This would be expected to bias the estimates towards predicting lower rates of disease progression and would tend to offset any bias resulting from the similar differential loss to follow-up in the RSS dataset.

7.7 To help quantify the possible impact of the first factor, we used the MLM to impute additional EDSS scores for patients in the BCMS dataset with relatively sparse follow-up, which we defined as any patients with a gap of more than 2 years between successive values (this used an identical BCMS dataset as used for the continuous Markov model). This would be expected to increase the weighting for patients with relatively slowly progressing disease and thus to compensate for the expected bias. We then re-estimated the transition probabilities for the Markov model using the same method as for the primary analysis. The result of this calculation, in contrast to our initial expectation, was to increase the mean rate of disease progression predicted for the untreated RSS cohort and thus to increase the estimated treatment effect (eg absolute treatment effect on the EDSS basis 0.28 (95% CIs 0.23, 0.34) compared with 0.12 (0.07, 0.17) for the primary analysis). For further results see table 2 and supp table 4 – please note that these results are only available for the Markov model.

7.8 Our tentative conclusion is that, if the differential patterns of attendance at clinics in the BCMS cohort compared to the RSS cohort are responsible for any bias in our estimates of the treatment effect, it is at most very small.

Appendix 8: results of analysis of changes over time

Multilevel model

8.1 The results of the analysis described at appendix 3 (para 3.2) are shown in figure A2 and summarised in the table below:

Year	Predicted EDSS progression using model fitted to BCMS data ("off treatment")	Predicted EDSS progression using model fitted to RSS data ("on treatment")	Difference
0	0	0	0
2	0.33 (0.30, 0.36)	0.64 (0.63, 0.64)	0.30 (0.27, 0.33)
4	0.69 (0.65, 0.72)	1.16 (1.16, 1.17)	0.48 (0.44, 0.51)
6	1.04 (1.00, 1.08)	1.60 (1.59, 1.61)	0.56 (0.52, 0.60)
8	1.39 (1.35, 1.44)	1.99 (1.99, 2.00)	0.60 (0.56, 0.65)
10	1.75 (1.69, 1.80)	2.36 (2.35, 2.37)	0.62 (0.56, 0.67)

There is a strong divergence of the two predictions in the first two years from baseline, suggestive of a strong initial treatment effect. After that the lines continue to diverge, but at a progressively slower rate, and between years 8 and 10 the two lines are almost parallel, suggesting that by this point the treatment effect is greatly attenuated[b].

Markov model: time-varying hazard ratios

8.1 The "implied hazard ratios" calculated by the method described at Appendix 3 (para 3.3) are shown in figure A3. The general picture is of a strong treatment effect in years 0-2

[b] One should not over-interpret these results because EDSS is not intended to be an ordinal scale and changes in mean EDSS at different points of the scale may not be equivalent.

(low hazard ratio) and then a rather smaller treatment effect for the remaining years, with probably little consistent variation between 2-year periods[c].

Markov model – direct estimation of a model using RSS data

8.5 The methods described in para 3.4 of appendix 3 were used to derive a parsimonious model with different hazard ratios for transitions starting at different EDSS values and at different times from baseline. The resulting best model had the following parameters:

Transition	Hazard ratios for:	
	Year 0-1	Years 1-9
Forward transitions starting at:		
EDSS 0	1	1
EDSS 1 to 6.5	0.67	0.59
EDSS 7 and above	1	1
Backward transitions starting at:		
EDSS 1 to 6.5	0.91	0.50
EDSS 7 and above	1	1

In year 1, forward transitions starting from EDSS 1-6 are retarded while backward transitions are hardly affected, resulting in a significant net reduction in the rate of disability progression. In the following years, both forward and backward transitions appear to be retarded. This could be interpreted as implying that the DMTs are reducing the variability in the disease process – or simply that what we have been interpreting as changes in the patient's underlying disability status are in some cases merely the result of recovery from concealed relapses, whose frequency is reduced as a result of the DMTs. There is still a reduction in the net rate of disability progression in years 2 onwards, because forward transitions are more frequent than backward transitions, but this net benefit from treatment is smaller than in year 1.

8.6 Repeating the analysis with the time-variant model described in appendix 6 (ie with different transition matrices for year 0 and for year 1 and following years) the estimated parameters were:

[c] At face value, the non-parametric estimates imply that the treatment effect for years 9-10 is stronger than in the three previous 2-year periods. Fitting a quadratic function to describe the hazard ratios from year 3 onwards gave a parameter for quadratic term, representing the downwards curvature seen in figure A2, which was just significant at a 95% threshold (p =0.047). However, it seems likely that this is a spurious result. Firstly, we cannot imagine any plausible biological mechanism to explain why the treatment effect should diminish and then increase again. Secondly, when the calculations are repeated using just patients with year 10 data (a "complete case" analysis) the curvature is still seen but is no longer significant at 95%.

Transition	Hazard ratios for:	
	Year 0-1	Years 1-9
Forward transitions starting at:		
EDSS 0	1	1
EDSS 1 to 6.5	0.71	0.77
EDSS 7 and above	1	1
Backward transitions starting at:		
EDSS 1 to 6.5	1.72	0.83
EDSS 7 and above	1	1

The average log likelihood was - 0.991 compared with – 0.996 for the simple model, ie using a different transition matrix for year 1 marginally improves the fit to the data. In this analysis, treatment with a DMT increases the probability of a backward transition (disease improvement) in the first year after initiation of treatment. Otherwise the interpretation of these estimated parameters is similar to that for the simple model described in para 8.5 above.

Conclusions

8.7 All three sets of results are consistent with the hypothesis that there is a strong initial treatment effect, followed by a period in which patient disability in treated patients increases, but at a rather slower rate than for untreated patients. To examine this further, we carried out a further unplanned sensitivity analysis comparing actual and expected disability progression starting not from the "true" baseline, but from year 1. As already noted in the main paper (para 29) this showed a smaller treatment effect than the primary analysis after adjusting for the shorter period of follow-up, especially on the EDSS basis (for instance the relative EDSS progression on the MLM is 77% (CIs 74%, 79%) with a year 1 baseline compared to 72% (69%, 74%) for the primary analysis). There is however still a substantial treatment effect on all measures after year 1 – see table 2 and supp table 4 for the details. This observation is, incidentally, further evidence that the observed treatment effect which we report on this paper could not be wholly explained as the result of some unconscious bias in the baseline EDSS assessments.

References to online appendices

1 National Institute of Health and Clinical Effectiveness Technology appraisal 32: disease modifying therapies for multiple sclerosis (NICE, February 2002)

2 Chilcott J, McCabe C, Tappenden P, Cooper NJ, Abrams K, Claxton K. Modelling the cost-effectiveness of interferon beta and glatiramer acetate in the management of multiple sclerosis. BMJ 2003; **326**: 522-6.

3 Department of Health. Cost effective provision of disease modifying therapies for people with multiple sclerosis. London: Health Service Circular (2002/004) London: Stationery Office; 2002.

4 MS Trust, personal communication 2013.

5 Orme M, Kerrigan J, Tyas D, Russell N, Nixon R. The effect of disease, functional status, and relapses on the utility of people with multiple sclerosis in the UK. Value Health 2007; **10**: 54-60.

6 Boggild M, personal communication 2012.

7 Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. University of York Centre for Health Economics, Discussion Paper 138 1995.

8 Palace J, Duddy M, Bregenzer T et al. Effectiveness and cost-effectiveness of interferon beta and glatiramer acetate in the UK Multiple Sclerosis Risk Sharing Scheme at 6 years: a clinical cohort study with natural history comparator. Lancet Neurology 2015;14:497-505 supplementary appendix

9 Scientific Advisory Group to the UK MS Risk Sharing Scheme. Analysis of the year 4 and year 6 data. Department of Health; 2015.

10 IMS Health. Utilities in Multiple Sclerosis patients (update July 2013): report for the MS Trust. London: IMS Health; 2013.

11 Scientific Advisory Group for the UK MS Risk Sharing Scheme Statistical analysis plan for analysis of the year 10 data. Department of Health; November 2015

12 Tremlett H, Paty D, Devonshire V. Disability progression in multiple sclerosis is slower than previously reported. Neurology 2006; **66**: 172–7.

13 Palace J, Bregenzer T, Tremlett H, et al. UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model. BMJ Open 2014; **4**: e004073.

14 Jackson CH, Sharples LS, Thompson SG, et al. Multistate Markov models for disease progression with classification error. J R Stat Soc 2003; **52**: 193–209.

15 Goldstein H. Multilevel Statistical Models (second edition). London: Edward Arnold; 1995.

16 Tilling K, Lawton M, Robertson N, et al. Modelling disease progression in relapsing remitting onset multiple sclerosis using multilevel models applied to longitudinal data from two natural history cohorts and one treated cohort. Health Technology Assessment 2016;**20:**(81).

17 Lawton M, Tilling K, Robertson N, et al. A longitudinal model for disease progression was developed and applied to multiple sclerosis. Journal of clinical epidemiology. 2015 Nov 30;68(11):1355-65.

18 Stata Corporation. College Station, Texas; 2007.

19 Leckie, G. and Charlton, C. runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. Journal of Statistical Software 2013: **52:** 1-40.

20 Tilling K, Sterne JA, Wolfe CD. Multilevel growth curve models with covariate effects: application to recovery after stroke. Stat Med. 2001; **20(5)**: 685-704.

21 Boggild M, Palace J, Barton P, et al. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. BMJ 2009; 339: b4677.

Figure A1: Distribution of patients over EDSS levels, comparing RSS observed values with values predicted for treated patients using the Markov model [and the "implied" hazard ratio]

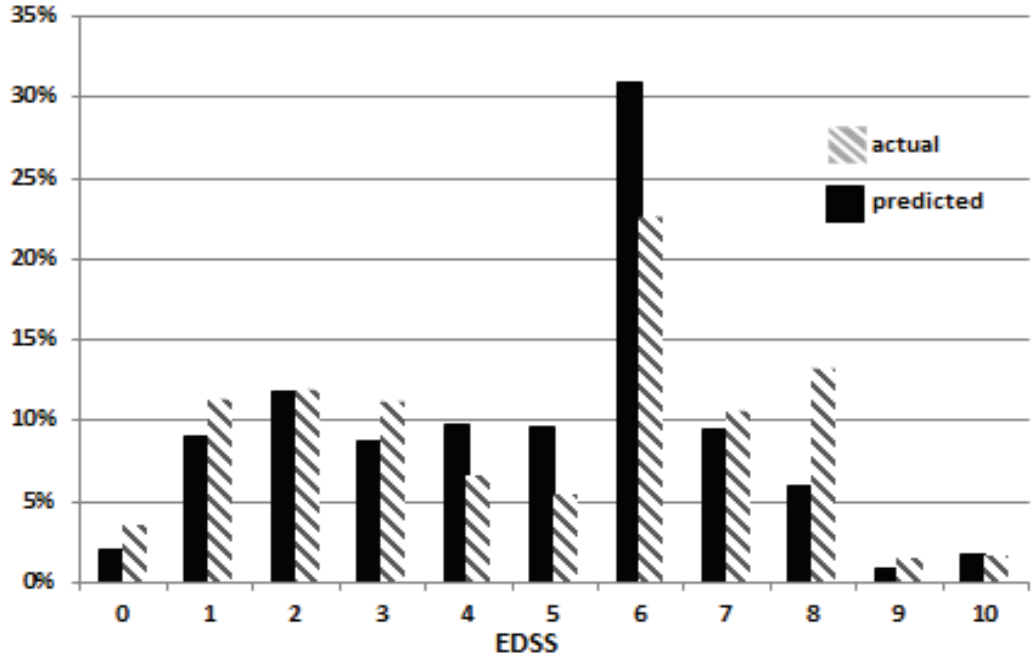


Figure A2: Variation of treatment effect with time - projected EDSS progression applying the multilevel model first to the RSS cohort and then to the untreated comparator control group

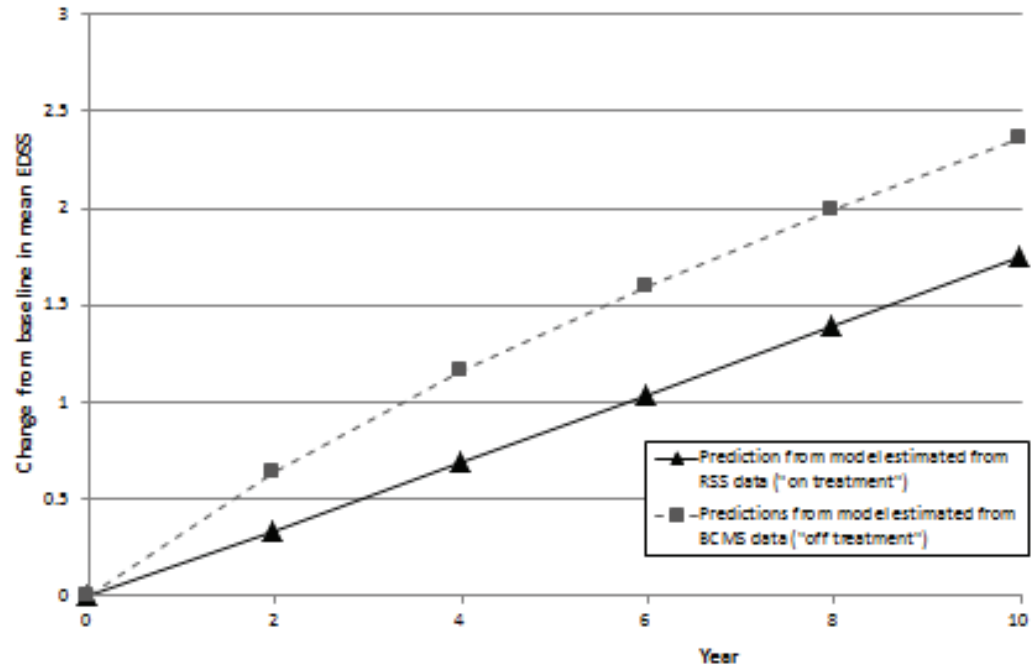
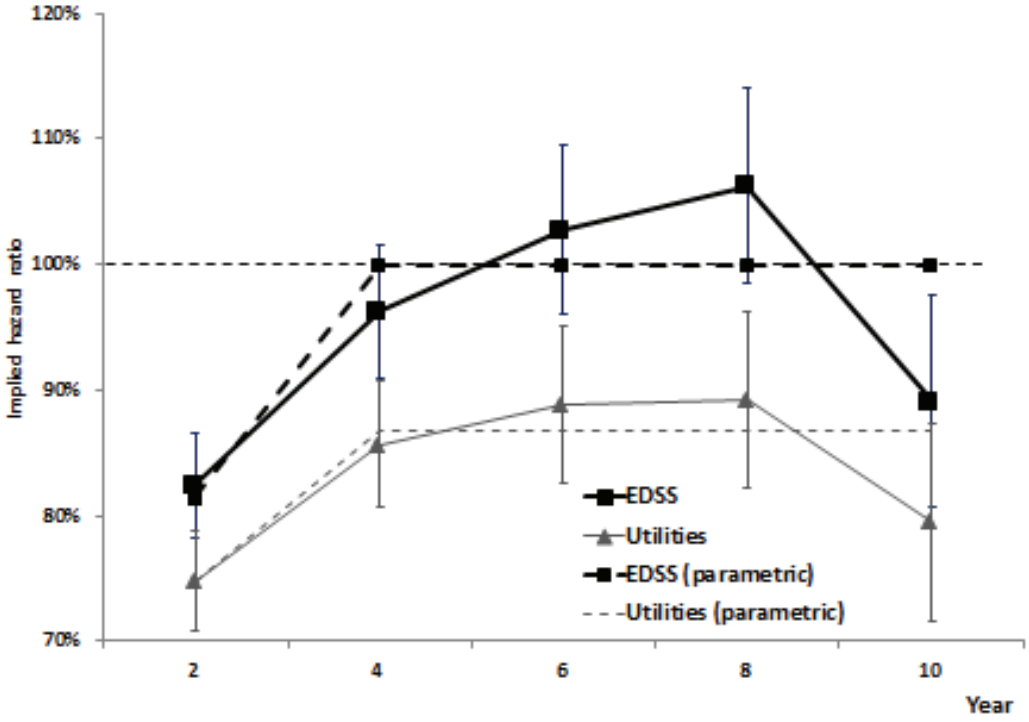


Figure A3: Variation of implied hazard ratios with time using the Markov model, for both utility and EDSS disability outcomes



Journal of Parkinson's Disease 5 (2015) 269–279
DOI 10.3233/JPD-140523
IOS Press

Research Report

Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort

Michael Lawton[a], Fahd Baig[b], Michal Rolinski[b], Claudio Ruffman[b], Kannan Nithi[c], Margaret T. May[a], Yoav Ben-Shlomo[a,1] and Michele T.M. Hu[b,c,1,*]
[a]*School of Social and Community Medicine, University of Bristol, Bristol, UK*
[b]*Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, University of Oxford, Oxford, UK*
[c]*Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK*

Abstract.
Background: Within Parkinson's there is a spectrum of clinical features at presentation which may represent sub-types of the disease. However there is no widely accepted consensus of how best to group patients.
Objective: Use a data-driven approach to unravel any heterogeneity in the Parkinson's phenotype in a well-characterised, population-based incidence cohort.
Methods: 769 consecutive patients, with mean disease duration of 1.3 years, were assessed using a broad range of motor, cognitive and non-motor metrics. Multiple imputation was carried out using the chained equations approach to deal with missing data. We used an exploratory and then a confirmatory factor analysis to determine suitable domains to include within our cluster analysis. K-means cluster analysis of the factor scores and all the variables not loading into a factor was used to determine phenotypic subgroups.
Results: Our factor analysis found three important factors that were characterised by: psychological well-being features; non-tremor motor features, such as posture and rigidity; and cognitive features. Our subsequent five cluster model identified groups characterised by (1) mild motor and non-motor disease (25.4%), (2) poor posture and cognition (23.3%), (3) severe tremor (20.8%), (4) poor psychological well-being, RBD and sleep (18.9%), and (5) severe motor and non-motor disease with poor psychological well-being (11.7%).
Conclusion: Our approach identified several Parkinson's phenotypic sub-groups driven by largely dopaminergic-resistant features (RBD, impaired cognition and posture, poor psychological well-being) that, in addition to dopaminergic-responsive motor features may be important for studying the aetiology, progression, and medication response of early Parkinson's.

Keywords: Parkinson's disease, Cohort studies, cluster analysis, factor analysis

INTRODUCTION

Parkinson's disease (PD) is a common neurodegenerative condition encompassing both motor and non-motor symptoms. Even within pathologically defined patient cohorts, there remains a spectrum of clinical features, treatment response and prognosis [1]. These differences in clinical phenotype may represent different PD subtypes, but there is no widely accepted consensus on the criteria for such groups. Clinically accurate sub-typing may result in improved delineation of aetiological mechanisms, better prognostic counselling, and improved targeting of disease modifying therapies.

[1]Joint Senior Authors.

*Correspondence to: Michele Hu, Department of Clinical Neurology, Level 3, West Wing, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK. Tel.: +44 01865 234337; Fax: +44 01865 234837; E-mail: michele.hu@ndcn.ox.ac.uk.

ISSN 1877-7171/15/$35.00

Attempts to sub-classify PD include "top-down" approaches which depend on an *a priori* assumption, such as the division of subjects by motor phenotype and age of onset [2, 3]. Unfortunately, this approach relies on accurate clinical observation to recognise patterns from all available variables, which is difficult given the breadth of clinical features. Recently, attempts at subtyping have employed data-driven, "bottom-up" approaches, allowing unexpected patterns or discriminating features to be determined [4]. Outcomes of the group characteristics depend heavily on the breadth and depth of the variables inputted into the models. In this study, we have used an approach with key methodological refinements including: 1) restriction to incident patients, to avoid the confounding effects of disease duration, with a far broader range of motor and non-motor assessments than most previously published studies 2) using factor analysis methods to reduce the large number of motor/non-motor variables into a smaller number of clinically important domains describing patient variability 3) using k-means cluster analysis with the inclusion of additional clinical features that may have not been already captured by the factor analysis.

MATERIALS AND METHODS

Patient selection with inclusion/exclusion criteria

PD patients diagnosed within the past 3.5 years were prospectively recruited as part of the Oxford Parkinson's Disease Centre (OPDC) cohort study from 11 hospitals across the Thames Valley covering a population of approximately 2.1 million (PD-Discovery, website: http://opdc.medsci.ox.ac.uk). Full details of this cohort are described elsewhere, [5] with participants being recruited between September 2010 and September 2014.

Patients were eligible for study inclusion if they met the UK Parkinson's Disease Society Brain Bank (UKPDBB) criteria for the diagnosis of idiopathic PD, as judged by a neurologist, with no atypical features to suggest an alternative diagnosis following systematic clinic assessment derived from the NIH PD-DOC study questionnaire (http://grants.nih.gov/grants/guide/rfa-files/RFA-NS-11-001.html). Patients with secondary parkinsonism due to head trauma or medication use, or features of atypical parkinsonism syndromes, such as multiple system atrophy, progressive supra nuclear palsy, corticobasal degeneration, dementia with Lewy bodies, or with significant documented postural BP drop on standardised measurement or significant urinary symptoms were excluded. Each patient was assigned a percentage probability that they met UKPDBB criteria for PD diagnosis by the research neurologist following the study visit. Date of symptom onset was recorded as the date the patient or their carer first became aware of motoric symptoms in relation to their PD, even if occurring on a mild or intermittent basis without initial obvious progression; for example hand tremor, reduced manual dexterity or arm swing. Date of diagnosis was recorded as the date the patient was first given a diagnosis of PD by their hospital specialist (neurologist or geratologist), with the subsequent delay from motoric symptom onset to diagnosis, and delay from date of diagnosis to first (baseline) research clinic visit calculated. Disease duration from motoric symptom onset to date of first (baseline) research clinic visit was also calculated.

Patient evaluation

A full description of the tests and assessments used to assess the Discovery cohort has been published [5, 6]. Assessments were done by the patient completing self-evaluating questionnaires at home and a clinic consultation conducted by a trained neurologist and a nurse. Where patients were taking dopaminergic medications, the assessment was carried out in the clinically-defined on-state. Medication use was recorded allowing the calculation of the levodopa equivalent daily dose (LEDD) [7]. Patient response to antiparkinson therapy was assessed using the physician-rated Clinical Global Impression of Change Scale (CGI-C) [8]. Included in the cluster analysis were: the Movement Disorders Society (MDS) revised Unified Parkinson's Disease Rating Scale (UPDRS part I and part III); 'Sniffin' Sticks 16-item odour identification test; Big Five Inventory – extraversion scale; Epworth Sleepiness Scale; REM Sleep Behaviour Disorder Screening Questionnaire; Leeds Anxiety and Depression Scale (LADS); Becks Depression Inventory (BDI); Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease; Honolulu Asia Aging Study Constipation Questionnaire; Montreal Cognitive Assessment; Phonemic and Semantic verbal fluency; Purdue Peg-board Test; the timed Get Up and Go test; Flamingo test; Orthostatic blood pressure measurement. We explicitly did not include age at onset as this is a demographic variable rather than a feature of PD. Age at onset could influence the phenomenology of PD through two mechanisms: (a) it may confound phenotypic variability due to age-related comorbidity so older patients will have worse motor

function unrelated to their PD and/or (b) it may be a proxy marker for different pathophysiological mechanisms which in turn alter the presenting features of PD. Adjusting for age would be helpful for the former but harmful for the latter as it would reduce the likelihood of identifying different sub-groups. Given the exploratory nature of the analysis, we therefore chose to see how any sub-types related to age in our analyses.

Standard protocol approvals, registrations, and patient consents

The study was undertaken with the understanding and written informed consent of each subject, with the approval of the local NHS ethics committee, and in compliance with national legislation and the Declaration of Helsinki.

Analysis dataset

Analysis was restricted to patients who were diagnosed within the previous 3.5 years and had a high probability of idiopathic PD ($\geq$90% clinician-determined) following careful, structured neurological assessment. Where available, we used the latest follow-up visit to determine the likelihood of PD ($n = 538$, 58.2% seen after 18 and $n = 170$, 18.4% seen after 36 months).

Dealing with missing data

Where questionnaire data were partially completed we used the mean score if 80% or more questions were answered within a questionnaire. We then carried out multiple imputation using the chained equation approach to create 10 imputed datasets.

Determining variables to include within the cluster analysis

Our first step was to carry out an exploratory factor analysis (EFA) within each imputed dataset. We determined the number of important factors, only retaining those with an eigenvalue >1. A promax (oblique) rotation was used and only variables with a loading modulus of $\geq$0.4 were deemed sufficiently important to carry over to the second step.

The second step involved a confirmatory factor analysis (CFA) using the multiply imputed data given the results from the EFA and examining the following goodness of fit statistics: Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA). A model was considered to fit the data well if CFI was $\geq$0.90, TLI $\geq$0.90 and RMSEA $\leq$0.06 [9]. We estimated factor scores for each individual from our CFA within each imputed dataset. At this stage we also considered other clinically important variables for the cluster analysis that were not found to load in any of our factors. Factor scores and other clinically important variables were combined using Rubin's rules [10] to construct a single dataset for carrying out the cluster analysis.

Cluster analysis

We then examined if any variables which did not load on the factor analysis had value in identifying subgroups by standardising them, so that they had equal weighting within the k-means cluster analysis and testing their inclusion in the cluster analysis. Ordered categorical and binary variables were weighted using the rules set out by Hennig et al. [11]. To determine the optimum number of clusters we carried out hierarchical clustering using the Ward algorithm [12] calculating the Calinski/Harabasz pseudo-F index [13] and the Duda/Hart pseudo-T-squared [14]. A higher value of Calinski/Harabasz pseudo-F index and a smaller value of the Duda/Hart pseudo-T squared indicate more distinct clustering. We considered models with between 2 to 5 clusters.

We then carried out k-means cluster analysis using the optimum number of clusters determined from the hierarchical analysis. To ensure convergence to the global maximum, we fitted the model using 500 random starts, and estimated the Calinski/Harabasz pseudo-F index stopping rule [13] to determine the optimal solution.

To test the utility of the sub-group classification, we examined the associations between the clusters with variables not included within the factor/cluster analysis, such as age at onset, time since diagnosis, and time since symptom onset, response to medication using the CGI-C, LEDD and the number of untreated individuals. We also examined the association between cluster membership and classification of PD patients into tremor dominant or postural instability/gait difficulty (PIGD), popularised by Jankovic [2] and updated by Stebbins et al. [15] for the MDS UPDRS.

To further test the reliability of the cluster solutions we applied a cross-validation approach where the data was randomly split into halves five times and the k-means cluster analysis repeated separately on each half. The number of individuals classified into the correct cluster was then determined. Hair et al. suggest that

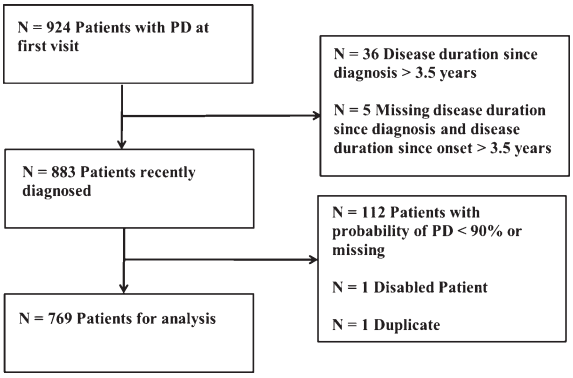


Fig. 1. Flow chart of patient entry into study.

a very stable cluster solution would lead to >90% being correctly classified, a stable cluster solution 80–90% being correctly classified and a somewhat stable cluster solution 75–80% being correctly classified [12].

Computing

STATA version 13 was used to carry out the multiple imputation, Exploratory Factor Analysis (EFA) and the k-means cluster analysis. The Confirmatory Factor Analysis (CFA) and estimation of factor scores was carried out within Mplus.

RESULTS

At the time of this analysis, OPDC had recruited 924 patients (see Fig. 1) but we excluded 154 subjects either because of disease duration (41), they had a prior PD probability of <90% clinically (112), or because of a concomitant neurological disorder leading to significant disability in addition to PD, such that assessment of motor function was invalid (1). One individual was subsequently found to be a duplicate and was dropped from the cluster analysis. This left 769 subjects (age of onset 64.8 years) for the analysis. The baseline characteristics are presented in Table 1. 12.7% of patients were untreated and the mean MDS-UPDRS part III was 26.3. The variables included in our factor analysis had between 0.3%–7.8% missing values.

Exploratory factor analysis (EFA)

We included 34 variables measuring motor and non-motor domains within our EFA. MDS-UPDRS part III was split into four domains (rigidity, bradykinesia, postural, and tremor) to enable better discrimination. We also included the part III question related to speech. Laterality of symptoms was derived from the difference in responses between corresponding questions related to the right- and left-side from the MDS UPDRS part III (see Supplementary Table 1 for more detail on how each variable was derived).

Within each imputed dataset we found four factors to have an eigenvalue greater than 1. The first factor was a mixture of variables measuring non-motor features mostly related to psychological well-being: LADS, BDI, QUIP, BFI neuroticism and apathy, fatigue and pain domains from MDS-UPDRS. The second factor captured motor features either from the MDS UPDRS (rigidity, bradykinesia, postural, speech) or quantified motor performance ('Get up and Go' test, the flamingo test and Purdue pegboard test). The third factor

Table 1
Basic baseline descriptives of patients

Variable	Observed N	Mean (sd; range) or *n* (%)
Female	769	261 (33.9%)
Ethnicity (non-white)	764	11 (1.4%)
Age onset (years)	765	64.77 (9.74; 28.17–87.45)
Disease duration from onset (years)	765	2.92 (1.86; 0.16–13.90)
Disease duration from diagnosis (years)	765	1.32 (0.96; 0.01–3.50)
Delay from first motoric symptom onset to diagnosis (years)	762	1.61 (1.64; 0–13.5)
MDS-UPDRS part I[a]	759	8.62 (5.16; 0–33)
MDS-UPDRS part II[a]	763	8.67 (6.13; 0–35)
MDS-UPDRS part III[a]	768	26.33 (11.00; 5–77)
MDS-UPDRS part IV[a]	767	0.26 (0.97; 0–11)
MDS-UPDRS total (parts I+II+III+IV)[a]	758	43.87 (17.85; 7–123)
MOCA (adjusted for education years)[a]	764	24.98 (3.36; 13–30)
Untreated	766	97 (12.7%)
Levodopa equivalent daily dose (mg)	762	284.38 (212.83; 0–1267.5)
Hoehn and Yahr: median (IQR[b]); mean (range)	768	2 (2-2); 1.84 (1–3)

[a]Changed denominator where 80% or more of questions were answered. [b]Inter-quartile range. Motor assessments (UPDRS and Hoehn and Yahr) were rated in the clinically-defined 'on medication' state for treated PD patients.

Table 2
Confirmatory factor analysis standardised factor loadings of variables selected from exploratory factor analysis

Variable	Factor 1 psychological well-being	Factor 2 Non-tremor motor	Factor 3 cognitive
UPDRS apathy	0.581		
UPDRS fatigue	0.675		
UPDRS pain	0.589		
BFI – neuroticism	0.529		
Leeds anxiety	0.718		
Leeds depression	0.756		
BDI	0.850		
QUIP	0.353		
UPDRS speech		0.452	
UPDRS rigidity		0.429	
UPDRS bradykinesia		0.560	
UPDRS postural		0.721	
Purdue peg board		−0.662	
Purdue assembly task		−0.656	
Get go		0.757	
Flamingo		0.600	
MOCA			0.778
MMSE			0.593
Phenomic fluency			0.622
Semantic fluency			0.727
CFI = 0.786			
TLI = 0.875			
RMSEA = 0.082			

CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, and RMSEA = Root Mean Square Error of Approximation. CFI, TLI and RMSEA are all measures of model fit.

captured cognition (MOCA, MMSE and phenomic and semantic fluency). A fourth factor captured constipation (from the UPDRS part I constipation question and Honolulu Asia Aging Study constipation questionnaire).

Confirmatory factor analysis (CFA)

Our EFA found very consistent results between the imputed datasets. For the CFA we could not estimate a constipation factor since a factor with only two variables is not identifiable. Our CFA of the remaining three factors fell slightly short of our pre-defined goodness of fit criteria with a CFI of 0.79, TLI of 0.88, and a RMSEA of 0.082. It is likely that our poor goodness of fit is due to the large number of variables in the first two factors [16] and our sample size [12]. However since we are only interested in calculating factor scores and not testing the validity of our structural model we kept the CFA as defined. We named factor 1, "psychological well-being", factor 2, "non-tremor motor" and factor 3, "cognitive" (Table 2). Within each factor, variable loadings varied from 0.35 to 0.85, 0.43 to 0.76, and 0.59 to 0.78 respectively.

The factor analysis did not capture a number of clinical features probably since they were not significantly correlated with any of the other variables in the analysis. Hence we decided to include any variable in our cluster analysis that was not loading into one of our factors. We did however exclude the other four BFI variables (since no other previous cluster analysis has looked at these personality traits) and the UPDRS constipation variable (since the other constipation variable was measuring the same trait) for the sake of parsimony.

Hierarchical and K-means cluster analysis

Supplementary Table 2 shows the statistics used to determine the optimum number of clusters from the hierarchical cluster analysis fitted using the Ward algorithm. Different conclusions on the optimum number of clusters would be drawn from different statistics. The Calinski/Harabasz pseudo-F index favoured a two cluster solution, and the Duda/Hart pseudo-T-squared a five cluster solution. This highlights the need for substantial researcher judgement on determining the optimum number of clusters and also the exploratory nature of cluster analysis. Because the two cluster solution appeared to discriminate patients mainly on disease severity with a poor and good group (Supplementary Figure 1), we chose to go forward with the exploratory five cluster solution as more helpful in describing the clinical heterogeneity between patients. Figure 2 shows the means values of each of the standardised variables within each cluster, all variables were coded such that positive indicates worse and negative better than average score. For the laterality variable positive is more bilateral than average and negative more unilateral than average. The groups were ordered in terms of size with the largest as the first. Table 3 shows the association between the clusters and ten variables not included within the factor analysis. There was moderate evidence of a difference in the mean duration of disease between clusters however in absolute terms the difference was negligible and hence we are confident these clusters are not an artefact of disease duration.

Patients in group 1 (25.4%) showed a milder form of PD and they also had a lower than average age at onset, a higher proportion of females, more drug naïve individuals and a lower LEDD. The second group (23.3%) comprised of individuals with worse than average non-tremor motor symptoms, cognitive features, smell, postural hypotension and with bilateral disease. They also had a higher than average age

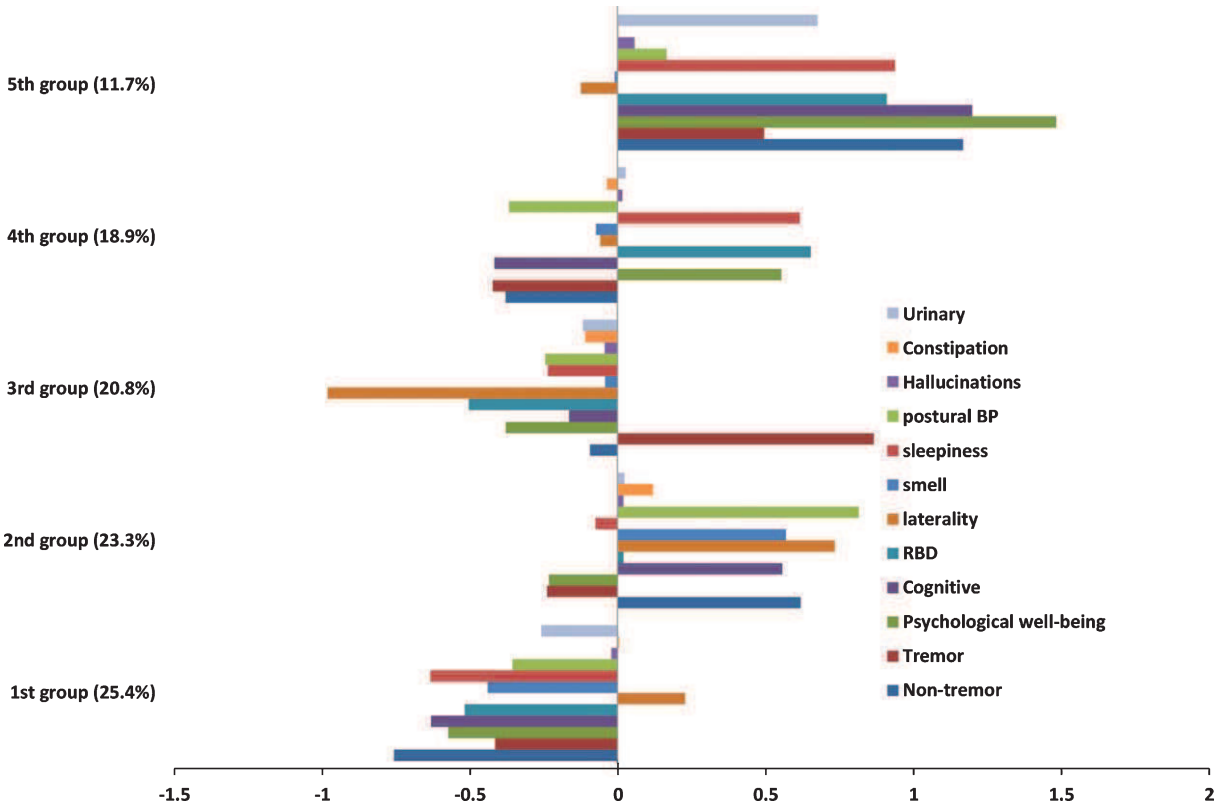


Fig. 2. Within cluster means of the standardised variables for the 5 cluster solution. Positive is worse than average and negative better than average. For laterality positive is more bilateral than average and negative more unilateral than average.

Table 3
Association of clusters with variables not included within the cluster analysis, along with a *p*-value derived from a hypothesis test that the variable is equally distributed (i.e. same mean or same proportion) amongst the five clusters. Note that these variables were derived from the complete case and there was some missingness associated with these variables

Variable (Hypothesis test statistic; *p*-value)	Total ($N=769$)	Cluster 1 ($N=195$, 25.4%)	Cluster 2 ($N=179$, 23.3%)	Cluster 3 ($N=160$, 20.8%)	Cluster 4 ($N=145$, 18.9%)	Cluster 5 ($N=90$, 11.7%)
Female[a] (12.1; $p=0.0166$)	261 (33.9%)	82 (42.1%)	45 (25.1%)	56 (35.0%)	49 (33.8%)	29 (32.2%)
Disease duration from onset[b] (2.5; $p=0.0423$)	2.9 (1.9)	2.7 (1.6)	2.8 (1.7)	3.0 (1.9)	3.1 (2.0)	3.3 (2.2)
Disease duration from diagnosis[b] (3.7; $p=0.0052$)	1.3 (1.0)	1.1 (0.9)	1.4 (1.0)	1.3 (1.0)	1.5 (1.0)	1.4 (0.9)
Age onset[b] (30.0; $p<0.0001$)	64.8 (9.7)	61.8 (8.7)	70.4 (7.7)	64.2 (9.7)	61.0 (9.3)	67.1 (10.8)
Age onset <50[a] (18.6; $p=0.0009$)	60 (7.8%)	20 (10.3%)	3 (1.7%)	16 (10.0%)	18 (12.5%)	3 (3.4%)
UPDRS motor phenotype[a] (104; $p<0.0001$)						
Tremor dominant	407 (53.8%)	118 (61.5%)	65 (37.1%)	132 (83.5%)	58 (40.6%)	34 (38.6%)
Indeterminate	95 (12.6%)	22 (11.5%)	24 (13.7%)	12 (7.6%)	25 (17.5%)	12 (13.6%)
Postural instability gait difficulty	254 (33.6%)	52 (27.1%)	86 (49.1%)	14 (8.9%)	60 (42.0%)	42 (47.7%)
[Clinician global impression of change (CGI-C)][a] (35.4; $p=0.0004$)						
Much or very much improved	354 (48.7%)	83 (46.1%)	96 (55.5%)	59 (39.6%)	77 (55.8%)	39 (44.8%)
Minimally improved	188 (25.9%)	49 (27.2%)	42 (24.3%)	34 (22.8%)	34 (24.6%)	29 (33.3%)
No change to much worse	124 (17.1%)	25 (13.9%)	29 (16.8%)	34 (22.8%)	19 (13.8%)	17 (19.5%)
No medication tried	61 (8.4%)	23 (12.8%)	6 (3.5%)	22 (14.8%)	8 (5.8%)	2 (2.3%)
Drug naíve[a] (21.7; $p=0.0002$)	97 (12.7%)	37 (19.1%)	14 (7.8%)	29 (18.2%)	11 (7.6%)	6 (6.7%)
LEDD total[b] (11.3; $p<0.0001$)	284.4 (212.8)	229.1 (191.8)	314.8 (186.5)	236.4 (214.8)	357.5 (251.1)	310.2 (185.7)
LEDD total on medication[bc] (6.0; $p=0.0001$)	328.8 (194.3)	287.5 (171.3)	341.7 (168.9)	295.9 (200.2)	387.1 (238.2)	336.7 (168.6)

[a]Chi-squared test. [b]Anova. [c]The LEDD restricted to those who are taking dopaminergic medication.

at onset and a lower proportion of females. Within this cluster over 49% were classified as PIGD, which was larger than the average proportion (34%). This cluster had higher than average LEDD and a higher proportion who responded well to therapy. The third group (20.8%) had patients with worse than average tremor scores but who were better than average in most of the other domains and with very unilateral disease. These individuals had similar average age at onset to the overall population and 84% of this cluster was classified as tremor dominant compared to 54% in the entire study population. This cluster had larger proportion of untreated individuals than the study population (18% versus 13%), and a lower than average LEDD. This cluster also had a lower proportion who responded well to PD therapy. The fourth group (18.9%) were marked by poor psychological well-being, RBD, and sleep problems. They also seemed to have better motor function, cognitive and postural hypotension than average with a lower than average age at onset, responded well to medication and were also on a higher than average LEDD. The fifth and smallest group (11.7%) were worse than average on almost all of the domains (except smell), showing a more severe form of PD. This group showed very severe psychological well-being which could be a secondary response to their fast progression or part of the clinical endophenotype. Within this cluster about 48% were classified as PIGD, very few individuals were untreated and they had a higher than average LEDD. The equivalent analysis for the two cluster solution is shown in Supplementary Table 3 for comparison. Supplementary Figure 2 shows the association between the UPDRS phenotype and the 5 cluster solution. It is interesting that although the third cluster is almost completely tremor dominant this cluster only includes about 32% of all the tremor dominant individuals. This highlights the differences in these approaches, that tremor dominant individuals have relatively more tremor problems compared to PIGD problems but do not necessarily have worse than average tremor.

Please note that the binary variable, hallucinations, and the categorical variables, urinary and constipation, have been scaled in a way so that they have equal weighting in the k-means cluster analysis when compared to the more continuous variables. However this does not mean that the distance of the within cluster means from the population average for these three variables can be interpreted in the same way as the other variables. Instead they should be considered relative to the other clusters. For instance the severity of urinary symptoms in the fifth cluster is worse that the severity in the fourth cluster. However we cannot be certain whether the severity of urinary symptoms in the fifth cluster is any less than the severity of sleepiness problems in the fifth cluster even though the within cluster mean of sleepiness is further from the population average.

Supplementary Table 4 shows the stability of the five cluster solution using our cross-validation approach. On average 73.8% of individuals were correctly classified which is close to the borders of a somewhat stable solution according to the Hair et al. criteria. The stability across the five split datasets was not consistent ranging from stable (83.7% correctly classified) to an unstable solution (64.5% correctly classified) so without an external validation it is difficult to determine the stability of our five cluster model. However the stability of the two cluster solution in the cross-validation is much better with on average 95.9% of individuals being correctly classified (Supplementary Table 5), a very stable solution which was consistent across the five split datasets. The apparent stability of the 2 cluster solution compared to the 5 cluster solution is not particularly surprising since there are 1 compared to 4 ways of incorrectly classifying an individual in the two and five cluster solutions respectively.

DISCUSSION

Our analyses suggest that there may be five subgroups of patients with recently diagnosed PD: 1) mild motor and non-motor disease, 2) poor posture, gait, cognition, smell and postural hypotension, 3) severe tremor, 4) poor psychological well-being, RBD and sleep, and 5) severe motor, non-motor and cognitive disease, with poor psychological well-being. Our initial approach used a bottom-up data-driven approach to group together individuals with similar symptoms with little a priori assumptions. However, one limitation of using a purely data-driven approach is that the choice of variables and their breadth will partially determine what factors are identified i.e. a badly measured domain, even though clinically important, will appear statistically less informative than another for which several scales have been included. We therefore felt it important to supplement the factor analysis with a second stage approach where we added nine other domains that had not emerged from the factor analysis. This combined approach has advantages over simply using a priori assumptions about the importance of UPDRS tremor and non-tremor sub-items in defining such phenotypes [2].

Depending on which statistical approach we chose, we could have decided that the two cluster solution was more appropriate. However this solution only seemed to group people essentially as good or bad across a range of disease severity measures (motor, non-motor, psychological well-being and cognition- see Supplementary Figure 1). The five cluster solution allowed for disease severity measures to be preferentially affected, for example group 3) above who score poorly on tremor motor measures with relative sparing of non-tremor motor, cognitive and psychological well-being measures. This may be a more valid representation of the PD disease spectrum encountered in routine clinical practice. Interestingly, the five cluster solution might be more clinically relevant as the two cluster solution found no evidence of an association with drug responsiveness, ($p = 0.13$, see Supplementary Table 3) whilst the five cluster solution did find strong evidence of an association between cluster grouping and levodopa response ($p = 0.0004$).

One caveat for this study is that levodopa responsiveness was assessed using the clinician-rated Clinical Global Impression of Change Scale (CGI-C). This retrospective questionnaire involves the clinician asking the patient (and carer) about their overall impression of motor response to previously trialled dopaminergic medications. The CGI-C therefore, is likely to be a less accurate measure of true levodopa response when compared to formal levodopa challenge testing for example. While we have performed this in a patient subgroup, unfortunately overall numbers are small due to practical purposes, and insufficient to extend to the more general cluster model being presented here. Caveats aside, it is interesting to note that cluster 2 (which resembles PIGD with impaired postural and cognitive function) has a good CGI-C medication response similar to cluster 4, despite these subjects being older. One of the difficulties in interpreting these results is that this group also have the fewest number of drug naïve patients, possibly because their parkinsonian motor features are more severe and disabling than those with tremor-dominant disease. Hence this may be biasing the proportions in the CGI-C results if we assume that drug naïve would show excellent response had they been treated. In addition this may also reflect a "ceiling effect" whereby milder tremor-dominant patients despite showing drug responsiveness can only improve to a more moderate degree than those with more severe disease. Lastly, cluster 2 has a higher LEDD so may have had the opportunity to demonstrate a bigger drug response compared to other clusters. The higher LEDD may also reflect the more severe motoric symptoms (bradykinesia, rigidity, and gait imbalance) experienced by this group, which are likely to be stronger determinants of disability than tremor symptoms, hence driving up the increased overall treatment doses. Future work will focus on comparing the accuracy of GCI-C versus formal levodopa challenge in assessing medication response and predicting progression in early PD.

At least two phenotypes defined in the current study, namely 4) poor psychological well-being, RBD and sleep and 5) severe motor, non-motor and cognitive disease with poor psychological well-being, would have been missed using conventional "top-down" PD classification models. These findings are novel and potentially of high clinical relevance, as they underline the importance of early non-motor symptoms such as RBD, anxiety, depression, apathy, pain and fatigue in underpinning the disease heterogeneity seen in early PD. To date, few studies using data-driven techniques have assessed the baseline importance of non-motor symptoms in such a large well-characterised incident PD cohort. This is particularly relevant given the increasingly acknowledged importance of non-motor symptoms over and above motor symptoms in determining patient-related quality of life and subsequent decline [17]. Symptoms such as RBD, which can manifest prior to the onset of motoric symptoms, are characterised pathologically by involvement of the locus coeruleus, subcoeruleus, pedunculopontine and serotonergic raphe nuclei [18].

The co-existence in group 5 of significantly worse scores in both motor and non-motor domains highlights the importance of the latter, and in particular psychological well-being which may or may not be secondary to a worse clinical evolution. This group had poor scores across both subjective and objective evaluations of motor, cognitive and other non-motor domains, thus excluding the possibility that they simply reflect a poor perception of personal well-being, which is a common occurrence in mood disorders such as anxiety and depression. Although every effort was made to exclude atypical Parkinsonism from our analysis, it is of course possible that a proportion of subjects in group 5 do not have PD but rather an atypical parkinsonian disorder such as multiple system atrophy or progressive supranuclear palsy. As we follow-up our subjects over the next 10 years, we will be able to determine if atypical features emerge in this subgroup and ultimately post-mortem pathological diagnosis should help clarify if they have a more rapidly progressive form of PD or atypical Parkinsonism.

Patients in group 4 scored particularly poorly on RBD measures, however function on motor and cognitive testing was good compared to other groups. This, together with the fact that RBD is a prodromal feature, which may have a 15-year latency before the emergence of motor symptoms [19] might suggest that RBD is a risk but not a prognostic marker for subsequent PD. Previous longitudinal studies have also shown that concomitant RBD was not associated with greater worsening of motor disability scores, cognition or depression, in patients with PD [20–22].

It is uncertain whether the phenotypic features of group 2, who are on average much older, merely reflect age-related co-morbidities such as poor posture, cognition, smell, postural hypotension, or are a distinct aetiological sub-group.

A recent study using principle component analysis [23] in a prevalent PD cohort found that a composite score of predominantly nondopaminergic (PND) features which are largely insensitive to dopaminergic medication (postural instability, gait difficulty, cognitive impairment, depressive symptoms, psychotic symptoms, excessive daytime somnolence and autonomic dysfunction) might provide a more accurate evaluation of disease severity and progression in PD. Our results support this finding and raise the important issue of how best to select patients for future disease-modifying or neuroprotective trials in PD.

If only certain sub-groups respond to a neuroprotective agent, existing trials are more likely to result in a false negative result and future trials will need far larger sample sizes to group according to baseline phenotype with concomitant cost implications. It is unclear as to whether the 12% of PD patients assigned to group 5, who appear severe across a range of motor and non-motor measures, should be selected as being the most likely to benefit from future disease modifying interventions, or might be least likely to benefit due to more advanced pathophysiology. Phenotypic differences seen across PD might therefore be a major contributing factor to the current lack of a convincing neuroprotective agent for this disease, despite multiple drug trials in this field.

Our results are consistent with several studies applying cluster methodology in PD [24–33] that found a milder disease group with a young age at onset, [25, 26, 28–33] a group with severe gait dysfunction and cognitive impairment [24, 27] and a tremor dominant group [28, 29, 31]. Most studies have found a rapid disease progression group with an older age at onset [25–32]. The relationship between cognitive function and impairments in gait, posture and non-tremor motor features in PD has been well documented in previous studies [34]. It seems that our finding of a group which has poor RBD, psychological well-being and sleepiness has not been found in previous studies, possibly because information about these features have not always been collected.

Although we argue against including age at onset in the cluster analysis we explored using it in our factor and cluster analysis as some of the previous studies have done [25, 27–29, 31, 32]. Including age at onset in the factor analysis would mean it loading on the non-tremor motor factor very weakly and hence would have made little difference to our estimated factor scores. If it was included at the cluster analysis stage we would have qualitatively found five groups of similar phenomenology.

These sub-groups have been derived from baseline visits, hence are not confounded by disease duration. Future evaluation will determine whether patients retain their initial clinical phenotype or whether changing from one clinical phenotype to another is an important marker of subsequent progression. We are currently undertaking an independent replication in collaboration with a second UK cohort (Tracking Parkinson's Disease) which used a very similar methodology (90% of variables are the same).

We cannot yet draw firm conclusions as to the prognostic value of these clusters, but with further follow-up we will determine whether this classification is of greater value than existing approaches which discriminate patients based on simpler baseline measures. We will also test whether these clusters have biological or clinical utility by comparing data on genotypes, biomarkers, including neuroimaging as well as responsiveness to drug therapy, and the onset of clinically-meaningful end points, such as motor fluctuations, dyskinesias, dementia, dependency or institutionalisation and long-term mortality.

ACKNOWLEDGMENTS

The authors would like to thank all subjects who have participated in this study. This study was funded by the Monument Trust Discovery Award from Parkinson's UK and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford, and the NIHR Clinical Research Network: Thames Valley and South Midlands. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

ETHICS APPROVAL

Local NHS ethics committee.

SUPPLEMENTARY MATERIAL

The supplementary tables and figures are available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-140523.

REFERENCES

[1] Selikhova M, Williams DR, Kempster PA, Holton JL, Revesz T, & Lees AJ (2009) A clinico-pathological study of subtypes in Parkinson's disease. *Brain*, **132**, 2947-2957.

[2] Jankovic J, McDermott M, Carter J, Gauthier S, Goetz C, Golbe L, Huber S, Koller W, Olanow C, Shoulson I, et al.(1990) Variable expression of Parkinson's disease: A base-line analysis of the DATATOP cohort. The Parkinson Study Group. *Neurology* **40**, 1529-1534.

[3] Wickremaratchi MM, Knipe MDW, Sastry B, Morgan E, Jones A, Salmon R, Weiser R, Moran M, Davies D, & Ebenezer L (2011) The motor phenotype of Parkinson's disease in relation to age at onset. *Mov Disord*, **26**, 457-463.

[4] van Rooden SM, Heiser WJ, Kok JN, Verbaan D, van Hilten JJ, & Marinus J (2010) The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Mov Disord*, **25**, 969-978.

[5] Szewczyk-Krolikowski K, Tomlinson P, Nithi K, Wade-Martins R, Talbot K, Ben-Shlomo Y, & Hu M (2013) The influence of age and gender on motor and non-motor features of early Parkinson's disease: Initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. *Parkinsonism Relat Disord*, **20**, 99-105.

[6] Rolinski M, Szewczyk-Krolikowski K, Tomlinson PR, Nithi K, Talbot K, Ben-Shlomo Y, & Hu MT (2014) REM sleep behaviour disorder is associated with worse quality of life and other non-motor features in early Parkinson's disease. *J Neurol Neurosurg Psychiatry*, **85**, 560-566.

[7] Tomlinson CL, Stowe R, Patel S, Rick C, Gray R, & Clarke CE (2010) Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov Disord*, **25**, 2649-2653.

[8] Schneider LS, Olin JT, Doody RS, Clark CM, Morris JC, Reisberg B, Schmitt FA, Grundman M, Thomas RG, & Ferris SH (1997) Validity and reliability of the Alzheimer's Disease cooperative study - Clinical global impression of change. *Alzheimer Dis Assoc Disord*, **11**, S22-S32.

[9] Tully PJ, Winefield HR, Baker RA, Turnbull DA, & de Jonge P (2011) Confirmatory factor analysis of the Beck Depression Inventory-II and the association with cardiac morbidity and mortality after coronary revascularization. *J Health Psychol*, **16**, 584-595.

[10] Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581-590.

[11] Hennig C, & Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *J R Stat Soc Ser C Appl Stat*, **62**, 309-369.

[12] Hair JF (2010) *Multivariate data analysis*, Prentice Hall, Upper Saddle River, NJ.

[13] Caliński T, & Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat Theory Methods*, **3**, 1-27.

[14] Duda RO, Hart PE, & Stork DG (2001) *Pattern Classification*, Wiley, New York.

[15] Stebbins GT, Goetz CG, Burn DJ, Jankovic J, Khoo TK, & Tilley BC (2013) How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale. *Mov Disord*, **28**, 668-670.

[16] Kenny DA, & McCoach DB (2003) Effect of the number of variables on measures of fit in structural equation modeling. *Struct Equ Modeling*, **10**, 333-351.

[17] Martinez-Martin P, Rodriguez-Blazquez C, Kurtis MM, & Chaudhuri K (2011) The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease. *Mov Disord*, **26**, 399-406.

[18] Braak H, Tredici KD, Rüb U, de Vos RA, Jansen Steur EN, & Braak E (2003) Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol Aging*, **24**, 197-211.

[19] Schenck CH, Boeve BF, & Mahowald MW (2013) Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: A 16-year update on a previously reported series. *Sleep Med*, **14**, 744-748.

[20] Gjerstad MD, Boeve B, Wentzel-Larsen T, Aarsland D, & Larsen JP (2008) Occurrence and clinical correlates of REM sleep behaviour disorder in patients with Parkinson's disease over time. *J Neurol Neurosurg Psychiatry*, **79**, 387-391.

[21] Lavault S, Leu-Semenescu S, Tezenas du Montcel S, Cochen de Cock V, Vidailhet M, & Arnulf I (2010) Does clinical rapid eye movement behavior disorder predict worse outcomes in Parkinson's disease? *J Neurol*, **257**, 1154-1159.

[22] Yoritaka A, Ohizumi H, Tanaka S, & Hattori N (2009) Parkinson's disease with and without REM sleep behaviour disorder: Are there any clinical differences? *Eur Neurol*, **61**, 164-170.

[23] van der Heeden JF, Marinus J, Martinez-Martin P, & van Hilten JJ (2014) Importance of nondopaminergic features in evaluating disease severity of Parkinson disease. *Neurology*, **82**, 412-418.

[24] Dujardin K, Defebvre L, Duhamel A, Lecouffe P, Rogelet P, Steinling M, & Destee A (2004) Cognitive and SPECT characteristics predict progression of Parkinson's disease in newly diagnosed patients. *J Neurol*, **251**, 1383-1392.

[25] Erro R, Vitale C, Amboni M, Picillo M, Moccia M, Longo K, Santangelo G, De Rosa A, Allocca R, Giordano F, Orefice G, De Michele G, Santoro L, Pellecchia MT, & Barone P (2013) The heterogeneity of early Parkinson's disease: A cluster analysis on newly diagnosed untreated patients. *PLoS One*, **8**, e70244.

[26] Gasparoli E, Delibori D, Polesello G, Santelli L, Ermani M, Battistin L, & Bracco F (2002) Clinical predictors in Parkinson's disease. *Neurol Sci*, **23**(Suppl 2), S77-S78.

[27] Graham JM, & Sagar HJ (1999) A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: Identification of three distinct subtypes. *Mov Disord*, **14**, 10-20.

[28] Lewis SJ, Foltynie T, Blackwell AD, Robbins TW, Owen AM, & Barker RA (2005) Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *J Neurol Neurosurg Psychiatry*, **76**, 343-348.

[29] Liu P, Feng T, Wang YJ, Zhang X, & Chen B (2011) Clinical heterogeneity in patients with early-stage Parkinson's disease: A cluster analysis. *J Zhejiang Univ Sci B*, **12**, 694-703.

[30] Post B, Speelman JD, de Haan RJ, & group CA-s (2008) Clinical heterogeneity in newly diagnosed Parkinson's disease. *J Neurol*, **255**, 716-722.

[31] Reijnders JSAM, Ehrt U, Lousberg R, Aarsland D, & Leentjens AFG (2009) The association between motor subtypes and psychopathology in Parkinson's disease. *Parkinsonism Relat Disord*, **15**, 379-382.

[32] Schrag A, Quinn NP, & Ben-Shlomo Y (2006) Heterogeneity of Parkinson's disease. *J Neurol Neurosurg Psychiatry*, **77**, 275-276.

[33] van Rooden SM, Colas F, Martinez-Martin P, Visser M, Verbaan D, Marinus J, Chaudhuri RK, Kok JN, & van Hilten JJ (2011) Clinical subtypes of Parkinson's disease. *Mov Disord* **26**, 51-58.

[34] Williams-Gray CH, Foltynie T, Brayne CE, Robbins TW, & Barker RA (2007) Evolution of cognitive dysfunction in an incident Parkinson's disease cohort. *Brain*, **130**, 1787-1798.

Supplementary Tables

Supplementary Table 1. Detailed description of the variables included in the analysis and how they were derived and analysed.

List of variables named	Explanation
UPDRS_apathy	Categorical from UPDRS part I (question 1.5)
UPDRS_hallucinations	Binary from UPDRS part I (question 1.2)
UPDRS_speech	Categorical UPDRS part III question 3.1
UPDRS_rigidity	Continuous rigidity questions from UPDRS part III (mean score 3.3)
UPDRS_bradykinesia	Continuous bradykinesia questions from UPDRS part III (mean score 3.2, 3.4, 3.5, 3.6, 3.7, 3.8, 3.14)
UPDRS_postural	Continuous postural questions from UPDRS part III (mean score 3.9, 3.10, 3.11, 3.12, 3.13)
UPDRS_tremor	Continuous tremor questions from UPDRS part III (mean score 3.15, 3.16, 3.17, 3.18)
UPDRS_laterality	Continuous. Absolute value of right-left questions from UPDRS part III
UPDRS_fatigue	Categorical from UPDRS part I (question 1.13)
UPDRS_pain	Categorical from UPDRS part I (question 1.9)
UPDRS_constipation	Categorical from UPDRS part I (question 1.11)
UPDRS_urinary	Categorical from UPDRS part I (question 1.10)
Sniffin	Continuous
MOCA	Continuous – adjusted for education years
MMSE	Continuous
Phenomic fluency	Continuous – age adjusted
Semantic fluency	Continuous – age adjusted
Purdue total	Continuous – Sum from the pegboard test using left hand, right hand and both hands
Purdue assembly	Continuous – Total from the assembly part of the purdue pegboard test.
Get go	Binary - Dichotomised into top quintile (average time)
Flamingo	Binary (dichotomised into bottom quintile)
BFI_extraversion	Continuous Big Five inventory total
BFI_agreeableness	Continuous Big Five inventory total
BFI_conscientiousness	Continuous Big Five inventory total
BFI_neuroticism	Continuous Big Five inventory total
BFI_openess	Continuous Big Five inventory total
ESS	Continuous Epworth sleepiness scale total

RBD	Continuous
Honolulu_Constipation	Categorical- presence of constipation
Leeds_anxiety	Continuous
Leeds_depression	Continuous
BDI	Continuous – Becks Depression Inventory
QUIP_all	Binary – presence of any one of the QUIP scores
Systolic_BP postural_drop	Continuous - (Mean of lying SBPs) – standing SBP

Supplementary Table 2. Statistics to determine the number of clusters from the Ward hierarchical clustering. A higher value of Calinski/Harabasz pseudo-F index indicates more distinct clustering and a smaller value of the Duda/Hart pseudo-T squared indicates more distinct clustering. Bold indicates most distinct cluster.

Number of clusters	Calinski/Harabasz pseudo-F	Duda/Hart pseudo T-squared
2	**98.37**	37.46
3	73.48	41.49
4	63.46	37.54
5	57.07	**25.17**

Supplementary Table 3. Association of clusters with variables not included within the cluster analysis, along with a p-value derived from a hypothesis test that the variable is equally distributed (i.e. same mean or same proportion) amongst the two clusters. Note that these variables were derived from the complete case and there was some missingness associated with them

Variable (Hypothesis test statistic; p-value)	**Total (N=769)**	**Cluster 1 (N=436, 56.7%)**	**Cluster 2 (N=333, 43.3%)**
Female [a] (10.4; p=0.0012)	261 (33.9%)	169 (38.8%)	92 (27.6%)
Disease duration onset[b] (1.9; p=0.1652)	2.9 (1.9)	2.8 (1.7)	3.0 (2.0)
Disease duration diagnosis[b] (8.5; p=0.0037)	1.3 (1.0)	1.2 (0.9)	1.4 (1.0)
Age onset[b] (85.4; p<0.0001)	64.8 (9.7)	62.1 (9.0)	68.3 (9.5)
Age onset <50 [a] (14.2; p=0.0002)	60 (7.8%)	48 (11.0%)	12 (3.6%)
UPDRS phenotype [a] (47.7; p<0.0001)			
Tremor dominant	407 (53.8%)	278 (64.7%)	129 (39.6%)
Indeterminate	95 (12.6%)	45 (10.5%)	50 (15.3%)
Postural instability gait difficulty	254 (33.6%)	107 (24.9%)	147 (45.1%)
Clinicians global impression of change [a] (5.7; p=0.1269)			
Much or very much improved	354 (48.7%)	194 (47.7%)	160 (50.0%)
Minimally improved	188 (25.9%)	102 (25.1%)	86 (26.9%)

No change to much worse	124 (17.1%)	68 (16.7%)	56 (17.5%)
No medication tried	61 (8.4%)	43 (10.6%)	18 (5.6%)
Drug naïve [a] (8.2; p=0.0042)	97 (12.7%)	68 (15.7%)	29 (8.7%)
LEDD total [b] (23.3; p<0.0001)	284.4 (212.8)	252.2 (209.0)	326.2 (210.8)
LEDD total on medication[bc] (14.4; p=0.0002)	328.8 (194.3)	302.8 (192.6)	359.9 (192.0)

[a]Chi-squared test

[b] Anova

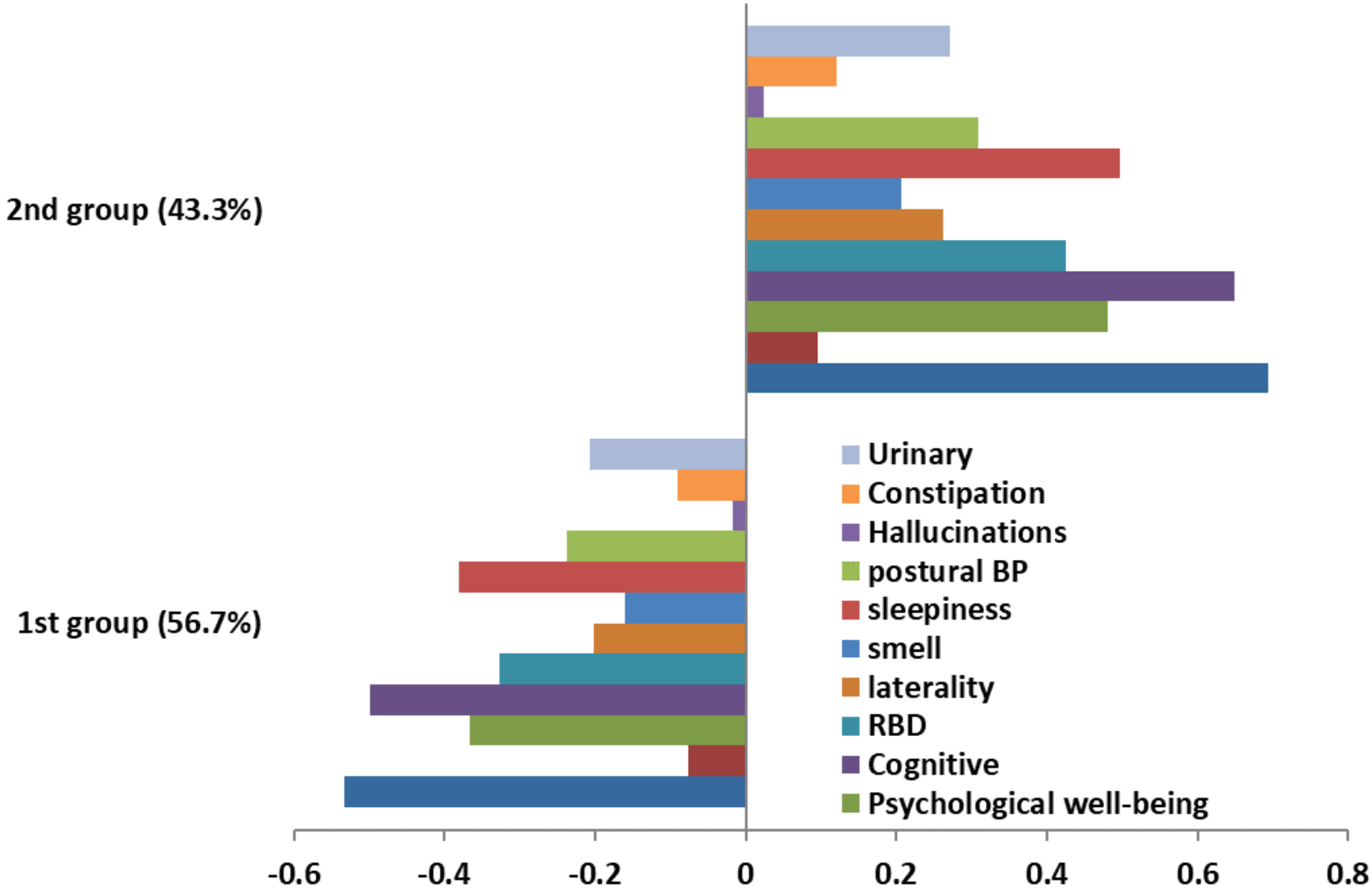
[c]The LEDD restricted to those who are taking dopaminergic medication

Supplementary Table 4. Stability of the five cluster solution using our cross-validation approach

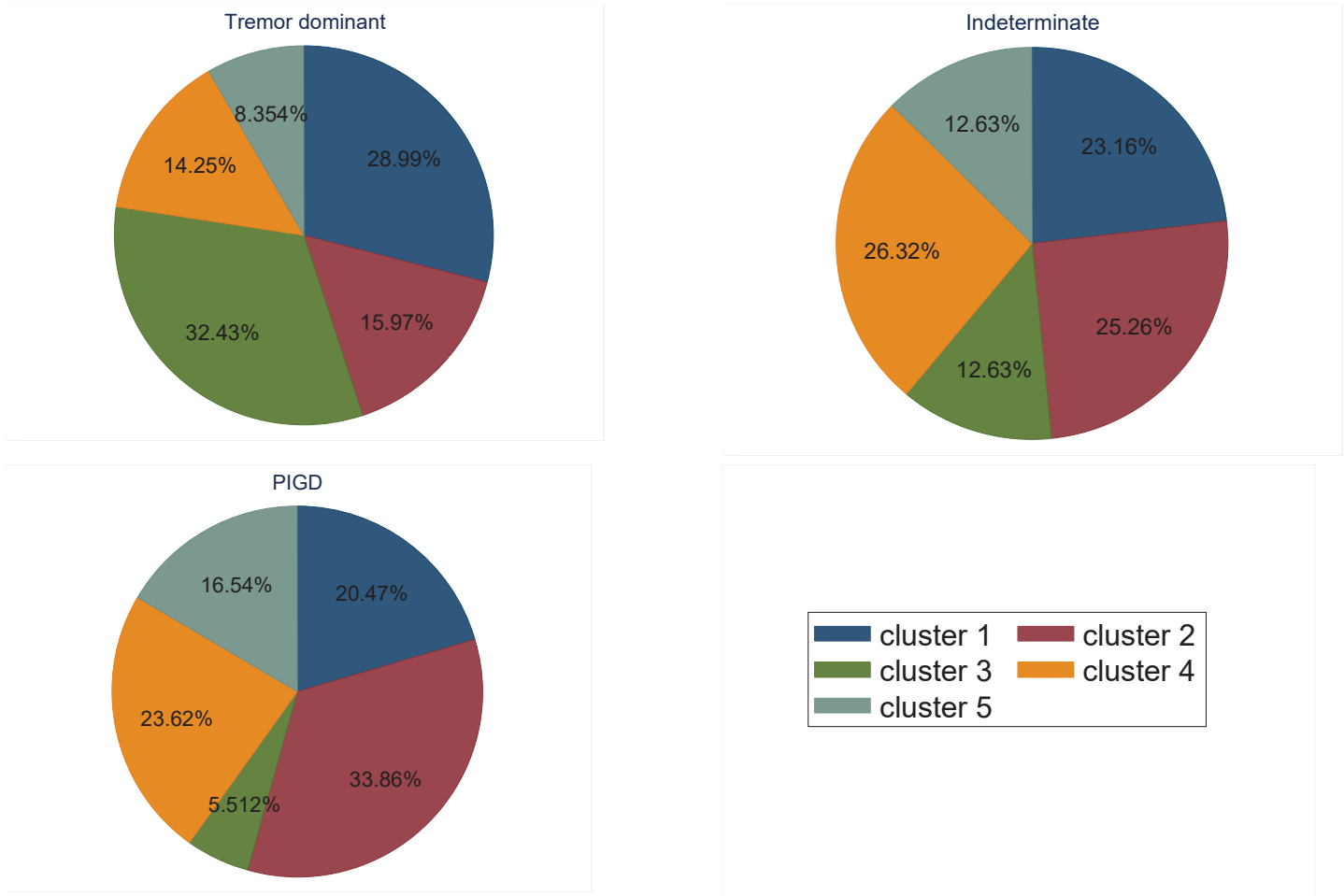
Split dataset	Number assigned to the same cluster (%)
1	644 (83.7%)
2	556 (72.3%)
3	510 (66.3%)
4	496 (64.5%)
5	631 (82.1%)
Average	567.4 (73.8%)

Supplementary Table 5. Stability of the two cluster solution using our cross-validation approach

Split dataset	Number assigned to the same cluster (%)
1	742 (96.5%)
2	713 (92.7%)
3	758 (98.6%)
4	754 (98.0%)
5	744 (96.7%)
Average	742.2 (96.5%)



Web Figure 1. Within cluster means of the standardised variables for the 2 cluster solution. Positive is worse than average. For laterality positive is more bilateral than average and negative more unilateral than average.



Web figure 2. Pie charts to visually show the association between the UPDRS phenotype (popularised by Jankovic et. al) and the 5 cluster solution.

Parkinsonism and Related Disorders 33 (2016) 96–101

Contents lists available at ScienceDirect

Parkinsonism and Related Disorders

journal homepage: www.elsevier.com/locate/parkreldis

Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease

Michael Lawton [a, *], Michele T.M. Hu [b, c], Fahd Baig [b, c], Claudio Ruffmann [b, c], Eilidh Barron [d], Diane M.A. Swallow [d], Naveed Malek [d], Katherine A. Grosset [d], Nin Bajaj [e], Roger A. Barker [f], Nigel Williams [g], David J. Burn [h], Thomas Foltynie [i], Huw R. Morris [j], Nicholas W. Wood [k], Margaret T. May [a], Donald G. Grosset [d], Yoav Ben-Shlomo [a]

[a] School of Social and Community Medicine, University of Bristol, United Kingdom
[b] Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, University of Oxford, United Kingdom
[c] Oxford Parkinson's Disease Centre, University of Oxford, United Kingdom
[d] Department of Neurology, Institute of Neurological Sciences, Queen Elizabeth University Hospital, Glasgow, United Kingdom
[e] Department of Neurology, Queen's Medical Centre, Nottingham, United Kingdom
[f] Clinical Neurosciences, John van Geest Centre for Brain Repair, Cambridge, United Kingdom
[g] Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, United Kingdom
[h] Institute of Neuroscience, University of Newcastle, United Kingdom
[i] Sobell Department of Motor Neuroscience, UCL Institute of Neurology, United Kingdom
[j] Department of Clinical Neuroscience, UCL Institute of Neurology, United Kingdom
[k] Department of Molecular Neuroscience, UCL Institute of Neurology, United Kingdom

ARTICLE INFO

Article history:
Received 13 April 2016
Received in revised form
8 September 2016
Accepted 23 September 2016

Keywords:
Olfaction
Sniffin' Sticks
University of Pennsylvania Smell Identification Test
Equating
Item Response Theory

ABSTRACT

Background: Impaired olfaction is an important feature in Parkinson's disease (PD) and other neurological diseases. A variety of smell identification tests exist such as "Sniffin' Sticks" and the University of Pennsylvania Smell Identification Test (UPSIT). An important part of research is being able to replicate findings or combining studies in a meta-analysis. This is difficult if olfaction has been measured using different metrics. We present conversion methods between the: UPSIT, Sniffin' 16, and Brief-SIT (B-SIT); and Sniffin' 12 and Sniffin' 16 odour identification tests.

Methods: We used two incident cohorts of patients with PD who were tested with either the Sniffin' 16 (n = 1131) or UPSIT (n = 980) and a validation dataset of 128 individuals who took both tests. We used the equipercentile and Item Response Theory (IRT) methods to equate the olfaction scales.

Results: The equipercentile conversion suggested some bias between UPSIT and Sniffin' 16 tests across the two groups. The IRT method shows very good characteristics between the true and converted Sniffin' 16 (delta mean = 0.14, median = 0) based on UPSIT. The equipercentile conversion between the Sniffin' 12 and 16 item worked well (delta mean = 0.01, median = 0). The UPSIT to B-SIT conversion showed evidence of bias but amongst PD cases worked well (mean delta = −0.08, median = 0).

Conclusion: We have demonstrated that one can convert UPSIT to B-SIT or Sniffin' 16, and Sniffin' 12 to 16 scores in a valid way. This can facilitate direct comparison between tests aiding future collaborative analyses and evidence synthesis.

1. Introduction

Impaired olfaction is an important non-motor feature of Parkinson's disease (PD). It is thought to be an early pre-clinical sign of PD [1] and can be used to help in the diagnosis of PD before the development of definite motor features [2,3]. Olfactory impairment may also be an early marker of other neurological diseases such as Alzheimer's disease [4], multiple sclerosis [5], idiopathic rapid eye movement sleep behaviour disorder [6], Huntington's disease [7], multiple system atrophy [8], progressive supranuclear palsy [9] and parkinsonism dementia complex seen in

* Corresponding author. Office G.04, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, United Kingdom.
E-mail address: Michael.Lawton@bristol.ac.uk (M. Lawton).

http://dx.doi.org/10.1016/j.parkreldis.2016.09.023

Guam [10]. Differences in olfactory dysfunction between neurological diseases may be helpful in the differential diagnosis [11] of parkinsonian disorders [12]. Detailed reviews of olfactory dysfunction in neurological disorders have been previously published [11,13].

Many research studies collect data on olfaction and an important aspect of high quality research is the ability to replicate findings from studies or undertaking systematic reviews with or without a meta-analysis to synthesise evidence and examine for heterogeneity. This is more difficult if olfaction has been measured using a different metric within the different studies leading to potentially artefactual differences. The ability to estimate scores on one test from scores on another test helps reduce this problem. Olfaction is often measured using smell identification tests such as Sniffin' Sticks [14] or the University of Pennsylvania Smell Identification Test (UPSIT) [15].

Both the Sniffin' [16] and UPSIT [17] tests have published normative data centiles stratified by age and gender allowing us to determine the olfactory changes that are likely to be caused by disease in addition to that due to the natural aging process. This is particularly important in PD which predominantly affect the older population. Whilst the published normative data for Sniffin' stratified age as 5–15; 16–35; 36–55; and >55, the UPSIT stratified using five year age bands up to 85 and above. The stratification method employed by UPSIT is arguably more sensible given that olfactory impairment rises dramatically between 65 and 80 years [18].

We aimed to create conversion tables from an UPSIT score to a standard Sniffin' 16 item odour identification score, between the Sniffin' 12 and 16 item odour identification versions and between the UPSIT and Brief Smell Identification test (B-SIT) using two large cohorts of individuals with PD to help researchers pool data in future collaborative studies. An additional useful by-product of our conversion is that we can convert the published age/gender stratified centiles for the UPSIT to equivalent Sniffin' scores.

2. Methods

2.1. Study populations

Data were available from two incidence cohorts of patients with PD. The Oxford Parkinson's Disease Centre Discovery cohort consists of individuals from 11 hospitals across the Thames Valley. Patients were recruited between study onset in September 2010 up to May 2015. Full details of this study are described in detail elsewhere [19]. Patients were eligible for study inclusion if they met the UK PD Brain Bank Criteria according to a neurologist with a special interest in PD. We included any individuals diagnosed within the last three and a half years and who were given a probability of PD $\geq$ 90% as rated by a clinician based on their clinical opinion. This was to try to eliminate the inclusion of similar conditions that have been incorrectly diagnosed as PD. All individuals in this study had their olfaction measured using the standard Sniffin' test.

Tracking Parkinson's is a large incidence cohort of patients with PD recruited from around the UK. Patients were recruited between February 2012 and May 2014 if they were diagnosed within the last 3.5 years and met Queen Square Brain Bank criteria. Full details of this study are described elsewhere [20]. Again we only included individuals who were given a probability of PD $\geq$ 90% as rated by a clinician. In this cohort, olfaction was initially measured using the UPSIT. However during the course of the study a difficulty arose in obtaining the UPSIT kits and the study was forced to switch to using the Sniffin' test instead. This means we have two groups of individuals within the same cohort completing different tests.

We also have a third dataset of subjects "Testing of olfaction in Parkinson's and controls" (TOPC) who undertook both tests (Sniffin' and UPSIT) concurrently so we could validate our conversion algorithms. This comprised of 128 subjects (61 PD and 67 controls) who were recruited as a convenience sample from the regional, West of Scotland, Movement Disorder Clinic. The order on which individuals took the two tests was randomised thus minimising any order effects, such as patients scoring worse on the second test due to fatigue.

All three studies had ethical approval and were undertaken with the understanding and written consent of each subject and in compliance with the declaration of Helsinki.

2.2. Olfaction tests

The UPSIT test has 40 items, where each item has one correct answer and three incorrect answers or "distractors". The test is a forced choice paradigm, that is, if an individual is unsure of an answer they are forced to guess a response hence a score of 25% on average would reflect random guessing. An UPSIT result is scored out of 40 where a higher score indicates better olfaction. There is also a reduced 12 item version [21] of the UPSIT called the Brief-Smell Identification Test (B-SIT), previously called the Cross-cultural Smell Identification Test (CC-SIT).

The standard Sniffin' test has 16 odour identification items, where each item has one correct answer and three incorrect answers or "distractors". Again the test is a forced choice paradigm. A Sniffin' result is scored out of 16 where a higher score indicates better olfaction. There is also a Sniffin' 12 item version [22] which is a subset of the 16 item version.

2.3. Statistical analysis

The first and simplest method of equating one scale to another is equipercentile equating with log-linear smoothing which matches scores on the two tests using their percentile ranks after first smoothing the distribution. This method requires that the two groups are equivalent in olfaction usually through design creating randomly equivalent groups or by carrying out both tests on the same population. In our case it would mean assuming the groups taking the Sniffin' and UPSIT tests are equivalent with regards to olfaction.

Our second method used Item Response Theory (IRT) which models individual's responses on the item level by fitting a series of latent variable models for each item. The power of the IRT approach is that we calibrated our model between groups with potentially different olfaction by using items that are common to both tests. We assumed that the two groups are linearly related by their olfaction and calculated a calibration slope and intercept between the two groups. After calibration we built the distribution of scores and then equated using equipercentile methods.

Both the equipercentile and IRT methods are described in detail by Kolen and Brennan [23] whilst the details of how we used the IRT method and the computing programs we used are discussed further in the Web appendix.

We used both methods to convert between the UPSIT and Sniffin' 16 item test. Since the Sniffin' 12 items is a subset of the Sniffin' 16 item and the B-SIT is a subset of the UPSIT they were carried out on the same population. Hence we only used the equipercentile method for the UPSIT to B-SIT and Sniffin' 12 to 16 item conversions. We used our validation dataset to test how well the conversions performed by comparing the concordance correlation coefficient [24] (a measure of agreement between two continuous variables) between true and equivalent results as well as the characteristics of the difference (or delta) between the true and equivalent.

We also converted the centile position stratified by age and gender from the UPSIT normative data charts to an equivalent Sniffin' score to provide more detailed normative comparative data. We used at or below the 15th centile as a cut-point for determining whether an individual has impaired olfaction corrected for age and gender as we have done in previous research [25]. There are some inconsistent and random fluctuations in the centiles (probably due to sample size issues) hence we used LOWESS techniques to smooth the cut-points before applying our conversion.

3. Results

3.1. Demographic and clinical data for Tracking Parkinson's and Oxford Discovery cohorts

Table 1 compares the data we have from the Tracking Parkinson's with 980 individuals who took the UPSIT test and 294 who took the Sniffin' test at the baseline visit. These two sub-groups of the Tracking Parkinson's cohort have a similar proportion of females, age when the testing took place, motor severity (measured by the Movement Disorder Society Unified PD Rating Scale or MDS-UPDRS part 3), disease severity (measured by Hoehn and Yahr stage) and cognitive impairment (measured by the education adjusted Montreal Cognitive Assessment or MoCA). However the UPSIT sub-group had slightly longer disease duration. This is not surprising given that the UPSIT sub-group would have been recruited first in the study, which would include both incident and some prevalent cases (up to 3.5 years), however the cases that are recruited later on in the centres would consist of mainly incident cases since the prevalent pool of cases would have already been recruited.

In the Oxford Discovery cohort we have 837 individuals who took the Sniffin' 16-item odour identification test at the baseline visit. When compared to the group who took the UPSIT test from the Tracking Parkinson's cohort they had slightly shorter disease duration, a similar proportion of females and similar age at testing. They also had worse motor severity, disease severity and more cognitive impairment. Comparing the Tracking Parkinson's Sniffin' subset and Oxford Discovery groups they show similar gender, age and cognitive impairment but Oxford Discovery has worse motor and disease severity and longer disease duration from diagnosis. Of paramount importance is that there is no evidence ($p = 0.12$) of a difference in Sniffin' scores between the Tracking Parkinson's subset and Oxford Discovery groups. We therefore pooled the Sniffin' data from the two cohorts for our UPSIT to Sniffin' 16 conversion. Web table 1 shows the demographic data from the TOPC validation study and Web Fig. 2 shows the distribution of UPSIT and Sniffin' 16 scores stratified by patient type. The correlation between the UPSIT and Sniffin' 16 scores was 0.81 in this sample.

3.2. UPSIT to Sniffin' 16 conversion

Table 2 shows the conversions from the UPSIT to a Sniffin' 16 equivalent using the two methods. In general, most UPSIT scores were grouped into 2 point values equivalent to 1 Sniffin' point but this could be as wide as 5 points for the (0–4) group using the IRT method. Table 3 presents the characteristics of these different conversions when tested on the TOPC validation data in which we compared an UPSIT predicted Sniffin' 16 to a true Sniffin' 16 score. The concordance correlation coefficient between the true and equivalent Sniffin' is very good and similar using both the equipercentile (0.79) and IRT methods (0.80). The difference between equipercentile predicted and true Sniffin' was acceptable although there was some evidence of under-prediction bias (positive mean delta). The individual IRT parameter estimates (*a, b, c*) for the UPSIT data and the combined Sniffin' data can be found in Web Tables 2 and 3 When using the IRT method we found that the calibration slope was 1.093 and the calibration intercept was 0.180. This is equivalent to saying that the individuals taking the UPSIT test had marginally better olfaction and also a slightly larger spread of olfaction when compared to the Sniffin' group. However mean olfaction that is 0.180 higher is small considering the groups are scaled to a mean of 0 and sd of 1. The validation of the IRT method

Table 1
Demographic and clinical data for Tracking Parkinson's and Discovery cohorts (restricted to recently diagnosed and probability of PD ≥ 90% at latest visit).

Variable	Tracking Parkinson's UPSIT data (N = 980): Mean (sd; range) or n(%)	Tracking Parkinson's Sniffin' data (N = 294): Mean (sd; range) or n(%)	P-value difference between two tracking Parkinson's groups[e]	Discovery Sniffin' data (N = 837) mean (sd; range) or n(%)	P-value difference between UPSIT and discovery group	P-value difference between two Sniffin' groups
Disease duration from diagnosis, years	1.38 (0.9; 0–3.5)	1.14 (0.9; 0–3.1)	<0.001[a]	1.28 (1.0, 0–3.5)	0.02[a]	0.02[a]
Female	347 (35.4%)	101 (34.4%)	0.76[c]	299 (35.7%)	0.89[c]	0.70[c]
Age at test	67.5 (9.1; 31.8–91.1)	67.6 (9.0; 38.1–88.3)	0.93[a]	67.3 (9.5; 32.2–90.5)	0.58[a]	0.62[a]
UPDRS 3	22.1 (11.6; 1–63)	22.1 (12.4; 1–74)	0.84[b]	26.4 (10.9; 5–77)	<0.001[b]	<0.001[b]
Hoehn and Yahr*			0.86[c]		<0.001[c,d]	<0.001[c,d]
0–1	508 (52.5%)	143 (49.8%)		193 (23.1%)		
2	417 (43.1%)	132 (46.0%)		581(69.4%)		
3+	43 (4.4%)	12 (4.2%)		63 (7.5%)		
MoCA adjusted	25.4 (3.3; 10–30)	25.4 (3.2; 10–30)	0.93[b]	25.0 (3.3; 13–30)	0.02[b]	0.07[b]
UPSIT score	19.6 (6.7; 3–37)	NA	NA	NA	NA	
Sniffin' 16 score	NA	7.5 (2.8; 0–15)	NA	7.2 (2.9; 1–15)	NA	0.12[a]
BSIT score	5.7 (2.2; 0–12)	NA	NA	NA	NA	
Sniffin' 12 score	NA	6.0 (2.4; 0–12)	NA	5.7 (2.5; 0–12)	NA	0.18[a]

UPDRS = Movement Disorder Society unified Parkinson's disease rating scale, MoCA = Montreal cognitive assessment, UPSIT = University of Pennsylvania Smell Identification Test, BSIT = Brief Smell Identification Test.

[a] T-test.
[b] Rank-sum test.
[c] Chi-squared test.
[d] In Tracking Parkinson's 1.5 changed to 1 and 2.5 changed to 2 for comparability between cohorts.
[e] One individual with both UPSIT and Sniffin' in Tracking Parkinson's was excluded from the test of differences between the two groups.

Table 2
Conversion table for different methods between the raw UPSIT scores and the equivalent Sniffin' 16 score.

Raw UPSIT score		Equivalent Sniffin' 16 score
Equipercentile method	IRT method	
0–3	0–4	0
4–6	5–6	1
7–8	7–8	2
9–10	9–10	3
11–13	11–12	4
14–15	13–14	5
16–17	15–16	6
18–20	17–18	7
21–22	19–21	8
23–24	22–23	9
25–27	24–25	10
28–29	26–27	11
30–32	28–30	12
33–34	31–32	13
35–36	33–35	14
37–38	36–37	15
39–40	38–40	16

on the TOPC data resulted in a delta that has a mean very close to zero and a median of zero showing that this conversion appears to have little evidence of bias. Web Fig. 3 shows graphically the degree of agreement between the true Sniffin' and the UPSIT equivalent Sniffin' using the two methods.

Comparison of these calibration estimates to the conversions carried out using the equipercentile method showed some agreement. Assuming these calibration estimates are correct implies that the olfaction was slightly different in the two populations and hence the assumptions for the equipercentile method do not hold. Considering these calibration estimates, individuals taking the UPSIT test seem to have slightly better olfaction when compared to the Sniffin'. In agreement with this the equipercentile method showed evidence of the difference in olfaction in the observed bias.

Table 4 shows the cut-points corresponding to the 15th centile of olfaction score stratified by age and gender from the UPSIT normative data. The table also shows the smoothed cut-points using LOWESS techniques and the equivalent Sniffin' score when applying our conversion chart from the IRT method in Table 2. This allows researchers to define a binary hyposmic group (Yes/No) based on poor olfaction ($\leq$15th centile) for each gender and different age groups which can be used in analyses testing predictors of hyposmia.

3.3. Sniffin' 12 to Sniffin' 16 conversion

In the conversion from Sniffin' 12 to 16 we are no longer bound by assuming the groups to be equal because they are identical. This means that we can use data from each visit in the Discovery cohort rather than only using the baseline data. The number of individuals eligible for analysis were 837, 564, and 275 from visits 1, 2, and 3 respectively from the Discovery cohort along with the 294 from the Tracking Parkinson's cohort. The 1970 observations of combined Sniffin' 16 data has a mean of 7.0 and s.d. of 2.8 whilst the combined Sniffin' 12 data has a mean of 5.6 and s.d. of 2.4. Web table 4 shows the conversion scores from Sniffin' 12 to a Sniffin' 16 equivalent and Table 3 shows the validation of this conversion using the TOPC data. With these two tests being so similar it is not surprising that the concordance between true and equivalent Sniffin' 16 was very high, 0.97, that the average delta between the two was so close to zero and the standard deviation of the delta was also low at 0.96. Web Fig. 4 shows graphically the degree of agreement using the true Sniffin' 16 and the Sniffin' 12 equivalent Sniffin' 16. It could be argued that the percentiles used in the equipercentile method should not include an individual more than once, re-running this method using only the baseline data from the Discovery cohort gave an identical conversion.

3.4. UPSIT to B-SIT conversion

Web table 5 shows the conversion scores from UPSIT to B-SIT and Table 3 shows the validation of this conversion. The concordance coefficient is relatively high, 0.82, however when looking at the delta there is some evidence of over-prediction bias (negative average delta) in our conversion, mean = −0.63 and median = −1. However if we stratify the delta by PD cases (mean delta = −0.08 and median = 0) and controls (mean delta = −1.13 and median = −1) there is only evidence of bias for the controls. Web Fig. 5 shows graphically the degree of agreement using the true B-SIT and the UPSIT equivalent B-SIT.

4. Discussion

We used two methods to equate scores on the UPSIT test to scores on the Sniffin' 16 smell identification test, scores on the Sniffin' 12 item to Sniffin' 16 item smell identification tests and also scores on the UPSIT and B-SIT tests.

It has been shown that the differences in olfaction between PD patients and controls is not related to any particular odour type [26]. This suggests that although our conversions have been created using only PD patients they could potentially be used for controls and/or other diseases where olfactory dysfunction is not related to particular odour types.

A previous paper reported that the correlation between the Sniffin' and UPSIT scores was 0.85 [14] which is similar to 0.81, the value we found in our TOPC data. Another reported that the test-retest correlation of the UPSIT was 0.9 [27] and was 0.86 in the Sniffin' [28]. These results are of a similar magnitude with our correlation between true and UPSIT equivalent Sniffin' 16 of 0.8. Both variability in test-retest performance and inadequate

Table 3
Validation of the different conversions in the Testing of olfaction in Parkinson's and controls (TOPC) validation dataset.

Analysis	Concordance between true score and converted equivalent score	Difference between true score and converted equivalent score mean (sd; range)	Difference between true score and converted equivalent score median (IQR)
Equipercentile method – converting UPSIT to Sniffin' 16	0.79	0.66 (2.38; −7 to 7)	1 (−1 to 2)
IRT method – converting UPSIT to Sniffin' 16	0.80	0.14 (2.42; −7 to 7)	0 (−1 to 2)
Equipercentile method – converting Sniffin' 12 to Sniffin' 16	0.97	0.01 (0.96; −2 to 2)	0 (−1 to 1)
Equipercentile method – converting UPSIT to BSIT	0.82	−0.63 (1.44; −4 to 2)	−1 (−2 to 0)

Table 4
Age and gender stratified 15th centile from UPSIT normative data included smoothed results and the equivalent Sniffin' results.

Age group	Males			Females		
	≤15th centile UPSIT	Smoothed[a] ≤15th centile UPSIT	Equivalent Sniffin'	≤15th centile UPSIT	Smoothed[a] ≤ 15th centile UPSIT	Equivalent Sniffin'
15–19	33	33	14	35	35	14
20–24	33	33	14	35	34	14
25–29	34	33	14	34	34	14
30–34	33	32	13	34	34	14
35–39	33	32	13	34	33	14
40–44	32	31	13	34	33	14
45–49	33	30	12	34	32	13
50–54	29	29	12	32	31	13
55–59	26	27	11	32	30	12
60–64	28	24	10	31	27	11
65–69	22	22	9	26	25	10
70–74	19	19	8	22	22	9
75–79	18	16	6	16	18	7
80–84	12	13	5	15	15	6
>=85	10	9	3	15	13	5

NB. For males 60–64 where the 15th centile is both a score of 28 and 29 we chose 28 which was more in keeping with the surrounding values.
[a] Smoothed using lowess techniques and a bandwidth of 0.7.

conversion may have contributed to the differences between the true and converted scores, though our results are consistent with the test-retest correlations.

There were a number of limitations to our work. The validation dataset we used was small and does not cover the entire range of scores for the two olfaction tests. Also if we had designed our two incidence cohorts with these conversions in mind it would have been better to randomise patients to receive either the UPSIT or the Sniffin' test. There are also clear differences between the Tracking and Discovery groups, especially in cognition which is related to olfaction, which could be the reason why the equipercentile method on the UPSIT to Sniffin conversion showed some evidence of bias and made it necessary to use the IRT method. Another consideration is that the UPSIT normative data was derived using a US version. The cohorts that we studied used a newer UK version adapted due to cultural differences as some smells in the US version were unfamiliar in the UK population. Despite this, the UK and US versions are still very similar, sharing 33 items with some changes to distractors.

Our UPSIT to B-SIT conversion had high concordance but some evidence of bias. However this disappeared when only considering the PD cases from the TOPC data. None of our other conversions showed evidence of difference in the delta when stratified by PD or Control. This could be because (a) this conversion is not valid; (b) the conversion is valid and the differential observation between PD cases and controls was a chance finding; or (c) our conversion is only valid for PD patients contradicting our belief that differences in olfaction between PD patients and controls is not related to any particular odour type.

The choice of what olfaction test to use in a study will be determined by several factors (i) time available and burden on participants (ii) cost of administering tests (iii) sample size. Another issue to consider is that shorter tests may be less sensitive (e.g. 40-item UPSIT versus 16 item Sniffin') thereby reducing the ability to differentiate between groups. However statistical power is also related to sample size and measuring the UPSIT on a large sample would take considerably more time than a quicker test like the B-SIT. In some circumstances one may be happy to trade-off sensitivity against increased sample size. Longer tests are also less likely to be affected by random measurement error and will therefore have greater reliability. The association between reliability and test length is most famously highlighted by the Spearman-Brown prediction formula [29] and has been modelled before in olfaction [27]. In olfactory tests this is emphasised by the fact that the test-retest correlation was 0.9 in the UPSIT and 0.71 in the B-SIT [27].

We created a valid and reliable conversion of UPSIT scores to Sniffin' scores and from Sniffin' 12 item to 16 item. Also we have arguably created a valid and reliable conversion from UPSIT to B-SIT scores for PD patients. These conversions will be used to merge olfaction data from the Oxford Discovery and Tracking Parkinson's cohorts to investigate the influence of baseline olfaction and hyposmia in predicting future cognitive and motor decline in these longitudinal cohorts of early PD. We believe that these conversion charts will facilitate more replication of research findings and greater data sharing across many neurological diseases and studies that measure olfaction using these tests.

Funding

The Oxford Discovery study was funded by the Monument Trust Discovery Award from Parkinson's UK (J-0901 and J-1403) and supported by the National Institute for Health Research (NIHR) (HMRWAJO4) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University Of Oxford, and the NIHR Clinical Research Network: Thames Valley and South Midlands.

The Tracking Parkinson's study was funded by Parkinson's UK (J-1101) and supported by the National Institute for Health Research (NIHR) DeNDRoN network, the NIHR Newcastle Biomedical Research Unit based at Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University, and the NIHR funded Biomedical Research Centre in Cambridge. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Author roles

ML: Data Analysis, manuscript writing and editing.

MTMH and DG: Study design, data collection, manuscript writing and editing.

FB, CR, EB, DMAS, NM: Data collection, manuscript editing.

KAG, NB, RAB, DJB, TF, and HRM: Study design, data collection and manuscript editing.

NW, NWW: Study design and manuscript editing.

MTM: Data analysis, manuscript editing.

YBS: Study design, data analysis and manuscript writing and editing.

Conflicts of interest

MA Lawton, M Hu, F Baig, C Ruffmann, E Barron, DMA Swallow, N Malek, KA Grosset, N Williams, N Wood, M May, Y Ben-Shlomo: No conflicts of interest.

N Bajaj has received payment for advisory board attendance from UCB, Teva Lundbeck, Britannia, GSK, Boehringer, and honoraria from UCB Pharma, GE Healthcare, Lily Pharma, Medtronic. He has received research grant support from GE Healthcare, Wellcome Trust, MRC and Parkinson's UK and royalties from Wiley.

RA Barker has received grants from Parkinson's UK, NIHR, Cure Parkinson's Trust, Evelyn Trust, Rosetrees Trust, MRC and EU along with payment for advisory board attendance from Oxford Biomedica and LCT, and honoraria from Wiley and Springer.

DJ Burn has received grants from NIHR, Wellcome Trust, GlaxoSmithKline Ltd, Parkinson's UK, and Michael J Fox Foundation. He has acted as consultant for GSK.

T Foltynie has received payment for advisory board meetings for Abbvie and Oxford Biomedica, and honoraria for presentations at meetings sponsored by Medtronic, St Jude Medical, Britannia and Teva pharmaceuticals.

H Morris reports grants from Parkinson's UK, grants from Medical Research Council UK, during the conduct of the study; grants from Welsh Assembly Government, personal fees from Teva, personal fees from Abbvie, personal fees from Teva, personal fees from UCB, personal fees from Boerhinger-Ingelheim, personal fees from GSK, non-financial support from Teva, grants from Ipsen Fund, non-financial support from Medtronic, grants from MNDA, grants from PSP Association, grants from CBD Solutions, grants from Drake Foundation, personal fees from Acorda, outside the submitted work; In addition, H Morris has a patent H. R. M is a co-applicant on a patent application related to C9ORF72 - Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140) pending.

DG Grosset has received payment for advisory board attendance from AbbVie, and honoraria from UCB Pharma, GE Healthcare, and Acorda.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.parkreldis.2016.09.023.

References

[1] G.W. Ross, H. Petrovitch, R.D. Abbott, C.M. Tanner, J. Popper, K. Masaki, L. Launer, L.R. White, Association of olfactory dysfunction with risk for future Parkinson's disease, Ann. Neurol. 63 (2008) 167–173.

[2] L. Silveira-Moriyama, A. Petrie, D.R. Williams, A. Evans, R. Katzenschlager, E.R. Barbosa, A.J. Lees, The use of a color coded probability scale to interpret smell tests in suspected parkinsonism, Mov. Disord. 24 (2009) 1144–1153.

[3] L. Silveira-Moriyama, M.D. Carvalho, R. Katzenschlager, A. Petrie, R. Ranvaud, E.R. Barbosa, A.J. Lees, The use of smell identification tests in the diagnosis of Parkinson's disease in Brazil, Mov. Disord. 23 (2008) 2328–2334.

[4] P.W. Schofield, H. Ebrahimi, A.L. Jones, G.A. Bateman, S.R. Murray, An olfactory 'stress test' may detect preclinical Alzheimer's disease, BMC Neurol. 12 (2012) 1–24.

[5] R.L. Doty, C. Li, L.J. Mannon, D.M. Yousem, Olfactory dysfunction in multiple sclerosis, N. Engl. J. Med. 336 (1997) 1918–1919.

[6] P. Mahlknecht, A. Iranzo, B. Hogl, B. Frauscher, C. Muller, J. Santamaria, E. Tolosa, M. Serradell, T. Mitterling, V. Gschliesser, G. Goebel, F. Brugger, C. Scherfler, W. Poewe, K. Seppi, Sleep Innsbruck Barcelona G. Olfactory dysfunction predicts early transition to a Lewy body disease in idiopathic RBD, Neurology 84 (2015) 654–658.

[7] F.W. Bylsma, P.J. Moberg, R.L. Doty, J. Brandt, Odor identification in Huntington's disease patients and asymptomatic gene carriers, J. Neuropsychiatry Clin. Neurosci. 9 (1997) 598–600.

[8] M. Abele, A. Riet, T. Hummel, T. Klockgether, U. Wullner, Olfactory dysfunction in cerebellar ataxia and multiple system atrophy, J. Neurol. 250 (2003) 1453–1455.

[9] L. Silveira-Moriyama, G. Hughes, A. Church, H. Ayling, D.R. Williams, A. Petrie, J. Holton, T. Revesz, A. Kingsbury, H.R. Morris, D.J. Burn, A.J. Lees, Hyposmia in progressive supranuclear palsy, Mov. Disord. 25 (2010) 570–577.

[10] J.E. Ahlskog, S.C. Waring, R.C. Petersen, C. Esteban-Santillan, U.K. Craig, P.C. O'Brien, M.F. Plevak, L.T. Kurland, Olfactory dysfunction in Guamanian ALS, parkinsonism, and dementia, Neurology 51 (1998) 1672–1677.

[11] M.D. Godoy, R.L. Voegels, R. Pinna Fde, R. Imamura, J.M. Farfel, Olfaction in neurologic and neurodegenerative diseases: a literature review, Int. Arch. Otorhinolaryngol. 19 (2015) 176–179.

[12] R. Katzenschlager, A.J. Lees, Olfaction and Parkinson's syndromes: its role in differential diagnosis, Curr. Opin. Neurol. 17 (2004) 417–423.

[13] R. Doty, Studies of olfactory dysfunction in major neurological disorders, Adv. Biosci. 93 (1994) 593–602.

[14] M. Wolfensberger, I. Schnieper, A. Welge-Lussen, Sniffin'Sticks (R): a new olfactory test battery, Acta Otolaryngol. 120 (2000) 303–306.

[15] R.L. Doty, P. Shaman, M. Dann, Development of the university-of-Pennsylvania smell identification test - a standardized microencapsulated test of olfactory function, Physiol. Behav. 32 (1984) 489–502.

[16] T. Hummel, G. Kobal, H. Gudziol, A. Mackay-Sim, Normative data for the "Sniffin' Sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects, Eur. Arch. Otorhinolaryngol. 264 (2007) 237–243.

[17] R.L. Doty, The Smell Identification Test™ Administration Manual, third ed., Sensonics, Inc, Haddon Heights, NJ, 1995.

[18] R.L. Doty, The olfactory system and its disorders, Semin. Neurol. 29 (2009) 74–81.

[19] K. Szewczyk-Krolikowski, P. Tomlinson, K. Nithi, R. Wade-Martins, K. Talbot, Y. Ben-Shlomo, M.T. Hu, The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort, Park. Relat. Disord. 20 (2014) 99–105.

[20] N. Malek, D.M. Swallow, K.A. Grosset, M.A. Lawton, S.L. Marrinan, A.C. Lehn, C. Bresner, N. Bajaj, R.A. Barker, Y. Ben-Shlomo, D.J. Burn, T. Foltynie, J. Hardy, H.R. Morris, N.M. Williams, N. Wood, D.G. Grosset, Tracking Parkinson's: study design and baseline patient data, J. Park. Dis. (2015) 947–959.

[21] R.L. Doty, A. Marcus, W.W. Lee, Development of the 12-item cross-cultural smell identification test (CC-SIT), Laryngoscope 106 (1996) 353–356.

[22] T. Hummel, C.G. Konnerth, K. Rosenheim, G. Kobal, Screening of olfactory function with a four-minute odor identification test: reliability, normative data, and investigations in patients with olfactory loss, Ann. Otol. Rhinol. Laryngol. 110 (2001) 976–981.

[23] M.J. Kolen, R.L. Brennan, Test Equating, Scaling, and Linking, second ed., Springer, 2004.

[24] L.I. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255–268.

[25] N. Malek, D.M. Swallow, K.A. Grosset, M.A. Lawton, C.R. Smith, N.P. Bajaj, R.A. Barker, Y. Ben-Shlomo, C. Bresner, D.J. Burn, T. Foltynie, H.R. Morris, N. Williams, N.W. Wood, D.G. Grosset, P.R. Investigators, Olfaction in Parkin single and compound heterozygotes in a cohort of young onset Parkinson's disease patients, Acta Neurol. Scand. 134 (4) (2015) 271–276.

[26] R.L. Doty, D.A. Deems, S. Stellar, Olfactory dysfunction in parkinsonism: a general deficit unrelated to neurologic signs, disease stage, or disease duration, Neurology 38 (1988) 1237–1244.

[27] R.L. Doty, D.A. McKeown, W.W. Lee, P. Shaman, A study of the test-retest reliability of ten olfactory tests, Chem. Senses 20 (1995) 645–656.

[28] A. Haehner, A.M. Mayer, B.N. Landis, I. Pournaras, K. Lill, V. Gudziol, T. Hummel, High test-retest reliability of the extended version of the "Sniffin' sticks", Test. Chem. Senses 34 (2009) 705–711.

[29] J.P. Guilford, Psychometric Methods, 2d ed., McGraw-Hill, New York, 1954.

Equating University of Pennsylvania Smell Identification Test scores and Sniffin' Scores in Patients with Parkinson's Disease

Supplemental web appendix

Methods

Statistical Analysis

To carry out the Item Response Theory method (IRT) we fitted a series of latent variable models for each item where olfaction is the latent variable. We modelled the probability of correctly answering an item on a test given the latent olfaction variable and a three parameter IRT model. These parameters are often described as the discrimination, difficulty and lower asymptote parameters. An example of an IRT model can be seen in web figure 1. Here θ is the latent olfaction variable; c is the lower asymptote parameter, or equivalently the probability of correctly answering the question even with very poor olfaction or by chance; b is the difficulty parameter or level of olfaction at which the probability is exactly half-way between 1 and the lower asymptote; a is the discrimination parameter or proportional to the slope where $\theta = b$. Within a group we scale the olfaction variable to have a mean of zero and standard deviation of 1.

After creating the IRT model we calibrated the two sets of models using the fact that the estimated parameters should be identical in the common items if the olfaction of the two groups is also identical. In this study we assumed that items were common if they have the

same answer. For instance the correct answer for item 11 on the Sniffin' and item 20 on the UPSIT is apple. Using this definition there are 13 common items between the Sniffin' and UPSIT. There are a number of methods that can be used to calibrate the items on two tests, in our case we used the Stocking-Lord characteristic curve method. Once we have fitted our models and calibrated the parameters between the two tests, then we can equate the overall scores on the two tests using either the true or observed score equating methods, in our case we used the observed score equating method. We decided to use the observed score method because theoretically its properties are easier to justify and it does not require extrapolation at the very low scores where the true score method is undefined.

The equipercentile equating was carried out using the equate library [1] in R 3.0.1. The IRT model fitting was carried out in BILOG-MG [2], the calibration using STUIRT [3] and the observed score equating using PIE [4].

Discussion

We used the observed score equating method as part of our IRT equating. However the other IRT equating method, true score equating, would have given slightly different conversions especially at the upper and lower ends of the scales. However when validating the true score equating on the TOPC data we found very similar results with a small bias between true and converted Sniffin' results (results not shown - available on request). Also of note is that as well as combining the Sniffin' data for the UPSIT to Sniffin' conversion we also tried each

method using the Sniffin' data from each cohort separately (results not included). This led to nearly identical results and for brevity we only included the combined analysis.

There are some limitations to this work in our application of the IRT method. Firstly there was evidence of a lack of fit in some questions but BILOG warns that the item fit statistic might be unreliable when the number of items is less than 20. This is exactly what we observed with a greater lack of fit in the Sniffin' test compared to the UPSIT. Secondly we had some slight convergence issues for the Sniffin' test which is likely to be due to the smaller number of items. Thirdly we made the tenuous assumption that the items on the Sniffin' and UPSIT tests are "common items" if the answers are the same. How a question functions in a test (which would affect estimates of the three IRT parameters) is likely to be related to not only the correct answer but also the three distractors which were not always the same across the two tests. Fourthly the different calibration methods in IRT gave different estimates for the calibration slope and intercept which is likely due to the problems around the "common items" assumption. However despite these limitations it is reassuring that the validation of our conversion showed good characteristics.

On a practical perspective, the equipercentile method was very simple to carry out within R. In contrast the time taken to learn and carry out IRT equating was considerably longer given the need to use multiple programs and the difficulty of exporting data in different formats between these programs.

References web appendix

[1] Albano A. Equate: Observed-Score Linking and Equating. (Version 2.0-3 R package) 2014. Available from http://CRAN.R-project.org/package=equate.
[2] Scientific Software International. BILOG-MG for Windows. (Version 3.0.2327.2) 2003. Available from http://www.ssicentral.com/.
[3] Kim S, Kolen MJ. STUIRT - A computer program for Scale Transformation under Unidimensional Item Response Theory Models. (Version 1.0) 2004. Available from http://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs
[4] Hanson B, Zeng L, Cui Z. PIE - A computer program for IRT Equating. (Windows Console version) 2004. Available from http://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Web table 1. Demographic and olfaction data for the Testing of olfaction in Parkinson's and controls (TOPC) validation dataset

Variable	Controls (N=67) mean (sd; range) or n (%)	PD (N=61) mean (sd; range) or n (%)
Female	36 (53.7%)	24 (39.3%)
Age at test	61.0 (13.2; 18.0 – 81.8)	66.8 (8.9; 46.5-81.6)
UPSIT	28.0 (6.9; 6-39)	16.7 (6.0; 7 – 38)
Sniffin' 16	11.6 (2.8; 5-16)	6.6 (3.3; 2-15)
BSIT	7.3 (2.2; 3- 11)	4.7 (2.1; 1-11)
Sniffin' 12	9.3 (2.4; 4-12)	5.3 (2.6; 2-11)
Order (took UPSIT first)	34 (50.8%)	33 (54.1%)

Web table 2. IRT parameter estimates for three-parameter logistic IRT model using UPSIT data of 980 individuals. The chi-square item fit p-value is a test between observed and predicted proportions so a small p-value represents lack of fit.

Item	Proportion correct	a (discrimination)	b (difficulty)	c (lower asymptote)	Chi-square item fit p-value
1	0.249	1.004	1.795	0.161	0.966
2	0.440	0.648	1.610	0.305	0.627
3	0.589	1.004	0.063	0.205	0.186
4	0.408	0.767	1.191	0.226	0.959
5	0.438	0.634	0.781	0.154	0.484
6 (mint)	0.543	1.113	0.394	0.253	0.528
7 (banana)	0.496	1.087	1.290	0.384	0.981
8 (clove)	0.522	1.291	0.683	0.316	0.901
9 (leather)	0.681	0.851	-0.291	0.246	0.217
10	0.442	1.169	1.075	0.290	0.967
11	0.758	0.835	-0.829	0.188	0.036
12	0.210	1.296	1.616	0.119	0.614
13	0.792	1.040	-0.895	0.203	0.008
14 (coffee)	0.558	0.500	0.278	0.198	0.689
15 (cinnamon)	0.430	1.146	1.534	0.342	0.694
16	0.383	0.906	1.007	0.176	0.672
17	0.545	0.918	0.248	0.196	0.599
18	0.476	0.945	0.965	0.293	0.571
19	0.632	0.757	-0.263	0.158	0.204
20 (apple)	0.533	0.790	1.115	0.376	0.993
21	0.664	0.580	-0.376	0.207	0.234
22 (turpentine)	0.233	0.935	2.447	0.191	0.655
23	0.612	0.811	-0.061	0.200	0.091
24 (liquorice)	0.331	1.404	1.214	0.197	0.831
25	0.195	0.861	3.124	0.175	0.961
26 (pineapple)	0.495	0.738	0.763	0.250	0.49
27	0.333	0.623	2.492	0.260	0.956
28 (orange)	0.532	0.859	0.502	0.252	0.478
29	0.519	0.816	0.627	0.263	0.992
30	0.445	0.880	0.888	0.228	0.652
31	0.388	0.611	1.470	0.214	0.71
32	0.382	0.981	1.563	0.279	0.764
33	0.730	1.135	-0.447	0.266	0.07
34	0.559	1.115	0.675	0.356	0.998
35	0.750	1.118	-0.637	0.216	0.033
36 (lemon)	0.291	0.803	2.477	0.241	0.218
37	0.403	0.487	3.013	0.337	0.953
38	0.536	0.956	0.250	0.183	0.222
39 (rose)	0.460	0.850	0.790	0.222	0.691
40	0.608	1.041	-0.043	0.196	0.263

Web Table 3. IRT parameter estimates for three-parameter IRT model using combined visit 1 Sniffin' data. The chi-square item fit p-value is a test between observed and predicted proportions so a small p-value represents lack of fit.

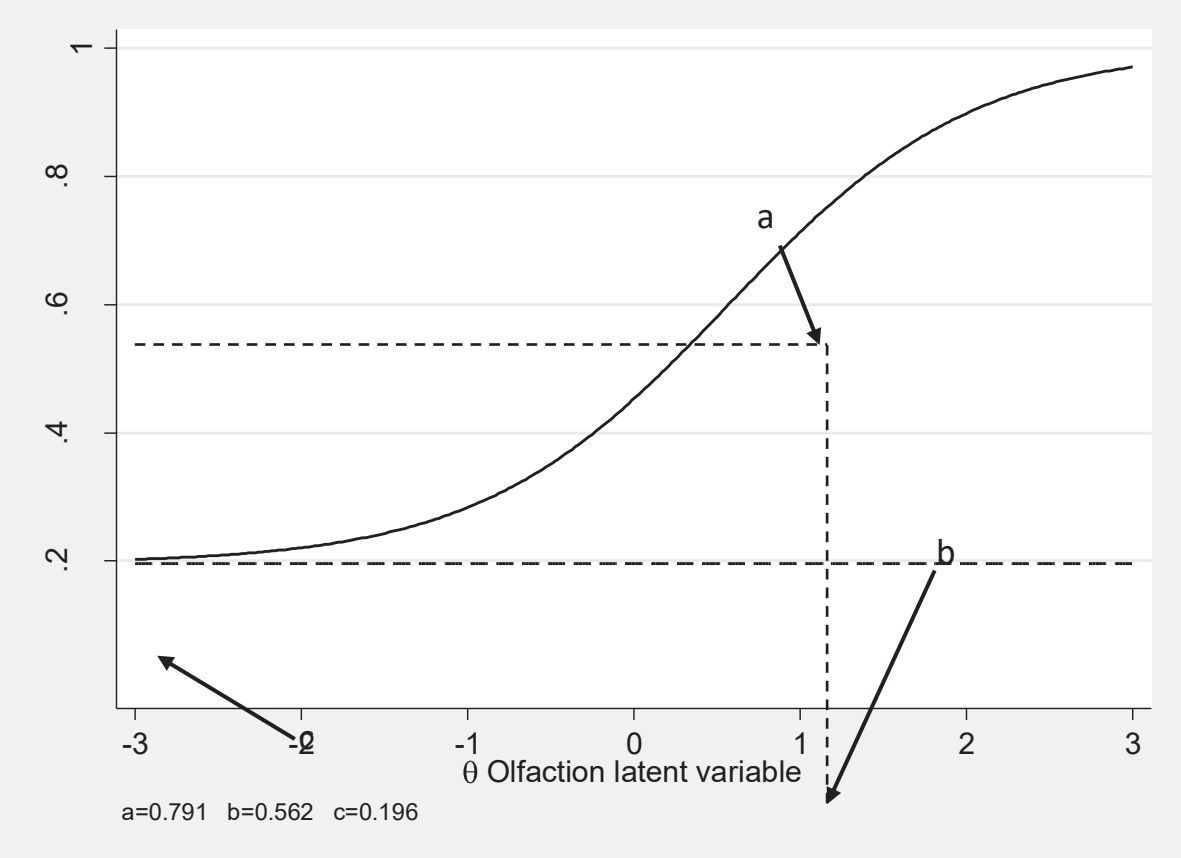
Item	Proportion correct	a (discrimination)	b (difficulty)	c (lower asymptote)	Chi-square item fit p-value
1 (orange)	0.714	0.747	-0.525	0.186	<0.001
2 (leather)	0.452	0.823	1.229	0.320	0.269
3 (cinnamon)	0.350	0.596	1.868	0.262	0.317
4 (mint)	0.677	0.790	-0.176	0.236	<0.001
5 (banana)	0.446	0.795	0.754	0.208	0.003
6 (lemon)	0.354	0.664	1.781	0.175	0.08
7 (liquorice)	0.469	1.666	0.924	0.218	0.04
8 (turpentine)	0.347	0.557	3.823	0.315	0.834
9	0.585	0.584	0.526	0.321	0.01
10 (coffee)	0.490	0.867	0.968	0.186	0.07
11 (apple)	0.190	1.098	2.861	0.154	0.427
12 (clove)	0.442	0.662	0.870	0.231	0.043
13 (pineapple)	0.364	0.562	1.283	0.171	0.031
14 (rose)	0.561	0.375	0.804	0.252	0.055
15	0.381	2.257	0.970	0.185	0.213
16	0.639	1.112	-0.166	0.195	<0.001

Web Table 4. Conversion from Sniffin' 12 score to Sniffin' 16.

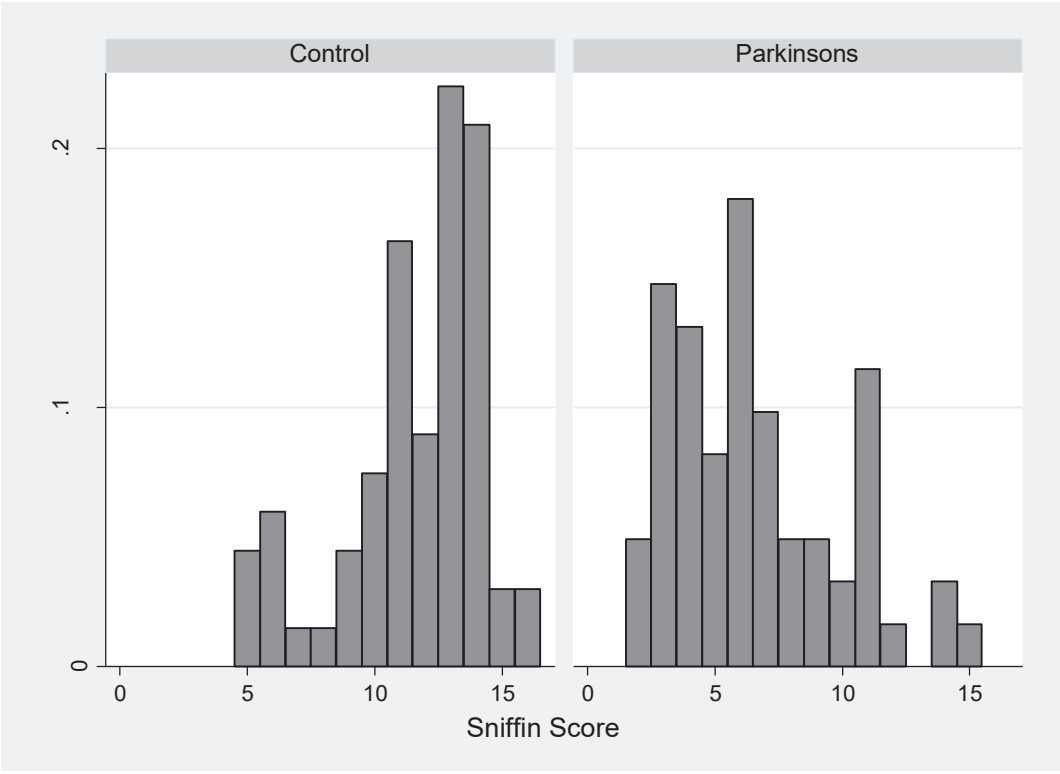
Raw Sniffin' 12 score	Equivalent Sniffin' 16 score
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	9
8	10
9	11
10	12
11	14
12	15

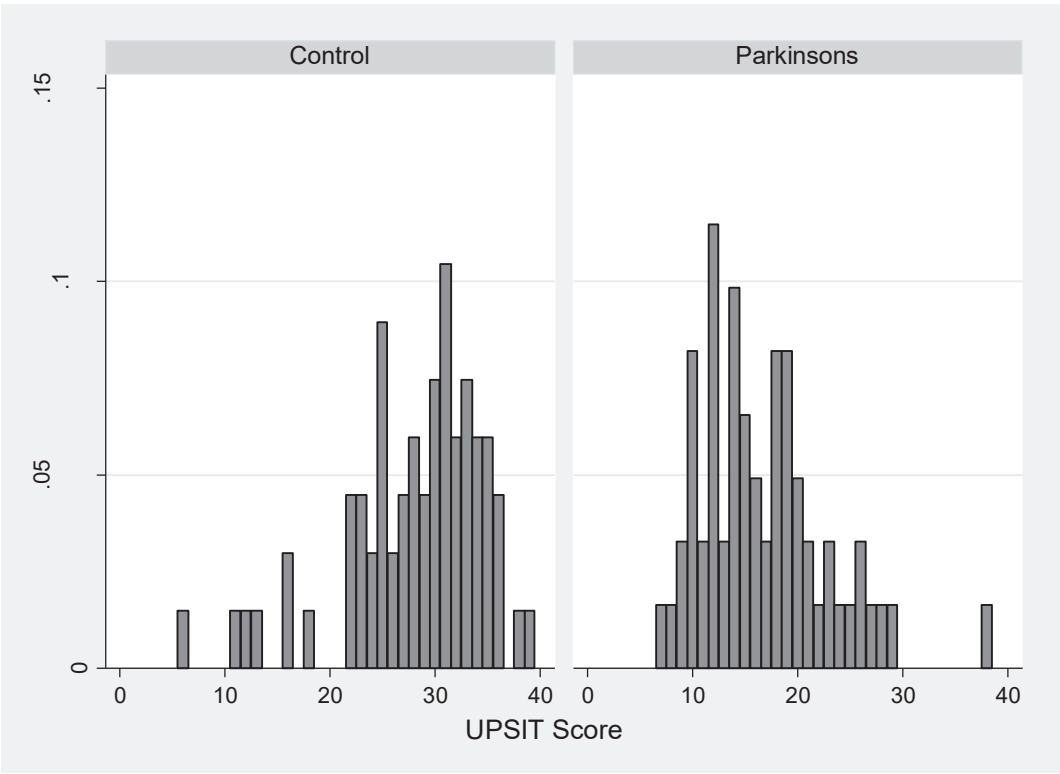
Web Table 5. Conversion from UPSIT score to B-SIT scores

Raw UPSIT score	Equivalent B-SIT score
0 - 4	0
5 - 7	1
8 - 10	2
11 - 12	3
13 - 15	4
16 - 18	5
19 - 21	6
22 - 24	7
25 - 28	8
29 - 31	9
32 - 35	10
36 - 38	11
39 - 40	12

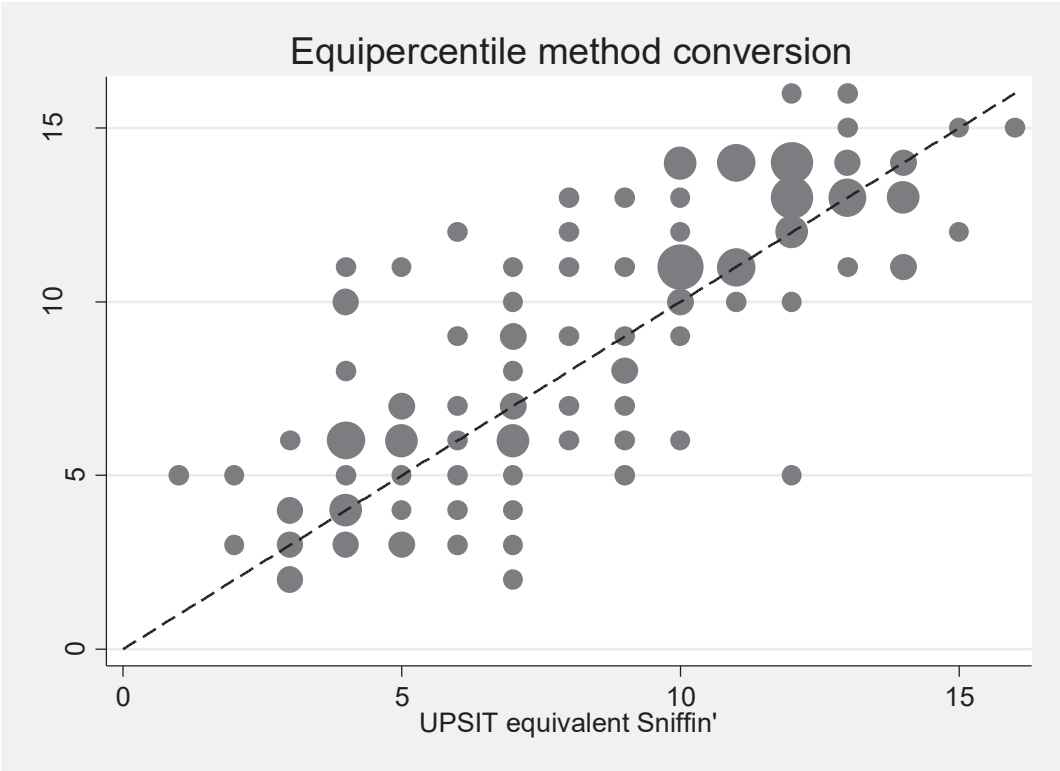


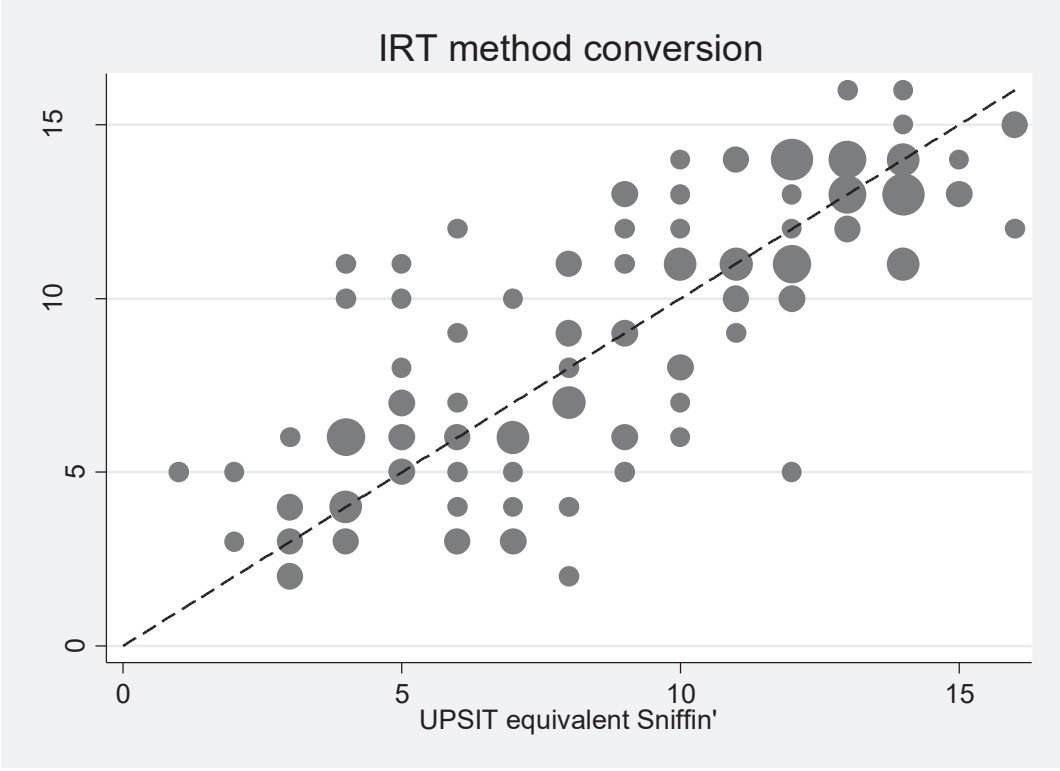
Web Figure 1. Example of a three parameter logistic IRT model. The dashed lines are added to help conceptualise the three parameters.



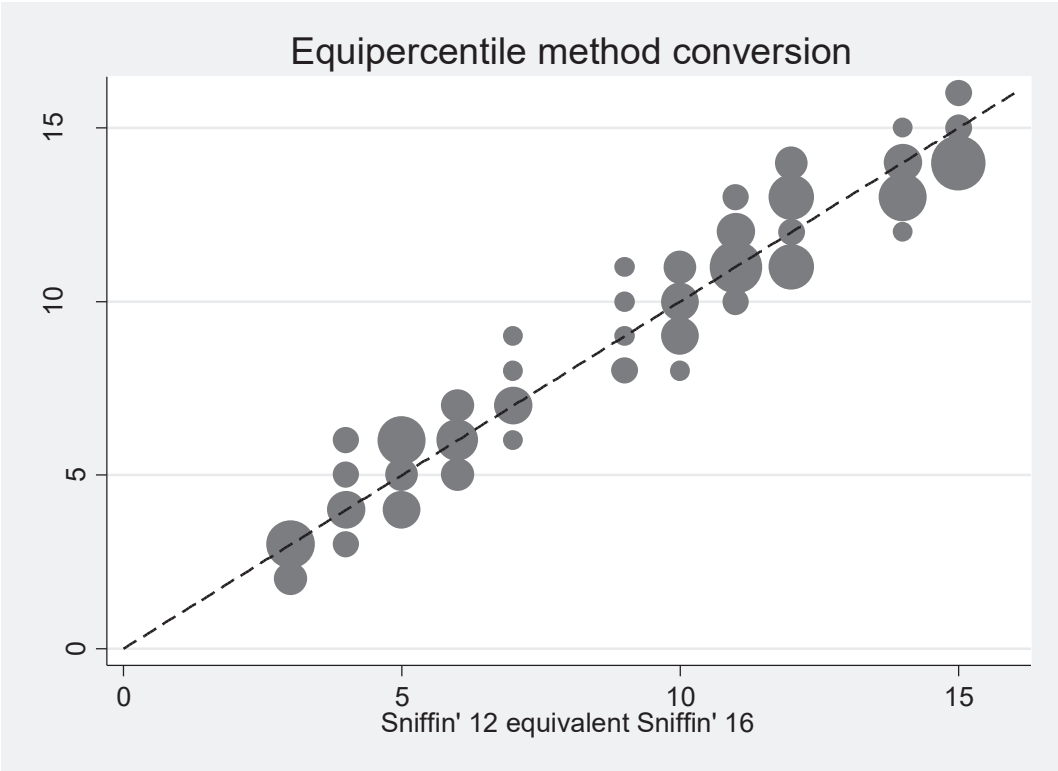


Web Figure 2. Distribution of Sniffin' 16 and UPSIT scores in the Testing of olfaction in Parkinson's and controls (TOPC) validation dataset, stratified by patient type.

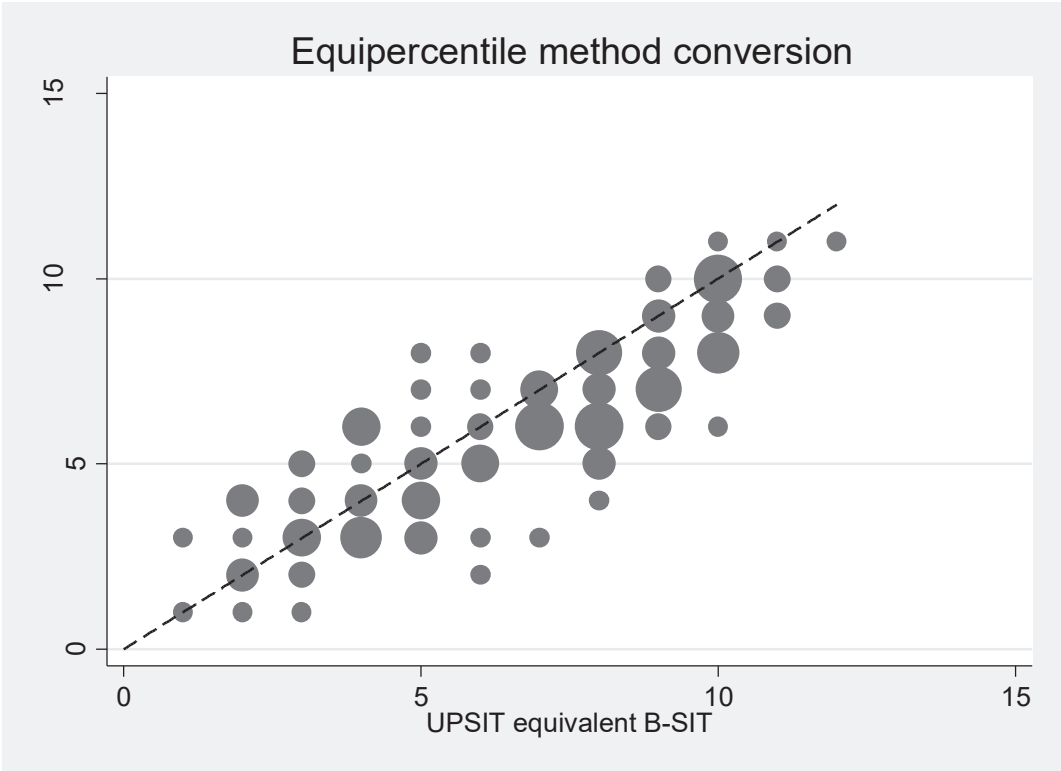




Web Figure 3. Agreement of true Sniffin' 16 and UPSIT equivalent Sniffin' 16 in the Testing of olfaction in Parkinson's and controls validation dataset using the two conversion methods. Size of dots are proportional to the number of individuals with that score.



Web Figure 4. Agreement of true Sniffin' 16 and Sniffin' 12 equivalent Sniffin' 16 in the Testing of olfaction in Parkinson's and controls validation dataset. Size of dots are proportional to the number of individuals with that score.



Web Figure 5. Agreement of true B-SIT and UPSIT equivalent B-SIT in the Testing of olfaction in Parkinson's and controls validation dataset. Size of dots are proportional to the number of individuals with that score.

OPEN ACCESS

RESEARCH PAPER

Developing and validating Parkinson's disease subtypes and their motor and cognitive progression

Michael Lawton,[1] Yoav Ben-Shlomo,[1] Margaret T May,[1] Fahd Baig,[2,3] Thomas R Barber,[2,3] Johannes C Klein,[2,3] Diane M A Swallow,[4] Naveed Malek,[5] Katherine A Grosset,[5] Nin Bajaj,[6] Roger A Barker,[7] Nigel Williams,[8] David J Burn,[9] Thomas Foltynie,[10] Huw R Morris,[11] Nicholas W Wood,[12] Donald G Grosset,[5] Michele T M Hu[2,3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jnnp-2018-318337).

For numbered affiliations see end of article.

Correspondence to
Michael Lawton, Department of Population Health Sciences, Canynge Hall, 39 Whatley Road, University of Bristol, Bristol BS8 2PS, UK; Michael.Lawton@bristol.ac.uk

DGG and MTMH contributed equally.

DGG and MTMH are joint senior authors.

Received 2 March 2018
Revised 5 June 2018
Accepted 13 June 2018

To cite: Lawton M, Ben-Shlomo Y, May MT, *et al*. *J Neurol Neurosurg Psychiatry* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jnnp-2018-318337

ABSTRACT

Objectives To use a data-driven approach to determine the existence and natural history of subtypes of Parkinson's disease (PD) using two large independent cohorts of patients newly diagnosed with this condition.

Methods 1601 and 944 patients with idiopathic PD, from Tracking Parkinson's and Discovery cohorts, respectively, were evaluated in motor, cognitive and non-motor domains at the baseline assessment. Patients were recently diagnosed at entry (within 3.5 years of diagnosis) and were followed up every 18 months. We used a factor analysis followed by a k-means cluster analysis, while prognosis was measured using random slope and intercept models.

Results We identified four clusters: (1) *fast motor progression* with symmetrical motor disease, poor olfaction, cognition and postural hypotension; (2) *mild motor and non-motor disease* with intermediate motor progression; (3) *severe motor disease*, *poor psychological well-being* and *poor sleep* with an intermediate motor progression; (4) *slow motor progression* with tremor-dominant, unilateral disease. Clusters were moderately to substantially stable across the two cohorts (kappa 0.58). Cluster 1 had the fastest motor progression in Tracking Parkinson's at 3.2 (95% CI 2.8 to 3.6) UPDRS III points per year while cluster 4 had the slowest at 0.6 (0.1–1.1). In Tracking Parkinson's, cluster 2 had the largest response to levodopa 36.3% and cluster 4 the lowest 28.8%.

Conclusions We have found four novel clusters that replicated well across two independent early PD cohorts and were associated with levodopa response and motor progression rates. This has potential implications for better understanding disease pathophysiology and the relevance of patient stratification in future clinical trials.

INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterised by a wide range of motor and non-motor features, for which there is no known cure. However, therapeutic strategies might soon be available with prolonged benefits that could affect the underlying pathogenesis, and hence delay or ultimately prevent the inexorable course of this disease. To date, none of the 16 drugs evaluated for PD disease modification have succeeded in phase III trials, with a further eight compounds currently in the discovery pipeline.[1] PD is an inherently complex disorder with known heterogeneity in terms of clinical presentation as well as rate of progression and risk of disease complications. The basis for this is only now starting to be understood, in terms of the role of genetic factors, for example, glucocerebrosidase gene mutations. The implications for future clinical trial design—if patient heterogeneity is ignored at baseline study selection, leading to potential confounds and misinterpretation of subsequent progression/complication rates—are highly significant.

Few naturalistic cohort studies in PD have been undertaken using large numbers of representative, community-ascertained patients, unselected on the basis of age or family history, and prospectively followed early from diagnosis. Such cohorts would more faithfully recapitulate disease evolution in the true-to-life populations encountered in clinical practice, where disease progression reflects both pathophysiology and any treatment effects, as reported in the CamPaIGN study.[2]

Data-driven approaches to delineate subtypes using cohorts of incident PD as well as cross-sectional studies[3–7] have hypothesised that there are different PD subtypes. Better defining these subtypes will be important for understanding the aetiology of the disease, discovering biomarkers related to prognosis and for stratified medicine, including the discovery and response to new medications.[8] In this study, we sought to better explore this aspect of PD using two large independent cohorts of newly diagnosed PD and in particular the number of distinct disease subtypes, their levodopa responsiveness and rate of motor and cognitive decline. This extends our previous work in this area using only one of the two cohorts (Discovery), without assessing levodopa responsiveness or the subsequent rate of motor and cognitive decline.[9]

MATERIALS AND METHODS

Patient populations

Tracking Parkinson's is a prospective cohort of recently diagnosed patients with PD who were recruited from around the UK between February 2012 and May 2014. Full details of this cohort and full inclusion/exclusion criteria have been published elsewhere.[10] The Oxford Parkinson's Disease Centre Discovery cohort (hereafter

referred to as Discovery) is also a prospective cohort of recently diagnosed patients with PD who were recruited from 11 hospitals in the Thames Valley region between September 2010 and January 2016. Full details of the Discovery cohort and full inclusion/exclusion criteria have also been published elsewhere.[11] In both studies, patients were defined as recently diagnosed if recruited within 3.5 years of diagnosis. In order to exclude patients with similar conditions that may have been incorrectly diagnosed as PD, we only included individuals in both cohorts if they had a probability of PD ≥90% as rated by a research neurologist/movement disorder specialist at their latest visit. Patients have been (and are continuing to be) followed up every 18 months. Both studies were funded by Parkinson's UK.

Patient evaluation

Assessments of patients were via self-completed questionnaires and from outpatient clinics using standardised and validated scales both at baseline and follow-up. Variables used in this analysis were those adopted in our original cluster analysis paper[9] and which were also included in the Tracking Parkinson's cohort, and these included the Movement Disorders Society (MDS) revised Unified Parkinson's Disease Rating Scale (UPDRS), where part III was measured in the 'on' state; Big Five Inventory; Epworth Sleepiness Scale; REM Sleep Behaviour Disorder Screening Questionnaire; Hospital Anxiety and Depression Scale; Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease; Honolulu Asia Aging Study Constipation Questionnaire; Montreal Cognitive Assessment (MoCA) adjusted for education years; Semantic verbal fluency (animals); Orthostatic blood pressure measurement; and Sniffin' 16 odour identification scores. The levodopa equivalent daily dose (LEDD) was calculated from medication use questionnaires using established formulae.[12] In addition, a levodopa challenge was carried out giving us a quantitative measure of response to medication (see online supplementary e-appendix for more details on methods).

Statistical analysis

We imputed missing data using the mean score if 80% or more questions were answered in any given test. Additionally, missing baseline data were imputed using the chained equations approach separately in the two cohorts. Factor analysis was used as a variable reduction technique on all the baseline phenotypic variables (details in online supplementary e-appendix). We then derived the clusters by using a k-means analysis of the factor scores and other baseline phenotypic variables not loading into one of the factors. Variables were standardised separately within each cohort to ensure that each variable had the same weighting within the k-means analysis. Further details are described in our previous publication.[9]

A discriminant analysis model was then fitted to the Tracking Parkinson's clusters and used to predict clusters within the Discovery cohort. These predicted clusters were compared with the k-means clusters in the Discovery cohort to determine the stability of our approach. We used the kappa statistic to compute the extent of agreement and adopted accepted guidelines[13] to determine the strength of this agreement.

We then carried out a between-cluster comparison of a range of clinical and demographic variables, which had not been used in the estimation of the clusters using analysis of variance and χ^2 tests. We modelled important disease-related variables (UPDRS III and MoCA scores) longitudinally using multilevel random slope and intercept models to estimate disease progression by cluster. A sensitivity analysis using pattern-mixture models was carried out to determine whether patients lost to follow-up may potentially have biased our disease progression estimates.[14]

RESULTS

We analysed data on 1601 patients in Tracking Parkinson's and 944 in Discovery (online web supplementary figure 1). Both cohorts had around 35% women, were predominantly white (>98%) and had an average age of diagnosis of about 66 years (see table 1). The disease duration from diagnosis was on average 1.2–1.3 years. Compared with Tracking Parkinson's, the

Table 1 Demographic and clinical characteristics at baseline for the patients in the two studies

Variable	Tracking Parkinson's cohort (n=1601) mean (SD) or n (%)	Discovery cohort (n=944) mean (SD) or n (%)	P-value difference between cohorts
Female	554 (34.6%)	334 (35.4%)	0.69*
Ethnicity (non-white)	28 (1.8%)	20 (2.1%)	0.51*
Age diagnosis (years)	65.9 (9.3)	65.9 (9.6)	0.92†
Disease duration from diagnosis (years)	1.3 (0.9)	1.2 (0.9)	0.03†
MDS-UPDRS part I‡	9.1 (5.2)	8.8 (5.1)	0.09†
MDS-UPDRS part II‡	9.5 (6.2)	8.7 (6.0)	<0.001†
MDS-UPDRS part III‡	22.3 (11.9)	26.4 (10.8)	<0.001†
MDS-UPDRS part IV‡	0.7 (1.7)	0.3 (1.1)	<0.001†
MDS-UPDRS total (all four parts)‡	41.8 (18.7)	44.2 (17.5)	0.002†
MoCA (adjusted for education years)‡	25.4 (3.4)	25.0 (3.3)	0.012†
Untreated	149 (9.3%)	119 (12.6%)	0.01*
LEDD (mg)	293 (205)	282 (212)	0.20†
LEDD (those on medication) (mg)	324 (190)	327 (194)	0.77†
Hoehn and Yahr§ median (IQR)	1 (1–2)	2 (2–2)	<0.001*

Motor assessments (UPDRS and Hoehn and Yahr) were rated in the clinically defined 'on medication' state for treated patients with PD.
*χ^2 test.
†T-test.
‡Changed denominator where 80% or more of questions were answered.
§In Tracking Parkinson's cohort, Hoehn and Yahr 1.5 and 2.5 were changed to 1 and 2, respectively, for comparison with Oxford Parkinson's Disease Centre Discovery cohort.
LEDD, levodopa equivalent daily dose; MDS, Movement Disorders Society; MoCA, Montreal Cognitive Assessment; UPDRS, Unified Parkinson's Disease Rating Scale.

Table 2 Confirmatory factor analysis within the Tracking Parkinson's cohort showing standardised factor loadings of variables selected from exploratory factor analysis and measures of model fit

Variable	Factor 1 Psychological well-being	Factor 2 Non-tremor motor
MDS-UPDRS I apathy	0.512	
MDS-UPDRS I fatigue	0.599	
MDS-UPDRS I pain	0.544	
BFI—neuroticism	0.459	
HADS anxiety	0.795	
HADS depression	0.863	
QUIP	0.307	
MDS-UPDRS III speech		0.420
MDS-UPDRS III rigidity subscore		0.535
MDS-UPDRS III bradykinesia subscore		0.769
MDS-UPDRS III postural subscore		0.609
CFI=0.909		
TLI=0.932		
RMSEA=0.063		

CFI, TLI and RMSEA are all measures of model fit.
BFI, Big five inventory; QUIP, Questionnaire for Impulsive-Compulsive Disorders in Parkinson's disease.
CFI, Comparative Fit Index; HADS, Hospital Anxiety and Depression Scale; MDS, Movement Disorders Society; RMSEA, root mean square error of approximation; TLI, Tucker-Lewis Index; UPDRS, Unified Parkinson's Disease Rating Scale.

Discovery cohort had more severe motor disease as measured by the UPDRS III and disease severity as measured by the Hoehn and Yahr or the sum score of UPDRS parts I–IV ($p<0.001$), and slightly worse average cognition as measured by the MoCA. However, the Tracking Parkinson's cohort had worse motor aspects of experiences of daily living (UPDRS II) and motor complications (UPDRS IV) and had fewer untreated patients.

Cluster analysis

In our factor analysis, we found two factors: a psychological well-being and a non-tremor motor factor (table 2), as we reported previously.[9] This shows that within our baseline phenotypic variables, we had multiple variables related to psychological well-being and to non-tremor motor function that were highly correlated. Using the statistics in web supplementary table 1 helped us decide that four clusters gave us an optimal solution. Figure 1 highlights the important features of the clusters and figure 2 shows the average of each of the standardised variables within each cluster for the Tracking Parkinson's and Discovery cohorts. The groups were arbitrarily ordered in terms of size for Tracking Parkinson's, but for the Discovery cohort they were ordered by similarity to the Tracking Parkinson's clusters. In general, the cluster patterns between the cohorts were fairly similar but with some differences (see below). Details of the clusters are available in web supplementary table 2, which shows all the scores from the different tests included in the cluster analysis and categorised scores using standard cut-points from the literature for easier clinical interpretation. More details of the factor and the cluster analysis can be found in the online supplementary e-appendix.

The following describes the clusters observed in Tracking Parkinson's (unless otherwise stated). The *fast motor progression* (1) cluster had less advanced motor features and psychological well-being but worse than average non-motor features such as blood pressure postural drop, olfaction and cognition with more symmetrical motor disease. However, within the Discovery cohort, the non-tremor motor was worse, rather than better than average, for this cluster. The *mild motor and non-motor disease* (2) cluster showed a milder form of the disease being better in most domains and was similar in the Discovery cohort analysis. The *severe motor disease*, *poor psychological well-being* and *poor sleep* (3) cluster was similar in the two cohorts and showed a more severe form of PD,

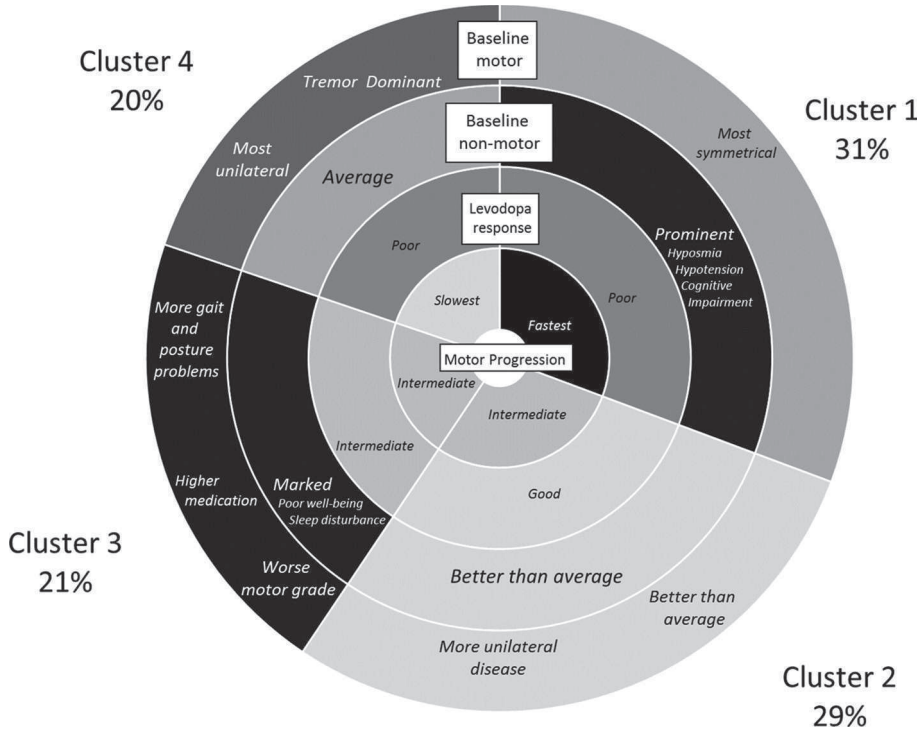


Figure 1 Important salient clinical features of the four clusters across the two cohorts where the percentages within each cluster are from the Tracking Parkinson's cohort.

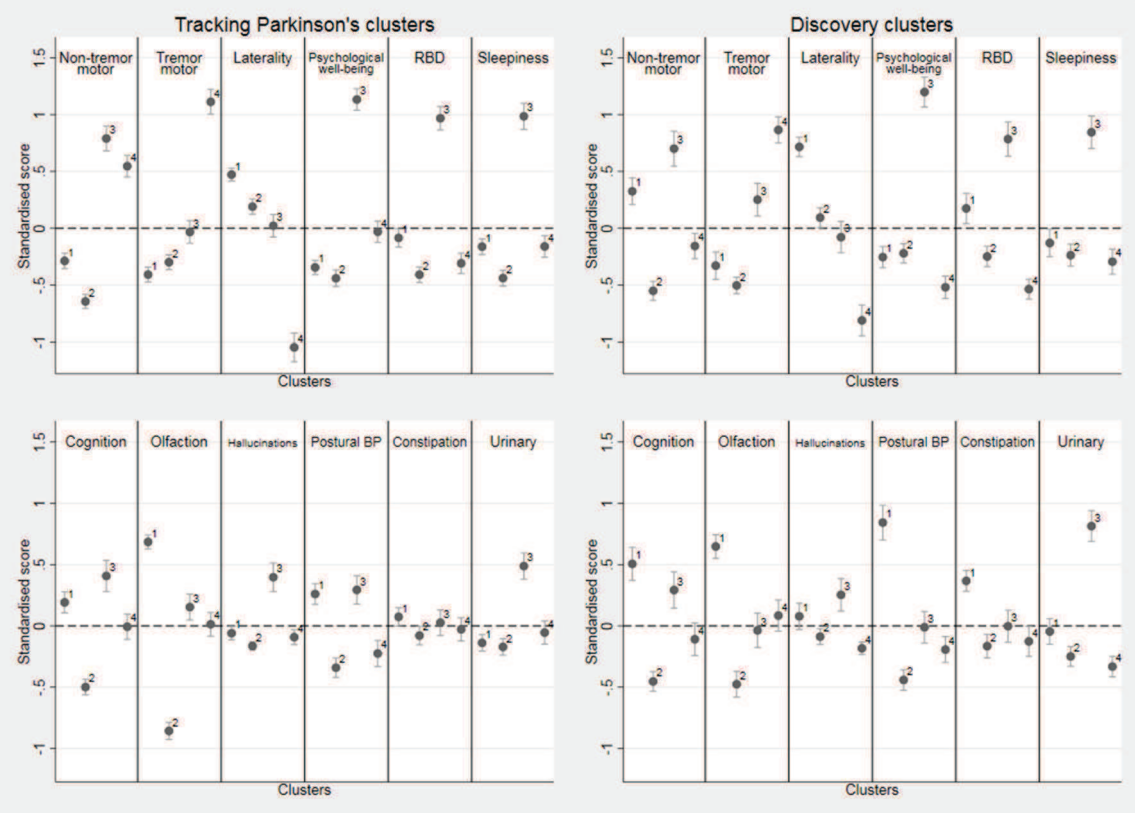


Figure 2 Within cluster means of the standardised variables for the four k-means cluster solution in both cohorts along with the 95% CI for the mean. Positive (above the dotted line) is worse than average and negative better than average. For laterality, positive is more bilateral than average and negative more unilateral than average. Note that hallucinations, constipation and urinary are categorical variables and were standardised using a slightly different method (see online supplementary appendix 1). In Tracking Parkinson's, cluster 1 n=493 patients, cluster 2 n=459, cluster 3 n=336 and cluster 4 n=313, while in Discovery, cluster 1 n=218, cluster 2 n=319, cluster 3 n=196 and cluster 4 n=211. BP, blood pressure; RBD, rapid eye movement sleep behaviour disorder.

especially in non-tremor motor features particularly bradykinesia and postural scores, worse psychological well-being and poor sleep and excessive daytime somnolence. The *slow motor progression* (4) cluster had severe tremor with unilateral disease and was similar in Discovery except for the fact that the non-tremor motor features were better than average in Discovery and worse than average in Tracking Parkinson's.

Web supplementary table 3 shows the agreement between the k-means clusters in Discovery and those predicted by the Tracking Parkinson's discriminant model. This reveals an overall agreement of 67.9% and a kappa value of 0.58 (95% CI 0.54 to 0.61) indicating moderate to substantial agreement.[13] The major inconsistency comes in the *mild motor and non-motor disease* (2) cluster where 110 (34.5%) individuals are wrongly predicted to be in the *fast motor progression* (1) cluster.

Clinical and demographic correlates of the clusters

The focus for the rest of this paper is on the clusters predicted from the larger Tracking Parkinson's model and applied to the Discovery cohort because future patients would be classified from their baseline measurements into predicted clusters. We found a modest difference in disease duration since diagnosis (maximum average difference 3.5 months) between the clusters in both cohorts (table 3) but did not regard this as being clinically important in terms of explaining differences in phenotype. There was evidence of differences in gender, age at diagnosis, motor phenotype, Hoehn and Yahr stage, and medication use at baseline across the four clusters in both cohorts ($p<0.001$ in all variables) (see table 3). The *mild motor and non-motor disease* (2) cluster had the highest proportion of women and youngest age at diagnosis, while the *fast motor progression* (1) cluster had the highest age at diagnosis. The *severe motor disease*, *poor psychological well-being* and *poor sleep* (3) cluster had the highest proportion with the postural instability gait difficulty (PIGD) phenotype while the *slow motor progression* (4) cluster had the highest proportion of tremor-dominant disease at baseline. The LEDD at baseline was highest in the *severe motor disease*, *poor psychological well-being* and *poor sleep* (3) cluster, which also had the smallest proportion of untreated patients.

Within the Tracking Parkinson's cohort, the L-dopa challenge was completed by 1021 (77.8%) out of 1313 patients who have had their 24-month visit. In the Discovery cohort, only 273 (35.5%) out of 770 patients completed the 18-month L-dopa challenge indicating a lack of power and potential selection bias in this data set. The mean percentage decrease in UPDRS III comparing pre with post challenge was greater in Tracking Parkinson's than in Discovery (32.1% vs 23.6%). The change was highest in the *mild motor and non-motor disease* (2) cluster and slightly lower than average in the *slow motor progression* (4) cluster within both cohorts. There was strong evidence of a difference in response to L-dopa across the clusters in Tracking Parkinson's ($p=0.002$), but not so strong in the smaller sample and potentially biased Discovery cohort ($p=0.06$).

Table 3 Association of clusters with variables not used in the cluster analysis, along with a p value derived from a hypothesis test that the variable is equally distributed (ie, same mean or same proportion) among the four clusters

	Tracking Parkinson's clusters						Discovery—clusters predicted from Tracking Parkinson's model					
Variable	P values	Total (N=1601)	Cluster 1 (N=493, 30.8%)	Cluster 2 (N=459, 28.7%)	Cluster 3 (N=336, 21.0%)	Cluster 4 (N=313, 19.6%)	P values	Total (N=944)	Cluster 1 (N=307, 32.5%)	Cluster 2 (N=167, 17.7%)	Cluster 3 (N=223, 23.6%)	Cluster 4 (N=247, 26.2%)
Women*	<0.001	554 (34.6%)	144 (29.2%)	204 (44.4%)	98 (29.2%)	108 (34.5%)	<0.001	334 (35.4%)	92 (30.0%)	87 (52.1%)	58 (26.0%)	97 (39.3%)
Disease duration from diagnosis†	<0.001	1.3 (0.9)	1.3 (0.9)	1.2 (0.9)	1.5 (0.9)	1.4 (0.9)	0.002	1.2 (0.9)	1.2 (0.9)	1.1 (0.9)	1.4 (0.9)	1.2 (0.9)
Age diagnosis†	<0.001	65.9 (9.3)	68.1 (8.1)	62.6 (9.3)	66.5 (9.8)	66.7 (9.2)	<0.001	65.9 (9.6)	67.6 (8.8)	62.7 (9.4)	67.0 (9.5)	65.1 (10.2)
Age diagnosis <50*	<0.001	98 (6.1%)	16 (3.2%)	51 (11.1%)	19 (5.7%)	12 (3.8%)	<0.001	60 (6.4%)	8 (2.6%)	18 (10.8%)	12 (5.4%)	22 (8.9%)
UPDRS motor phenotype*‡												
Tremor dominant	<0.001	741 (48.0%)	194 (40.8%)	241 (54.9%)	92 (28.3%)	214 (70.6%)	<0.001	510 (54.7%)	129 (43.0%)	98 (59.0%)	90 (40.7%)	193 (78.5%)
Indeterminate		196 (12.7%)	61 (12.8%)	54 (12.3%)	41 (12.6%)	40 (13.2%)		115 (12.3%)	44 (14.7%)	22 (13.3%)	28 (12.7%)	21 (8.5%)
Postural instability gait difficulty		606 (39.3%)	221 (46.4%)	144 (32.8%)	192 (59.1%)	49 (16.2%)		308 (33.0%)	127 (42.3%)	46 (27.7%)	103 (46.6%)	32 (13.0%)
Hoehn and Yahr stage*												
0–1.5	<0.001	808 (51.4%)	259 (53.6%)	298 (66.2%)	110 (33.4%)	141 (45.5%)	<0.001	216 (23.0%)	76 (24.9%)	60 (35.9%)	21 (9.5%)	59 (24.0%)
2–2.5		685 (43.6%)	211 (43.7%)	147 (32.7%)	181 (55.0%)	146 (47.1%)		660 (70.2%)	208 (68.2%)	103 (61.7%)	178 (80.2%)	171 (69.5%)
3		79 (5.0%)	13 (2.7%)	5 (1.1%)	38 (11.6%)	23 (7.4%)		64 (6.8%)	21 (6.9%)	4 (2.4%)	23 (10.4%)	16 (6.5%)
Untreated*	<0.001	149 (9.3%)	33 (6.7%)	69 (15.0%)	12 (3.6%)	35 (11.2%)	0.001	119 (12.6%)	35 (11.4%)	28 (16.8%)	14 (6.3%)	42 (17.0%)
LEDD total†	<0.001	293 (205)	304 (195)	245 (202)	361 (204)	272 (203)	<0.001	282 (212)	292 (196)	242 (206)	345 (225)	241 (209)
LEDD total on medication†§	<0.001	324 (190)	327 (183)	289 (188)	375 (195)	309 (188)	<0.001	327 (194)	333 (173)	293 (191)	368 (213)	297 (193)
Levodopa challenge†												
Percentage change	0.002	32.1 (22.8)	30.6 (23.0)	36.3 (24.0)	31.9 (21.7)	28.8 (20.9)	0.06	23.6 (15.2)	22.1 (15.5)	29.4 (16.7)	23.4 (16.0)	22.5 (12.3)

*χ^2 test.
†Analysis of variance.
‡Changed denominator where 80% or more of questions were answered.
§The LEDD restricted to those who are taking dopaminergic medication.
Cluster 1 is *fast motor progression*; cluster 2 is *mild motor and non-motor disease*; cluster 3 is *severe motor disease*, *poor psychological well-being* and *poor sleep*; cluster 4 is *slow motor progression*.
LEDD, levodopa equivalent daily dose; UPDRS, Unified Parkinson's Disease Rating Scale.

Table 4 Comparison of per-year progression rates within the two cohorts using the two approaches: multilevel random slope and intercept models (MLMs) versus pattern-mixture models (PMMs)

		Tracking Parkinson's cohort		Discovery cohort	
	Cluster	MLM slope estimate (95% CI)	PMM slope estimate (95% CI)	MLM slope estimate (95% CI)	PMM slope estimate (95% CI)
MDS-UPDRS III	1	3.16 (2.76 to 3.55)	3.08 (2.70 to 3.45)	2.76 (2.30 to 3.22)	2.66 (2.20 to 3.13)
	2	2.56 (2.18 to 2.95)	2.62 (2.23 to 3.01)	2.25 (1.63 to 2.86)	2.29 (1.72 to 2.87)
	3	2.48 (1.99 to 2.97)	2.66 (2.02 to 3.31)	1.81 (1.26 to 2.37)	1.79 (1.13 to 2.46)
	4	0.61 (0.11 to 1.11)	0.65 (0.09 to 1.21)	1.61 (1.08 to 2.15)	1.67 (1.04 to 2.30)
	P values	<0.001	<0.001	0.007	0.04
MoCA adjusted	1	−0.16 (−0.26 to −0.06)	−0.20 (−0.32 to −0.09)	−0.19 (−0.30 to −0.07)	−0.21 (−0.33 to −0.09)
	2	−0.02 (−0.12 to 0.08)	−0.04 (−0.12 to 0.04)	−0.10 (−0.25 to 0.05)	−0.09 (−0.24 to 0.05)
	3	−0.22 (−0.34 to −0.09)	−0.31 (−0.50 to −0.13)	−0.27 (−0.41 to −0.14)	−0.34 (−0.53 to −0.14)
	4	−0.04 (−0.17 to 0.08)	−0.10 (−0.29 to 0.08)	−0.17 (−0.30 to −0.04)	−0.20 (−0.34 to −0.06)
	P values	0.04	0.017	0.41	0.26
MDS-UPDRS II	1	1.63 (1.46 to 1.81)	1.61 (1.44 to 1.78)	1.43 (1.22 to 1.64)	1.44 (1.21 to 1.67)
	2	1.25 (1.08 to 1.42)	1.32 (1.13 to 1.51)	1.01 (0.73 to 1.28)	0.94 (0.68 to 1.19)
	3	1.51 (1.29 to 1.74)	1.68 (1.33 to 2.02)	1.25 (1.01 to 1.49)	1.41 (1.08 to 1.74)
	4	1.14 (0.92 to 1.37)	1.32 (1.02 to 1.62)	1.25 (0.99 to 1.51)	1.34 (1.04 to 1.63)
	P values	0.001	0.06	0.13	0.02

Cluster 1 is *fast motor progression*; cluster 2 is *mild motor and non-motor disease*; cluster 3 is *severe motor disease*, *poor psychological well-being* and *poor sleep*; cluster 4 is *slow motor progression*.
MDS, Movement Disorders Society; MoCA, Montreal Cognitive Assessment; UPDRS, Unified Parkinson's Disease Rating Scale.

Comparison of prognosis by clusters between Tracking Parkinson's and Discovery

In Tracking Parkinson's, 1421 (88.8%), 1154 (72.1%) and 204 (12.7%) have had 18-month, 36-month and 54-month assessment visits, respectively, with a median follow-up time of 3.0 years (IQR 1.8–3.2). In Discovery, 770 (81.6%), 490 (51.9%), 230 (24.4%) and 39 (4.1%) have had 18-month, 36-month, 54-month and 72-month assessment visits, respectively, with a median follow-up time of 3.0 years (IQR 1.5–4.4). All of the progression rates by cluster and cohort are shown in table 4. There was evidence of a significant difference in progression rates for the UPDRS III across clusters in Tracking Parkinson's ($p<0.001$) and in Discovery ($p=0.007$). The same pattern of was seen in both cohorts. The *fast motor progression* (1) cluster had the fastest progression: 3.2 UPDRS III points per year in Tracking Parkinson's and 2.8 points per year in Discovery, while the *slow motor progression* (4) cluster had the slowest motor progression, although the estimate for progression in the *slow motor progression* (4) cluster was markedly slower in Tracking Parkinson's (0.6 UPDRS III points per year) than Discovery (1.6 points per year) and with hardly any overlap across the 95% CIs (see figure 3). Repeating the analysis using the UPDRS part II score (web supplementary figure 2), we found the same clusters in Tracking Parkinson's with the fastest and slowest progression; however, in the Discovery cohort, we found no evidence of difference in progression rates.

Cognitive decline, as measured by the MoCA, was fastest in the *severe motor disease*, *poor psychological well-being* and *poor sleep* (3) cluster in both cohorts (figure 4), but overall there was no significant difference in cognitive progression rates across clusters (Tracking Parkinson's $p=0.04$; Discovery $p=0.41$).

Repeating our analyses using pattern-mixture models showed little difference in progression rate estimates (table 4), providing evidence that withdrawal has not biased our estimates. Adjusting the slope and intercept for baseline LEDD in our UPDRS III models, an attempt to see the effect that treatment had on progression rates, we found very similar rates (results not included).

No significant differences in motor UPDRS III progression were found between conventional tremor, PIGD and mixed clusters (web supplementary figure 3), although in Tracking ($p<0.001$), there was some evidence to suggest that those in the PIGD cluster have faster cognitive decline (web supplementary figure 4).

DISCUSSION

Our analyses identified four phenotypic subgroups among patients recently diagnosed with PD: (1) *fast motor progression* with symmetrical motor disease, poor olfaction, cognition and postural hypotension; (2) *mild motor and non-motor disease* with intermediate motor progression; (3) *severe motor disease* (prominent bradykinesia/postural impairment), *poor psychological well-being* (mood, apathy, pain, fatigue) and *poor sleep* with intermediate motor progression; (4) *slow motor progression* with tremor-dominant, unilateral disease. The kappa statistic showed that the clusters calculated within the Discovery cohort were relatively stable compared with those predicted using the Tracking Parkinson's cohort model even though some baseline characteristics differed significantly between the cohorts.

Our analysis has taken into account the five points recommended for studies using cluster analysis.[6] (1) Our sample of patients with PD were all recently diagnosed and hence had more similar disease duration than other cross-sectional studies. (2) We used two sample populations of patients who have been well phenotyped across a wide a range of important domains. (3) We have taken into account the limitations of k-means by (a) using hierarchical clustering prior to the analysis to determine the number of clusters, (b) standardising all the variables so they have equal weighting and (c) using 500 random starts to prevent the selection of local optima. (4) We have looked at independent associations between our clusters with clinically meaningful variables such as response to L-dopa challenge and disease progression. (5) We have validated our approach using a second cohort collected using a nearly identical methodology.

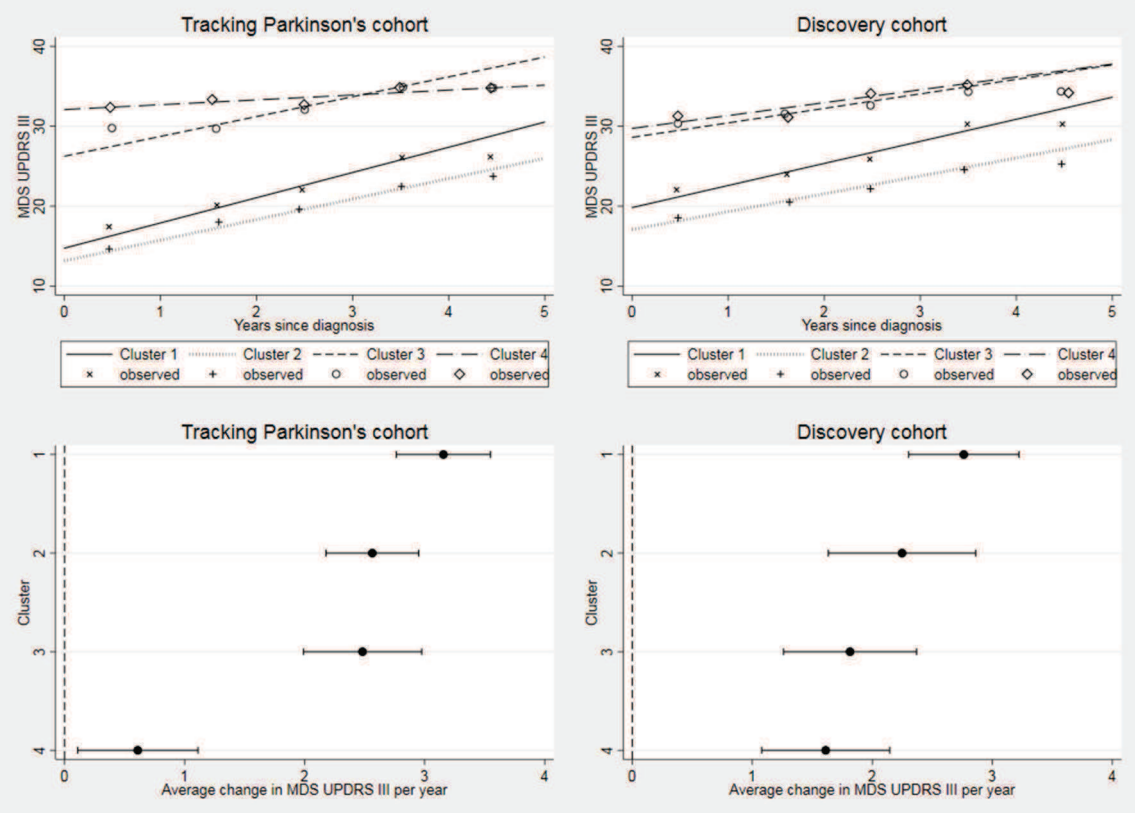


Figure 3 Longitudinal follow-up in MDS-UPDRS part III by cohort. Difference between clusters progression rates p<0.001 in Tracking Parkinson's and p=0.007 in Discovery. Changed denominator where 80% or more of questions were answered. Observed data were split into yearly bins (0–1, 1–2, 2–3, 3–4 and 4–5 years) and the means plotted. MDS, Movement Disorders Society; UPDRS, Unified Parkinson's Disease Rating Scale.

Our previous paper reported five clusters in the Discovery cohort. The clusters in our new analysis are qualitatively similar although two of the original clusters (a) poor psychological well-being, rapid eye movement sleep behaviour disorder and sleep, and (b) severe motor and non-motor disease with poor psychological well-being have now merged into a single cluster (cluster 3). Our clusters are consistent with other similar studies in PD, which generally find a group with milder symptoms and a younger age at onset[3 5 15–22] (our second cluster). Three studies also found a tremor-dominant group[17 18 20] (our fourth cluster) and most studies find a group with more severe symptoms or rapid disease progression[3 4 15–22] (our first and third clusters). Importantly, we have now demonstrated different rates of motor progression across our baseline-defined clusters, with a mean annual deterioration in UPDRS III scores varying significantly from 0.6 to 3.2 points (in Tracking Parkinson's) between those with slowest and fastest progression. Interestingly, we also found, in keeping with another study[3], that poor cognition and postural hypotension predicted faster motor progression.

What is the clinical relevance of these findings?

Stratification, or defining different subcategories, is key to better understanding disease mechanisms and kinetics in PD, predicting disease course and ultimately delivering personalised management strategies. The emerging focus of PD trial design is on early motor disease, including novel immunomodulatory therapies that require intensive and invasive monitoring. Traditionally, little account has been taken of disease heterogeneity in early PD when selecting patients for randomised, placebo-controlled studies. However, our results show that baseline phenotype is associated with variable rates of subsequent motor progression, although confounded by potential medication response effects. The mean difference in UPDRS motor scores between the fastest and slowest motor progression subtypes in Tracking Parkinson's was 2.6 points, equivalent to the primary hierarchical endpoint of several studies, including the ADAGIO study.[23] Recruitment without taking into account heterogeneity and potential sources of recruitment bias may lead to less efficient designs, though there are various trade-offs between the cost of selecting patient subgroups, the sample size required for demonstrating a reduction in disease progression and increasing the length of follow-up.

Strengths and limitations

This study has used two of the largest PD incidence cohorts worldwide. In addition, the methods were designed collaboratively with similar variables being collected using almost identical inclusion criteria, though the source of recruitment differed. While this may impact on the frequency of the subtypes of PD, it should not influence the consistency of the clusters or the within-cluster progression rates. It is possible that some patients will turn out to have other parkinsonian disorders, such as multiple system atrophy, despite only including those with a diagnostic probability of ≥90% at the latest visit, especially in the fast progression cluster. We had little missing data and we used imputation methods to reduce any bias. The association we found with levodopa response (which was analysed as relative change) may simply reflect the fact that the second cluster has milder disease, and since our estimates of motor function is carried out in the 'on' state, we would expect those with mild motor disease to be those who respond well to the medication. We are also limited by the proportion who completed the L-dopa

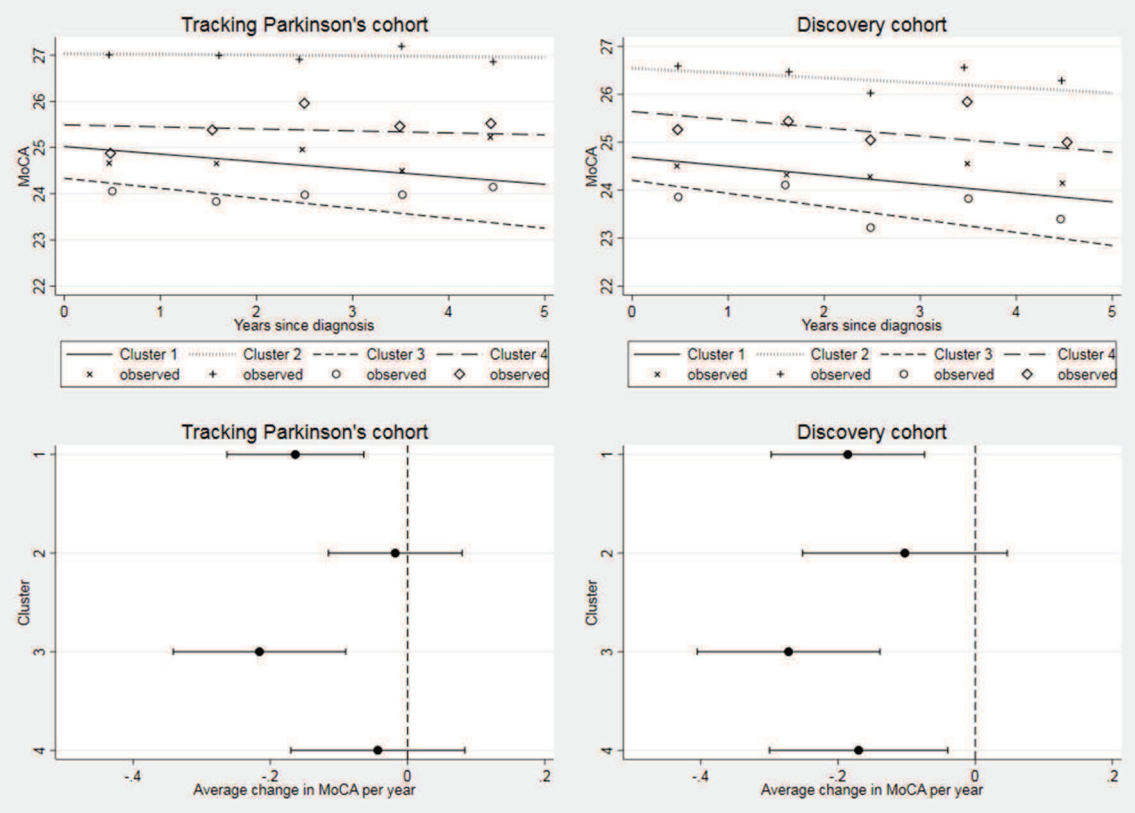


Figure 4 Longitudinal follow-up in Montreal Cognitive Assessment (MoCA) by cohort. Difference between cluster progression rates p=0.04 in Tracking Parkinson's and p=0.41 in Discovery. Changed denominator where 80% or more of questions were answered. Observed data were split into yearly bins (0–1, 1–2, 2–3, 3–4 and 4–5 years) and the means plotted.

challenge although the vast majority of those missing this data in the Tracking Parkinson's cohort is due to them either not taking levodopa as part of their normal medication regime or not reaching the 24-month time point. Levodopa response is also composed of both short-duration and long-duration responses.[24] Our levodopa challenge only measures the short-term response and our pre-dose scores are largely determined by the long-duration response. Also, the long-duration response is typically much larger than the short-duration response.

We used non-statistical criteria to help judge the best number of clusters, as the optimal number of clusters can differ depending on which statistic is the primary focus. Each cohort has its strengths and weaknesses. Tracking Parkinson's is larger with more centres from a wider area of the UK population. The Discovery cohort used fewer clinicians to assess participants and had lower inter-rater variability. Discovery had more disabling disease and slightly worse cognitive function at baseline. Each cohort may have a slightly different mix of patients, but this will also occur in patients recruited for different clinical trials.

The major limitation in this analysis is that most of our data are restricted to the first 3 years of follow-up due to the studies being ongoing and patients not yet reaching 4.5 years of follow-up. We suspect this has reduced our power to detect differences between the clusters. The associations we saw between clusters and progression rates could be due to non-linearity of growth rates; however, non-linearity cannot be tested until the vast majority of patients have four or more observations.

We took a pragmatic perspective where disease progression estimates reflected both pathophysiology and treatment effects. An alternative approach is measurement of the untreated (underlying) progression of subtypes, which reduces potential confounding effects of dopaminergic therapy in modifying disease progression, and has been applied elsewhere.[25 26] Accordingly the generalisability of our method may be limited if different treatment regimes were used in other clinical settings.

Neuropathological characterisation of the patient clusters at post mortem would help to address the question of the distribution and loads of α-synuclein, tau, vascular and amyloid pathology in driving both baseline clinical phenotype and subsequent motor and cognitive progression throughout the disease evolution of PD.[27] It is intriguing to speculate whether patients in cluster 1, who have the fastest motor progression, prominent baseline non-motor symptoms, more symmetrical disease and a poor levodopa response, are defined by prominent cerebrovascular or amyloid pathologies. Clear delineation of patient subtypes is likely to introduce other potential therapeutic targets and lifestyle interventions to the clinical trials arena that look beyond pure α-synuclein-driven pathology. To date, a total of 345 subjects with PD (195 Tracking Parkinson's, 150 Discovery cohort) have signed up to the nationally funded Parkinson's UK Brain Donation programme, with six brains now available for neuropathological characterisation to begin to address these issues.

CONCLUSION

We have found four clusters that replicate across two large independent cohorts of newly diagnosed patients with PD and which are associated with different responses to levodopa and motor progression rates. Future work should examine the reasons for these differences, and with longer follow-up and using growth mixture models, we should be able to identify more easily patient

groups with different progression rates and how this relates to their baseline characteristics. This will also allow us to determine the robustness and clinical use of stratifying patients early in the disease course with better defined endpoints.

Author affiliations
[1]Department of Population Health Sciences, University of Bristol, Bristol, UK
[2]Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, University of Oxford, Oxford, UK
[3]Oxford Parkinson's Disease Centre, University of Oxford, Oxford, UK
[4]Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
[5]Department of Neurology, Institute of Neurological Sciences, Glasgow, UK
[6]Department of Neurology, Queen's Medical Centre, Nottingham, UK
[7]Clinical Neurosciences, John van Geest Centre for Brain Repair, Cambridge, UK
[8]Cardiff University, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff, UK
[9]Faculty of Medical Sciences, Newcastle University, Newcastle, UK
[10]Sobell Department of Motor Neuroscience, UCL Institute of Neurology, London, UK
[11]Department of Clinical Neuroscience, UCL Institute of Neurology, London, UK
[12]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK

Acknowledgements We would like to thank the anonymous reviewers for their useful comments and all patients who have participated in this study.

Contributors ML: analysis and interpretation of the data, writing of the manuscript. YB-S: study concept and design, analysis and interpretation of the data, revision of the manuscript. MTY: analysis and interpretation of the data, revision of the manuscript. FB: acquisition of data, revision of the manuscript. TRB: acquisition of data, revision of the manuscript. JCK: acquisition of data, revision of the manuscript. DMAS: acquisition of data, revision of the manuscript. NM: acquisition of data, revision of the manuscript. KAG: study concept and design, acquisition of data, revision of the manuscript. NB: study concept and design, acquisition of data, revision of the manuscript. RAB: study concept and design, acquisition of data, revision of the manuscript. NW: study concept and design, revision of the manuscript. DJB: study concept and design, acquisition of data, revision of the manuscript. TF: study concept and design, acquisition of data, revision of the manuscript. HRM: study concept and design, acquisition of data, revision of the manuscript. NWW: study concept and design, revision of the manuscript. DGG: study concept and design, acquisition of data, revision of the manuscript. MT-MH: study concept and design, acquisition of data, revision of the manuscript.

Funding The Oxford Discovery study was funded by the Monument Trust Discovery Award from Parkinson's UK and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford, and the NIHR Clinical Research Network: Thames Valley and South Midlands. The Tracking Parkinson's study was funded by Parkinson's UK and supported by the National Institute for Health Research (NIHR) DeNDRoN network, the NIHR Newcastle Biomedical Research Unit based at Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University, and the NIHR-funded Biomedical Research Centre in Cambridge.

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests NB has received payment for advisory board attendance from UCB, Teva Lundbeck, Britannia, GSK, Boehringer and honoraria from UCB Pharma, GE Healthcare, Lily Pharma and Medtronic. He has received research grant support from GE Healthcare, Wellcome Trust, MRC and Parkinson's UK and royalties from Wiley. RAB received grants from Parkinson's UK, NIHR, Cure Parkinson's Trust, Evelyn Trust, Rosetrees Trust, MRC and EU along with payment for advisory board attendance from Oxford Biomedica and LCT, and honoraria from Wiley and Springer. DJB received grants from NIHR, Wellcome Trust, GlaxoSmithKline Ltd, Parkinson's UK and Michael J Fox Foundation. TF received payment for advisory board meetings for Abbvie and Oxford Biomedica, and honoraria for presentations at meetings sponsored by Medtronic, St Jude Medical, Britannia and Teva pharmaceuticals. HRM has received grants from Parkinson's UK, grants from Medical Research Council UK, during the conduct of the study; grants from Welsh Assembly Government, personal fees from Teva, personal fees from Abbvie, personal fees from Teva, personal fees from UCB, personal fees from Boerhinger-Ingelheim, personal fees from GSK, non-financial support from Teva, grants from Ipsen Fund, non-financial support from Medtronic, grants from MNDA, grants from PSP Association, grants from CBD Solutions, grants from Drake Foundation and personal fees from Acorda, outside the submitted work. In addition, HRM has a patent and is a co-applicant on a patent application related to C9ORF72—Method for diagnosing a neurodegenerative disease (PCT/GB2012/052140) pending. DGG received payment for advisory board attendance from AbbVie and honoraria from UCB Pharma, GE Healthcare and Acorda.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Espay AJ, Brundin P, Lang AE. Precision medicine for disease modification in Parkinson disease. *Nat Rev Neurol* 2017;13:119–26.
2 Williams-Gray CH, Mason SL, Evans JR, *et al*. The CamPaIGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. *J Neurol Neurosurg Psychiatry* 2013;84:1258–64.
3 Fereshtehnejad SM, Romenets SR, Anang JB, *et al*. New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. *JAMA Neurol* 2015;72:863–73.
4 Graham JM, Sagar HJ. A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. *Mov Disord* 1999;14:10–20.
5 van Rooden SM, Colas F, Martínez-Martín P, *et al*. Clinical subtypes of Parkinson's disease. *Mov Disord* 2011;26:51–8.
6 van Rooden SM, Heiser WJ, Kok JN, *et al*. The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. *Mov Disord* 2010;25:969–78.
7 Vu TC, Nutt JG, Holford NH. Progression of motor and nonmotor features of Parkinson's disease and their response to treatment. *Br J Clin Pharmacol* 2012;74:267–83.
8 Latourelle JC, Beste MT, Hadzi TC, *et al*. Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol* 2017;16:908–16.
9 Lawton M, Baig F, Rolinski M. Parkinson's disease subtypes in the Oxford Parkinson Disease Centre (OPDC) discovery cohort. *J Parkinsons Dis* 2015;5:269–79.
10 Malek N, Swallow DM, Grosset KA. Tracking Parkinson's: study design and baseline patient data. *J Parkinsons Dis* 2015;5:947–59.
11 Szewczyk-Krolikowski K, Tomlinson P, Nithi K, *et al*. The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. *Parkinsonism Relat Disord* 2014;20:99–105.
12 Tomlinson CL, Stowe R, Patel S, *et al*. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov Disord* 2010;25:2649–53.
13 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
14 Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol Methods* 1997;2:64–78.
15 Erro R, Vitale C, Amboni M, *et al*. The heterogeneity of early Parkinson's disease: a cluster analysis on newly diagnosed untreated patients. *PLoS One* 2013;8:e70244.
16 Gasparoli E, Delibori D, Polesello G, *et al*. Clinical predictors in Parkinson's disease. *Neurol Sci* 2002;23(Suppl 2):s77–78.
17 Lewis SJ, Foltynie T, Blackwell AD, *et al*. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *J Neurol Neurosurg Psychiatry* 2005;76:343–8.
18 Liu P, Feng T, Wang YJ, *et al*. Clinical heterogeneity in patients with early-stage Parkinson's disease: a cluster analysis. *J Zhejiang Univ Sci B* 2011;12:694–703.
19 Post B, Speelman JD, de Haan RJ, *et al*. Clinical heterogeneity in newly diagnosed Parkinson's disease. *J Neurol* 2008;255:716–22.
20 Reijnders JS, Ehrt U, Lousberg R, *et al*. The association between motor subtypes and psychopathology in Parkinson's disease. *Parkinsonism Relat Disord* 2009;15:379–82.
21 Schrag A, Quinn NP, Ben-Shlomo Y. Heterogeneity of Parkinson's disease. *J Neurol Neurosurg Psychiatry* 2006;77:275–6.
22 Fereshtehnejad SM, Zeighami Y, Dagher A, *et al*. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain* 2017;140:1959–76.
23 Rascol O, Fitzer-Attas CJ, Hauser R, *et al*. A double-blind, delayed-start trial of rasagiline in Parkinson's disease (the ADAGIO study): prespecified and post-hoc analyses of the need for additional therapies, changes in UPDRS scores, and non-motor outcomes. *Lancet Neurol* 2011;10:415–23.
24 Chan PL, Nutt JG, Holford NH. Modeling the short- and long-duration responses to exogenous levodopa and to endogenous levodopa production in Parkinson's disease. *J Pharmacokinet Pharmacodyn* 2004;31:243–68.
25 Chan PL, Nutt JG, Holford NH. Levodopa slows progression of Parkinson's disease: external validation by clinical trial simulation. *Pharm Res* 2007;24:791–802.
26 Holford NH, Chan PL, Nutt JG, *et al*. Disease progression and pharmacodynamics in Parkinson disease—evidence for functional protection with levodopa and other treatments. *J Pharmacokinet Pharmacodyn* 2006;33:281–311.
27 Selikhova M, Williams DR, Kempster PA, *et al*. A clinico-pathological study of subtypes in Parkinson's disease. *Brain* 2009;132(Pt 11):2947–57.

E-appendix

Developing and validating Parkinson's Disease Subtypes and their motor and cognitive progression

E-appendix Methods

Patient evaluation

The Tracking Parkinson's cohort began measuring olfaction using the University of Pennsylvania Smell Identification Test (UPSIT) before changing to the Sniffin' sticks 16 item odour identification test when there became a difficulty in obtaining the UPSIT kits. The Discovery cohort only measured olfaction using the Sniffin' sticks 16 item odour identification test. We used IRT methods to convert the UPSIT scores to equivalent Sniffin' 16 scores[1]. We also used equipercentile methods to convert Leeds Anxiety and Depression scale into the more commonly used Hospital Anxiety and Depression scale.

We consider the L-dopa challenge as a percentage change by dividing the difference in pre and post dose total MDS-UPDRS III score measurements by the pre dose measure. A pragmatic levodopa challenge test was performed only in consenting patients who were already taking levodopa medication by the time of their 18 month (Discovery) or 24 month (Tracking Parkinson's) visit. Patients were asked to omit their usual levodopa dose approximately 12 hours before the morning challenge test. Patients who were also taking levodopa agonist medication, MAO-B or COMT inhibitors were also asked to omit these 12 hours before the challenge test (or 24 hours before if taking once daily dopamine agonist formulations). During the levodopa challenge, the patient was given their usual dose of oral levodopa with peripheral dopa-decarboxylase inhibitor, rather than a supramaximal standard dose of levodopa, and the MDS-UPDRS III performed at baseline and 1 hour later to assess response by a trained neurologist.

Statistical Analysis

Since k-means cluster analysis is not a statistical model per-say (it does not measure the uncertainty in any model estimates as it is just an algorithm) so it is not possible to use Rubin's rules to collate the 10 imputed datasets into one model. So for simplicity we used the data from our 10 multiply imputed datasets to create one single dataset (after carrying out the confirmatory factor analysis which is a statistical model) by taking the average for each variable across all 10 datasets. Also note that the amount of missing data we had was small and unlikely to bias our results in anyway. After taking into account those individuals who answered between 80-100% of a questionnaire in Tracking Parkinson's we had between 0.9%-5.3% missing baseline data (although the BFI and Sniffin' scores had ~10% missing data because they were collected at six months post baseline and were hence affected by drop-out) whilst in Discovery we had between 0.4% - 4.8% missing data.

Our factor analysis consisted of first an exploratory factor analysis in the Discovery cohort followed by a confirmatory factor analysis (CFA) in the Tracking Parkinson's cohort. In the CFA we examined the following goodness of fit statistics: Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA). A model was considered to fit the data well if CFI was ≥ 0.90, TLI ≥ 0.90 and RMSEA ≤ 0.08 The same algorithm that produced the factor scores in Tracking Parkinson's was used in Discovery to ensure comparability between these variables across the cohorts.

Pattern-mixture models were estimated using a Structural Equation Modelling approach within Mplus. We constrained all the variances of the outcomes equal across the clusters and

time-points. The variances and covariance of the latent (random) intercept and slope were constrained to be equal across the clusters. Our pattern-mixture model was defined such that all patients withdrawing were considered the same, i.e withdrawing after visit 2 was considered to have the same effect on the intercept and slope as withdrawing after baseline. The numbers withdrawn in Tracking Parkinson's was 302/1601 (18.9%) and within Discovery 233/944 (24.7%).

E-appendix Results

Exploratory factor analysis (EFA)

In Discovery using the eigenvalue criteria we found 2 factors in each of the ten imputed datasets. The two factors identified were identical to the psychological well-being and non-tremor motor factors from the original paper EFA [2] (ignoring unavailable variables not in both datasets, that is the Purdue tests, the Get Up and Go Test and the flamingo test) except that the non-tremor motor factor also included the MoCA and semantic fluency variables.

Confirmatory factor analysis (CFA)

The CFA in the Tracking Parkinson's cohort using the variables from the EFA only met one of our pre-defined goodness of fit criteria with a CFI of 0.83, TLI of 0.87, and a RMSEA of 0.078. Although the cognitive and motor variables in the second factor might be highly correlated we thought that clinically it did not make sense to include them within the same factor. Dropping the two cognitive variables from this factor improved the goodness of fit with a CFI of 0.91, TLI of 0.93, and a RMSEA of 0.063. Therefore we excluded the cognitive variables from this factor with main manuscript Table 2 displaying the results from the resulting CFA. We named factor 1 "psychological well-being" and factor 2 "non-tremor motor" matching our original paper. The factor loadings varied from 0.31 – 0.86 and 0.42 – 0.77 in the two factors respectively. At this stage we excluded the other four BFI variables not loading into a factor and the UPDRS constipation variable for the sake of parsimony, which is the same approach used in our original paper. We also excluded the semantic fluency variable since we thought that the MoCA was a better measure of global cognitive function.

Cluster Analysis – choosing number of clusters

Web table 1 shows the statistics we used to determine the optimum number of clusters which pointed to the two and five cluster solutions in both Tracking Parkinson's and Discovery. So we used criteria other than these fit statistics to decide which was the optimum number of clusters.

We considered, initially, the agreement between our k-means clusters in the Discovery cohort and the clusters predicted from our Tracking Parkinson's discriminant model. The two cluster solution had excellent overall agreement (91.5%) and a kappa statistic consistent with "almost perfect" agreement (0.83) however because this only stratified individuals into a good and bad group this was not regarded as clinically that informative. The five cluster solution had the same overall agreement (67.9%) as the four cluster solution and a higher kappa statistic 0.60 compared to 0.58. However these kappa statistics are almost equivalent and both close to the borders of what would be considered "moderate" to "substantial" agreement. We chose the four cluster solution because it is more parsimonious than the five cluster.

Comparison of prognosis by clusters between Tracking Parkinson's and Discovery

We looked at what would happen if we relaxed the assumptions in our pattern-mixture model such that variances of the outcomes are equal across time-points or clusters and the variances/covariances of the random effects are equal across clusters. So we also fitted the

following models for UPDRS III and compared commonly used goodness of fit statistics like AIC, BIC as well as likelihood ratio tests.

1. Variances and covariance of random effects are different for the four clusters
2. Variances of the outcomes are different at each time-point
3. Combining assumptions for models 1 and 2 above.
4. As model 3 but variances of the outcomes are also different within each cluster

Using the goodness of fit statistics in Tracking PD we would select model 3 over our standard PMM model. However within all models the largest difference in mean progression rate (compared to standard PMM model) in any cluster was 0.22 UPDRS III points per year so almost identical to our default model. Using the same model (which was also favoured by AIC and likelihood ratio tests) in Discovery the largest difference in mean progression rate in any cluster was 0.12 UPDRS III points per year and the difference between clusters p-value increased from 0.04 to 0.10. However if we used the BIC to select a model in Discovery we would have selected the default model. Hence we are confident that the assumptions we made in our model has not made any impact on our progression rate estimates. We will further explore these issues along with non-linearity in a future paper when we have more longitudinal data available.

Web Table 1. Statistics to determine the number of clusters from the Ward hierarchical clustering. A higher value of Calinski/Harabasz pseudo-F index indicates more distinct clustering and a smaller value of the Duda/Hart pseudo-T squared indicates more distinct clustering. Bold indicates most distinct cluster.

	Tracking Parkinson's cohort		Discovery cohort	
Number of clusters	Calinski/Harabasz pseudo-F	Duda/Hart pseudo T-squared	Calinski/Harabasz pseudo-F	Duda/Hart pseudo T-squared
2	**163.4**	75.3	**96.2**	32.3
3	116.1	73.2	69.6	39.6
4	99.8	42.4	58.9	26.0
5	92.4	**37.2**	53.8	**18.1**

Web Table 2. Scores from each test within the four k-means clusters from both cohorts at baseline using the imputed data. Where standard cut-points exist in literature categorised scores are given as well as their total test scores. For standalone questions from the MDS-UPDRS scales the questions were dichotomised at 1 or above.

	Tracking Parkinson's clusters					Discovery clusters				
Variable	**Total N=1601**	**Cluster 1 N=493**	**Cluster 2 N=459**	**Cluster 3 N=336**	**Cluster 4 N=313**	**Total N=944**	**Cluster 1 N=218**	**Cluster 2 N=319**	**Cluster 3 N=196**	**Cluster 4 N=211**
UPDRS speech[a]	782 (48.8%)	251 (50.9%)	136 (29.6%)	233 (69.3%)	162 (51.8%)	445 (47.1%)	126 (57.8%)	95 (29.8%)	109 (55.6%)	115 (54.5%)
UPDRS rigidity	3.7 (2.9)	3.1 (2.7)	2.6 (2.2)	4.8 (3.2)	5.1 (2.9)	5.3 (2.7)	6.0 (2.9)	4.4 (2.3)	6.2 (2.8)	5.2 (2.5)
UPDRS bradykinesia	10.8 (7.0)	8.9 (5.8)	7.0 (4.9)	15.0 (7.3)	15.0 (6.5)	13.0 (6.5)	15.0 (6.1)	9.9 (5.2)	16.3 (6.9)	12.6 (5.9)
UPDRS Postural	2.6 (2.3)	2.2 (1.8)	1.6 (1.6)	3.8 (2.7)	3.2 (2.4)	2.6 (2.2)	3.1 (2.1)	1.7 (1.6)	3.8 (2.8)	2.1 (1.7)
UPDRS tremor	4.6 (3.8)	3.1 (2.9)	3.5 (2.7)	4.5 (3.5)	8.9 (3.8)	5.0 (3.8)	3.8 (3.4)	3.1 (2.5)	5.9 (3.8)	8.3 (3.2)
Percentage UPRDS III due to tremor	21.8 (17.6)	18.7 (18.7)	24.7 (20.2)	16.3 (12.9)	28.4 (12.0)	19.4 (14.2)	13.1 (11.5)	16.9 (14.8)	18.9 (11.7)	30.3 (11.8)
Laterality	6.4 (4.4)	4.3 (2.8)	5.5 (3.2)	6.3 (4.1)	11.0 (5.0)	7.1 (4.3)	4.0 (2.8)	6.7 (3.4)	7.4 (4.2)	10.5 (4.3)
Laterality (dichotomised unilateral[b])	1154 (72.1%)	290 (58.8%)	325 (70.8%)	245 (72.9%)	294 (93.9%)	725 (76.8%)	107 (49.1%)	257 (80.6%)	160 (81.6%)	201 (95.3%)
UPDRS apathy[a]	499 (31.2%)	123 (24.9%)	98 (21.4%)	184 (54.8%)	94 (30.0%)	194 (20.6%)	32 (14.7%)	43 (13.5%)	91 (46.4%)	28 (13.3%)
UPDRS fatigue[a]	1242 (77.6%)	365 (74.0%)	327 (71.2%)	307 (91.4%)	243 (77.6%)	670 (71.0%)	155 (71.1%)	220 (69.0%)	184 (93.9%)	111 (52.6%)
UPDRS pain[a]	894 (55.8%)	233 (47.3%)	220 (47.9%)	270 (80.4%)	171 (54.6%)	753 (79.8%)	173 (79.4%)	247 (77.4%)	179 (91.3%)	154 (73.0%)
BFI neuroticism	23.3 (6.4)	22.1 (5.8)	22.1 (6.5)	26.4 (6.3)	23.7 (6.3)	22.3 (6.5)	20.9 (6.6)	21.6 (5.9)	26.4 (6.0)	20.9 (6.3)
HADS anxiety	5.3 (4.2)	3.9 (3.0)	4.0 (3.5)	9.3 (4.3)	5.1 (3.8)	4.6 (3.8)	3.4 (2.8)	4.1 (3.1)	8.2 (4.2)	3.1 (2.8)
HADS anxiety (dichotomised[c])	442 (27.6%)	66 (13.4%)	76 (16.6%)	215 (64.0%)	85 (27.2%)	197 (20.9%)	21 (9.6%)	50 (15.7%)	109 (55.6%)	17 (8.1%)
HADS depression	4.6 (3.4)	3.6 (2.6)	3.2 (2.6)	8.2 (3.1)	4.4 (2.9)	4.4 (3.4)	3.7 (2.6)	3.8 (2.8)	8.0 (3.4)	2.8 (2.5)

HADS depression (dichotomised[c])	310 (19.4%)	37 (7.5%)	29 (6.3%)	194 (57.7%)	50 (16.0%)	168 (17.8%)	15 (6.9%)	34 (10.7%)	102 (52.0%)	17 (8.1%)
QUIP	350 (21.9%)	89 (18.1%)	87 (19.0%)	113 (33.6%)	61 (19.5%)	204 (21.6%)	30 (13.8%)	71 (22.3%)	74 (37.8%)	29 (13.7%)
RBD	4.8 (3.1)	4.5 (2.8)	3.5 (2.3)	7.8 (3.0)	3.8 (2.5)	4.8 (3.1)	5.3 (3.1)	4.0 (2.5)	7.2 (3.2)	3.2 (2.0)
RBD (dichotomised[d])	714 (44.6%)	213 (43.2%)	119 (25.9%)	283 (84.2%)	99 (31.6%)	422 (44.7%)	119 (54.6%)	111 (34.8%)	148 (75.5%)	44 (20.9%)
ESS	6.8 (4.5)	6.0 (3.5)	4.8 (3.3)	11.1 (4.8)	6.0 (3.8)	7.6 (4.5)	7.0 (4.0)	6.5 (3.9)	11.3 (4.5)	6.3 (3.6)
ESS (dichotomised[e])	295 (18.4%)	48 (9.7%)	28 (6.1%)	180 (53.6%)	39 (12.5%)	230 (24.4%)	41 (18.8%)	51 (16.0%)	113 (57.7%)	25 (11.8%)
MoCA	25.4 (3.4)	24.7 (3.3)	27.1 (2.3)	24.0 (4.0)	25.4 (3.1)	25.0 (3.3)	23.3 (3.3)	26.5 (2.4)	24.0 (3.5)	25.4 (3.2)
MoCA[f]- Normal	1211 (75.6%)	339 (68.8%)	428 (93.2%)	208 (61.9%)	236 (75.4%)	671 (71.1%)	113 (51.8%)	283 (88.7%)	112 (57.1%)	163 (77.3%)
MoCA[f]- MCI	183 (11.4%)	77 (15.6%)	18 (3.9%)	46 (13.7%)	42 (13.4%)	145 (15.4%)	45 (20.6%)	27 (8.5%)	47 (24.0%)	26 (12.3%)
MoCA[f]- Demented	207 (12.9%)	77 (15.6%)	13 (2.8%)	82 (24.4%)	35 (11.2%)	128 (13.6%)	60 (27.5%)	9 (2.8%)	37 (18.9%)	22 (10.4%)
Sniffin'	7.7 (2.9)	5.7 (1.9)	10.2 (2.2)	7.2 (2.9)	7.6 (2.5)	7.1 (2.9)	5.3 (2.1)	8.5 (2.7)	7.2 (2.8)	6.9 (2.7)
Sniffin' – hyposmic[g]	1180 (73.7%)	439 (89.0%)	251 (54.7%)	256 (76.2%)	234 (74.8%)	743 (78.7%)	193 (88.5%)	232 (72.7%)	143 (73.0%)	175 (82.9%)
UPDRS hallucinations[a]	141 (8.8%)	31 (6.3%)	11 (2.4%)	83 (24.7%)	16 (5.1%)	105 (11.1%)	32 (14.7%)	23 (7.2%)	44 (22.4%)	6 (2.8%)
Systolic postural drop	4.2 (13.6)	7.8 (13.1)	-0.4 (11.8)	8.3 (14.7)	1.2 (13.1)	6.3 (15.9)	19.6 (16.8)	-0.7 (12.1)	6.1 (14.5)	3.2 (12.4)
Constipation[h]	529 (33.0%)	176 (35.7%)	115 (25.1%)	144 (42.9%)	94 (30.0%)	459 (48.6%)	153 (70.2%)	117 (36.7%)	103 (52.6%)	86 (40.8%)
UPDRS urinary[a]	942 (58.8%)	262 (53.1%)	243 (52.9%)	261 (77.7%)	176 (56.2%)	594 (62.9%)	141 (64.7%)	168 (52.7%)	177 (90.3%)	108 (51.2%)

[a]Dichotomised individual UPDRS questions at 1 or more
[b]Dichotomised Laterality at a difference of four or more between left side and right side
[c]Dichotomised HADS at 8 or more
[d]Dichotomised RBD at 5 or more
[e]Dichotomised ESS at 11 or more

[f]Categorised MoCA at ≤ 21 = Dementia, 22-23 = MCI, or 24+ = Normal
[g]Dichotomised Sniffin' at or below the 15th centile by age and gender group
[h]Dichotomised Constipation <1 bowel movement per day or laxative use

Web Table 3. Agreement of patients in Discovery classified in each cluster: Kmeans on Discovery vs predicted clusters from Tracking Parkinson's discriminant analysis model

	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Kmeans 1	160	0	44	14
Kmeans 2	110	154	22	33
Kmeans 3	6	3	157	30
Kmeans 4	31	10	0	170
Overall agreement 67.9%				
Kappa (95% CI): 0.58 (0.54, 0.61)				

WEB FIGURE LEGENDS

Web figure 1. Flow chart for entry into this analysis

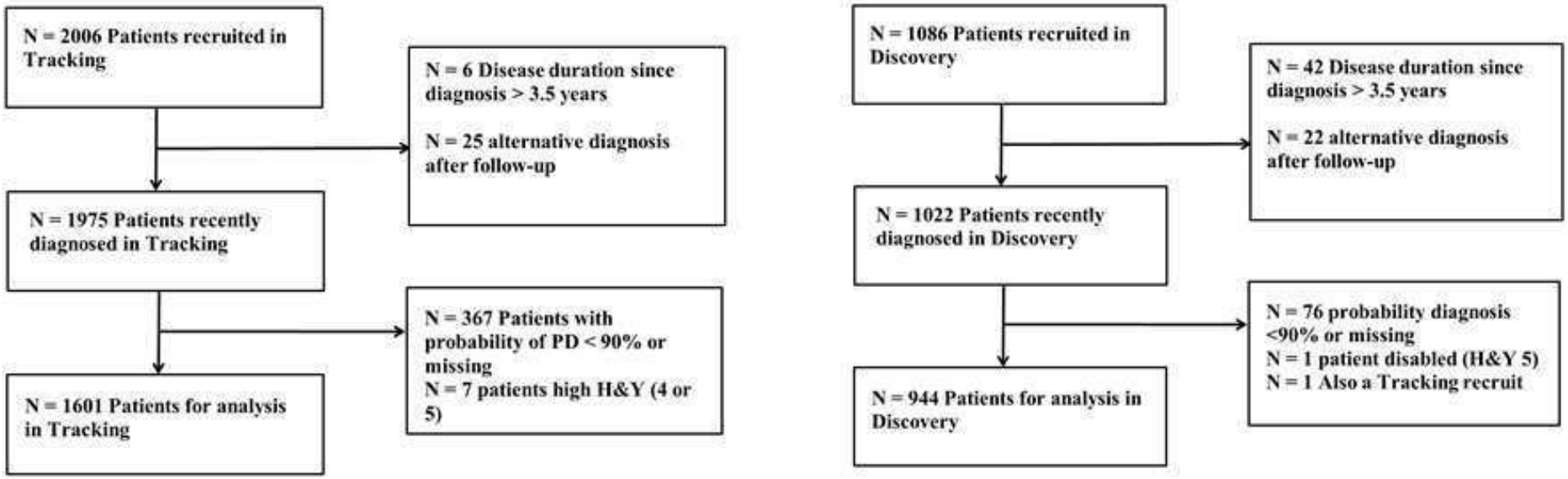
Web Figure 2. Longitudinal follow up in MDS-UPDRS part II by cohort Difference between clusters progression rates p=0.001 in Tracking Parkinson's and p=0.13 in Discovery. Changed denominator where 80% or more of questions were answered Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.

Web Figure 3. Longitudinal follow up in MDS-UPDRS part III by cohort looking at conventional clusters (TD, PIGD, mixed). Difference between clusters progression rate p=0.21 in Tracking Parkinson's and p=0.95 in Discovery. Changed denominator where 80% or more of questions were answered. Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.

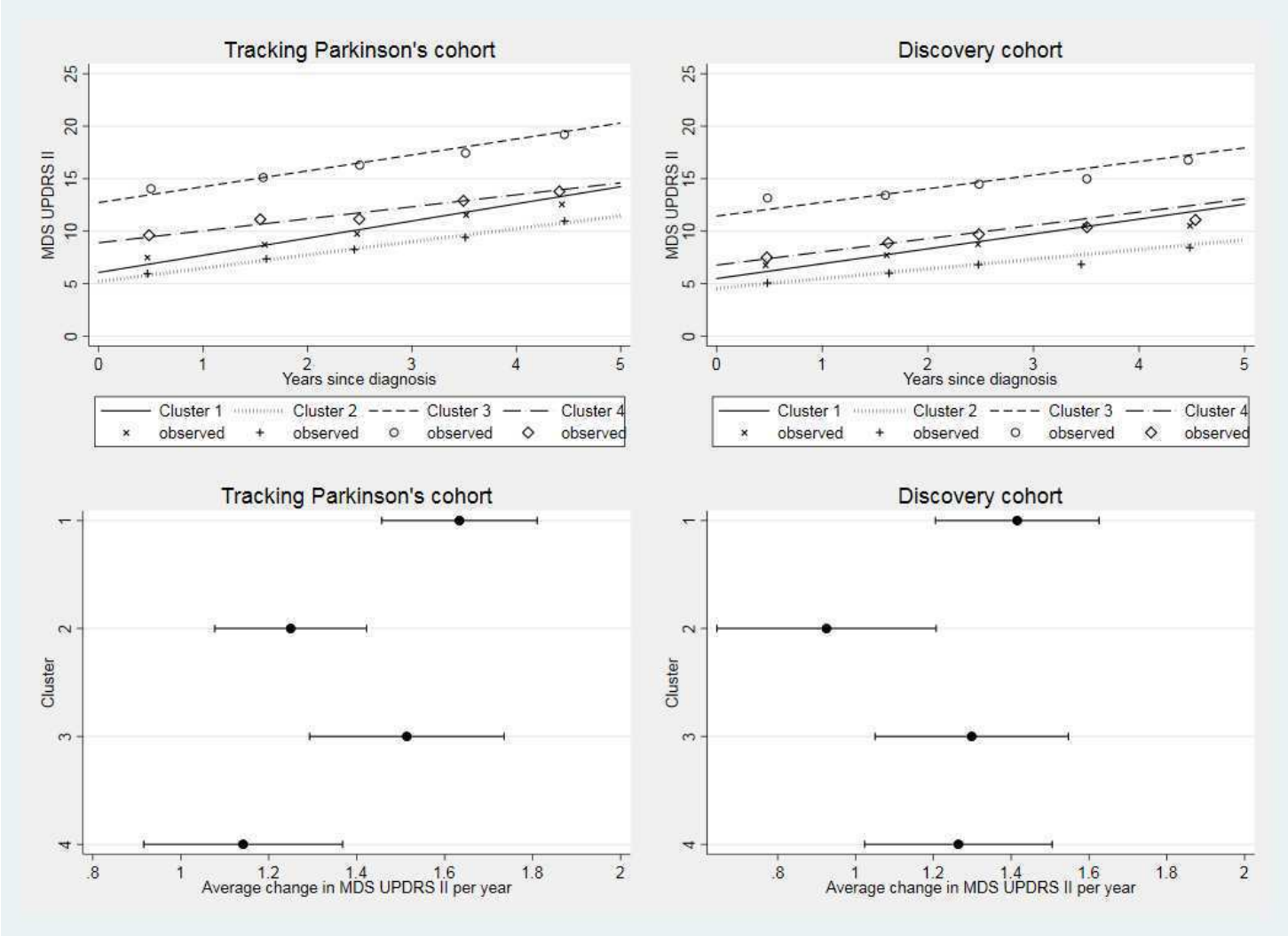
Web Figure 4. Longitudinal follow up in MoCA by cohort looking at conventional clusters (TD, PIGD, mixed). Difference between clusters progression rate p<0.001 in Tracking Parkinson's and p=0.33 in Discovery. Changed denominator where 80% or more of questions were answered. Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.

eReferences

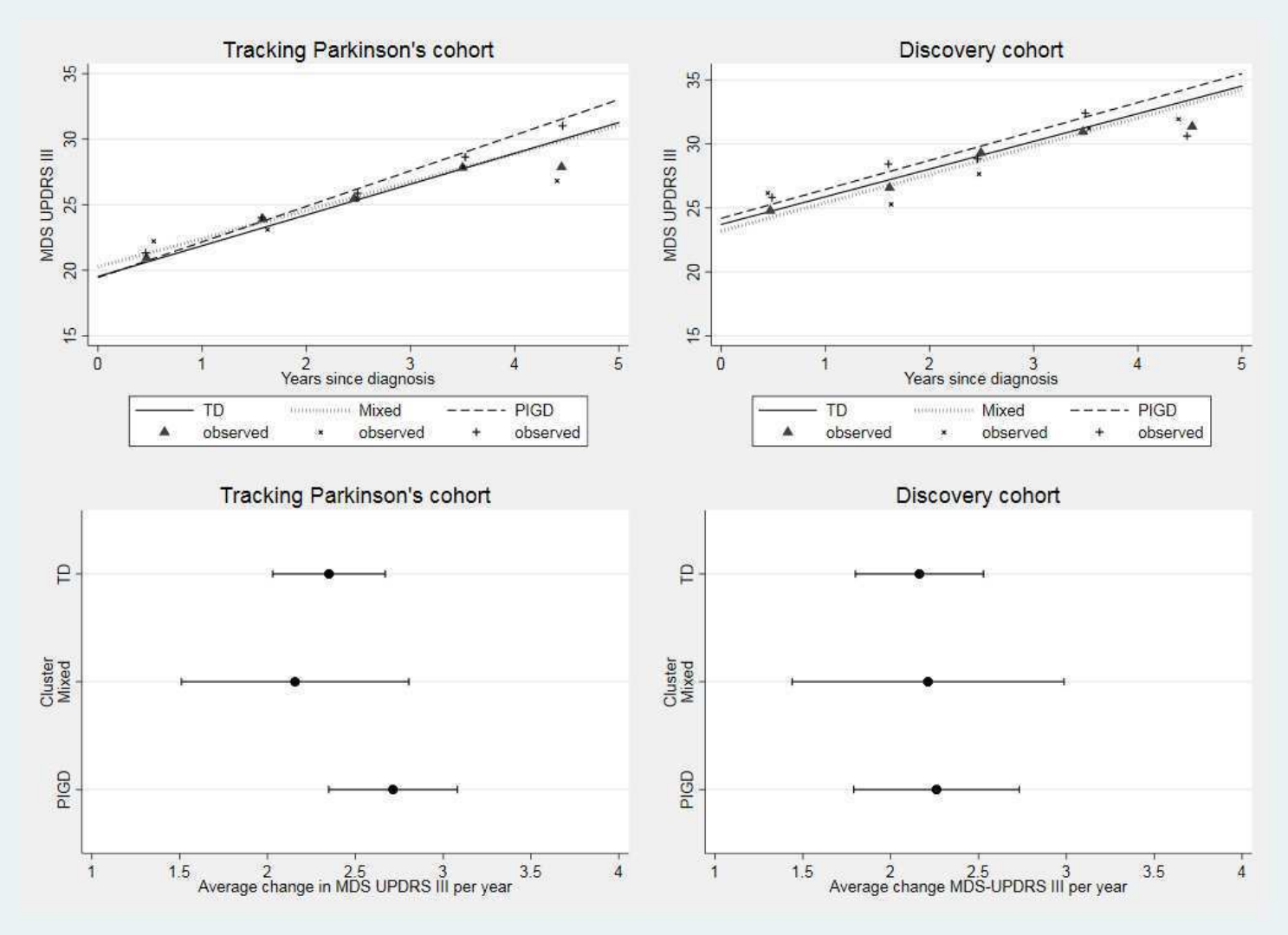
1. Lawton M, Hu MT, Baig F, et al. Equating scores of the University of Pennsylvania Smell Identification Test and Sniffin' Sticks test in patients with Parkinson's disease. *Parkinsonism Relat Disord.* 2016.
2. Lawton M, Baig F, Rolinski M, et al. Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort. *J Parkinsons Dis.* 2015;5(2):269-279.



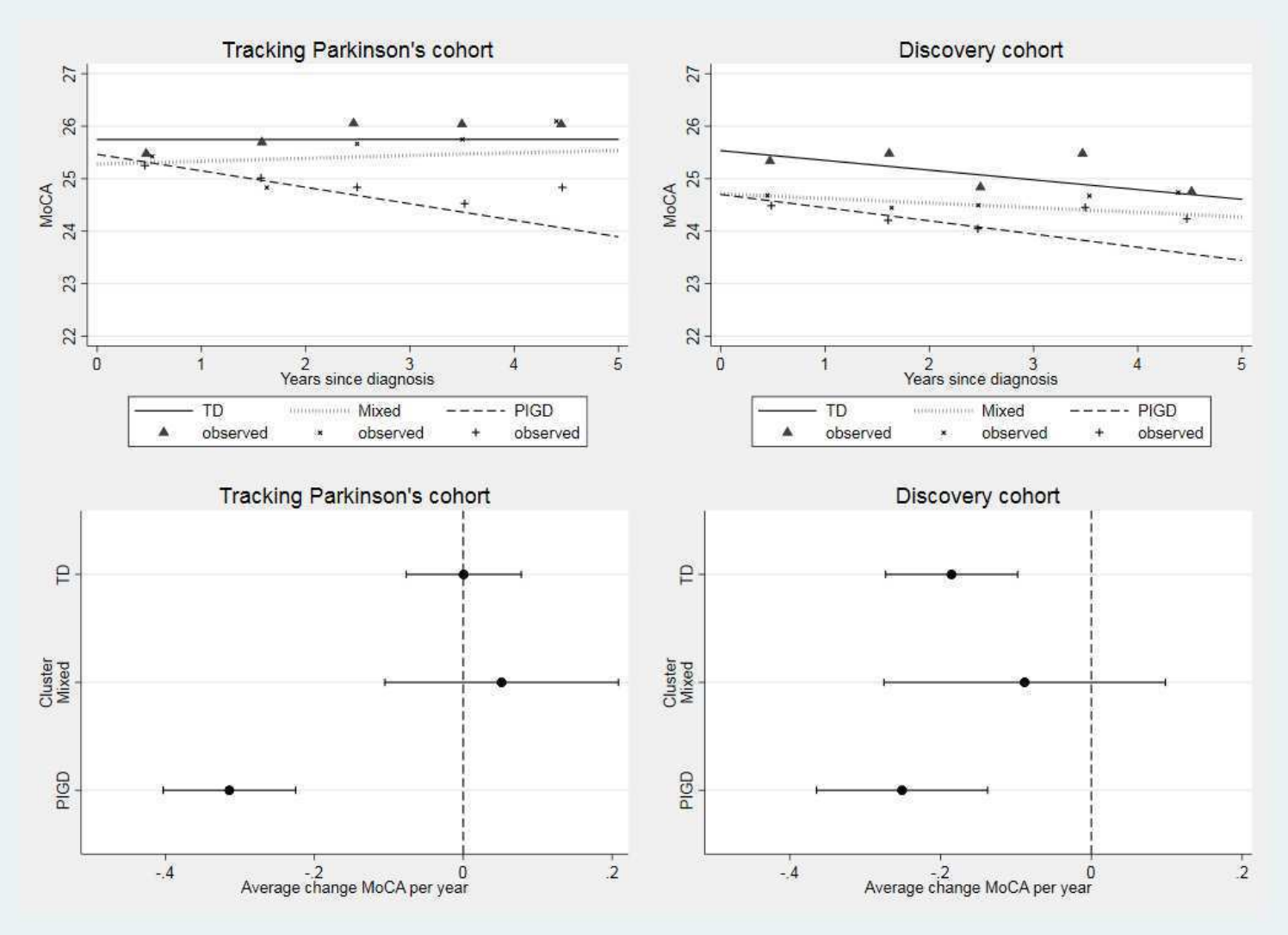
Web figure 1. Flow chart for entry into this analysis



Web Figure 2. Longitudinal follow up in MDS-UPDRS part II by cohort Difference between clusters progression rates p=0.001 in Tracking Parkinson's and p=0.13 in Discovery. Changed denominator where 80% or more of questions were answered Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.



Web Figure 3. Longitudinal follow up in MDS-UPDRS part III by cohort looking at conventional clusters (TD, PIGD, mixed). Difference between clusters progression rate p=0.21 in Tracking Parkinson's and p=0.95 in Discovery. Changed denominator where 80% or more of questions were answered. Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.



Web Figure 4. Longitudinal follow up in MoCA by cohort looking at conventional clusters (TD, PIGD, mixed). Difference between clusters progression rate $p<0.001$ in Tracking Parkinson's and $p=0.33$ in Discovery. Changed denominator where 80% or more of questions were answered. Observed data was split into yearly bins (0-1,1-2,2-3,3-4 and 4-5 years) and the means plotted.