PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	UK Multiple Sclerosis Risk-sharing Scheme: a new natural history dataset and an improved Markov model
AUTHORS	Palace, Jacqueline; Bregenzer, Thomas; Tremlett, Helen; Oger, Joel; Zhu, Feng; Boggild, Mike; Duddy, Martin; Dobson, Charles

VERSION 1 - REVIEW

REVIEWER	David Henry Institute for Clinical Evaluative Sciences, ICES
REVIEW RETURNED	01-Nov-2013

GENERAL COMMENTS	Cost-effectiveness analysis is widely used to establish eligibility, usage criteria and prices of drugs. In the case of drugs with high unit acquisition costs and uncertainty about long term outcomes some form of risk sharing between the manufacturer and the purchaser is becoming commoner. These risk sharing agreements seem to take a number of forms ranging from simple (a 'price/volume trade off') to more complex (a price determined by a modeled cost/QALY threshold). This paper is an example of the latter. The UK risk sharing scheme is one of the most prominent examples of the latter. The performance of this scheme will be of great interest internationally. When I first read the title of the paper I had a moment of hope that we might learn something of the internal workings of the scheme and some account of how it actually worked. However, I think that the operational details of this scheme, and others, are not going to be very transparent – despite that fact that it is public money at risk. Therefore, I was a little disappointed to find a technical paper that describes the search for and performance of a population based of untreated MS patients as a basis for improved modeling of outcomes in the UK risk sharing scheme. Transparency in the model is welcome but how this impacts on decision making as it affects eligibility, usage and price is what we need to hear about. As to the description of the model, this is nicely done. The paper is very clearly written, easy to read and relatively easy to understand for a reasonably knowledgeable individual. The figures display the degree of fit well. I think there is sufficient (not excessive) detail. As a non-specialist I can't judge the more technical aspects, but this is a skilled group of collaborators and I don't have serious doubts about what they have done. The Abstract is entirely free of data. I realize that it is hard to provide meaningful data about the model outputs in an Abstract, but some details of the cohort and some description of disease progression in the observed and modeled cohorts would be welcome.

	For me the main issues belong properly with the editors – but I will comment anyway. I don’t think this paper, in its present form, will be of great interest to the majority of readers. Only a small relatively specialized group will care. However the authors could turn this into a more interesting paper for general readers by doing two things: 1) Explaining much more clearly to general readers how the risk sharing scheme could work – accepting that many details may be confidential. They could use simple examples to explain how a higher than acceptable cost/QALY might result in a lower price 2) Can they use their continuous Markov model with age at onset (apparently the best fit) and some indicative costs plus a range of hazard ratios for treatment effect to indicate what prices might be reasonable – under pre-specified assumptions? Again this is illustrative rather than definitive.

REVIEWER	Butzkueven, Helmut University of Melbourne, Medicine
REVIEW RETURNED	04-Nov-2013

GENERAL COMMENTS	It’s a complex task to review this paper, because the biggest problems, in my mind, are not with the approach as outlined, but with the premise of the UK risk share scheme, namely that it is feasible to validly compare long-term disability outcomes of a contemporary interferon/glatiramer (DMD) treated cohort with a historical control dataset acquired in a different country 10-20 years earlier (as appropriate UK datasets don’t exist. In my view, the authors present a very good (maybe even the best) solution to a really complex problem. Most of my review therefore addresses the shortcomings of the brief. I would like some more of these problems to be discussed as limitations, and if possible modeled and explored in greater detail. Homogeneity of data collection across space and time and centres EDSS scores are derived from Kurtzke Functional System Scores and an ambulation score, which becomes relevant from EDSS 4. In the last 15 years, EDSS measurement has become much more standardized due to the proliferation of clinical trials and the increased testing and certification of many trial physicians in the “Neurostatus” EDSS format. However, this format is also gradually changing. Examples include the change in the underlying definitions of assignment of a pyramidal KFS of 3 versus 2, the belated recognition of intermittent bladder self-catheterisation as a symptomatic therapy (reducing the impact of “catheterization “ in the bladder and bowel score form 4 to 2), KFS adjustment for visual and bladder/bowel scores and the many different versions for the crucial EDSS definitions of 6 and 6.5. Given that this is the primary outcome measure of all the analyses, how would the analysis ensure/validate consistency of EDSS assessment across time and space? This should be discussed as another limitation. I would strongly suggest that the EDSS calculations are validated from the collected KFS/Ambulation indices so that at least one consistent rule for EDSS calculation can be applied, even if the KFS assignment definitions differ. The data acquisition similarities or differences are not sufficiently discussed in this paper. Are there guidelines for EDSS assessment provided for the physicians of the BCMS and RSS centres? How do

these differ? Are patients routinely asked to perform a 500m walk, for instance, (leading to much more accurate ambulation scores) and do these instructions differ between the UK and BC? Is EDSS competency using Neurostatus testing (or similar) demanded in the UK? What is the typical appointment time in Canada and the UK centres used for deriving the EDSS in clinical practice?

Another example of the difficulty of this approach is the markedly increased recognition of the NMO phenotype, which carries a much worse prognosis. This would have been significantly mixed with MS (especially before the advent of MRI scans) in BC but much less so in the contemporary UK cohort. Although it might only represent 2-5% of the BC cohort, the much more rapid disability progression of this subset could influence comparability.

Selection bias

To my mind, the biggest difference between the two cohort studies might be selection bias. The BC dataset is a multi-decade, 4 centre (apologies if that's wrong, but I think talking to Prof Tremlett in the past she said the data comes from 4 centres, Vancouver Hospital being by far the largest), long-term MS cohort study. The UK RSS study is explicitly a treatment inception cohort. Especially in the UK, where there is a long-term culture of low DMD use, people actually selected for DMD initiation are likely to have much more severe MS (eg more severe relapses, motor relapses/ baseline pyramidal dysfunction, multiple cord lesions, high cerebral MRI lesion load and activity) than people who "algorithmically" become "eligible" for DMD treatment at some point during their natural history follow-up in a longitudinal cohort study. One way to address this would be to compare, between the two datasets, the relapses that define entry into the cohort study, namely relapse measures that might serve as proxies for severity (methylprednisolone treatment, hospital admission, stated relapse severity and duration, pyramidal system involvement). One indicator that this might be a big problem is the EDSS difference "at starting point" between the two cohorts of a whole EDSS point (Table 1), namely 2 versus 3 (or 3.5) in the BC and RSS cohorts, respectively. This may not sound like much, but in the contemporary MSBase cohort, an EDSS score of 2.0 at 8 years disease duration is the exact median (50th centile) whereas an EDSS score of 3.0 for that same disease duration is the 66th centile for the 8213 EDSS scores available at that disease duration.

Data presentation for the scientific communication

The transition matrices for the BC cohort should be shown. This would then allow scientific comparison with other datasets with longitudinal natural history data, eg the Italian dataset and the EDMUS collaboration.

The Welsh dataset validation results should be shown. As this comprises an opportunity for assessing comparability of EDSS transitions between the BC and a UK dataset, it would be useful to show the transition matrices for this dataset.

Alternative modeling ideas within the Markov model/transition probability

The "continuous" Markov modeling described is not really continuous. The approach is to assign the temporally closest EDSS score to a time "t" that is required for the model. Another alternative would be to use all the available EDSS scores across 12-month post-baseline intervals and average them to create a "mean" annualized EDSS for modeling purposes. In many analyses, pyramidal dysfunction is a strong independent predictor of subsequent disability progression, certainly stronger than age. In my opinion, the author should include in this manuscript exploration of a

	model adjusted for the presence or absence of a pyramidal lead KFS contribution to the EDSS. Other approaches that could be considered to answer the question in complementary ways This is of course outside the scope of the RSS brief, but there are many other approaches that might address some of the issues raised above, and should probably compliment the proposed analysis, acknowledging that no one solution is likely to be best. One option would be use of the RSS cohort on its own. Then all between-cohort biases would be eliminated. The starting point for the modeling would be the entry into RSS with first DMD treatment initiation. If KFS/EDSS scores, relapses and their treatment, pregnancies, and DMD commencement/cessation dates are recorded for this cohort, then the influence of DMD exposure on EDSS change (eg at 4,6,8 years) could be estimated within cohort, as it would inevitably contain groups of patients with variable treatment exposures post initial DMD commencement. Another option would be to focus on the occurrence of individual persistent disability progression events (using time-to-event modeling), and in long duration studies one can specify 12 or 24 or 36 month (ie long-term) confirmation of disability progression. The effect of DMD exposure time on these progression events could be modeled within cohort, or a propensity-adjusted model with one/one matching could be used to compare between the BC ad RSS cohorts. A difficulty with the proposed cost-benefit analysis As the actual analysis plan is not presented, it is difficult to gauge the RSS proposal but it might be worth a comment in the manuscript. The cost of MS increases dramatically with high disability states (6.5 and up). In the early years of the proposed RSS cohort study, few patients would transition to these high EDSS states, ie. the DMD-mediated reduction in EDSS transitions would largely take place in the EDSS 0-4 part of the scale, where the costs are small. Most of the established cost-benefit models extrapolate early reduction in disability progression (as demonstrated in the Phase III studies) to cost-savings several decades hence, when extrapolated time at 6.5 or above for patients is reduced. How does this reconcile with the RSS approach? Minor comments My strongest criticism for the authors would be the title. An actual detailed analysis plan is not presented. What is presented are excellent within BC cohort validations of the proposed model. As an example, I can't find a definition for the "RSS" cohort versus the "RSS analysis cohort", as shown in Table 1. The UK datasets are being derived from a large number of centres across the UK, already more than 5610 patients are enrolled, whereas the BC cohort is a 4 centre study and the "eligible" dataset for comparison purposes comprises only 898 individuals. Will the RSS dataset be subjected to repeated random sampling of 900 or so to assess consistency? Comparability of populations: One concern is that UK and BC populations could be different. In particular, differences in proportion of black /black admixed heritage (a population mostly reported to have much worse disability outcomes) could significantly alter results. Are these data collected and can they be adjusted for? There are other large natural history cohorts in existence in the French EDMUS collection and the Italian dataset. I believe both would, if approached by the UK RSS group, most likely help to

	supply additional natural history datasets. I know that the "EDSS density" of these datasets is high, and they might therefore serve as additional useful validation datasets if approached.

VERSION 1 – AUTHOR RESPONSE

Reviewer Name David Henry
Institution and Country University of Toronto, Toronto, Canada
Please state any competing interests or state 'None declared': None

Cost-effectiveness analysis is widely used to establish eligibility, usage criteria and prices of drugs. In the case of drugs with high unit acquisition costs and uncertainty about long term outcomes some form of risk sharing between the manufacturer and the purchaser is becoming commoner. These risk sharing agreements seem to take a number of forms ranging from simple (a 'price/volume trade off') to more complex (a price determined by a modeled cost/QALY threshold). This paper is an example of the latter. The UK risk sharing scheme is one of the most prominent examples of the latter. The performance of this scheme will be of great interest internationally.
We thank the reviewer for their supportive comments

When I first read the title of the paper I had a moment of hope that we might learn something of the internal workings of the scheme and some account of how it actually worked. However, I think that the operational details of this scheme, and others, are not going to be very transparent – despite that fact that it is public money at risk.

Therefore, I was a little disappointed to find a technical paper that describes the search for and performance of a population based of untreated MS patients as a basis for improved modeling of outcomes in the UK risk sharing scheme. Transparency in the model is welcome but how this impacts on decision making as it affects eligibility, usage and price is what we need to hear about.
We realise the title was misleading and suggested a broader remit than this work covers and have now modified the title. In the introductions we have given more detail of the RSS scheme.
As to the description of the model, this is nicely done. The paper is very clearly written, easy to read and relatively easy to understand for a reasonably knowledgeable individual. The figures display the degree of fit well. I think there is sufficient (not excessive) detail. As a non-specialist I can't judge the more technical aspects, but this is a skilled group of collaborators and I don't have serious doubts about what they have done.
Thank you for these comments.

The Abstract is entirely free of data. I realize that it is hard to provide meaningful data about the model outputs in an Abstract, but some details of the cohort and some description of disease progression in the observed and modeled cohorts would be welcome.
The abstract structure has been modified according to the BMJ open guidelines and more data has been added.

For me the main issues belong properly with the editors – but I will comment anyway. I don't think this paper, in its present form, will be of great interest to the majority of readers. Only a small relatively specialized group will care. However the authors could turn this into a more interesting paper for general readers by doing two things:

1) Explaining much more clearly to general readers how the risk sharing scheme could work – accepting that many details may be confidential. They could use simple examples to explain how a

higher than acceptable cost/QALY might result in a lower price
The referee is correct in that the individual target for each drug is confidential between the drug company and the department of health, although all agreements will be based upon the same target of £36,000 per QALY over a 20 year time span.
Nonetheless, we have added more detail in the introduction regarding the actual RSS and agree this background makes its clearer to general readers why the work we present here is important. We have explained that a positive deviation would lead to a price reduction and given an example of how this works by mentioning the average price reduction at the onset of the study.

2) Can they use their continuous Markov model with age at onset (apparently the best fit) and some indicative costs plus a range of hazard ratios for treatment effect to indicate what prices might be reasonable – under pre-specified assumptions? Again this is illustrative rather than definitive.

The individual companies each have a confidential agreement with the Department of Health regarding how their current costings relate to the agreed NICE target (eg companies will vary in the 20 year cost per QALY target but it will be a maximum of £36,000 - ie they may target a lower cost per QALY over the 20 years to allow them less risk of a price adjustment). We feel therefore that giving an explicit example in the paper will mislead readers because some companies may fall below the target in the future and not require a price adjustment if their target is lower than £36,000. However we think a good example is what happened at the onset of the scheme which led to price reductions of the drugs to meet their individual targets and we have now added the average reduction at onset of the scheme required to fall within the target in the introduction.
For the reviewers interest: The published 20 year cost per QALY was around £70,000 for the drugs combined (Chilcott J, BMJ 2003;326:522) and had to reduce to a maximum of £36,000 and this led to the average reduction in price at the onset of the scheme was 13.7%.
Reviewer Name Butzkueven, Helmut
Institution and Country
University of Melbourne, Medicine
Please state any competing interests or state 'None declared': None

It's a complex task to review this paper, because the biggest problems, in my mind, are not with the approach as outlined, but with the premise of the UK risk share scheme, namely that it is feasible to validly compare long-term disability outcomes of a contemporary interferon/glatiramer (DMD) treated cohort with a historical control dataset acquired in a different country 10-20 years earlier (as appropriate UK datasets don't exist. In my view, the authors present a very good (maybe even the best) solution to a really complex problem.
We are grateful that the reviewer recognises the challenges of assessing long term effectiveness of treatments and we do aim to produce a pragmatic 'best ' (and agree not perfect) solution and hope this may be relevant to other future treatments in MS and other chronic diseases. This paper is focused (as the authors notes) only on the predictive model used and on a better natural history dataset.
We agree that there is no ideal (untreated) comparator cohort. There are pros and cons of using an historic dataset (as we have here) vs a contemporary dataset. This is discussed in the paper and we have broadened this discussion based on the comments below.

Most of my review therefore addresses the shortcomings of the brief.
I would like some more of these problems to be discussed as limitations, and if possible modeled and explored in greater detail.
We appreciate the thoughtful comments and have opted to discuss as limitations, a suggested by the reviewer. More detailed responses and rationale are below

Homogeneity of data collection across space and time and centres EDSS scores are derived from Kurtzke Functional System Scores and an ambulation score, which becomes relevant from EDSS 4. In the last 15 years, EDSS measurement has become much more standardized due to the proliferation of clinical trials and the increased testing and certification of many trial physicians in the “Neurostatus” EDSS format. However, this format is also gradually changing. Examples include the change in the underlying definitions of assignment of a pyramidal KFS of 3 versus 2, the belated recognition of intermittent bladder self-catheterisation as a symptomatic therapy (reducing the impact of “catheterization “ in the bladder and bowel score form 4 to 2), KFS adjustment for visual and bladder/bowel scores and the many different versions for the crucial EDSS definitions of 6 and 6.5. Given that this is the primary outcome measure of all the analyses, how would the analysis ensure/validate consistency of EDSS assessment across time and space? This should be discussed as another limitation. I would strongly suggest that the EDSS calculations are validated from the collected KFS/Ambulation indices so that at least one consistent rule for EDSS calculation can be applied, even if the KFS assignment definitions differ.

The data acquisition similarities or differences are not sufficiently discussed in this paper. Are there guidelines for EDSS assessment provided for the physicians of the BCMS and RSS centres? How do these differ? Are patients routinely asked to perform a 500m walk, for instance, (leading to much more accurate ambulation scores) and do these instructions differ between the UK and BC? Is EDSS competency using Neurostatus testing (or similar) demanded in the UK? What is the typical appointment time in Canada and the UK centres used for deriving the EDSS in clinical practice?

The reviewer is correct in this limitation of the EDSS score measurement. Unfortunately neither dataset has access to the KFS.

However, a strength of our study is the consistency of EDSS scorers in each jurisdiction (BC and the UK). We believe this minimizes ‘intra’ observer variation. For instance, in the UKRSS, we have purposely aimed for the same EDSS scorers to continue over time because we are more interested in whether there is a true change or not over time. When EDSS scorers handed over to new scorer they were asked to do hand over clinics jointly to ensure they scored in the same way. All EDSS scores are performed in UK centres where the neurologist has an MS interest and thus experienced in performing EDSS. Similarly, the BCMS clinics had a consistent and limited number of experienced MS neurologists measuring the EDSS, with the same five core neurologists contributing to the bulk of the EDSS scores captured in this study. Neither of the cohorts had EDSS scores collected under strict trial conditions. We have added this limitation into the discussion.

Another example of the difficulty of this approach is the markedly increased recognition of the NMO phenotype, which carries a much worse prognosis. This would have been significantly mixed with MS (especially before the advent of MRI scans) in BC but much less so in the contemporary UK cohort. Although it might only represent 2-5% of the BC cohort, the much more rapid disability progression of this subset could influence comparability.

This is an important point. The reviewer is correct in that NMO is an emerging concept, only being well-recognized in the last 5 years or so. The UKRSS cohort completed recruitment in 2005 and before the NMO antibody assay was available (initial NMO antibodies were only just described and the modern NMO criteria were published in 2006). So, we think if NMO patients have been included they have an equal chance of being in both cohorts. However, while it is still not known what the true prevalence of NMO is, the recent ‘Atlas of MS’ initiativehttp://www.msif.org/includes/documents/cm_docs/2013/m/msif-atlas-of-ms-2013-report.pdf?f=1) has estimated the pooled prevalence as 1 in 100,000 (rising to 2.6 in 100,000 in countries such as Japan). The current prevalence of MS in Canada / UK is around 200 in 100,000. Therefore we estimate 1/200 (0.5%) NMO patients may have been misclassified as MS. We do not think this will substantially affect findings. Nonetheless, we have included this as a study limitation in the discussion.

Selection bias

To my mind, the biggest difference between the two cohort studies might be selection bias. The BC dataset is a multi-decade, 4 centre (apologies if that's wrong, but I think talking to Prof Tremlett in the past she said the data comes from 4 centres, Vancouver Hospital being by far the largest), long-term MS cohort study. The UK RSS study is explicitly a treatment inception cohort. Especially in the UK, where there is a long-term culture of low DMD use, people actually selected for DMD initiation are likely to have much more severe MS (eg more severe relapses, motor relapses/ baseline pyramidal dysfunction, multiple cord lesions, high cerebral MRI lesion load and activity) than people who "algorithmically" become "eligible" for DMD treatment at some point during their natural history follow-up in a longitudinal cohort study. One way to address this would be to compare, between the two datasets, the relapses that define entry into the cohort study, namely relapse measures that might serve as proxies for severity (methylprednisolone treatment, hospital admission, stated relapse severity and duration, pyramidal system involvement). One indicator that this might be a big problem is the EDSS difference "at starting point" between the two cohorts of a whole EDSS point (Table 1), namely 2 versus 3 (or 3.5) in the BC and RSS cohorts, respectively. This may not sound like much, but in the contemporary MSBase cohort, an EDSS score of 2.0 at 8 years disease duration is the exact median (50th centile) whereas an EDSS score of 3.0 for that same disease duration is the 66th centile for the 8213 EDSS scores available at that disease duration.

These are all very important and relevant points and very difficult to address.

We have included relapse data in Table 1. However, we do not have detailed information related to relapse severity. With respect to the differences in baseline EDSS, we do not consider this impacting our model which is calculating the transition probabilities between EDSS states, and different baseline EDSS would only matter if baseline EDSS is directly associated with disease progression. Therefore we have looked at the rates of disease progression stratified by baseline EDSS in both datasets (BCMS and RSS) but didn't identify remarkable differences. The respective data will be presented in the results paper and therefore haven't been shown here.

Data presentation for the scientific communication The transition matrices for the BC cohort should be shown. This would then allow scientific comparison with other datasets with longitudinal natural history data, eg the Italian dataset and the EDMUS collaboration.

This is a good suggestion and we have inserted the transition matrices as an appendix.

The Welsh dataset validation results should be shown. As this comprises an opportunity for assessing comparability of EDSS transitions between the BC and a UK dataset, it would be useful to show the transition matrices for this dataset.

We have now removed any references to the Welsh datatset from the paper. With hindsight we recognize that it was an error to include the Welsh dataset to validate this model. Further, the way in which we had reported the findings was factually incorrect. In reality, our approaches were different – as we have already noted in this paper, we needed to censor the BCMS data once DMTs became available because of selection bias and the Welsh data was actually collected entirely after DMTs became available.Therefore the use of a contemporary dataset of untreated patients may introduce all kind of bias.

Alternative modeling ideas within the Markov model/transition probability The "continuous" Markov modeling described is not really continuous. The approach is to assign the temporally closest EDSS score to a time "t" that is required for the model.

We apologise for not being clear in our explanation of the continuous Markov model. This re-allocation was described for the discrete model as a way of "annualizing" the data and in Appendix 2 a way of graphic representation of the 'observed EDSS' at any time t was described . The continuous model does not need any such adjustments as every time "t" at which an EDSS was recorded is entered into the model, but for a graphic representation of observed EDSS over time some assumptions are needed. We have re-phrased the respective paragraph to hopefully make it clearer.

Another alternative would be to use all the available EDSS scores across 12-month post-baseline intervals and average them to create a “mean” annualized EDSS for modeling purposes. In many analyses, pyramidal dysfunction is a strong independent predictor of subsequent disability progression, certainly stronger than age. In my opinion, the author should include in this manuscript exploration of a model adjusted for the presence or absence of a pyramidal lead KFS contribution to the EDSS.

This is an interesting idea that would be worth exploring but unfortunately we do not have access to the KFS in either dataset.

Other approaches that could be considered to answer the question in complementary ways This is of course outside the scope of the RSS brief, but there are many other approaches that might address some of the issues raised above, and should probably compliment the proposed analysis, acknowledging that no one solution is likely to be best.

One option would be use of the RSS cohort on its own. Then all between-cohort biases would be eliminated. The starting point for the modeling would be the entry into RSS with first DMD treatment initiation. If KFS/EDSS scores, relapses and their treatment, pregnancies, and DMD commencement/cessation dates are recorded for this cohort, then the influence of DMD exposure on EDSS change (eg at 4,6,8 years) could be estimated within cohort, as it would inevitably contain groups of patients with variable treatment exposures post initial DMD commencement.

Another option would be to focus on the occurrence of individual persistent disability progression events (using time-to-event modeling), and in long duration studies one can specify 12 or 24 or 36 month (ie long-term) confirmation of disability progression. The effect of DMD exposure time on these progression events could be modeled within cohort, or a propensity-adjusted model with one/one matching could be used to compare between the BC ad RSS cohorts.

These are interesting alternatives, but we are limited by the original goals of the UK RSS, including data collection. The data that has been and is being collected within the RSS does not contain the level of detail outlined above.

Further, using a ‘within cohort’ approach is not without its own limitations – especially in a disease such as MS and an outcome such as the EDSS. This approach may not sufficiently account for the variable and unpredictable individual progression profiles of MS patients.

Propensity score adjustments can be a useful alternative when there are a very large number of baseline covariates which is not the case in our study. When we have previously applied propensity score adjustments in similar situations, the findings have been virtually identical to conventional covariate adjustments (Shirani, A, JAMA. 2012 Jul 18;308(3):247-56) More importantly, it is physically impossible to do in the current study – the untreated cohort resides in Canada, the treated in the UK – because of historical data agreements and data privacy requirements, data is not allowed to leave Canada and the UK data cannot be transferred to Canada. The two datasets have never been and will never be allowed to merge. By virtue of this, it is not possible to calculate the propensity for treatment when the untreated and treated cohorts are physically separate. Finally, matching on propensity scores usually would mean that a comparatively small number of ‘cases’ (i.e. ‘treated’ patients) is matched to ‘controls’ (i.e. ‘untreated’ patients) which implies that the total number of ‘controls’ to choose from is sufficiently large– this condition clearly is not met here.

A difficulty with the proposed cost-benefit analysis As the actual analysis plan is not presented, it is difficult to gauge the RSS proposal but it might be worth a comment in the manuscript. The cost of MS increases dramatically with high disability states (6.5 and up). In the early years of the proposed RSS cohort study, few patients would transition to these high EDSS states, ie. the DMD-mediated reduction in EDSS transitions would largely take place in the EDSS 0-4 part of the scale, where the costs are small.

Most of the established cost-benefit models extrapolate early reduction in disability progression (as demonstrated in the Phase III studies) to cost-savings several decades hence, when extrapolated time at 6.5 or above for patients is reduced. How does this reconcile with the RSS approach?
This is very important and the whole crux of the RSS. The cost per QALY calculations for short term treatments were up to £500,000 over 2 years but it was recognised that this might be reduced when calculated over the longer-term when effect on higher EDSS states occurred (e.g. delay/prevention of people going into wheelchairs). Thus this is why the target cost effectiveness was modelled over 20yrs. We have added this into the introduction.
Minor comments
My strongest criticism for the authors would be the title. An actual detailed analysis plan is not presented. What is presented are excellent within BC cohort validations of the proposed model. As an example, I can't find a definition for the "RSS" cohort versus the "RSS analysis cohort", as shown in Table 1. The UK datasets are being derived from a large number of centres across the UK, already more than 5610 patients are enrolled, whereas the BC cohort is a 4 centre study and the "eligible" dataset for comparison purposes comprises only 898 individuals. Will the RSS dataset be subjected to repeated random sampling of 900 or so to assess consistency?
Thank you for these relevant comments. The reviewer is correct and we have altered the title. The table now has the definition of the RSS full and analysis cohorts. We did not consider repeated random sampling as we are not going to directly merge and compare data from BCMS and RSS databases.

Comparability of populations: One concern is that UK and BC populations could be different. In particular, differences in proportion of black /black admixed heritage (a population mostly reported to have much worse disability outcomes) could significantly alter results. Are these data collected and can they be adjusted for?
We do not have access to individual patient-level data on ethnicity. However, we have now included a broader discussion on the ethnic mix in British Columbia vs the UK which were highly comparable. Of note, over the respective study time-periods, 30.2% of the BC population were 'British' and 0.5% from Africa.
There are other large natural history cohorts in existence in the French EDMUS collection and the Italian dataset. I believe both would, if approached by the UK RSS group, most likely help to supply additional natural history datasets. I know that the "EDSS density" of these datasets is high, and they might therefore serve as additional useful validation datasets if approached.
Confavreux's much published dataset only reported EDMUS scores, not EDSS scores. We are not aware of a prospectively collected long-term untreated large Italian cohort. However, the agreement to change a natural history dataset has required more than two years of negotiation with each pharmaceutical company having to do corporate risk assessments before they have signed a revised contract with the department of health.