PEER REVIEW HISTORY

BMJ Medicine publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Non-Pharmaceutical Interventions and COVID-19 Burden in the United States
AUTHORS	Ziaeian, Boback; Ahlers, Michael; Aralis, Hilary; Tang, Wilson; Sussman, Jeremy; Fonarow, G

VERSION 1 - REVIEW

REVIEWER 1	Alvi, Mohammed. Competing Interest: None
REVIEW RETURNED	12-Nov-2021

GENERAL COMMENTS	Very interesting study overall, emphasizing the importance and usefulness of non pharmaceutical interventions in saving countless lives. Well designed models and collection of data gives meaningful results.

REVIEWER 2	Meyerowitz-Katz, Gideon. Competing Interest: I have previously published a short letter on a different topic with a team that included the senior author of this paper.
REVIEW RETURNED	30-Nov-2021

GENERAL COMMENTS	In general, I think this is a useful and interesting addition to the literature. It is more robust than many published evaluations of NPIs for COVID-19, and in general seems well-done. As an epidemiologist, the statistical approach appears well-constructed and carefully implemented. However, I would recommend a detailed statistical review as I am not deeply familiar with the methodology used here. As with all COVID-19 papers, I would strongly urge the authors to provide a few key points in lay language at the start of the paper, and I would recommend that the editors allow and encourage this to improve readability for the public. My primary concerns about the paper relate to two points. I think these should be discussed further and/or analyzed in additional analyses: 1. The source data for this paper, as with all similar papers, is problematic. It is well-established that case counts, particularly at the start of the pandemic, were a poor proxy for the true number of cases in an area. There are several ways that the authors could deal with this issue (which is inherent to any dataset of US case data), but one approach might be to divide the dataset into first vs second wave and compare the results. Given the far greater adequacy of testing later in 2020 when compared to March/April, this might demonstrate which associations are robust to testing capacity and which are based more on the testing adequacy of an area. If the

	authors do not wish to add further analyses to the paper, the underlying issues with the data should be more fully discussed as they are an inherent limitation that may make the analysis incorrect. The statement currently in the discussion, that: "underreporting of cases may have biased our results towards the null", is not necessarily true, as it depends entirely on which places underreport cases and for which reasons they do so. 2. It is likely that there is a great deal of between-state variability which impacts COVID-19 spread and is not captured in this analysis. This might be possible to analyse to some extent by reviewing state-level characteristics such as socio-economic indicators, and controlling for these in the regression models, but is otherwise problematic for the analysis. This is true of all such papers, of course, and is simply an issue with ecological studies in general. It might be worthwhile to either run a sensitivity analysis including some covariates, or perhaps outline this more fully in the discussion with reference to specific states and why they may differ with respect to reported cases and deaths for reasons other than NPIs. In addition to these two fairly major issues, I would note that the authors have not addressed several critiques that I am certain they will face. The first one is a common argument, that NPIs may reduce cases in the short but not the long-term. I understand this is not the purpose of the paper, which is to review whether NPIs may reduce cases/deaths immediately after implementation, but it is important to discuss this point nevertheless. I think the discrepancy between <10 and >10 person orders for public gatherings is very interesting as well. This may, as the authors note, imply that the less restrictive orders are less effective. It may also imply that this analysis is not able to capture the true impact that these policies have on COVID-19 cases/deaths because of various confounding factors. One could also argue that this is evidence that the public response to orders, rather than the orders and their enforcement, is the key intervention, and that more restrictive orders are effective simply not because they change a specific behaviour but because they signal a change in behaviour to the public generally. These are all arguments that have been made about similar studies in the past, and might be worth considering. As a final point, I think the "unverified assumptions" in terms of the lead time between policy announcement, enforcement, and adoption is an interesting matter. The authors have assumed that there is some lag between the implementation and impact, which all seem reasonable. However, it might be useful, and possible, to test these assumptions statistically. One could, for example, run a Bayesian analysis taking account of the informative prior with regards to the length of time between intervention and impact, to see whether the proposed lag is reasonable or not. This might be a very valuable addition to the paper, and would also be a novel analysis that I do not believe anyone else has done. One more, additional final point (apologies) is that it would be useful to double-check the data source. I note the authors have used The Atlantic's tracking project for case data – personally, I would cross-check these figures with other published numbers (i.e. from the CDC) to ensure that they are correct.

REVIEWER 3	Odugbemi, Babatunde. Competing Interest: None
REVIEW RETURNED	06-Dec-2021

GENERAL COMMENTS	This is an important study which evaluates the effectiveness of non-pharmaceutical intervention (NPI) policies against COVID-19 in the United States. Its main strength is that it is based on real world data. The methods are sound and sufficiently detailed. Associations between implementation of specific NPIs and case and mortality velocities were assessed using generalized linear models. The manuscript is well structured and clearly written. The study shows that the NPIs, particularly stay at home orders, were associated with a reduced burden of COVID-19. One area which would need clarification is that it appears the use of NPI policy data stopped on January 29, 2021 even though the study used case and mortality data up to March 7, 2021. Additionally, there should be more discussion around the finding that “gathering bans with limits greater than 10 were insufficient or exacerbated COVID-19 spread.

REVIEWER 4	Riley, Richard. Keele University, School of Medicine. Competing Interest: None
REVIEW RETURNED	12-Jan-2022

GENERAL COMMENTS	This is an interesting study, examining the impact of various national interventions on covid19 case velocity. I do not have specific expertise in the time series and breakpoints modelling methods being used, but I do not identify any general statistical issues. Of course – as noted by a reviewer – it is hard (impossible) to disentangle the national decisions from other changes that happen at around the same time (including the availability of tests, masks, personal decisions etc), and so making causal inferences is very difficult. However, if this is appropriately acknowledged I think the article adds value to the literature and might generate some discussion and debate.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Mohammed Alvi
Comments to the Author
Very interesting study overall, emphasizing the importance and usefulness of non-pharmaceutical interventions in saving countless lives. Well-designed models and collection of data gives meaningful results.

We thank Dr. Mohammed Alvi for his review of our work.

Reviewer: 2
Dr. Gideon Meyerowitz-Katz
Comments to the Author
In general, I think this is a useful and interesting addition to the literature. It is more robust than many published evaluations of NPIs for COVID-19, and in general seems well-done.

We thank Dr. Gideon Meyerowitz-Katz for his review of our work.

As an epidemiologist, the statistical approach appears well-constructed and carefully implemented. However, I would recommend a detailed statistical review as I am not deeply familiar with the methodology used here. As with all COVID-19 papers, I would strongly urge the authors to provide a few key points in lay language at the start of the paper, and I would recommend that the editors allow and encourage this to improve readability for the public.

Thank you. We have addressed statistical concerns as commented in further detail below. We have rewritten the introduction.

My primary concerns about the paper relate to two points. I think these should be discussed further and/or analyzed in additional analyses:

1. The source data for this paper, as with all similar papers, is problematic. It is well-established that case counts, particularly at the start of the pandemic, were a poor proxy for the true number of cases in an area. There are several ways that the authors could deal with this issue (which is inherent to any dataset of US case data), but one approach might be to divide the dataset into first vs second wave and compare the results. Given the far greater adequacy of testing later in 2020 when compared to March/April, this might demonstrate which associations are robust to testing capacity and which are based more on the testing adequacy of an area. If the authors do not wish to add further analyses to the paper, the underlying issues with the data should be more fully discussed as they are an inherent limitation that may make the analysis incorrect. The statement currently in the discussion, that: "underreporting of cases may have biased our results towards the null", is not necessarily true, as it depends entirely on which places underreport cases and for which reasons they do so.

Thank you. We agree undercounting of COVID-19 cases is a concern. To minimize overinterpretation of cases reported on a given day, we used 7-day averages of cases in interpretations of case velocities. Furthermore, arbitrary decisions to create time cut points for pandemic waves have the potential to introduce bias. We believe the breakpoint analysis is a more blinded approach to segmenting shifts in case and death counts. The growth in testing capacity was approximately linear during the duration of the study compared to exponential type of growth and decline for cases/deaths. Differences in case ascertainment week to week are unlikely to make a large contribution to velocities (1st derivative of absolute counts). The limitation of testing capacity and case ascertainment has been further emphasized in the manuscript. We have removed the statement "underreporting of cases may have biased our results towards the null" and instead replaced with:

"The early period of the COVID-19 pandemic in the U.S. likely suffered from lower rates of case-ascertainment and differences in testing capacity between states. Although we are unable to explicitly control testing capacity and policy by state in the statistical models, the availability of diagnostic tests grew linearly during the period of analysis and therefore would be unlikely to explain shifts in either case or death velocities."

2. It is likely that there is a great deal of between-state variability which impacts COVID-19 spread and is not captured in this analysis. This might be possible to analyze to some extent by reviewing state-level characteristics such as socio-economic indicators, and controlling for these in the regression models, but is otherwise problematic for the analysis. This is true of all such papers, of course, and is simply an issue with ecological studies in general. It might be worthwhile to either run a sensitivity analysis including some covariates, or perhaps outline this more fully in the discussion with reference to specific states and why they may differ with respect to reported cases and deaths for reasons other than NPIs.

We agree with the reviewers that additional factors that may impact COVID-19 spread are not captured in this analysis. As the reviewer notes, the limitation in our modeling approach is not only in the variables that can be ascertained, but also by the number of states (50) available for analysis. To avoid overfitting models, we wanted to evaluate the contribution of adoption and discontinuation of NPIs in conjunction over time. We focused on factors that were most likely to have changed over time to explain differences in case burden between states. As socio-economic factors would be expected to vary less over the period of our study (less than 15 months), they would likely influence the intercept of our models but not changes in case velocity on a week by week basis.

We acknowledge that lower socio-economic status may associate with a higher COVID-19 burden (https://pubmed.ncbi.nlm.nih.gov/32707661/, https://www.nejm.org/doi/full/10.1056/NEJMsa2011686, https://www.nejm.org/doi/full/10.1056/NEJMp2023616), even at the state level. To address this, we have also now noted accountability for socio-economic status as a limitation of our model.

We would also like to address in this comment from Dr. Gideon Meyerowitz-Katz additional factors that we considered incorporating into our model. Additional variables that we considered included average temperature, lived population density, K-12 school closures, public university closures, and socialized benefits. We assessed that these variables contributed more instability to the modeling than true variability capture.

With regards to socialized benefits, we considered both the duration and amount of unemployment benefit that was made available in each U.S. state. What we found was that The Coronavirus Aid, Relief, and Economic Security Act (CARES Act), which was passed by the U.S. Congress and signed into U.S. law on March 27, 2020, homogenized unemployment benefits in the U.S. for most of the duration of our study both in terms of length (up to 39 weeks additional unemployment via the Pandemic Emergency Unemployment Compensation) and amount of benefit (additional $600 per week to individuals on unemployment insurance via Pandemic Unemployment Assistance program) (https://www.congress.gov/116/bills/hr748/BILLS-116hr748enr.pdf). Thus, we elected against incorporating it in our model.

In addition to these two fairly major issues, I would note that the authors have not addressed several critiques that I am certain they will face. The first one is a common argument, that NPIs may reduce cases in the short but not the long-term. I understand this is not the purpose of the paper, which is to review whether NPIs may reduce cases/deaths immediately after implementation, but it is important to discuss this point nevertheless.

Thank you. We have included the following sentence in the discussion section:

"Finally, although several of the NPIs studied were associated with a decrease in COVID-19 burden over the period of study, we acknowledge that the effect of NPIs on the total number of COVID-19 cases and deaths during the ongoing COVID-19 pandemic is not known."

I think the discrepancy between <10 and >10 person orders for public gatherings is very interesting as well. This may, as the authors note, imply that the less restrictive orders are less effective. It may also imply that this analysis is not able to capture the true impact that these policies have on COVID-19 cases/deaths because of various confounding factors. One could also argue that this is evidence that the public response to orders, rather than the orders and their enforcement, is the key intervention, and that more restrictive orders are effective simply not because they change a specific behaviour but because they signal a change in behaviour to the public generally. These are all arguments that have been made about similar studies in the past and might be worth considering.

We agree with Dr. Gideon Meyerowitz-Katz that what we have studied is the association of given policies that were oriented toward altering human population behavior with subsequent COVID-19 burden. This is distinct from studying human population behavior itself and the burden of COVID-19.

To address this comment, we have included the following sentence in the discussion section:

"It is also possible that different NPIs may have been associated with increases or decreases in COVID-19 burden due to behavioral changes linked to the NPI but not specifically addressed by it. For example, indoor public gathering bans with maximums greater than 10 may not inherently be ineffective in decreasing burden of COVID-19, but rather elicit a different generalized public response especially relative to indoor public gathering bans with a maximum of 10 or fewer."

As a final point, I think the "unverified assumptions" in terms of the lead time between policy announcement, enforcement, and adoption is an interesting matter. The authors have assumed that there is some lag between the implementation and impact, which all seem reasonable. However, it might be useful, and possible, to test these assumptions statistically. One could, for example, run a Bayesian analysis taking account of the informative prior with regards to the length of time between intervention and impact, to see whether the proposed lag is reasonable or not. This might be a very valuable addition to the paper, and would also be a novel analysis that I do not believe anyone else has done.

We appreciate this statement and suggestion from Dr. Gideon Meyerowitz-Katz. The unverified assumptions related to lag between policy enactment and our window of observation for the outcome of interest (breakpoint occurrence) have the potential to be enormously impactful. One analysis plan we originally considered entailed inclusion of variable lag times using a Bayesian approach, very similar to the method suggested. In considering this approach we found this produced a higher than ideal uncertainty in the results, considering the relationship of limited data points (only 50 states with a limited number of policy adoptions and discontinuations) relative to a wide range of different lag time windows to consider optimizing over. Therefore, we opted for a hypothesis-driven approach of a fixed set of lag times that were motivated by our review of the literature instead.

Nevertheless, since our unverified assumptions of lead times is such an important topic to consider, we now present a sensitivity analysis for review and to be included as a published supplementary appendix to the manuscript. This sensitivity analysis involved replicating the model results that were presented in **Table 2** and **Table 3** for two additional lag times for both cases and deaths in unadjusted (**eTable 1**)and adjusted (**eTable 2**) models . The two lag times that we selected for this sensitivity analysis are shifted by 7 days prior and 7 days subsequent, respectively, to our original hypothesis-driven base case (a +/- 7 day shift in the lag time interval's proximity to the week being examined for case/death velocity change). In this limited exploration we have found that this sensitivity analysis strengthens our manuscript and is supportive of our original hypothesis-driven approach of lag times. We have incorporated this analysis into the manuscript in the methods, results, and discussion sections and included a supplementary appendix for your review.

One more, additional final point (apologies) is that it would be useful to double-check the data source. I note the authors have used The Atlantic's tracking project for case data – personally, I would cross-check these figures with other published numbers (i.e. from the CDC) to ensure that they are correct.

Gideon Meyerowitz-Katz

Thank you. The COVID Tracking Project was the most timely and comprehensive dataset of COVID-19 cases and deaths in the U.S. at the time our study was initiated. Given the varied reporting of COVID-19 data from U.S. states, that the COVID Tracking Project included sourcing to every datapoint it captured provided us the confidence to use this data in analysis. The COVID Tracking Project also included manual rechecks and verifications of its data for enhanced accuracy.

The COVID Tracking Project has been cited in more than 1,000 papers including in *Nature* for its case and death counts. Additionally, The Johns Hopkins University of Medicine Coronavirus Resource Center utilized The COVID Tracking Project for case and death count data from the start of the pandemic through March 2, 2021, our period of study, in its popularized dashboard.

As a considered alternative dataset, we evaluated the COVID-19 Case Surveillance Public Use Data With Geography dataset from the Centers for Disease Control and Prevention (CDC). We examined case counts over our period of study from two large U.S. states, California and New York. We found that the CDC dataset had 51.8% fewer cases documented for the state of California and 42.4% less cases for the state of New York.

In summary, we have elected to continue with The COVID Tracking Project as our primary source of data.

Reviewer: 3
Dr. Babatunde Odugbemi
Comments to the Author
This is an important study which evaluates the effectiveness of non-pharmaceutical intervention (NPI) policies against COVID-19 in the United States. Its main strength is that it is based on real world data. The methods are sound and sufficiently detailed. Associations between implementation of specific

NPIs and case and mortality velocities were assessed using generalized linear models. The manuscript is well structured and clearly written. The study shows that the NPIs, particularly stay at home orders, were associated with a reduced burden of COVID-19. One area which would need clarification is that it appears the use of NPI policy data stopped on January 29, 2021 even though the study used case and mortality data up to March 7, 2021.

We thank Dr. Babatunde Odugbemi for his review of our work. It is correct that the use of NPI policy data stopped on January 29, 2021. We chose to include COVID-19 case and mortality data up to March 7, 2021 for optimization of identifying breakpoints at the end of the study period. We excluded case breakpoints occurring after February 5, 2021 and death breakpoints occurring after February 12, 2021. This was necessary to align the policy data (which extended through January 29, 2021) with the appropriate lag time for identification of potential NPI impacts. Our unit of analysis was the state-week with week referring to calendar week. For each NPI, we determined the week in which it was adopted or discontinued and whether breakpoints occurred during one of the relevant subsequent weeks specified by our pre-established lag times. For an NPI adoption or discontinuation on January 29, 2021, the lag time implied that death breakpoints occurring as late as the week between February 7, 2021 and February 12, 2021 were considered. Case breakpoints occurring as late as the week between January 31, 2021 and February 5, 2021 were considered. Thus, all breakpoints identified after February 12, 2021 (for deaths) and February 5, 2021 (for cases) were ignored.

We have now included the following sentence in the Methods section. “We excluded breakpoints after February 5, 2021 for cases and February 12, 2021 for deaths to align with the data of NPI adoption and discontinuation dates which extended through January 29,2021, plus anticipated lag time.”

Additionally, there should be more discussion around the finding that “gathering bans with limits greater than 10 were insufficient or exacerbated COVID-19 spread.

Thank you. We refer the editors and reviewer to comments made by Dr. Gideon Meyerowitz-Katz as well in which we addressed a related comment by including the following sentence in the discussion section:

“Different NPIs may have been associated with increases or decreases in COVID-19 burden due to behavioral changes linked to the NPI but not specifically addressed by it. For example, indoor public gathering bans with maximums greater than 10 may not inherently be ineffective in decreasing burden of COVID-19, but rather elicit a different generalized public response especially relative to indoor public gathering bans with a maximum of 10 or fewer).”

We have also changed the sentence above to the following sentence to remove implications of causation, and explicitly state that we found an association. “Overall, we found that gathering bans with limits greater than 10 were insufficient or were associated with exacerbation of COVID-19 spread.”

Reviewer: 4
Prof. Richard Riley, Keele University
Comments to the Author
This is an interesting study, examining the impact of various national interventions on covid19 case velocity. I do not have specific expertise in the time series and breakpoints modelling methods being used, but I do not identify any general statistical issues. Of course – as noted by a reviewer – it is hard (impossible) to disentangle the national decisions from other changes that happen at around the same time (including the availability of tests, masks, personal decisions etc), and so making causal inferences is very difficult. However, if this is appropriately acknowledged I think the article adds value to the literature and might generate some discussion and debate.

Richard Riley, Prof of Biostatistics

We thank Dr. Richard Riley for his review of our work.

Regarding appropriately acknowledging that our analysis assesses for association and not for causation, we would refer Dr. Riley to our comments addressing similar critique from the editors. In paragraph #5 of the Discussion section, we state “Our model does not account for national recommendations and policies...” to acknowledge this inherent limitation.

Furthermore, we have inserted “association” within the titles of Table 2 and Table 3 to minimize suggestion of causation.

VERSION 2 – REVIEW

REVIEWER 4	**Riley, Richard.** Keele University, School of Medicine. Competing Interest: None
REVIEW RETURNED	06-May-2022

GENERAL COMMENTS	The response and revision is very clear, and I only have a few remaining comments: 1) Please define the meaning of the phrase “mutually adjusted models” – this is not defined in the methods section. 2) I could not see any explanation of how adjustment factors were chosen for the ‘mutually adjusted’ models, and whether any causal principles (DAGs) were part of the reasoning. 3) Previously we asked the authors to acknowledge that they could not disentangle the national decisions from other changes happening at the time, for example the availability of masks. In response they do add a brief note in paragraph 5 of the discussion to say ““Our model does not account for national recommendations and policies...”, but this is quite buried, and I was expecting this limitation to be more prominent, for example in the abstract conclusion. 4) Based on the mutually adjusted model results, the authors emphasise those that are statistically significant. However, it is difficult to disentangle particular strategies from others, as they are now considered jointly (in a multivariable model). Even those that are not statistically significant are potentially contributing in a multifaceted manner, and therefore it also worth drawing attention to the direction of effects (and still the uncertainty, width of CIs) for those that are not ‘significant’. Indeed, some has p-values only just above 0.05, and so it seems arbitrary to draw attention to just those statistically significant. Nearly all ORs are > 1. Worth considering the issue of Table 2 fallacy as described here: https://academic.oup.com/aje/article/177/4/292/147738

VERSION 2 – AUTHOR RESPONSE

The response and revision is very clear, and I only have a few remaining comments:

1) Please define the meaning of the phrase “mutually adjusted models” – this is not defined in the methods section.

Thank you, we have clarified our manuscript. The term mutually adjusted model refers to the analyses where all NPIs were adjusted in our ordinal logistic regression model for case or death velocities.

2) I could not see any explanation of how adjustment factors were chosen for the ‘mutually adjusted’ models, and whether any causal principles (DAGs) were part of the reasoning.

We did not prespecify a DAG in our models. Our interest was principled in identifying common NPI policy interventions and their temporal associations with case and death velocities during the pandemic. We believe our modeling approach allows for evaluating temporal associations appropriately while also considering several public health related NPIs that were deployed over time in various U.S. states. The natural history of COVID allows for modeling expected lags from transmission, diagnosis, hospitalization, and death which can strengthen evidence for the potential utility of various NPIs. This approach is similar to that of analyses of state variation in cigarette taxation that demonstrates differential reductions in cardiovascular and cancer mortality.

3) Previously we asked the authors to acknowledge that they could not disentangle the national decisions from other changes happening at the time, for example the availability of masks. In response they do add a brief note in paragraph 5 of the discussion to say ““Our model does not account for national recommendations and policies...”, but this is quite buried, and I was expecting this limitation to be more prominent, for example in the abstract conclusion.

Thank you. We agree this is a limitation and have highlighted this in the methods and conclusions sections of the abstract now.

4) Based on the mutually adjusted model results, the authors emphasise those that are statistically significant. However, it is difficult to disentangle particular strategies from others, as they are now considered jointly (in a multivariable model). Even those that are not statistically significant are potentially contributing in a multifaceted manner, and therefore it also worth drawing attention to the direction of effects (and still the uncertainty, width of CIs) for those that are not ‘significant’. Indeed, some has p-values only just above 0.05, and so it seems arbitrary to draw attention to just those statistically significant. Nearly all ORs are > 1. Worth considering the issue of Table 2 fallacy as described here: https://academic.oup.com/aje/article/177/4/292/147738

Thank you. We have de-emphasized discussion of statistical significance in the manuscript and focused on the direction of ORs in broader context of all models and NPIs we investigated, particularly with additional writing in the results section.