PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A retrospective cohort study of the relationship between systolic blood pressure variability and multiple sclerosis disability
AUTHORS	Goldman, Myla; Min, Seulgi; Mason Lobo, Jennifer; Sohn, Min-Woong

VERSION 1 – REVIEW

REVIEWER	Farren Briggs Case Western Reserve University, United States
REVIEW RETURNED	23-Oct-2019

GENERAL COMMENTS	Pg4 line 3 Cardiovascular comorbidities are not only associated with poor MS outcomes, there are also more prevalent in persons with MS. This is an important point to acknowledge in the introduction: PMID: 30014877 DOI: 10.1016/j.msard.2018.07.011; PMID: 18728060 DOI: 10.1177/1352458508092263. What is the distribution of PDDS? Pg6 – Methods Within STATA, there is gologit that is a partial proportional odds model that relaxes the need for proportionality, that might be worth exploring rather than grouping PDDS. Table 1 Is there any information on MS disease duration – this is very important to note? Subtype? Hypertensive medication – wouldn't 'managed' BP influence variation? Why only these comorbidities? What about hyperlipidemia, COPD, type 1 diabetes? What about MS therapies? Was DMTs considered in the models? Did you adjust for disease duration? Table 1 It is a bit surprising to see patients with 6-10+ visits ("number of measures") in a single year at a primary care or specialty clinic. Can you describe further who are the patients? Are they representative of the general MS population? What is the average time between SBP measurements (hours, days, months)? Concerns with the fact that many hypertensives were excluded (Table 2)? Why is diabetes in Table 2 but not Table 1? Table 3/Table e-1 Any violations of the proportionality assumptions? It is clear there is some residual positive confounding when not including the comorbid conditions, as the magnitude of effect for SBP is reduced from 3.8 to 3.5 and 5.5 to 5.2. I would recommend switching these two tables, or combining them into one table.

	Results Please do not interpret ORs as more likely, as it implies probability, and readers who are not as adept with the differences between ORs and RRs might not appreciate the use of “more likely”, because it is not clear was likely means. Thus, I might suggest a more precise interpretation. Table 4 Add to the footnote what the reference category was for the outcome Table 5 What is the distribution for PDDS? Was it normal? Multi-directionality models I’m concern with the time-interval between SBP measures across all analyses. Variability across a short window would be concerning… averages can be misleading is many observations are from a specific period and fairly uniform with a few from other time points. How did you address this? This point diminishes enthusiasm for the manuscript and the possible interpretations.

REVIEWER	Peter Watson University of Cambridge UK
REVIEW RETURNED	13-Nov-2019

GENERAL COMMENTS	Systolic Blood Pressure Variability is Associated with Increased Multiple Sclerosis Disability bmjopen-2019-034355 My main comment here is that there are just too many regression analyses carried out which add very little, if anything, to addressing the declared research aim of this paper, namely the relationship between systolic blood pressure (SBP) variability and disability group as measured by PDDS score group. As far as I can see you just need the two multiple regressions with SBP variability as outcome and PDDS score group as one of a set of predictor variables. These two analyses using one of the many regression formulations reported are presented in the paper looking at PDDS group's relationship with SBP variability before and after the survey (Page 6, lines 42-49 and Tables e-3 and e-4 on pages 24 and 25). I don't see the point in presenting any of the three different regression analyses (the proportional odds and two additional sensitivity analyses consisting of a binary logistic regression and a multiple linear regression with raw PDDS score as the outcome described and presented on pages 6 and 8) when they appear to be answering the same question and giving very similar results in Tables 3-5 (Pages 17-19) as in the regressions whose results are presented in Tables e-3 and e-4 on pages 24 and 25 of this paper looking at SBP variability before and after the survey and mentioned by me earlier as the only analyses worth reporting. Your declared interest (Page 4, lines 30-34) is in the relationship between disability as measured by the PDDS group with the SBP variability. You only need, therefore, to test this hypothesis using the regressions reported in Tables e-3 and e-4 which have SBP as the continuous outcome and PDDS group as a predictor with confounders added as considered necessary as other predictors in the regression. This would also reduce the number of tables in the

	paper and give a more focussed, less confused and much shorter results section removing the rather cumbersome motivation of the choice of cut-offs for the PDDS scores to recode the PDDS score groups. You could use each of the nine PDDS score groups (Page 4, lines 52-55) or pool together score groups to boost to a meaningful sample size and/or using medical/clinical rationale to justify pooling PDDS groups (see next comment). Page 6, lines 3-28 I would prefer if the pooling of values to form the PDDS score groups used as the outcome variable in the logistic regression models was motivated by medical reasons rather than statistical such as satisfying a model assumption since there is a danger here of the data influencing the values taken by the outcome variable used in the regression analyses, in this case in the definition of the PDDS groups being compared. Ideally these PDDS groups should be decided apriori before any analysis is carried out. Could you, therefore, motivate the PDDS score 3 or above as the cut-off used in the binary logistic regression using medical rationale? I would also suggest that if the Proportional Odds assumption is not satisfied you can use a multinomial logistic regression which models more than three groups in the outcome variable and which does not require the proportional odds assumption hence requires no pooling of PDDS score groups although you could pool PDDS groups with small numbers of people in them. Agresti describes the multinomial approach in his books but as I mention earlier I don't think you need to use any logistic regressions to address your research question presented on page 4, lines 30-34. Page 5, line 51, page 8, line 8. This is a small point but I think the standard terminology is "multiple" regression rather multivariable regression or multivariable analysis. Page 6, lines 3-6. Did other groupings satisfy the Proportional Odds assumption and, if so, why were the reported particular groupings chosen? Page 6, line 32. I don't see a mention of a sleep disturbances predictor in Table e-1 as claimed in the text here. Page 7, lines 20-34. I don't really understand precisely what you mean by a "tertile". The tertiles seem to be (Page 8, line 10) SBP variability groups but how are these groups obtained e.g. what cut-offs used? Is the tertile a form of percentile? I am also not clear what the purpose of defining the tertiles is. It appears that the tertiles are there to describe relationships with average SBP (Page 7, lines 20-34) and also to relate to PDDS disability groups (Page 8, lines 8-17) which are in any case tested using the regressions where the more informative raw SBP variability, as an outcome, is used and related to the PDDS groups pre and post survey. Page 8, lines 3-5. What are the rammifications of the included subjects being less hypertensive and more depressed than those who did not submit more than the requisite 3 blood pressure measures in 12-months (Page 7, lines 8-12)? Does this imply people are not missing completely at random so that if one had imputed missing values for these people and included them in the study one might have got a different set of results? Page 8, lines 42-43 and elsewhere. Is the "gradient relationship"

	reported briefly here a test of a linear trend or some other sort of trend or is it just the semi-partial correlation between the predictor and outcome variable?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. Pg4 line 3. Cardiovascular comorbidities are not only associated with poor MS outcomes, they are also more prevalent in persons with MS. This is an important point to acknowledge in the introduction: PMID: 30014877 DOI: 10.1016/j.msard.2018.07.011; PMID: 18728060 DOI: 10.1177/1352458508092263. What is the distribution of PDDS?

Revised as suggested (See Page 4). The distribution of PDDS is shown in Table 1. The mean crude (unadjusted) PDDS score was 2.22 ± 1.89 overall (median 2), 1.52 ± 1.95 in SBPv Tertile 1, 2.73 ± 1.70 in Tertile 2, and 2.42 ± 1.86 in Tertile 3. Over a quarter of all patients had PDDS score 4 or higher, indicating severe disability.

2. Pg6 – Methods. Within STATA, there is gologit that is a partial proportional odds model that relaxes the need for proportionality, that might be worth exploring rather than grouping PDDS.

We appreciate this suggestion. We have already tried to fit the model using "gologit2" to fit a generalized ordered logit model on the original PDDS scores. None of the SBPv tertiles and many other variables did not show any significant associations with the PDDS score due to the small sample sizes in each group. The relationship did not emerge until we combined PDDS scores into smaller groups.

3. Table 1. Is there any information on MS disease duration – this is very important to note? Subtype? Hypertensive medication – wouldn't 'managed' BP influence variation? Why only these comorbidities? What about hyperlipidemia, COPD, type 1 diabetes? What about MS therapies? Was DMTs considered in the models? Did you adjust for disease duration?

We did not have data on MS disease duration, DMTs and other MS therapies, and hypertension medications. We identified many comorbidities including CVD, PVD, sleep disturbance, hypothyroidism, diabetes, hyperlipidemia. Most conditions were quite rare for our sample patients and we were not able to include them in our multiple regressions. For example, we identified only six patients with diabetes in the entire sample of n = 218, of whom only 2 were included in the analytic sample (This WAS reported in Table 2. This row was deleted now. See the response to #4 below). Likewise, we had 8 patients with sleep disturbance and 3 patients with hypothyroidism. We recognized this as a limitation in the manuscript. See page 13.

4. Table 1. It is a bit surprising to see patients with 6-10+ visits ("number of measures") in a single year at a primary care or specialty clinic. Can you describe further who are the patients? Are they representative of the general MS population? What is the average time between SBP measurements (hours, days, months)? Concerns with the fact that many hypertensives were excluded (Table 2)? Why is diabetes in Table 2 but not Table 1?

These subjects included in this project were part of an original cohort that were selected from the MS clinic population for participation in unrelated clinical research and represent "all comers" to an

Academic MS Clinic. There were 52 patients who had 6 or more BP readings and 40 had 3 – 5 readings during one year before the PDDS survey. Those with 6 or more readings had significantly lower mean SBP (119.3±11.9 mmHg vs 130.2±12.3 mmHg, p < 0.001) compared to those with 3 to 5 readings, but they were not different in within-subject SBP standard deviation (9.9±3.5 mmHg vs 10.0±5.7 mmHg, p = 0.872). On all other characteristics such as demographics, BMI, and comorbidities, these two groups were not significantly different. The average time between SBP measurements were 25.5 ± 34.5 days (median = 14 days, interquartile range = 7 – 42 days) in the entire sample. Within-subject mean number of days between measures were shorter for patients with 6 or more readings than those with 3 – 5 readings (30.1±22.6 vs 61.1±38.1, p < 0.001). We are do not have a clear explanation for why a larger proportion of patients with a hypertension diagnosis had fewer BP readings leading to exclusion. However, in our study cohort we had 30% with a hypertension diagnosis, which is in line with the incidence reported in the MS population in general and which we believe is a meaningful proportion to contribute to our findings and understanding of the SBP variability and MS-disability as well as the relationship between hypertension diagnosis and SPB variability. We had only 2 patients with diabetes in the study sample, which we showed in Table 2. We dropped diabetes from Table 2 now (see #3 above).

5. Table 3/Table e-1. Any violations of the proportionality assumptions? It is clear there is some residual positive confounding when not including the comorbid conditions, as the magnitude of effect for SBP is reduced from 3.8 to 3.5 and 5.5 to 5.2. I would recommend switching these two tables, or combining them into one table.

We replaced Table 3 with Table e-1 as suggested. As we discussed in the Method section, both models (those in Table 3 and Table e-1) did not violate the proportionality assumptions. The likelihood-ratio test for the proportionality of odds was chi-square = 3.35 with p = 0.500 overall and the Brant test showed that the proportionality assumption specifically related to the SBP variability tertiles was not violated (p = 0.508).

6. Results. Please do not interpret ORs as more likely, as it implies probability, and readers who are not as adept with the differences between ORs and RRs might not appreciate the use of “more likely”, because it is not clear was likely means. Thus, I might suggest a more precise interpretation.

We appreciate the suggestion. We revised the manuscript as suggested. See the abstract and the results section (p. 9).

7. Table 4. Add to the footnote what the reference category was for the outcome

Revised as suggested. The Table was moved to the online appendix, Table e-1.

8. Table 5. What is the distribution for PDDS? Was it normal?

PDDS scores are NOT normally distributed (p < 0.007 using Sharipo-Wilk W test).

9. Multi-directionality models. I’m concern with the time-interval between SBP measures across all analyses. Variability across a short window would be concerning… averages can be misleading is many observations are from a specific period and fairly uniform with a few from other time points. How

did you address this? This point diminishes enthusiasm for the manuscript and the possible interpretations.

Before the PDDS survey, the average time between SBP measurements were 25.5 ± 34.5 days (median = 14 days, interquartile range = 7 – 42 days). After the survey, the average time between SBP measurements were 180.3 ± 97 (median = 173, IQR = 99 – 263 days). The spacing between measures in the post period was larger. Time-interval between SBP measures across all analyses were 33.4 ± 57.6 days (median = 21, IQR = 1 – 36). This suggests that fewer BP readings were taken during one year after the PDDS survey and they were spread out in time much more than readings taken before. To test whether variability across a short window may be systematically different from that across longer time intervals, we computed SBP standard deviations using only those measures taken within two days (e.g., 0 or 1 day interval between measure) and those with 2 or more days apart. We could compute both SDs for only 50 patients, and they were not statistically different (n = 50; 10.1±5.3 mmHg vs 9.9±4.0 mmHg, p = 0.447). This suggests that measures taken in uneven intervals between readings do not necessarily lead to biased estimation of SBP variability.

Reviewer: 2

10. My main comment here is that there are just too many regression analyses carried out which add very little, if anything, to addressing the declared research aim of this paper, namely the relationship between systolic blood pressure (SBP) variability and disability group as measured by PDDS score group. As far as I can see you just need the two multiple regressions with SBP variability as outcome and PDDS score group as one of a set of predictor variables. These two analyses using one of the many regression formulations reported are presented in the paper looking at PDDS group's relationship with SBP variability before and after the survey (Page 6, lines 42-49 and Tables e-3 and e-4 on pages 24 and 25).

I don't see the point in presenting any of the three different regression analyses (the proportional odds and two additional sensitivity analyses consisting of a binary logistic regression and a multiple linear regression with raw PDDS score as the outcome described and presented on pages 6 and 8) when they appear to be answering the same question and giving very similar results in Tables 3-5 (Pages 17-19) as in the regressions whose results are presented in Tables e-3 and e-4 on pages 24 and 25 of this paper looking at SBP variability before and after the survey and mentioned by me earlier as the only analyses worth reporting.

Your declared interest (Page 4, lines 30-34) is in the relationship between disability as measured by the PDDS group with the SBP variability. You only need, therefore, to test this hypothesis using the regressions reported in Tables e-3 and e-4 which have SBP as the continuous outcome and PDDS group as a predictor with confounders added as considered necessary as other predictors in the regression. This would also reduce the number of tables in the paper and give a more focussed, less confused and much shorter results section removing the rather cumbersome motivation of the choice of cut-offs for the PDDS scores to recode the PDDS score groups. You could use each of the nine PDDS score groups (Page 4, lines 52-55) or pool together score groups to boostto a meaningful sample size and/or using medical/clinical rationale to justify pooling PDDS groups (see next comment).

We appreciate the suggestion. Because PDDS scores violated the normality assumption (See Critique #8 above), we decided to keep the regression in Table e-1 as our main regression (See Critique #5 above) and move all other regressions to the online appendix (See our response to the Critique #11 below).

11. Page 6, lines 3-28 I would prefer if the pooling of values to form the PDDS score groups used as the outcome variable in the logistic regression models was motivated by medical reasons rather than statistical such as satisfying a model assumption since there is a danger here of the data influencing the values taken by the outcome variable used in the regression analyses, in this case in

the definition of the PDDS groups being compared. Ideally these PDDS groups should be decided apriori before any analysis is carried out. Could you, therefore, motivate the PDDS score 3 or above as the cut-off used in the binary logistic regression using medical rationale? I would also suggest that if the Proportional Odds assumption is not satisfied you can use a multinomial logistic regression which models more than three groups in the outcome variable and which does not require the proportional odds assumption hence requires no pooling of PDDS score groups although you could pool PDDS groups with small numbers of people in them. Agresti describes the multinomial approach in his books but as I mention earlier I don't think you need to use any logistic regressions to address your research question presented on page 4, lines 30-34.

We determined the PDDS groups according to the medical significance of each group. Moderate disability includes persons who have limits with daily activities or need assistance during an attack. Severe disability includes persons who needs cane and other forms of support in walking. We revised the manuscript to make this clearer (See p. 7). We moved the logistic regression to the online appendix (Table e-1) as per your suggestion in Critique #10 above.

12. Page 5, line 51, page 8, line 8. This is a small point but I think the standard terminology is "multiple" regression rather multivariable regression or multivariable analysis.

We changed “multivariable” to “multiple” throughout the manuscript.

13. Page 6, lines 3-6. Did other groupings satisfy the Proportional Odds assumption and, if so, why were the reported particular groupings chosen?

Due to the small sample size (N = 92), we only tested tertiles and quartiles. The latter violated the proportional odds assumption in many regressions we presented in the manuscript, while tertiles did not. That was the reason we chose tertile groupings.

14. Page 6, line 32. I don't see a mention of a sleep disturbances predictor in Table e-1 as claimed in the text here.

We dropped this variable (“sleep disturbances”) from the manuscript.

15. Page 7, lines 20-34. I don't really understand precisely what you mean by a "tertile". The tertiles seem to be (Page 8, line 10) SBP variability groups but how are these groups obtained e.g. what cut-offs used? Is the tertile a form of percentile? I am also not clear what the purpose of defining the tertiles is. It appears that the tertiles are there to describe relationships with average SBP (Page 7, lines 20-34) and also to relate to PDDS disability groups (Page 8, lines 8-17) which are in any case tested using the regressions where the more informative raw SBP variability, as an outcome, is used and related to the PDDS groups pre and post survey.

Tertiles are defined as three equal-sized groups. We used the SBP coefficient of variation (SBPCV) to divide the study sample (n=92) into tertiles whose SBPCV ranges are .01 - .064 for Tertile 1, .065 – 0.087 for Tertile 2, and 0.089 – 0.172 for Tertile 3. See p. 6 for the revision.

16. Page 8, lines 3-5. What are the ramifications of the included subjects being less hypertensive and more depressed than those who did not submit more than the requisite 3 blood pressure measures in 12-months (Page 7, lines 8-12)? Does this imply people are not missing completely at

random so that if one had imputed missing values for these people and included them in the study one might have got a different set of results?

We do not believe that the exclusion of more depressed patients have any ramifications on our results. In regard to those with lower BP being included – if anything this may make our findings more conservative, because people with higher BP tends to have higher variability in SBP and so our findings may be biased toward the null if there is any. This was a retrospective cohort study and is intended as a first step in identifying and understanding the relationship between sBPV and MS disability. As with all developing lines of research, our findings will have to be confirmed in larger, prospectively designed studies. This work, supports that this larger studies should be pursued.

17. Page 8, lines 42-43 and elsewhere. Is the "gradient relationship" reported briefly here a test of a linear trend or some other sort of trend or is it just the semi-partial correlation between the predictor and outcome variable?

It was a test of a linear trend. The P-value for linear trend is now reported in the abstract and in the results section (p. 9).

VERSION 2 – REVIEW

REVIEWER	Farren Briggs Case Western Reserve University United States
REVIEW RETURNED	23-Dec-2019

GENERAL COMMENTS	I appreciate the authors revisions. I do have one more point that needs be discussed, and that is the generalizability of the study and the potential for selection bias by requiring >=3 SBP measures within 12 months prior PDDS measurement.. this is possibly why the expected sex bias for MS (2.5:1 F:M) is not observed which may results in males with more active/severe disease being over-rperesented in the sample. The presence of selection bias is also illustrated by the higher disability among female patients than male patients (Table 3), as the vast majority of studies have shown male patients experience more severe disease.

REVIEWER	Peter Watson University of Cambridge UK
REVIEW RETURNED	10-Jan-2020

GENERAL COMMENTS	A retrospective cohort study of the relationship between systolic blood pressure variability and multiple sclerosis disability bmjopen-2019-034355.R1 I think the proportional odds model is correctly used since (a) the PO assumption has been tested and found to hold (Page 8, lines 13-20) and (b) the results are correctly described in that one is comparing odds of being in a higher disability group for each tertile compared to the reference tertile (1). It may be worth adding in a reference for the proportional odds model for those unfamiliar with its use e.g. to Agresti, A. (1996) An introduction to categorical data analysis Wiley. The main thing I would add here is that I assume that the three disability groups which have been decided mathematically so that

	they satisfy the proportional odds assumption first and foremost make psychological and medical sense as distinctive groups. Page 8, lines 35-37. You mention here that the model(3) whose results are presented in Table e-2 on page 26 is identical to the ordinal logistic model in Table 3 (page 20) other than that the outcome, PDDS, is measured continuously instead of grouped. The same predictors are used in both models so they both address the same question of whether there is a relationship between PDDS and SBP and give the same results. I still, therefore, don't see why you need to fit both the proportional odds ordinal logistic model and the multiple regression model or in other words why you assess disability score as both grouped and continuous. I would have thought the multiple regression (whose results are in Table e-2 on page 26) would be sufficient as it does not rely on arbitrary groupings suggested mathematically rather than apriori through psychological or medical means and, since DPPS is continuous, contains more information than grouping this continuous variable. Page 10, line 26. Should read "ordinal logistic regression" rather than "multiple regression".

VERSION 2 – AUTHOR RESPONSE

Reviewer 1

I do have one more point that needs be discussed, and that is the generalizability of the study and the potential for selection bias by requiring >=3 SBP measures within 12 months prior PDDS measurement. This is possibly why the expected sex bias for MS (2.5:1 F:M) is not observed which may results in males with more active/severe disease being over-represented in the sample. The presence of selection bias is also illustrated by the higher disability among female patients than male patients (Table 3), as the vast majority of studies have shown male patients experience more severe disease.

We revised the limitations section to add the potential for selection bias arising from requiring >=3 SBP measures within 12 months prior PDDS measurement. See page 13.

Reviewer 2

I think the proportional odds model is correctly used since (a) the PO assumption has been tested and found to hold (Page 8, lines 13-20) and (b) the results are correctly described in that one is comparing odds of being in a higher disability group for each tertile compared to the reference tertile (1). It may be worth adding in a reference for the proportional odds model for those unfamiliar with its use e.g. to Agresti, A. (1996) An introduction to categorical data analysis Wiley.

We added the reference as suggested. See the reference number 22.

The main thing I would add here is that I assume that the three disability groups which have been decided mathematically so that they satisfy the proportional odds assumption first and foremost make psychological and medical sense as distinctive groups.

We revised the text as suggested. See page 7.

Page 8, lines 35-37. You mention here that the model(3) whose results are presented in Table e-2 on page 26 is identical to the ordinal logistic model in Table 3 (page 20) other than that the outcome, PDDS, is measured continuously instead of grouped. The same predictors are used in both models

so they both address the same question of whether there is a relationship between PDDS and SBP and give the same results. I still, therefore, don't see why you need to fit both the proportional odds ordinal logistic model and the multiple regression model or in other words why you assess disability score as both grouped and continuous. I would have thought the multiple regression (whose results are in Table e-2 on page 26) would be sufficient as it does not rely on arbitrary groupings suggested mathematically rather than apriori through psychological or medical means and, since DPPS is continuous, contains more information than grouping this continuous variable.

We dropped the multiple linear regression models from the paper. We did not want to impose a linearity assumption to the ordinal data (PDDS scores) and we could identify an ordinal logistic regression model that did not violate the proportionality assumption.

Page 10, line 26. Should read "ordinal logistic regression" rather than "multiple regression".

We revised the text as suggested. See page 9.