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)	Compress the Curve: A Cross Sectional Study of Variations in COVID-19 Infections Across California Nursing Homes
AUTHORS	Gopal, Ram; Han, Xu; Yaraghi, Niam

VERSION 1 – REVIEW

REVIEWER	Charlene Harrington Professor University of California San Francisco, USA I have published a new paper on COVID infections in CA nursing homes. Harrington, C., Ross, L., Chapman, S., Halifax, E., Spurlock, B, and Bakerjian, D. 2020. Nursing staffing and coronavirus infections in California nursing homes. Policy, Politics, & Nursing Practice. DOI: 10.1177/1527154420938707.
REVIEW RETURNED	03-Aug-2020

GENERAL COMMENTS	Compress the Curve BMJ Open This is an interesting article that addresses a timely topic of COVID-19 infections in nursing homes. Specific comments: I would suggest a change in the title since this is not really an “observational study” – at least not from a US perspective. I would just call it Variations in COVID-19 Infections in California Nursing Homes The introduction and methods sections are clearly written. Limitations One limitation in the data is that it uses 2017 data which is over two years prior to the outbreak. This should be listed as a limitation since 2019 data were available at the time of the study. Because the authors are using a variable constructed called an “inflation score” and the “self reported indicator of quality” from 2017 data, the study lost a number of nursing homes from the data set (493 facilities) which would otherwise been available. Another limitation of the data is the potential inaccuracy in reporting that the authors should mention. It should be noted that 86 percent of CA nursing homes are for-profit and these nursing homes were probably more likely to under-report their infection rates and deaths than other nursing homes for fear of losing residents and money.

Data Sources and Descriptions

Also the authors should check the data source, but the California data only listed facilities that had 11 or more infections (not 10 or more as stated in the paper on p. 7 and throughout the paper).

Table 1 lists the sources and descriptions of study variables. On the Certified beds, this label should say number of beds certified to provide Medicare and Medicaid (not just Medicaid).

I am not sure why the study would include whether or not a facility has a family council.

In terms of the inflation score, how many years was it possible to have a significant discrepancy? The inflation score is not a standard measure but one developed by the authors. It would be good to add a little more description to how this score was calculated.

Also the authors might want to explain the quality rating – it was a CMS calculated score based on a summary of a number of self-reported quality measures such as the rate of rate of pressure ulcers.

On the staffing rating description on Table 1, it should be pointed out that the rating is based on a combination of RN hours per resident day and the total nursing hours per resident day. The label should say a self-reported measure of staffing hours (not features). In 2018, CMS did correct the staffing data so that the source was payroll data rather than self-reported and thus more accurate.

Methods

One feature of the analysis is that it uses four methods, a zero inflated bivariate poisson and a zero inflated double poisson model. These are then compared to a neural networks and a support vector machine with radial basis function methods. While the analysis argues that the NN and the SVM-RBF models useful, it is not clear whether these last two analyses added information not found and reported in the poisson models. If they do not add value to the methods, perhaps they could be omitted because the average reader like myself has no understanding of these advanced models.

Findings

The findings from this study were very interesting and useful. One thing the paper does not point out are the complex relationships between the main variables in the model. For one thing, for-profit NHs generally have lower nurse staffing, more deficiencies, are larger in size, and probably have a greater likelihood of inflating their ratings. It makes sense that they were found to be more likely to have larger numbers of COVID infected residents and staff.

The findings on the inflation of self-reported quality ratings were very useful. It is logical that NHs that inflate their quality ratings are more likely to be having quality problems which are probably related to low staffing and high deficiencies.

	Discussion It seems to me that in spite of the CMS reporting problems related to the quality measures and the staffing data (which was corrected in 2018), at least the US has a system for standardized data collection of ownership, size, etc. and an overall quality rating system which can be used by residents, families, discharge planners and others in contrast to the UK and other EU countries. Even though the US has had its national NH rating system in effect since 2009, the federal and state governments failed to make NHs residents a priority for protection during the pandemic and the government failed to utilize the data they had to target those high risk NHs. The government should have targeted the poor quality NHs immediately during the pandemic for oversight and intervention to prevent infections and deaths. No doubt they could have save thousands of lives. Since the NH ownership, size, and ratios did have predictive value, it seems like the authors' findings should be a strong endorsement for using provider data and a similar quality rating system in the UK and other industrialized countries. Countries like the UK, where NH care is largely provided by for-profit NHs and chains, have had serious quality problems. (See Armstrong and Armstrong 2019. Privatization of Care book by Rutledge and many NH comparison studies of the US and UK). The UK and EU countries also had a large proportion of COVID-19 deaths in nursing homes and perhaps they too could have saved lives. I think this should be added to the discussion.

REVIEWER	Raymond Bond Ulster University, UK
REVIEW RETURNED	16-Sep-2020

GENERAL COMMENTS	The paper is very interesting and is well written. The paper points out that for-profit nursing homes are much more likely to have large outbreaks of COVID-19. This result alone is an important finding. The researchers analyse data from 713 care facilities in the state of California. A number of points to consider in the review of the manuscript: - I would not say that the findings are limited to California but rather the findings are limited by the fact that the study only analysed data from California. - It is difficult to understand the abstract when 'CMS database' is not defined - Further description of what self reported measures are and what CMS inspections are - and how the data is collected - would help the clarity of the introduction - Whilst you simply reference to another article to explain ZIBP method - I believe it is best to have your own short justification and description of this method and what it is and how you use it - Please present what occupancy rate is and how it is calculated on page 9 line 29. - Whilst you have supplementary material, the paper should have details on how you trained and evaluated the machine learning models and any other models.

	- Supplementary material could also show the confusion matrices which provide clarity in the results - I am unsure if you are predicting large outbreaks or classifying - of course a classification can be a prediction - a very minor point - When you say 'Interestingly, self-reported quality ratings are associated with larger size of outbreaks.' - do you mean 'higher self-reporting quality ratings...' - Figures 1,2 and 3 need figure descriptions - Table 2 and table S1 should clarify what the estimate/coefficient is, this could be for example log of odds ratios or odds ratios etc. Overall this paper looks interesting but would benefit from more clarity and detail. I also think the paper needs to present more limitations of the work in the discussion section.

REVIEWER	Derek Young University of Kentucky
REVIEW RETURNED	06-Oct-2020

GENERAL COMMENTS	Comments on Compress the Curve: An Observational Study of Variations in COVID-19 Infections Across California Nursing Homes by Gopal, Han, and Yaraghi (bmjopen-2020-042804) Summary In this article, the authors study why some nursing homes are more susceptible to larger COVID-19 outbreaks. This research is an observational study of all nursing homes in the state of California up until 05/01/20. Overall Comments Overall, the authors present an interesting analysis of these COVID-19 data. While the study is restricted to the state of California, this type of modeling provides a good foundation for exploring similar models for other states as well as for a possible national-level model, assuming that comparable datasets can be assembled. Comments 1. Software for performing estimation of zero-inflated double Poisson (ZIDB) and zero-inflated bivariate Poisson (ZIBP) regression models is not standard. The authors should state how they performed estimation (e.g., clearly maximum likelihood estimation was performed, but via what type of algorithm) as well as which software language they used. Moreover, for the sake of transparency, the code for their analysis should be made available even if the data cannot be shared. 2. While the correlation between the number of infected staff and the number of infected residents is definitely a tenable assumption, and this is backed-up with a statistical analysis near the end of the manuscript, including a simple jittered scatterplot of these bivariate counts will be a highly-effective visualization. 3. A sensitivity analysis of the proposed models should also include fitting their non-zero-inflated counterparts; i.e., a bivariate Poisson regression model and a double Poisson regression model. 4. Model selection criteria are calculated and reported in Table 2. However, no interpretation of those

	results is given in the paper. The authors should provide a couple of sentences interpreting those results. 5. On page 11 and for Figure 2, what is meant by ”lift” of the ZIBP models? Please clarify. 1 6. Consider proposing in the discussion a spatial random effect to account for spatial dependencies, which could be synthesized with any temporal data made available in the future for which a zero-inflated spatio-temporal model would be appropriate

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer 1

This is an interesting article that addresses a timely topic of COVID-19 infections in nursing homes.

Response: Thank you very much for supporting our research and providing us with such detailed feedback. We have incorporated all your suggestions in this revision. Please find our detailed response to your comments below.

Specific comments:

I would suggest a change in the title since this is not really an “observational study” – at least not from a US perspective. I would just call it Variations in COVID-19 Infections in California Nursing Homes

Response: We appreciate the suggestion. We have now changed the title in the revised manuscript. The new title is now “Compress the Curve: Variations in COVID-19 Infections Across California Nursing Homes” The introduction and methods sections are clearly written.

Response: We appreciate your affirmation and will endeavored to be clear in our exposition.

Limitations:

One limitation in the data is that it uses 2017 data which is over two years prior to the outbreak. This should be listed as a limitation since 2019 data were available at the time of the study. Because the authors are using a variable constructed called an “inflation score” and the “selfreported indicator of quality” from 2017 data, the study lost a number of nursing homes from the data set (493 facilities) which would otherwise been available.

Response: Thank you very much for raising these points. We have now clearly identified the limitations of our study in the last paragraph of the Discussion section. It is worth noting that to obtain all the key variables needed for our analysis, we had to merge various datasets with different primary keys and structures. The 493 facilities were dropped mainly because of missing values and inconsistencies in primary keys.

Another limitation of the data is the potential inaccuracy in reporting that the authors should mention. It should be noted that 86 percent of CA nursing homes are for-profit and these nursing homes were probably more likely to under-report their infection rates and deaths than other nursing homes for fear of losing residents and money.

Response: Thank you for raising this important limitation. We have now noted this in the last paragraph of the Discussion section in the revised manuscript as well as in the Article Summary section.

Data Sources and Descriptions:

Also the authors should check the data source, but the California data only listed facilities that had 11 or more infections (not 10 or more as stated in the paper on p. 7 and throughout the paper).

Response: Thank you for pointing this out. We used the wording “more than 10” throughout the paper, which is consistent with “11 or more” as you pointed out.

Table 1 lists the sources and descriptions of study variables. On the Certified beds, this label should say number of beds certified to provide Medicare and Medicaid (not just Medicaid).

Response: Thanks for catching this. We have now fixed it.

I am not sure why the study would include whether or not a facility has a family council.
Response: The variable family council is included in our analysis since it is one of many features in the CMS rating data differentiating nursing homes. While almost all nursing homes have resident councils, only 20 percent of nursing homes have existing family councils. We included this in our analysis with the contention that family councils imply closer coordination and higher engagement with the families of the residents. We have included this explanation in section Data Sources and Study Variables in the revised manuscript.
In terms of the inflation score, how many years was it possible to have a significant discrepancy? The inflation score is not a standard measure but one developed by the authors. It would be good to add a little more description to how this score was calculated.
Response: The inflation score is calculated based on the method described in Han et al.,[1] in which likely inflators are predicted for a consecutive 5-year period. We aggregated the result and used the number of years a nursing home is predicted to be a likely inflator to be the overall inflation score for a nursing home. Therefore, an honest nursing home will have an inflation score of 0 while an inflating nursing home can have an inflation score between 1 to 5, with 5 being the most severe. In our dataset, 19.25% of nursing homes were inflating their scores and some of these had a score of 5 indicating that they inflated their scores in all 5 years. We have included these details in Data Sources and Study Variables section of the revised manuscript.
Also the authors might want to explain the quality rating – it was a CMS calculated score based on a summary of a number of self-reported quality measures such as the rate of pressure ulcers.
On the staffing rating description on Table 1, it should be pointed out that the rating is based on a combination of RN hours per resident day and the total nursing hours per resident day. The label should say a self-reported measure of staffing hours (not features). In 2018, CMS did correct the staffing data so that the source was payroll data rather than self-reported and thus more accurate.
Response: Thank you for pointing this out. We have created a new section called Description of CMS' Nursing Home Compare System in which we provide a summary of quality, staffing and inspection ratings and how the overall rating is calculated by CMS. We have also revised Table 1 and included the description of staffing per your recommendation.
Methods:
One feature of the analysis is that it uses four methods, a zero inflated bivariate Poisson and a zero inflated double Poisson model. These are then compared to neural networks and a support vector machine with radial basis function methods. While the analysis argues that the NN and the SVM-RBF models useful, it is not clear whether these last two analyses added information not found and reported in the Poisson models. If they do not add value to the methods, perhaps they could be omitted because the average reader like myself has no understanding of these advanced models.
Response: Machine learning models such as NN and SVM are popular in data analytics and have been shown to perform well in predictions in different application settings. In our setting, these machine learning models can be potentially useful for prediction purposes in case our problem domain encompasses highly nonlinear relationships not captured in our baseline ZIBP model. Our thought process was to include these methods in the analysis to ascertain possible nonlinearities and the predictive power of these machine learning models. We are also cognizant of the fact that even if these methods prove to be highly predictive, they suffer from the problem of "explainability" as they are black-box methods that unlike our approach cannot offer detailed insights on the impacts of various independent variables on COVID-19 infections. However, in case their predictive power were far superior to our regression models, they could be of practical use to predict COVID-19 cases in nursing homes. Our results show that our ZIBP model outperforms SVM and that the predictive ability of the NN is only modestly better than ZIBP model. That is, the application and comparison of these machine learning models with the results of the ZIBP model confirms that not only the ZIBP model has the advantage of "explainability", but it also offers competitive predictive performance. We have included the above discussions in the Discussion section of the revised manuscript.
Findings:

The findings from this study were very interesting and useful. One thing the paper does not point out are the complex relationships between the main variables in the model. For one thing, forprofit NHs generally have lower nurse staffing, more deficiencies, are larger in size, and probably have a greater likelihood of inflating their ratings. It makes sense that they were found to be more likely to have larger numbers of COVID infected residents and staff.

The findings on the inflation of self-reported quality ratings were very useful. It is logical that NHs that inflate their quality ratings are more likely to be having quality problems which are probably related to low staffing and high deficiencies.

Response: Thank you very much for these useful insights. These comments prompted us to explore the literature a bit more and have now included these points in the Discussion section in the revised manuscript.

Discussion:

It seems to me that in spite of the CMS reporting problems related to the quality measures and the staffing data (which was corrected in 2018), at least the US has a system for standardized data collection of ownership, size, etc. and an overall quality rating system which can be used by residents, families, discharge planners and others in contrast to the UK and other EU countries.

Even though the US has had its national NH rating system in effect since 2009, the federal and state governments failed to make NHs residents a priority for protection during the pandemic and the government failed to utilize the data they had to target those high risk NHs. The government should have targeted the poor quality NHs immediately during the pandemic for oversight and intervention to prevent infections and deaths. No doubt they could have save thousands of lives.

Since the NH ownership, size, and ratios did have predictive value, it seems like the authors' findings should be a strong endorsement for using provider data and a similar quality rating system in the UK and other industrialized countries. Countries like the UK, where NH care is largely provided by for-profit NHs and chains, have had serious quality problems. (See Armstrong and Armstrong 2019. Privatization of Care book by Rutledge and many NH comparison studies of the US and UK). The UK and EU countries also had a large proportion of COVID-19 deaths in nursing homes and perhaps they too could have saved lives. I think this should be added to the discussion.

Response: Thank you very much for the insightful feedback. We agree that an important takeaway from this research is the importance of data collection and transparency. Our research was made possible because of the availability of key information on COVID-19 infections in nursing homes in the US and publicly available data such as ownership, size, staffing, and key performance measures. Access to such data is invaluable in both understanding and taking preventive action to curb the COVID-19 infections in nursing homes. As such we hope that other industrialized nations take necessary steps to collect and disseminate such information to protect and safeguard the vulnerable residents in long-term care facilities. We have included the above paragraph in Discussion section.

In closing, we would like to extent our most sincere appreciation for your excellent feedback and comments. We have tried our best to incorporate all your suggestions in this revision and believe that the paper is now much better as a result. We hope that you also find our efforts in improving the paper to be fruitful.

Response to Reviewer 2

The paper is very interesting and is well written.

The paper points out that for-profit nursing homes are much more likely to have large outbreaks of COVID-19. This result alone is an important finding. The researchers analyze data from 713 care facilities in the state of California.

Response: We are very glad that you have found our research interesting and appreciate your support and constructive feedback. In the following, we provide detailed response to your suggestions.

A number of points to consider in the review of the manuscript:

1. I would not say that the findings are limited to California but rather the findings are limited by the fact that the study only analyzed data from California.

Response: Thank you for the suggestion. You are completely right. We have now modified this statement in Article Summary section as follows: The findings of the study are limited by the fact that the study was conducted with data only from California.

2. It is difficult to understand the abstract when 'CMS database' is not defined

Response: We have revised the abstract and instead of referring to “CMS database” we state: “data on ratings and governance features of nursing homes provided by CMS”. We have also added a new section in the paper to provide detailed description of the CMS rating system.

3. Further description of what self-reported measures are and what CMS inspections are - and how the data is collected - would help the clarity of the introduction

We thank the reviewer for the comments. In the revised manuscript, we have added a new section called Description of CMS’ Nursing Home Compare System in which we provide detailed description of the CMS’s nursing home compare system, CMS inspections, and quality and staffing ratings. We have also added a reference for “CMS database on ratings and governance features” which links directly to the URL to CMS data.

4. Whilst you simply reference to another article to explain ZIBP method - I believe it is best to have your own short justification and description of this method and what it is and how you use it
Response: Thank you very much for your suggestion. We have added a paragraph to the Statistical Analysis section to provide more justification about the application of the model. To ensure transparency and replicability of our study, we uploaded both data and the SAS code for estimating the model on a publicly accessible repository. They are now available via the following links:

Data:
https://figshare.com/articles/dataset/Data_for_COVID19_in_California_nursing_homes/13148813
Code:
https://figshare.com/articles/software/SAS_code_for_BMJ/13179875

5. Please present what occupancy rate is and how it is calculated on page 9 line 29.

Response: Thank you for questioning on this variable. In our analysis, occupancy rate is calculated by using the number of enrolled patients divided by the number of certified beds of the nursing home. We have added this clarification on page 9.

6. Whilst you have supplementary material, the paper should have details on how you trained and evaluated the machine learning models and any other models.

Response: Following your suggestion, we have added a paragraph to provide more details on the training and evaluation of our machine learning models in the Statistical Analysis section.

7. Supplementary material could also show the confusion matrices which provide clarity in the results

Response: Thank you very much for this great suggestion. We have added two new tables in the Supplementary Appendix (Tables S4 and S5) in which we present the confusion matrices for NN and SVM models.

8. I am unsure if you are predicting large outbreaks or classifying - of course a classification can be a prediction - a very minor point

Response: Our intention is to predict large outbreaks. As you have rightly pointed out, classification can be a prediction.

9. When you say 'Interestingly, self-reported quality ratings are associated with larger size of outbreaks.' - do you mean 'higher self-reporting quality ratings...'

sResponse: Exactly. We have slightly revised this sentence to make it clear. It now reads "higher self-reported quality ratings are associated with larger size of outbreaks."

10. Figures 1,2 and 3 need figure descriptions

Response: The title and descriptions of the figures are now available in the manuscript right after the References section.

11. Table 2 and table S1 should clarify what the estimate/coefficient is, this could be for example log of odds ratios or odds ratios etc.

Response: Thank you very much for pointing this important issue out. We have provided this clarification both in the main text and also in the table notes for both Table 2 and Table S1.

Overall this paper looks interesting but would benefit from more clarity and detail. I also think the paper needs to present more limitations of the work in the discussion section.

Response: In closing, we would like to thank you again for your great suggestions and feedback. We hope that you find the new version of our manuscript satisfactory.

Response to Reviewer 3

Summary

In this article, the authors study why some nursing homes are more susceptible to larger COVID19 outbreaks. This research is an observational study of all nursing homes in the state of California up until 05/01/20.

Overall Comments

Overall, the authors present an interesting analysis of these COVID-19 data. While the study is restricted to the state of California, this type of modeling provides a good foundation for exploring similar models for other states as well as for a possible national-level model, assuming that comparable datasets can be assembled.

Response: Thank you very much for your very helpful feedback. In this revision, we have tried our best to implement all your suggestions. In the following, we provide detailed responses to your comments.

Comments

1. Software for performing estimation of zero-inflated double Poisson (ZIDB) and zero-inflated bivariate Poisson (ZIBP) regression models is not standard. The authors should state how they performed estimation (e.g., clearly maximum likelihood estimation was performed, but via what type of algorithm) as well as which software language they used. Moreover, for the sake of transparency, the code for their analysis should be made available even if the data cannot be shared.

Response: We have used NLMIXD procedure in SAS software to develop the maximum likelihood functions ourselves. Following your suggestion, we have uploaded the code for these methods. In addition to that, we have also uploaded the de-identified data, so all the results can be replicated using the code. Both the code and data are available on Figshare.com via the following links:

Data:

https://figshare.com/articles/dataset/Data_for_COVID19_in_California_nursing_homes/13148813

Code:

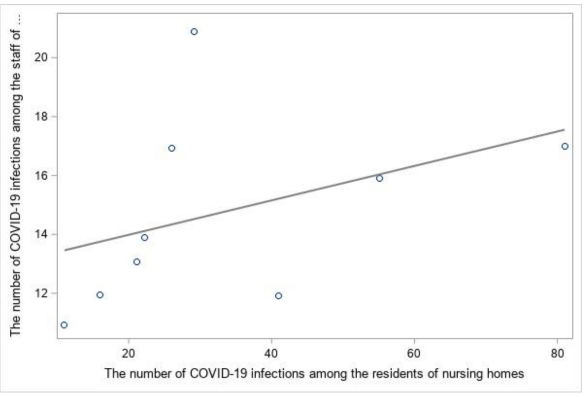
https://figshare.com/articles/software/SAS_code_for_BMJ/13179875

We have provided these detailed and the links to data and code in Statistical Analysis section of the revised manuscript.

2. While the correlation between the number of infected staff and the number of infected residents is definitely a tenable assumption, and this is backed-up with a statistical analysis near the end of the manuscript, including a simple jittered scatterplot of these bivariate counts will be a highlyeffective visualization.

Response: Thank you very much for this great suggestion. Since only 22% of the nursing homes report at least one infection among either their staff or residents (reported in Table 1), a scatter plot does not help with visual representation of the correlations as most of the points will be concentrated around zero points. The other issue is that California censors the data and does not report the exact number of infections unless it is more than 10. This will lead to even a lower number of cases for which we can exactly know the number of infections among both staff and residents. Therefore, to better show the correlations among the infections, we thought the better way would be to provide a table and a scatter plot together. We have added a new section called "Correlation Between Infections Among Staff and Residents" in the Supplementary Appendix and provided a detailed discussion. For your convenience, we reproduce the table, figure and the corresponding discussion below.

"we report the number and percentage of nursing homes with and without infections among their staff and residents in Table S6. We can observe that 91.75% of nursing homes with no infections among their residents did not have infections among their staff either. On the other hand, 54.21% of nursing homes that had at least one infection among their residents, also had at least one infection among their staff. In Figure S4, we show the scatter plot of number of infections among staff and residents for only those nursing homes that experienced a large outbreak among both their staff and residents. We can observe a strong correlation between the number of infections among staff and residents."



		INFECTIONS AMONG STAFF (%)	
		0	>=1
INFECTIONS AMONG RESIDENTS (%)	0	556 (91.75%)	50 (8.25%)
	>=1	49 (45.79%)	58 (54.21%)

3. A sensitivity analysis of the proposed models should also include fitting their non-zero-inflated counterparts; i.e., a bivariate Poisson regression model and a double Poisson regression model.

Response: Thank you very much for your suggestion. We have now provided the estimates for Double and Bivariate Poisson models in Table S3 of the Supplementary Appendix. The corresponding discussions are provided in Bivariate and Double Poisson Estimates section of the supplementary appendix.

4. Model selection criteria are calculated and reported in Table 2. However, no interpretation of those results is given in the paper. The authors should provide a couple of sentences interpreting those results.

Response: Thank you very much for reminding us of this important issue. We have now added an interpretation in the Results section of the revised manuscript.

5. On page 11 and for Figure 2, what is meant by "lift" of the ZIBP models? Please clarify.

Response: Thank you for questioning on this part. We use lift as a measure for the ability of the model at predicting or classifying cases with respect to random selection. Lift shows how much better our model works compared to a random selection model. We have added this explanation in Results section of the revised paper.

6. Consider proposing in the discussion a spatial random effect to account for spatial dependencies, which could be synthesized with any temporal data made available in the future for which a zeroinflated spatio-temporal model would be appropriate.

Response: Thank you very much for this great suggestion. We have added this suggestion as a possible area for future research in Discussion section.

In closing, we would like to thank you again for your constructive feedback. Your suggestions have helped us to significantly improve the paper. We hope that the new version of our manuscript is up to your high standards.

VERSION 2 – REVIEW

REVIEWER	charlene Harrington University of California San Francisco CA, USA
REVIEW RETURNED	12-Nov-2020

GENERAL COMMENTS	The authors have answered all my questions and comments. I did note that on P.13 line 47, the word "enrolled" should be eliminated because it is not necessary and appropriate.