Appendix 1 EDSS data extraction from the BCMS database

For the discrete Markov model and for tabular display of annual data the EDSS data were extracted and processed as follows. If the baseline EDSS fell when it was not the ‘usual’ yearly visit period, then subsequent EDSS scores could be lost. To maximize the number of EDSS transitions per patient, but keeping the yearly (+/- 3 calendar months) interval, EDSS scores were also ‘individualized’ as follows: the baseline EDSS became the baseline year, within that year, we searched to find the optimal new, individualized baseline for which the patient would, over the coming years have the most number of yearly EDSS scores. For some patients, this new individualized baseline date would not coincide with a clinic visit and would therefore not have an EDSS score present. When a baseline EDSS was required (e.g. for the discrete Markov model), data was also ‘shifted’ such that each patient would have a baseline EDSS. This lag between baseline (‘eligibility’) and EDSS was considered consistent with clinical practice in that it is not unusual to have a lag time between a patient becoming eligible for treatment and actual treatment initiation.

Appendix 2 Algorithms to forecast EDSS distributions at any given time

The ‘msm’ algorithm allows the EDSS distribution to be forecasted at any time t.19 To define what is the actual EDSS at a given time t (i.e. not necessarily when an observation was recorded) ‘msm’ offered two variants: (i) the last observation carried forward (LOCF) for each individual patient, and (ii) a ‘midpoint interpolation’ algorithm in which the EDSS state for a given patient at time t was taken as the score closest in time to the actually observed EDSS. Suppose an individual was observed in EDSS states Sr-1 and Sr at two consecutive times tr-1 and tr, and we wanted to estimate ‘observed’ proportions at a time t between tr-1 and tr. LOCF then meant that individuals were assumed to be in state Sr-1 at time t, the same state as they were at tr-1. Midpoint interpolation meant if t ≤ (tr-1 + tr) / 2, the midpoint of tr-1 and tr, the state at t was assumed to be Sr-1., otherwise Sr. Option (i) would be more appropriate if EDSS values were always measured immediately after each transition. Option (ii) would be more appropriate if EDSS values were measured at mixed time intervals (‘fixed or random’). Option (ii) was considered to mimic the clinical setting more closely and is also appears more applicable in progressive diseases.19 Therefore all results presented in this paper were based on this ‘midpoint interpolation’ approach.

For the continuous Markov model we limited the range of transitions which can be regarded as ‘instantaneous’ to +/- 3, which meant that at any time t an instantaneous progression (or improvement) into another EDSS state was only possible when not exceeding three consecutive states (an instantaneous transition from 1 to 2, 3, or 4 was possible, for example, whereas 1 to 5 was not). This restriction was recommended to avoid computationally inefficient modelling of hazard rates which were virtually 0.

Appendix 3 Validation of the models

The following validation techniques were applied when evaluating the different Markov models.
The most straightforward method for the discrete Markov model consisted of applying the transition matrix (and the 2nd, 3rd etc. power) to the vector of the baseline EDSS distribution in the BCMS reference database, calculating the forecasted EDSS distribution for t=1, 2, 3... years and comparing against the actual EDSS. A similar validation was performed in the continuous Markov model, as described by Jackson.1.

As a second validation we divided the BCMS dataset randomly into two subsets of equal size, using one half to derive the model separately for the two subgroups, and then assessing the goodness of fit in the other half of the dataset. Although Jackson19 emphasises that “Assessing the goodness of fit of this class of models [...] is worth further research” we decided to use a classical mean square prediction error (weighted root mean square over years of the prediction error in the average quantity shown, weighted by the number of patients contributing data in the given year) to compare competing models. Moreover, the computed likelihood itself as a result of the maximum likelihood algorithm was used to rank the different one and two covariate models.