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 Cross-sectional Study of Self-Rated Health among Older Adults: A Comparison of China and the United States
AUTHORS	Xu, Dongjuan; Arling, G; Wang, Kefang

VERSION 1 – REVIEW

REVIEWER	Hanna Falk Institute of Health and Care Sciences, Sahlgrenska Academy at the University of Gothenburg, Sweden
REVIEW RETURNED	14-Dec-2018

GENERAL COMMENTS	The authors make a strong point in that although previous research has demonstrated many factors influencing self-rated health in the older population, these studies are largely limited to samples in one country. With the global aging of the population, cross-national comparison provides a unique perspective to gain knowledge on the similarities and differences in the aging experience. It is evident that the differences between China and the US as countries/cultures are massive which point torward the difficulties making this kind of comparbale study. It is well established in the litterature that our perception of health is contingent on our sociocultural context in which we are embedded. In order to address these differences the authors controlled for age, gender, educational level, currently working, living arrangements, number of children, functional limitations, self-reported memory, chronic conditions, mental health, and health-related behaviors. The authors also point out in the background section that there is emerging evidence that there is an association between humility and better self-rated (i.e. more humbleness leads to poorer self-rated health). Humility is a virtue and prominent feature in the Chinese culture, which is not the case in the US – rather the opposite. How the authors address these cultural differences needs to be further clarified. AIM 1) Whether factors (demographic, cognitive, physical, social, and mental) influencing self-rated health among older Chinese were similar to those among older Americans (ordered logistic regression models). 2) Whether there is a significant cross-national difference in self-rated health between China and the US after controlling those available influencing factors (ordered logistic regression models).

	METHODS Data and Samples – Needs to be clarified. Confusing description of datasets. Measures – How was educational level made comparable between countries? In the results section it becomes evident that you combine “good”, “very good”, or “excellent” health versus “fair” or “poor” self-rated health in order to make a point regarding the differences between China and the US. This needs to be addressed in the methods section. STATISTICS Chi-square or t-tests were used to evaluate the statistical significance of differences between the US and China. Ordered logistic regression models were conducted to investigate factors influencing self-rated health among older adults in the US and China respectively, as well as to see whether there was a cross-national difference on self-rated health between China and the US after controlling those available influencing factors. “Proportional odds assumption to test whether the assumption (?) was the same across the five-category self-rated health variable” – Needs to be clarified and moved to the results section of the manuscript. “A generalized ordered logistic model yielded similar results, indicating that the estimated difference in self-rated health between the US and China were not influenced by violations of the proportional odds assumption” – This is also results! Move to the results section. Sensitive analyses (?) using age groups and groups based on number of children showed similar results (?) – Similar to what? Needs to be clarified and moved to the results section of the manuscript. The authors state that they repeated all the analyses using the 2012-2013 HRS and CHARLS data sets to test the sensitivity of our. Why? If this was done in order to confirm that the 2014-2015 results, this needs to be mover to the results section of the manuscript. Or delete it. RESULTS Did you introduce the influencing factors stepwise into the logistic regression models or is Table 2 only showing the fully adjusted models? Needs to be clarified. Please put the OR, P and confidence interval in brackets after each statement pertaining to the results so that the reader easily can see how each factor influence self-rated health. Those who had more ADLs limitations, poorer self-reported memory, worse mental health, and chronic health conditions, had lower self-rated health. Factors including gender, number of living children, IADLs limitations, and ever smoking were not associated with self-rated health in either China or the US. Factors positively influencing self-rated health in China – living alone (OR 1.25).

	Factors positively influencing self-rated health in the US – younger age (OR 1.02), currently working (OR 1.31), higher educational level (OR 1.48), better recall summary score (OR 1.02), and not drinking/drinking less (OR 1.41). In the overall ordered logistic regression model when controlling for all factors mentioned in Table 2, older adults in China were much more likely to rate their health as being poor compared to older adults in US when the self-rated health was dichotomized into good (including "good", "very good", and "excellent") or poor (including "fair" and "poor") (OR=4.88, 95% CI: 4.06-5.86). DSICUSSION The authors state that independence and privacy is highly valued in the US family, while in China living with adult children is more normative because they are expected to take care of their elders, and elder parents are expected to provide grandchild care and that this is supported by their demographic findings. To me it seems strange that this is what they choose to highlight first in their discussion section since the only factor that positively influenced self-rated health in China was living alone (OR 1.25). The striking difference in (good/poor) self-rated health between China and the US is very interesting. Possible explanations put forth by the authors as to why a larger proportion of Chinese older adults rate their health as poor compared to Americans include under reporting of chronic conditions, lower education and health literacy, limited access to health care and sociocultural desirability (i.e. humbleness). In contrast, older adults in the US may be reluctant to see their health as poor for various reasons. The authors especially put forth their fear of losing independence or becoming a burden. However, since the self-rated health response options were collapsed into two categories with the alternative "fair" being put in the poor category this also might contribute to the striking difference between countries. This needs to be further elaborated in the discussions section.

REVIEWER	Miao Cai and Rhonda BeLue St. Louis University
REVIEW RETURNED	24-Dec-2018

GENERAL COMMENTS	Overview: The manuscript used two representative nationwide survey data in the US and China to compare self-reported health between the two countries. The authors showed that China respondents reported much worse health status than their US counterparts. Comments: 1. My biggest criticism to this manuscript is their outcome variable self-rated health. Since people have very different interpretation on "good" health in American and Chinese culture, how can the authors compare the outcome between the two countries? what is the practical meaning of comparing self-reported health between Chinese and American citizens?

	2. In the results of the abstract, I suggest the authors add significant factors for self-reported health in the two countries since it is one of the two major objectives of this study. 3. Page 4, Line 39-42, can the authors provide specific numbers and years in the US and China 4. The authors merged the two databases and estimated the model in one ordinal logistic regression by adding a country index. In this statistical model, the authors are implicitly assuming other covariates had the same effect on the outcome across the two countries. However, in the Introduction, the authors claimed a lot of differences between the two countries. Would two separate models or adding interactions between country index and covariates be more appropriate given the authors' introduction? 5. On page 8, line 33-42, the authors claimed that "a test of proportional odds assumption was performed; demonstrating the assumption did not hold across the five categories.". The authors should specify the name of the test, test statistics and P-value. In addition, if the proportional odds assumption is not met, they shouldn't have used ordered logistic regression. 6. The odds ratio of ordinal logistic regression should be interpreted as the cumulative odds of reporting good health status. The interpretation in this study looks more like the interpretation of a binary logistic regression results. 7. Page 8 line 37-38, a generalized ordered logistic regression seems to be a confusing term to me. From my understanding, an ordered logistic regression is equivalent to proportional odds model. Is the "generalized ordered logistic regression" a proportional odds model or some other model? 8. Page 8 line 42-47, it is very unlikely that all the estimates, standard errors and P-values are exactly the same in your sensitivity analysis and main model. If the number of tables did not excel the limit, I recommend the authors also include the sensitivity analysis results. 9. Page 12 line 31-42, the explanation of Chinese elders having a lower education level seems to be contradictory to the results. If older Chinese elders had lower health literacy, wouldn't they report better health status than American elders since they are dismissive of their disease? 10. Given that Chinese older respondents had a significantly high odds ratio of 4.88, can it be that they are really worse in physical/psychological health, instead of just self-reported health?

REVIEWER	Alexis Santos The Pennsylvania State University
REVIEW RETURNED	28-Dec-2018

GENERAL COMMENTS	Each of the No marked above has been referenced to in the review provided by me. I believe the authors need to indicate how they measured each variable and the manner they dealt with missing values in the text. This is a major roadblock to replication of this study. I have discussed in detail the need to reshape the outcome, perform a binary logistic regression rather than an ordered regression. This is reinforced by the fact that the authors STATE that the proportional odds assumption is not met (by the model and the data).

	Review of: Self-Rated Health among Older Adults: A Comparison of China and the United States Abstract Remove the words recent and large from the objective section, these are good qualities of the data but it should suffice by saying a nationally representative sample. There are some concerns with other sections of the abstract but I will address them within the manuscript in my line-by-line review; the same goes for the Strengths and limitations section presented in page 3. Introduction The first statement of the second paragraph (lines 17-23) is misleading. There are multiple examples in the literature where Self-Reported Health is studied using samples from different countries (cross-national analyses). To mention one: 1. Hardy, Melissa A., Acciai, Francesco, and Reyes, Adriana M. How health conditions Translate into Self-Ratings: A comparative study of older adults across Europe. Journal of Health and Social Behavior. The SHARE dataset has provided the ability to do these cross-national comparisons. Probably, a best way to frame this, is to say that China-US studies are not common and state WHY this comparison is important. The seeds are there, for example social support and support networks are expected to better help those in China in comparison to US, however this expectation has not been met by the aging of persons with only one child etc. Maybe just do away with this statement and focus on the China-US comparison, it is enough of a justification to do it if well established. I would say, no need to mention recent or large in the description of the dataset. Just state it is a nationally representative sample of Chinese and US older population - the message will get through. In lines 29-31 of page 5 of the manuscript, please finish the sentence in a manner similar to this one “national differences of self-rated health between China and the US after controlling for potential covariates (or confounders)”. Methods Data and Sample Please include a sentence on how you reached the final sample size (what was the initial size?) and what you did with missing values (list-wise deletion?). Measures Self-Rated Health Did you include the variable as a 1-5 in the regression model? This may be problematic, conventional studies go with the dichotomization of the variable and simply fit a logistic binary regression model. At least you should say whether the results vary by specification, in my experience it does. The 1-5 and ordered regression specification may not be the best way to approach this outcome. You can decide how to do it, but if so - provide some robustness checks on whether the results differ if you dichotomize this. Right now it is unclear if some of the associations may be influenced by the scale. Is Self-Rated Health really an ordered outcome? Sociodemographic and family structure

	You mention age and gender (sex?) in your first sentence. But you don't say how you operationalized them. Is age continuous? If so, did you include a square term in your model to control for potential curvilinear trends? If it was categorized then the associations will show you the shape of the associations in the magnitude of the OR or Coefficients. I don't know if calling Male/Female dichotomy a Gender variable is appropriate, maybe it is, but most recently gender encompasses more than this dichotomy. Can you refer to it as Sex? Functional limitations Why is ADL doe as a summary score of 0-3 and IADLs a binary when you have the variables to create a 0-4 variable. It seems quite inconsistent here, either dichotomize both (and you need to provide previous literature that operationalized it that way) or just add them both as continuous. It may trigger a reader to ask this if this is allowed to go through as it is. Cognition Here we have an interesting case of a variable that would really exemplify my concerns with the operationalization of SRH as a 1-5 variable. In the outcome 1 is worse and 5 is better; however, in this variable of Self-Reported Memory lower is better and higher is worse. If the outcome is dichotomized I wouldn't have a problem, but the outcome is ordered and this order may just provide a counterintuitive association if one does not reads carefully. If the author decides to use the 1-5 scale, they should flip it (and say it in the text) so that things go in the same direction. So 1 would be poor and 5 excellent in this "flipping" of the variable. Regarding the recall summary score - is it associated/correlated with the other cognition variable? Why are both being included? As it stands now it seems they were just thrown in because they were available. Chronic Conditions and Mental Health I would recommend separating these two sections. But I'll comment on it, as it stands. Why these eight chronic conditions? It comes to mind that these conditions would be associated with functional limitations ADL or IADL (also cognition). I am hard pressed to see why the models are going to include SO many variables that encompass the same factors or capture the same effects. At least the authors should provide VIFs (OLS of SRH as a 1-5 with variables included in the model) or a similar indicator that indicate these variables are not introducing redundancy in the models. My hunch is that they will, and are doing it. Health-related behaviors These are bound to be associated at least with Chronic Conditions (see comment above). Again, I'm concerned the model is over specified and some variables are introducing redundancy that may hide the REAL association of the variable of interest. Statistical Analysis Sentence 2, why ordered logistic regression models. See my comments above about this. It is clear there is a difference even if we dichotomize by poor/fair as 1 and the rest as 0. We are interested in poor/fair not the other way around. I've found the dichotomization is much better specification, and if you see more recent work on SRH they are moving away from the Ordered Logistic or other Multinomial specifications. A look at recent work by Anna Zajacova could be helpful in seeing the vast number of papers published using this specification.

	My biggest concern is not the model specification but the WAY the data are analyzed. It is very difficult to see what is gained from joining both datasets. It would suffice to perform the regression analysis on both datasets and compare the coefficients seeing whether the Confidence Intervals derived from the associations have any overlap or not. If this is the standard way you need to state it and provide reference for example to the rescaling of weights and how things were treated to be a valid sample design. Right now this is the weakest section of the paper. What is really being captured by adding a dummy indicating China/US - that the data come from a different dataset not necessarily what is being said it measured in the model. I think just specifying the FULLY specified model for both countries and comparing effects would suffice for the purposes of the paper. Take away the overall model, and just discuss the US China models and then differences/similarity in associations. 1. If you dichotomize you deal with the issue of non-proportional odds of the association. 2. If you fit the models and compare the Coefficients across models, without joining the datasets, you deal with issue of a flawed data design. Which may render your finding invalid. It is good that the way age and number of children is operationalized differently in the sensitivity analyses are included here, but it needs to be said also in the measurements section (as stated before). I have trouble with the phrase “statistical significance was accepted”. Could this be phrased as “Statistical significance was established at the 95% level ($p < 0.05$). Results Something to note here, is that Chinese seem to have better indicators regarding conditions than US older adults, but their mental health and cognitive markers are different. This may lay work for additional analyses in the future. The authors should consider my previous comments regarding model specifications. The results section is well written, but it is flawed because it does not meets the proportional odds assumption, a key assumption for these kind of models. If the author makes the revisions laid before I think they will end up with a stronger paper and a great section with results that are easily interpretable. I would include a whole section discussing what is found here in the regression models, despite it being the most important part of the paper its discussion is really superficial. Again, if the author follows some of my recommendations this results section will be really great to read. Discussion The first sentence here (and in the Conclusions) still is anchored in the recent dataset, although timeliness of analysis is great there is amazing science being done with recently recovered datasets (NHANES I, NHANES II, NHANES III - to mention some). The big element here is that you have two nationally representatives datasets that collect data in a way that makes this paper possible. If the results found in the initial models are found in the new model specification (which may be the case) I would review this section to be sure it is consistent with new findings (if any). Conclusion See above, not further comments.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

The authors make a strong point in that although previous research has demonstrated many factors influencing self-rated health in the older population, these studies are largely limited to samples in one country. With the global aging of the population, cross-national comparison provides a unique perspective to gain knowledge on the similarities and differences in the aging experience.

It is evident that the differences between China and the US as countries/cultures are massive which point torward the difficulties making this kind of comparbale study. It is well established in the litterature that our perception of health is contingent on our sociocultural context in which we are embedded. In order to address these differences the authors controlled for age, gender, educational level, currently working, living arrangements, number of children, functional limitations, self-reported memory, chronic conditions, mental health, and health-related behaviors. The authors also point out in the background section that there is emerging evidence that there is an association between humility and better self-rated (i.e. more humbleness leads to poorer self-rated health). Humility is a virtue and prominent feature in the Chinese culture, which is not the case in the US – rather the opposite. How the authors address these cultural differences needs to be further clarified.

RESPONSE: There are no variables related to cultures in the harmonized data sets. As we discussed in the manuscript, culture differences likely contributed to the difference in self-rated health between China and the US. However, we cannot offer direct evidence for this statement.

AIM

1) Whether factors (demographic, cognitive, physical, social, and mental) influencing selfrated health among older Chinese were similar to those among older Americans (ordered logistic regression models).
2) Whether there is a significant cross-national difference in self-rated health between China and the US after controlling those available influencing factors (ordered logistic regression models).

METHODS

Data and Samples – Needs to be clarified. Confusing description of datasets.

RESPONSE: We further clarified the data and samples. We added the initial sample size and used list-wise deletion for handling missing data. “Initially, 10,374 older adults in the US and 5,751 older adults in China reported their health status. Number and percentage of missing data is presented in Appendix Table A. We used list-wise deletion for handling missing data.”

Measures – How was educational level made comparable between countries?

RESPONSE: As we described in the measure of education, the three categories of educational level were not the same in these two countries. The China Health and Retirement Longitudinal Study (CHARLS) was designed to be comparable with the Health and Retirement Study (HRS). Although education is not exactly comparable between the Harmonized CHARLS and RAND HRS, it is defined as closely as possible by the Gateway to Global Aging Data.

In the results section it becomes evident that you combine “good”, “very good”, or “excellent” health versus “fair” or “poor” self-rated health in order to make a point regarding the differences between China and the US. This needs to be addressed in the methods section.

RESPONSE: Regarding the ordered logistic regressions, we will respond to all 3 reviewers’ comments here. First, we further explained the rationale to use ordered logistic regression to analyze the selfrated health, instead of binary logistic regression. Ordered logistic regressions take advantage of the full information of five-category of self-rated health and provides one set of coefficients.
Second, the proportional odds assumption held for most variables. Only a few variables did not meet the proportional odds assumption. Our test of proportional odds assumption found that independent variables include country (the US vs. China) (chi2=89.39, P<0.001), educational level (chi2=32.77, P<0.001), hypertension (chi2=19.66, P<0.001), diabetes (chi2=19.83, P<0.001), stoke (chi2=15.34, P=0.002), and arthritis (chi2=19.66, P<0.001) did not meet the proportional odds assumption.
Third, for the variables that failed the proportional odds assumption, we further analyzed the data using logistic regression models with four different ways (Models 1-4) to bifurcate the scale of self-rated health. The outcome variables in the four logistic regression models were:

- Model 1: self-rated health (“excellent” vs. “very good”, “good”, “fair” and “poor”)
- Model 2: self-rated health (“excellent” and “very good” vs. “good”, “fair” and “poor”)
- Model 3: self-rated health (“excellent”, “very good”, and “good” vs. “fair” and “poor”)
- Model 4: self-rated health (“excellent”, “very good”, “good”, and “fair” vs. “poor”)

In the methods section, we now state, “We used ordered logistic regression analysis as our primary statistical approach. This approach takes advantage of the full five-category of self-rated health in the analyses rather than collapsing the categories into a binary indicator. Ordered logistic regression provides one set of coefficients under the assumption that the association between an independent variable and each pair of outcome groups is the same (this is called the proportional odds assumption). Our test of proportional odds assumption found that some independent variables, including country (the US vs. China), educational level, hypertension, diabetes, stoke, and arthritis, did not meet the proportional odds assumption. In order to test the sensitivity of the results for variables violating the proportional odds assumption, we ran additional regression models (reported in Appendix Table B) with four different ways to bifurcate the scale of self-rated health.”
Fourth, we reported the results from the ordered logistic regression models (Table 2). For those variables violating the proportional odds assumption, we further reported the results from the four logistic regression models in the Appendix, Table B. Because the key independent variable is country (the US vs. China), we provided Table 3 to show cross-country differences in selfrated health with the four bifurcated self-rated health variables.
We revised the results section. We now state, “The odds of having better versus poorer health was almost 5 times greater in American older adults than those in China (OR=4.88, 95% CI:
4.06-5.86, Table 2). Because of the issue with the proportional odds assumption, we performed sensitivity analysis with alternative models. When shifting comparison pivot point down the selfrated health scale, we found the odds ratios range from 3.98 to 7.92 in the logistic regression models (Table 3). For example, the odds of having the combined “good”, “very good”, or “excellent” health versus “fair” or “poor” health was 7 times greater in American older adults than those in China (OR=7.03, 95% CI: 5.41-9.12, Table 3).”
Table 3 Differences in self-rated health between the US and China in logistic regression models

US vs. China Outcome variable: self-rated health	OR	Lower	Upper	P
Model 1: “excellent” vs. “very good”, “good”, “fair” and “poor”	7.92	5.19	12.09	<0.001
Model 2: “excellent” and “very good” vs. “good”, “fair” and “poor”	3.98	3.18	4.98	<0.001
Model 3: “excellent”, “very good”, and “good” vs. “fair” and “poor”	7.03	5.41	9.12	<0.001
Model 4: “excellent”, “very good”, “good”, and “fair” vs. “poor”	4.24	3.14	5.73	<0.001

Note: OR=odds ratio

Fifth, for the covariates violating the proportional odds assumption, we provided Appendix Table B to present the logistic regression results in the US and China respectively.

Appendix Table B. Results of logistic regression models for variables violating proportional odds assumption in the US and China respectively

Model 1	Model 2	Model 3	Model 4	
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
US				
Education				
High-school	0.86 (0.61, 1.20)	1.01 (0.85, 1.21)	1.50 (1.26, 1.79)	1.13 (0.84, 1.50)
Some college or college and above	1.01 (0.73, 1.40)	1.28 (1.07, 1.53)	1.87 (1.56, 2.25)	1.12 (0.82, 1.52)
Hypertension	0.51 (0.41, 0.63)	0.65 (0.57, 0.74)	0.77 (0.66, 0.91)	0.85 (0.64, 1.13)
Diabetes	0.57 (0.41, 0.79)	0.52 (0.45, 0.60)	0.65 (0.56, 0.76)	0.79 (0.61, 1.01)
Stroke	0.60 (0.39, 0.92)	0.73 (0.59, 0.91)	0.81 (0.67, 0.99)	0.53 (0.40, 0.72)
Arthritis	0.50 (0.41, 0.62)	0.62 (0.54, 0.70)	0.74 (0.63, 0.87)	0.71 (0.51, 1.00)
China				
Education				
Upper secondary & vocational training	0.65 (0.14, 2.88)	0.75 (0.46, 1.24)	1.33 (0.68, 2.60)	1.62 (0.99, 2.64)
Tertiary	0.42 (0.17, 1.05)	1.19 (0.66, 2.13)	0.97 (0.52, 1.78)	-

0.76 0.72 0.74 0.75			
Hypertension			
(0.40, 1.48)	(0.56, 0.91)	(0.60, 0.91)	(0.63, 0.90)
1.32 0.76 0.67 0.96			
Diabetes			
(0.50, 3.48)	(0.47, 1.22)	(0.46, 0.98)	(0.74, 1.25)
1.10 0.70 0.53 0.51			
Stroke			
(0.20, 6.09)	(0.33, 1.48)	(0.31, 0.90)	(0.35, 0.75)
0.55 0.49 0.62 0.84			
Arthritis			
(0.30, 1.04)	(0.38, 0.63)	(0.50, 0.77)	(0.70, 1.01)

Notes:

1. Outcome variable in logistic regression models:
    - Model 1: self-rated health (“excellent” vs. “very good”, “good”, “fair” and “poor”)
    - Model 2: self-rated health (“excellent” and “very good” vs. “good”, “fair” and “poor”)
    - Model 3: self-rated health (“excellent”, “very good”, and “good” vs. “fair” and “poor”)
    - Model 4: self-rated health (“excellent”, “very good”, “good”, and “fair” vs. “poor”)
2. Because only 3 older adults with tertiary education level reported excellent health, the odds ratio was not estimated.
3. All models included socio-demographics (age, sex, educational level, currently working), family structure (living arrangements, number of children), functional limitations (ADLs and IADLs), cognition (self-reported memory, a total recall summary score), chronic conditions (high blood pressure, diabetes, cancer, lung disease, heart problem, stroke, psychiatric problems, and arthritis), mental health, and health-related behaviors (ever drinking and ever smoking).

STATISTICS

Chi-square or t-tests were used to evaluate the statistical significance of differences between the US and China.

Ordered logistic regression models were conducted to investigate factors influencing self-rated health among older adults in the US and China respectively, as well as to see whether there was a cross-national difference on self-rated health between China and the US after controlling those available influencing factors.

“Proportional odds assumption to test whether the assumption (?) was the same across the fivecategory self-rated health variable” – Needs to be clarified and moved to the results section of the manuscript.

“A generalized ordered logistic model yielded similar results, indicating that the estimated difference in self-rated health between the US and China were not influenced by violations of the proportional odds assumption” – This is also results! Move to the results section.

Sensitive analyses (?) using age groups and groups based on number of children showed similar results (?) – Similar to what? Needs to be clarified and moved to the results section of the manuscript。

The authors state that they repeated all the analyses using the 2012-2013 HRS and CHARLS data sets to test the sensitivity of our. Why? If this was done in order to confirm that the 20142015 results, this needs to be mover to the results section of the manuscript. Or delete it.

RESPONSE:: Thank you for the suggestion. We further clarified these points and moved them to the results section of the manuscript. See our response above regarding the proportional odds assumption.

RESULTS
Did you introduce the influencing factors stepwise into the logistic regression models or is Table 2 only showing the fully adjusted models? Needs to be clarified.

RESPONSE: Table 2 shows the fully adjusted models. We did not introduce the influencing factors stepwise into the regression models.

Please put the OR, P and confidence interval in brackets after each statement pertaining to the results so that the reader easily can see how each factor influence self-rated health.

Those who had more ADLs limitations, poorer self-reported memory, worse mental health, and chronic health conditions, had lower self-rated health.
Factors including gender, number of living children, IADLs limitations, and ever smoking were not associated with self-rated health in either China or the US.

Factors positively influencing self-rated health in China – living alone (OR 1.25).

Factors positively influencing self-rated health in the US – younger age (OR 1.02), currently working (OR 1.31), higher educational level (OR 1.48), better recall summary score (OR 1.02), and not drinking/drinking less (OR 1.41).

In the overall ordered logistic regression model when controlling for all factors mentioned in Table 2, older adults in China were much more likely to rate their health as being poor compared to older adults in US when the self-rated health was dichotomized into good (including “good”,
“very good”, and “excellent”) or poor (including “fair” and “poor”) (OR=4.88, 95% CI: 4.065.86).

RESPONSE: Thank you! We put OR, P or confidence interval in brackets when it is appropriate as suggested by the reviewer. When we summarized the similarities between the US and china, we did not attempt to insert the OR, P and confidence intervals because we felt so much detail would distract readers. Readers can easily find the corresponding OR, P and confidence interval in Table 2.
DISCUSSION
The authors state that independence and privacy is highly valued in the US family, while in China living with adult children is more normative because they are expected to take care of their elders, and elder parents are expected to provide grandchild care and that this is supported by their demographic findings. To me it seems strange that this is what they choose to highlight first in their discussion section since the only factor that positively influenced self-rated health in China was living alone (OR 1.25).

RESPONSE:: The associations between living arrangement and self-rated health were different in the US and China. Older adults in China living alone rated their health better than those living with spouse/partner (OR=1.25, P=0.043); however, no significant difference was found between these two living arrangements in older Americans (OR=0.96, P=0.528). In contrast, older adults in the US living with others rated their health worse compared to those living with spouse/partner (OR=0.85, P=0.049); however, no significant difference was found between these two living arrangements in Chinese older

adults (OR=1.06 P=0.605). In the paragraph, we fully discussed possible reasons for the findings related to living arrangements.

The striking difference in (good/poor) self-rated health between China and the US is very interesting. Possible explanations put forth by the authors as to why a larger proportion of Chinese older adults rate their health as poor compared to Americans include under reporting of chronic conditions, lower education and health literacy, limited access to health care and sociocultural desirability (i.e. humbleness). In contrast, older adults in the US may be reluctant to see their health as poor for various reasons. The authors especially put forth their fear of losing independence or becoming a burden. However, since the self-rated health response options were collapsed into two categories with the alternative "fair" being put in the poor category this also might contribute to the striking difference between countries. This needs to be further elaborated in the discussions section.

RESPONSE: Actually, "fair" was not put in the poor category. We used the full five-categories of self-rated health in the analyses rather than a binary indicator, because information is lost when collapsing categories. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

Reviewer: 2

Overview: The manuscript used two representative nationwide survey data in the US and China to compare self-reported health between the two countries. The authors showed that China respondents reported much worse health status than their US counterparts.

Comments:

1. My biggest criticism to this manuscript is their outcome variable self-rated health. Since people have very different interpretation on "good" health in American and Chinese culture, how can the authors compare the outcome between the two countries? what is the practical meaning of comparing self-reported health between Chinese and American citizens?

RESPONSE: We agree, "people have very different interpretation on 'good' health in American and Chinese culture." We attempted to arrive at differences in perception by controlling for objective health status factors and other potential covariates or confounders. The perception of good health seems to be the main difference, since a striking difference in self-rated health remains after introduction of controls variables.

2. In the results of the abstract, I suggest the authors add significant factors for self-reported health in the two countries since it is one of the two major objectives of this study.

: Thank you for your suggestion. We added it.

3. Page 4, Line 39-42, can the authors provide specific numbers and years in the US and China

RESPONSE: We now state, "The prevalence of hypertension is higher in the US (46.9% vs. 38.6%) among adults aged 45 to 75 years old during 2011-2012".

4. The authors merged the two databases and estimated the model in one ordinal logistic regression by adding a country index. In this statistical model, the authors are implicitly assuming other covariates had the same effect on the outcome across the two countries. However, in the Introduction, the authors claimed a lot of differences between the two countries. Would two separate models or adding interactions between country index and covariates be more appropriate given the authors' introduction?

RESPONSE: We analyzed the data using two separate models, as shown in Table 2. We included many covariates in the model. So adding interactions between country index and covariates is not workable. As the reviewer suggested, we ran two different regression models (Table 2) to investigate whether factors influencing self-rated health among older Chinese were similar to those among older Americans. In order to investigate whether there was a significant cross-national difference on self-rated health, we merged the two databases and added a country variable (the US vs. China) while at the same time controlling for sociodemographics, family structure, functional limitations, cognition, chronic conditions, mental health, and health-related behaviors (Table 2).

5. On page 8, line 33-42, the authors claimed that "a test of proportional odds assumption was performed; demonstrating the assumption did not hold across the five categories.". The authors should specify the name of the test, test statistics and P-value. In addition, if the proportional odds assumption is not met, they shouldn't have used ordered logistic regression.

RESPONSE: We further clarified this in the method section. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

6. The odds ratio of ordinal logistic regression should be interpreted as the cumulative odds of reporting good health status. The interpretation in this study looks more like the interpretation of a binary logistic regression results.

RESPONSE: We revised the manuscript with the interpretation of cumulative odds as the reviewer suggested. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

We now state, "The odds of having better versus poorer health was almost 5 times greater in American older adults than those in China (OR=4.88, 95% CI: 4.06-5.86, Table 2). Because of the issue with the proportional odds assumption, we performed sensitivity analysis with alternative models. When shifting comparison pivot point down the self-rated health scale, we found the odds ratios range from 3.98 to 7.92 in the logistic regression models (Table 3). For

example, the odds of having the combined “good”, “very good”, or “excellent” health versus “fair” or “poor” health was 7 times greater in American older adults than those in China (OR=7.03, 95% CI: 5.41-9.12, Table 3).”

7. Page 8 line 37-38, a generalized ordered logistic regression seems to be a confusing term to me. From my understanding, an ordered logistic regression is equivalent to proportional odds model. Is the “generalized ordered logistic regression” a proportional odds model or some other model?

RESPONSE: We further clarified this in the method section. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

8. Page 8 line 42-47, it is very unlikely that all the estimates, standard errors and P-values are exactly the same in your sensitivity analysis and main model. If the number of tables did not excel the limit, I recommend the authors also include the sensitivity analysis results.

RESPONSE: We provided additional tables and described the sensitivity results in the text.

9. Page 12 line 31-42, the explanation of Chinese elders having a lower education level seems to be contradictory to the results. If older Chinese elders had lower health literacy, wouldn’t they report better health status than American elders since they are dismissive of their disease?

RESPONSE: Because of the lower health literacy, Chinese older adults may not realize they have the disease, but also they may have limited capacity to obtain, process, and understand health information and services. As a result they may be less likely to receive treatments, have poor self-management of chronic conditions and experience more severe symptoms. For example, Lu et al. (2018) found that compared to the US, China had a higher proportion of patients with severe hypertension (10.5% vs. 4.5%) and lower rates of hypertension treatment (46.8% vs. 77.9%) and control (20.3% vs. 54.7%) among population aged 45-75 years old, even though the prevalence of hypertension was lower in China.

10. Given that Chinese older respondents had a significantly high odds ratio of 4.88, can it be that they are really worse in physical/psychological health, instead of just self-reported health?

RESPONSE: We do not exclude this possibility. Although we controlled all available variables related to psychical/psychological health in the models, the difference in self-rated health between the US and Chinese older adults is still large. Unmeasured health status variables may account for the difference.

Reviewer: 3

I believe the authors need to indicate how they measured each variable and the manner they dealt with missing values in the text. This is a major roadblock to replication of this study. I have discussed in detail the need to reshape the outcome, perform a binary logistic regression rather than an ordered regression. This is reinforced by the fact that the authors STATE that the proportional odds assumption is not met (by the model and the data).

RESPONSE: We clarified the way we dealt with missing values in the text. Regarding the regression models, we further clarified in the method section the rationale to use ordered logistic regression. We also provided additional tables to present results from the binary logistic regression models for

variables that violated the proportional odds assumption. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

Review of: Self-Rated Health among Older Adults: A Comparison of China and the United States

Abstract
Remove the words recent and large from the objective section, these are good qualities of the data but it should suffice by saying a nationally representative sample. There are some concerns with other sections of the abstract but I will address them within the manuscript in my line-byline review; the same goes for the Strengths and limitations section presented in page 3.

RESPONSE: We removed the words recent and large from the manuscript.

Introduction
The first statement of the second paragraph (lines 17-23) is misleading. There are multiple examples in the literature where Self-Reported Health is studied using samples from different countries (cross-national analyses).
To mention one:
1. Hardy, Melissa A., Acciai, Francesco, and Reyes, Adriana M. How health conditions Translate into Self-Ratings: A comparative study of older adults across Europe. Journal of Health and Social Behavior.
The SHARE dataset has provided the ability to do these cross-national comparisons. Probably, a best way to frame this, is to say that China-US studies are not common and state WHY this comparison is important. The seeds are there, for example social support and support networks are expected to better help those in China in comparison to US, however this expectation has not been met by the aging of persons with only one child etc. Maybe just do away with this statement and focus on the China-US comparison, it is enough of a justification to do it if well established.

RESPONSE: We revised the introduction section as the reviewer suggested.

I would say, no need to mention recent or large in the description of the dataset. Just state it is a nationally representative sample of Chinese and US older population - the message will get through. In lines 29-31 of page 5 of the manuscript, please finish the sentence in a manner similar to this one "national differences of self-rated health between China and the US after controlling for potential covariates (or confounders)".

RESPONSE: We removed the words recent and large from the manuscript. We revised the sentences as the reviewer suggested.
Methods
Data and Sample
Please include a sentence on how you reached the final sample size (what was the initial size?) and what you did with missing values (list-wise deletion?).

RESPONSE: We added the initial sample size and used list-wise deletion for handling missing data. "Initially, 10,374 older adults in the US and 5,751 older adults in China reported their health status. Number and percentage of missing data is presented in Appendix Table A. We used list-wise deletion for handling missing data."

Measures

Self-Rated Health
Did you include the variable as a 1-5 in the regression model? This may be problematic, conventional studies go with the dichotomization of the variable and simply fit a logistic binary regression model. At least you should say whether the results vary by specification, in my experience it does. The 1-5 and ordered regression specification may not be the best way to approach this outcome. You can decide how to do it, but if so - provide some robustness checks on whether the results differ if you dichotomize this. Right now it is unclear if some of the associations may be influenced by the scale. Is Self-Rated Health really an ordered outcome?

RESPONSE: We included self-rated health as a 1-5 in the ordered regression model. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

Sociodemographic and family structure
You mention age and gender (sex?) in your first sentence. But you don't say how you operationalized them. Is age continuous? If so, did you include a square term in your model to control for potential curvilinear trends? If it was categorized then the associations will show you the shape of the associations in the magnitude of the OR or Coefficients. I don't know if calling Male/Female dichotomy a Gender variable is appropriate, maybe it is, but most recently gender encompasses more than this dichotomy. Can you refer to it as Sex?

RESPONSE: age was a continuous variable. We checked the square term of age and found it was not significant. Then we dropped the square term of age in the models. "In sensitivity analyses, we categorized age (65-74, 75-84, and 85+) and found that age groups were not association with self-rated health in China, while in the US, compared to those aged 65-74, older adults aged 75-84 and aged 85+ reported better health respectively." We used sex in the text.

Functional limitations
Why is ADL doe as a summary score of 0-3 and IADLs a binary when you have the variables to create a 0-4 variable. It seems quite inconsistent here, either dichotomize both (and you need to provide previous literature that operationalized it that way) or just add them both as continuous. It may trigger a reader to ask this if this is allowed to go through as it is.

RESPONSE: Both ADLs and IADLs were used as categorical variables. The ADLs variable was used a four-level categorical variable instead of a binary variable to show the trends in OR estimates. As ADLs impairment increased, older adults reported worse health in both countries.
There were no significant associations between IADLs and self-rated health regardless the IADLs variable was used as a binary variable or as a four-level categorical variable like the ADLs variable. The following table is part of Table 2.

Self-rated health	China			US		
	OR	P	95% CI	OR	P	95% CI
ADLs						
1	0.52	<0.001	0.40 0.67	0.48	<0.001	0.40 0.57
2	0.35	<0.001	0.24 0.51	0.25	<0.001	0.18 0.33
3	0.31	0.001	0.16 0.62	0.23	<0.001	0.14 0.40
IADLs	0.91	0.288	0.76 1.08			
	0.90	0.302	0.74 1.10			

Cognition

Here we have an interesting case of a variable that would really exemplify my concerns with the operationalization of SRH as a 1-5 variable. In the outcome 1 is worse and 5 is better; however, in this variable of Self-Reported Memory lower is better and higher is worse. If the outcome is dichotomized I wouldn't have a problem, but the outcome is ordered and this order may just provide a counterintuitive association if one does not reads carefully. If the author decides to use the 1-5 scale, they should flip it (and say it in the text) so that things go in the same direction. So 1 would be poor and 5 excellent in this "flipping" of the variable. Regarding the recall summary score - is it associated/correlated with the other cognition variable? Why are both being included? As it stands now it seems they were just thrown in because they were available.

RESPONSE: We labeled each category of the cognition variable in Table 2. The correlation coefficients between self-reported memory and recall summary score were -0.209 in Chinese data set and -0.225 in the US data set. Moreover, we calculated the variance inflation factor (VIF) to check for multicollinearity. The VIF values of all variables in the model ranged from 1.01 to 2.66. As a rule of thumb, a variable whose VIF values are greater than 10 may be a concern and merit further investigation. One of our objectives was to investigate national differences in self-rated health between the US and China after controlling for potential covariates. Therefore, we included all available variables in the data sets that might contribute to a cross-national difference.

Chronic Conditions and Mental Health
I would recommend separating these two sections. But I'll comment on it, as it stands. Why these eight chronic conditions? It comes to mind that these conditions would be associated with functional limitations ADL or IADL (also cognition). I am hard pressed to see why the models are going to include SO many variables that encompass the same factors or capture the same effects. At least the authors should provide VIFs (OLS of SRH as a 1-5 with variables included in the model) or a similar indicator that indicate these variables are not introducing redundancy in the models. My hunch is that they will, and are doing it.

RESPONSE: We separated these two sections as suggested by the reviewer. In the two datasets, information of these eight chronic conditions were available. The VIF values of all variables in the model ranged from 1.01 to 2.66. Because VIF values were all less than 10, multicollinearity is not a concern in the analyses. In addition, the highest correlating coefficient among all variables was only 0.473 between education and recall summary score.

Health-related behaviors
These are bound to be associated at least with Chronic Conditions (see comment above). Again, I'm concerned the model is over specified and some variables are introducing redundancy that may hide the REAL association of the variable of interest.

RESPONSE: We calculated the variance inflation factor (VIF) to check for multicollinearity. The VIF values of all variables in the model ranged from 1.01 to 2.66. Because VIF values were all less than 10, multicollinearity is not a concern in the analyses. One of our objectives was to investigate a national difference in self-rated health between the US and China after controlling for potential covariates. Therefore, we included all available variables in the data sets that might contribute to a cross-national difference.

Statistical Analysis
Sentence 2, why ordered logistic regression models. See my comments above about this. It is clear there is a difference even if we dichotomize by poor/fair as 1 and the rest as 0. We are interested in

poor/fair not the other way around. I've found the dichotomization is much better specification, and if you see more recent work on SRH they are moving away from the Ordered Logistic or other Multinomial specifications. A look at recent work by Anna Zajacova could be helpful in seeing the vast number of papers published using this specification.

RESPONSE: Thank you for your suggestion. Please refer to Pages 3-5 for our detailed response regarding the analyses of self-rated health.

My biggest concern is not the model specification but the WAY the data are analyzed. It is very difficult to see what is gained from joining both datasets. It would suffice to perform the regression analysis on both datasets and compare the coefficients seeing whether the Confidence Intervals derived from the associations have any overlap or not. If this is the standard way you need to state it and provide reference for example to the rescaling of weights and how things were treated to be a valid sample design. Right now this is the weakest section of the paper. What is really being captured by adding a dummy indicating China/US - that the data come from a different dataset not necessarily what is being said it measured in the model. I think just specifying the FULLY specified model for both countries and comparing effects would suffice for the purposes of the paper. Take away the overall model, and just discuss the US China models and then differences/similarity in associations.
1. If you dichotomize you deal with the issue of non-proportional odds of the association. 2. If you fit the models and compare the Coefficients across models, without joining the datasets, you deal with issue of a flawed data design. Which may render your finding invalid.

RESPONSE: The purpose of the two separate models for China and the US (Table 2) was to investigate whether covariates influencing self-rated health among older Chinese were similar to those among older Americans. As the reviewer suggested, we investigated the cross-national difference in self-rated health between China and the US by merging the two datasets, adding a country variable (the US vs. China), and controlling for sociodemographics, family structure, functional limitations, cognition, chronic conditions, mental health, and health-related behaviors (Table 2).

It is good that the way age and number of children is operationalized differently in the sensitivity analyses are included here, but it needs to be said also in the measurements section (as stated before). I have trouble with the phrase "statistical significance was accepted". Could this be phrased as "Statistical significance was established at the 95% level ($p < 0.05$).

RESPONSE: We revised the sentence as the reviewer suggested.

Results
Something to note here, is that Chinese seem to have better indicators regarding conditions than US older adults, but their mental health and cognitive markers are different. This may lay work for additional analyses in the future. The authors should consider my previous comments regarding model specifications. The results section is well written, but it is flawed because it does not meets the proportional odds assumption, a key assumption for these kind of models. If the author makes the revisions laid before I think they will end up with a stronger paper and a great section with results that are easily interpretable. I would include a whole section discussing what is found here in the regression models, despite it being the most important part of the paper its discussion is really superficial. Again, if the author follows some of my recommendations this results section will be really great to read.

RESPONSE: We revised the manuscript as the reviewer suggested.

Discussion
The first sentence here (and in the Conclusions) still is anchored in the recent dataset, although timeliness of analysis is great there is amazing science being done with recently recovered datasets (NHANES I, NHANES II, NHANES III - to mention some). The big element here is that you have two nationally representatives datasets that collect data in a way that makes this paper possible. If the results found in the initial models are found in the new model specification (which may be the case) I would review this section to be sure it is consistent with new findings (if any).

RESPONSE: We revised the first sentence in the discussion and conclusion sections as the reviewer suggested. We provided additional tables and result description from the logistic regression models in the results section.

Conclusion
See above, not further comments.

RESPONSE: Thank you very much!