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

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

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]

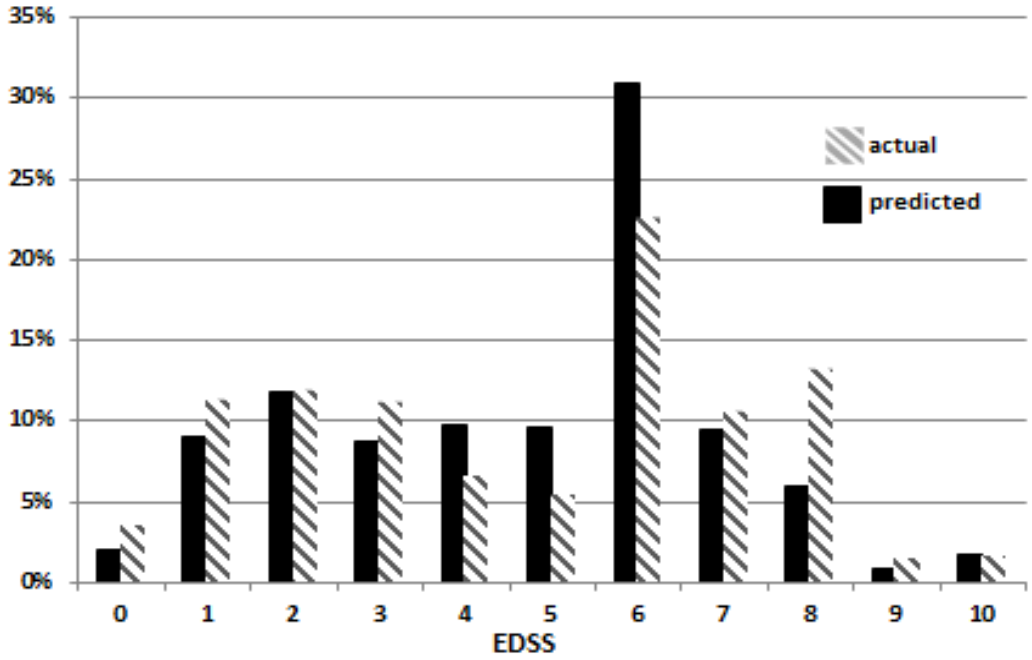

Figure A2: Variation of treatment effect with time - projected EDSS progression applying the multilevel model first to the RSS cohort and then to the untreated comparator control group

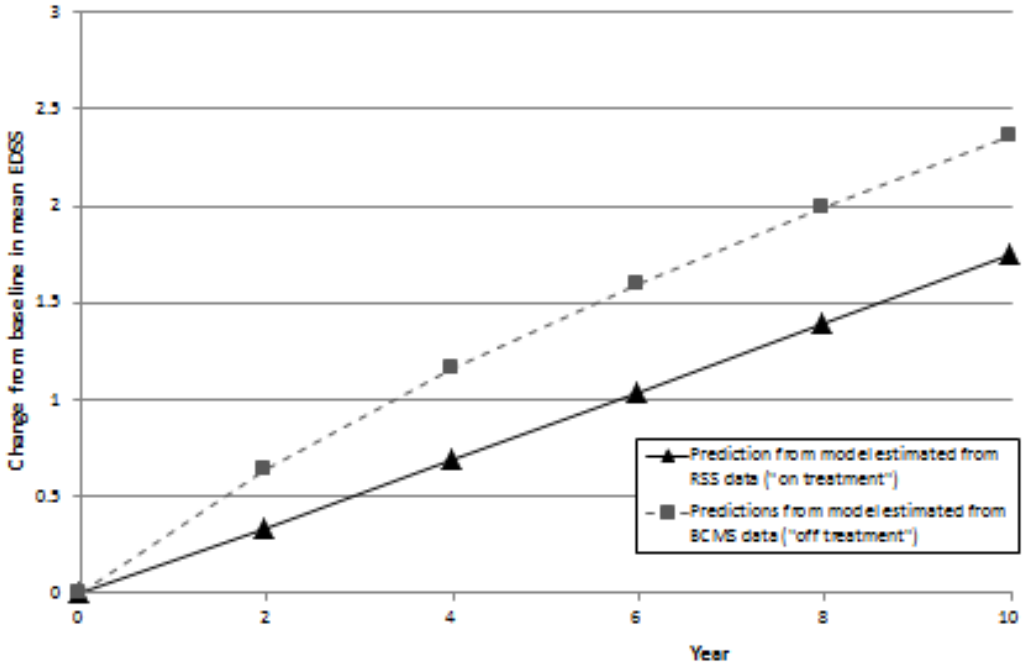

Figure A3: Variation of implied hazard ratios with time using the Markov model, for both utility and EDSS disability outcomes