Response to Reviewer Comments

Title: Spatial and temporal epidemiology of SARS-CoV-2 virus lineages in Teesside, UK, in 2020: effects of socio-economic deprivation, weather, and lockdown on lineage dynamics
Manuscript: https://doi.org/10.1101/2022.02.05.22269279 version 5
Date: 28/05/2024

Dear Prof Guegan,

Thank you for giving us the opportunity to submit a revised draft of our manuscript to PCI Infections. We are grateful to the reviewers for providing comments on our manuscript. These comments were extremely useful, as the reviewers have picked up on several areas requiring greater clarity and detail. Our responses to each comment and changes to the manuscript are detailed below.

Many thanks,

The authors

The reviewer comments are in black *italic* text, and we have numbered them for ease of reference.

Our written response to each comment is in orange.

Changes made to the paper are in blue text. All line numbers are given relative to the new version of the document (the May resubmission).

Reviewer 1

The authors present a local study of the number of positive SARS-CoV-2 tests aiming at relating the number of positive tests to external factors like weather conditions, social deprivation, and NPI.

While there is some descriptive value to this work, I believe it has at least two fundamental flaws:

1. *Homogeneity of exposure and collinearity issues: While they advocate that using a small area allows for fine-grained analysis which may be lacking in other studies, interventions like NPI (either lockdown or local tiers) are applied in the whole area with no differentiation. It would have been different if they had included other areas in the UK but they did not. Or, if they had chosen to use a proxy of mobility reduction (phone data, or Google mobility index) but again they did not use such data. Hence, the robustness of their data can be questioned as some results may just be coincidental. Also, in some cases, they just had to discard important variables due to collinearity like "lockdown 2" and Tier 3 for B.1.1.7, which is unfortunate as one could hypothesize that those measures may have delayed the emergence of B1.1.7.*
   [Response]
   We thank the reviewer for this comment, and the second one, as they have highlighted a major oversight on our part: we did not adequately explain our study aims. The fact that the national and local interventions were applied universally across Teesside is not a problem in the context of our study aims. The reason that we looked at these two types of intervention in a specific spatial context, is because we wanted to compare the effectiveness of the interventions that were applied based on the current situation in the present context (cases in Teesside), to those that were applied based on a different context (cases in England). Using a relatively small

geographic area allows us to match the scale of the local interventions to the local cases, and using a relatively isolated sub-region like Teesside means that we can be confident that neighbouring effects are minimised. We recognise that we did not properly explain our study aims, or why we chose to use the specific location of Teesside, and so we have made changes to the paper in several places to improve this.

Regarding the use of other areas, there were some practical factors that would have made it extremely time consuming to extend our study area. There was a great deal of data sourcing and manual curation required, across many different datasets (e.g. IMD calculation, weather data sourcing and processing, sourcing the school/college/university term dates, collecting all the specific local scale tier levels and timings, tidying the postcode shapefiles etc.). Unfortunately, we did not have sufficient funding to extend the timeframe of the project to accommodate any further work in this area. It would also have been difficult for us to identify and acquire the PCR sequence data from another suitable sub-region, as we are less familiar with other areas of England and would lack the necessary contacts that could provide the required data (several of the study authors were involved in sequencing the PCR tests from Teesside).

Regarding mobility data, we unfortunately did not have a budget that could stretch to using proprietary commercial data, like O2 Motion. Google Mobility Reports are not available in the correct spatial grouping or scale for our case data, as they are organised by the larger local authority areas, which do not have complementary boundaries with postcode districts. This means that we would have had to incorporate information from residents outside of our study area or exclude some of the residents within it. We would also have had to collapse the data from the different local authorities together (due to overlapping boundaries making it impossible to match up information), which would have removed any spatial variation and reduced utility. Incorporating data on mobility would also not help us with our study aims, as we are more interested in highlighting the overall outcome in relation to the policy context, regardless of the mechanisms that drive them.

Regarding collinearity, after receiving these reviews we re-analysed our datasets using several alternative models (please see our response to the final comment from this reviewer (5)). The new final models suffer less from collinearity issues and are improved fits to the data. We have also improved the process by which we dropped variables due to high VIF (variance inflation factor) values, by including redundancy analysis and variable clustering plots, which we were able to use to confirm the redundancy of dropped variables.

[Changes]

- Re-wrote the second paragraph of the introduction, to more clearly link the discussion of relevant literature to our study aims, including this new text from line 82: “The UK government introduced a heterogenous “tier” system of restrictions in England during the second wave, which were applied on local scales and were intended to be more responsive and appropriate to the local disease context (UK Government, 2020c). While analysis of the tier system across the entirety of England has revealed the more stringent tiers to be more effective than the lower ones (Davies et al., 2021; Laydon et al., 2021), there has been very little consideration of these tiers, or comparison with the national-level restrictions, within their specific geographic and community contexts.”
- Re-wrote start of the final paragraph of the introduction, to state our study aims more clearly: “Our aim was to investigate how a range of variables that can influence COVID-19 cases affected positive tests in the Teesside sub-region of North East England during 2020. We wanted to understand how the national and local government interventions introduced over the course of this year affected the population of this sub-region, which has high levels of deprivation and is reasonably geographically isolated. Examining cases in his context also allows us make policy

recommendations aiming to improve outcomes in future epidemics, for this and other vulnerable populations in the UK."

- Added a new subsection to the methods section of the paper (Location and timeframe) that gives some background information about Teesside and helps to justify our focus on this area.
- Added a new section (3.2) to the supplementary materials file that includes: some text explaining the collinearity/redundancy analysis process, a table (S10) of the outputs of the redundancy analyses for each dataset (the R2 with which each variable can be predicted from all other variables), and variable clustering plots (S23 and S25) for each dataset (hierarchical cluster analysis using both squared Pearson and Spearman correlation).
- We have added the collinearity issue as another limitation in the discussion section, from line 430: "Thirdly, collinearity between temporal variables was present in almost all of the GLMM models. While we were able to assess and correct for this using a systematic approach, our results must be viewed in the context of the variables that were dropped. It is possible that some of the effects we see in our simplified models are actually being driven by those variables that were not included."
- Added some text to the final limitation in the discussion section, from line 435: "We also did not have access to mobility data of a sufficient spatial resolution to incorporate into our models (due to cost). The covariates used in our models were therefore surrogates for the underlying mechanisms associated with disease transmission and spread. However, this is less of an issue in the context of our study, as we are more interested in highlighting the overall impact on cases in Teesside in relation to the local and national restriction policies (and other covariates), rather than the specific mechanisms that may be driving these relationships."

The second flaw may be the most problematic.

2. *They do not account for the fact that COVID-19 is a communicable disease. They do not account in their model for the difference in transmissibility between variants and that epidemic spreading has a non-linear behavior. Other researchers managed to perform studies similar to the one the authors tried to complete (Paireau et al BMC Inf Dis. 2023 doi: 10.1186/s12879-023-08106-1, or Collin et al. Inf J Biostat 2023 doi: 10.1515/ijb-2022-0087) and found results exactly opposite to the ones presented here. The difference with the other studies is that the scale of the study is larger, and they used models to account for the epidemiological spreading of the diseases. This is not just a methodological difference, it's a fundamental flaw that totally invalidates their results.*
   [Response]
   We acknowledge that assessment of covariates (such as weather and non-pharmaceutical interventions) on a communicable disease, in a general manner, would require a different approach that incorporates transmission (such as using reproduction number rather than case counts as the response variable). However, what we attempted to understand was much more specific; we wanted to understand these relationships within a particular spatial and temporal context, one that we could link back to the specific policy decisions that were made at the time. We are interested not only in the effect of these covariates, but also how that related to the specific decisions made by the UK government, both in terms of the timing and duration of interventions, but also in terms of the geographic scale and location to which they were applied. Once again, we now see that this was not communicated effectively in the paper and that the language we used in places was unclear and potentially misleading, which has now been corrected. In addition to the previously mentioned changes in the introduction, we have also made some changes in the discussion to make a stronger link to the specific study aims and avoid generalising beyond the scope of our study.

Regarding variant transmissibility, this is unknown for most of the lineages that are within our dataset. Only one of the lineages (B.1.1.7, also called the alpha variant) has a documented increased transmission rate, and this only showed up at the end of the study period, which means it would not be likely to have a large effect on many of the models. The separate lineages models also have small sample sizes that unfortunately would not likely accommodate any further parameters.

Regarding the non-linear behaviour of the disease, we have now expanded our frequentist GLMM models to include comparison of the basic model with several other more complicated ones, including a restricted cubic spline for time (using both 3 and 4 knots), and a smooth spline for time, which should be able to account for any non-linearity of cases over time. These models were compared using validation figures created with the DHARMa R package (QQ plot, residuals vs predicted, outlier check, dispersion check, residuals vs time, and ACF), alongside a plot of the observed values vs fitted values from the model. These spline models did not prove to be a good fit for the datasets, and we have instead swapped to modelling them with an autoregressive term (AR1) (please see our response to the final comment from this reviewer for further details on this).

[Changes]

- Changed the wording throughout the paper to better reflect the modelling approach that we used, e.g., changed "spread" and "transmission/transmissibility" in the context of our research to "number of cases" or "positive tests".
- Expanded our interpretation of the differing effects of the government interventions to include more context-specific points, including this new text from line 498: "People experiencing greater socio-economic deprivation in the UK have been shown to experience increased exposure to high infection risk activities permitted during the lockdowns, and that this varied slightly over time between different lockdowns/restrictions (Beale et al., 2022). Further research is needed to understand the factors affecting lockdown success in different communities, particularly ones with high levels of deprivation, such as Teesside."
- Altered the wording of a sentence in the final paragraph of the discussion, from line 549, now reads: "While further research is needed to investigate the factors affecting lockdown success in different communities, we feel confident to make several recommendations regarding future epidemic policy responses in local/regional contexts, based on both ours and other's findings".
- Removed the second policy recommendation from the final paragraph of the discussion ("avoid issuing hospitality promotions"), as it was too general and did not relate to our study findings.
- Added a new policy recommendation to the final paragraph of the discussion, from line 554: "Secondly, interventions applied at the local/regional scale are less effective if they are less strict or applied later (Davies et al., 2021; Torres et al., 2022), therefore all tier levels (not just the highest) should be stringent and they should be imposed early. It is also imperative that local restrictions are communicated clearly and effectively with the public, and that the rules are simple, so as to facilitate adherence (Smith et al., 2022)."
- Reanalysed the datasets using several alternative specifications for time in the models, including smooth and restricted cubic splines (to account for non-linearity in cases over time), and an autoregressive term of order 1.

There are also a number of additional major issues, the main ones being:

3. *They do not mention the potential heterogeneity of community testing which can vary for example in time and by socio-economic factors. Factors like implementation delay (particularly*

during the first year of the pandemic), or pandemic fatigue (during the second wave) may greatly bias the representativeness of the real number of cases. They seem to pool pillar 1 (hospital based testing) and pillar 2 (community testing) as if they had the same meaning, timing, and value which can only bias their results. They could have provided analysis on pillar 1 only to limit the issue, but they would have lack power.

[Response]

We acknowledge that we neglected to discuss the potential for changes in testing capacity to affect case numbers over time, and we have now added some further text to the discussion to account for this. We do not believe our dataset is as affected by this as the full testing datasets may be, as we only looked at the subset of samples that had been whole genome sequenced, and the sequencing capacity did not increase at the same rapid rate as PCR and antigen testing capacity. This can be confirmed by comparing the magnitude of the peaks in the first and second waves in our dataset to those of the full testing dataset (see an example here for 4 of the communities in Teesside: https://ourworldindata.org/grapher/uk-daily-new-covid-cases?time=earliest..2020-12-31&country=Stockton-on-Tees~Redcar+and+Cleveland~Middlesbrough~Hartlepool), which shows the two waves are far more similar in our dataset than the full one.

We appreciate that there will be some differences between the two pillars, but we believe that they are similar enough in terms of the PCR testing to combine in our analyses. We found a positive correlation between the number of cases of each lineage recorded by each pillar, which suggested that any bias did not affect the detection of different lineages. We are therefore fairly confident that disease severity, testing location, and testing population did not bias lineage detection between the pillars. This is a useful check of our approach as our dataset only included positive PCR tests that had been genome-sequenced (data from the UK's genomic sequencing dataset). We also included the different testing methods of the two pillars as a limitation in our discussion, acknowledging the potential for bias. However, we accept that we could give further details to justify our approach and recognise other potential sources of bias.

The pillar 2 PCR testing during 2020 was predominantly of symptomatic individuals, though it began to also be used for asymptomatic testing of suspected cases and high-risk settings (confirming lateral flow results, elective care settings, care homes, contacts of confirmed cases etc.) from the autumn in 2020 (https://www.gov.uk/government/publications/pcr-testing-for-sars-cov-2-during-the-covid-19-pandemic). This is not so dissimilar from the situation for pillar 1, which focussed on symptomatic patients and staff in hospitals and care homes, but also included asymptomatic staff (e.g. contacts of confirmed cases). The main distinction between the pillars was the locations of the test sample collection and the laboratories that were carrying out the test processing (https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes/coronavirus-covid-19-scaling-up-our-testing-programmes). Pillar 1 tests were collected in hospitals and social care settings and processed in hospital or government facilities, while pillar 2 tests were collected in regional and local testing centres, at-home using postal kits, and in social care settings, and were processed in commercial and university facilities. We understand that there will be some differences in the access to tests between the pillars, but we think this bias should have been minimised by the availability of local testing centres within suburbs, mobile testing centres visiting remote or deprived communities, and at-home postal testing kits (all of which were free and strongly encouraged). We also think that the inclusion of both symptomatic and asymptomatic people in both pillars will minimise differences in the timing of testing in relation to the timing of infection.

We have added some of these details to the paper to address these concerns.

[Changes]
- The first part of the 'Data Collation' section, from line 164, now reads: "The UK government introduced two 'pillars' of COVID-19 testing, which utilised polymerase chain reaction (PCR) to identify positive cases. Pillar 1 testing was of staff and patients in hospitals and care homes, mainly focussing on symptomatic individuals, but also included asymptomatic staff (e.g. contacts of confirmed cases). Pillar 2 was community testing of symptomatic individuals, which also began to include asymptomatic testing of suspected cases and high-risk situations (e.g. confirming lateral flow results, elective care settings, care homes, and contacts of confirmed cases) from autumn 2020 (Dept. Health & Social Care, 2021; UK Health Security Agency, 2023)."
- Added some text to the limitations section of the discussion, from line 422: "It is also possible that the timing of and access to tests could have differed between the two pillars, however, we believe the availability and promotion of local testing facilities and free at-home postal test kits, and the fact that both pillars included asymptomatic and symptomatic individuals, should minimise any related bias."
- Added some text to the paragraph in the discussion covering the NPIs, from line 491: "Other research has also found no effect or positive effects of non-pharmaceutical interventions on case counts, and it has been suggested that this could represent an association with increases in testing capacity or with changes in testing policy (Giudici et al., 2023; Lison et al., 2023). However, this is effect will be minimised in our dataset (total sequenced PCR tests rather than total positive PCR and/or antigen tests) by the sequencing capacity of the COG-UK consortium, which did not increase at the same rate as testing capacity in 2020."

4. *Their data includes both hospital-based and community testing. That kind of testing will occur at different moments over the natural history of the disease. One will be further in time from the infectious event than the other. However, they choose to apply the same 2-week lag to all temporal variables. They do not justify this choice nor explain how it may potentially bias your results.*
[Response]
We have addressed the concerns relating to the timing of testing and the timing of infection in the response to the previous comment. We believe that the testing regimes of the pillars (both focussing on symptomatic but also including high risk asymptomatic individuals) are similar enough to analyse using the same time delay, especially as we aggregated cases into weekly totals.
We chose to use a 14-day delay period for two reasons:
i) One of the studies we referenced (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.28.2001401) included a sensitivity analysis using different length delays. They found that the impact of NPIs did not become apparent until after 14 days: "For time periods 1–7 and 8–14 days, the IRR values were above 1, indicating a positive association between cases/death and the intervention variable. For periods starting 15 days onwards the IRR was generally below 1 suggesting a negative association between the outcome and the intervention. This pattern probably reflects the time lag between exposure, latency and disease detection, so that the impact of interventions only kicks in after what is effectively a lag period of 14 days."
ii) We aggregated cases in our dataset by 7-day periods. If we had used a 7-day delay, this could potentially allow situations where a case at the beginning of a week is considered to be separated from the initiation of an NPI by 7 days, when in fact it may be considerably fewer (and too close in time to demonstrate a relationship).
We have now clarified this reasoning.

[Changes]
- Added this text to line 267 of the paper: "Because we aggregated case numbers and summarised temporal variables by week, a one-week delay could have artificially assumed a greater separation between cases and temporal events than actually happened. Additionally, the sensitivity analyses in Hunter et al. (2021) demonstrated that non-pharmaceutical interventions in the UK did not show an impact on numbers of cases until after 14 days."

5. *I'm unsure about how they deal with the dependency of the observations when using glmm. Observations at t+1 strongly depend on observations at time t due to the communicable nature of COVID-19. Hence, they should not be able to directly use glmm on these data. Also, the homoscedasticity assumption may be wrong in this case, which bias at least the error estimates and the confidence intervals.*

[Response]
After receiving the reviewer comments we went back and reanalysed all of the datasets using several additional methods in the GLMMs: a restricted cubic spline for time (using both 3 and 4 knots) (implemented via the "splines" R package); a smooth spline for time (implemented via the "mgcv" R package); an AR1 term (autoregressive order 1) for time (without any grouping); and an AR1 term for time grouped by postcode. The R package that we used to create the GLMMs (glmmTMB) can incorporate autoregressive terms of order 1 (https://cran.r-project.org/web/packages/glmmTMB/vignettes/covstruct.html#the-ar1-covariance-structure) as well as the splines. Each of these new models (and the original models) were compared using validation figures created with the DHARMa R package (QQ plot, residuals vs predicted, outlier check, dispersion check, residuals vs time, and ACF), alongside a plot of the observed values vs fitted values from the model. For each dataset, the AR1 model that was grouped by postcode proved to be the best fit. This model structure accounts for the temporal autocorrelation present in the data, across time, for each postcode and fits the data well (see the observed vs fitted values across week and postcode for the all-cases model in figure S3). The validation figures for all of the different model specifications for each dataset can be found in the analysis repository (https://doi.org/10.25405/data.ncl.23815077). We have also included some example validation figures for these alternative models in the supplementary material (section 3.2).

[Changes]
- Reanalysed the datasets using several alternative specifications for time in the models, including smooth and restricted cubic splines, and an autoregressive term of order 1.
- Changed the description of the GLMMs in the methods section of the paper to include the new AR1 term, from line 251: "We included an autoregressive term of order 1 (AR1) for week of year for each postcode district to account for temporal auto-correlation (non-independence) of cases over time within each separate district."
- Added validation figures to section 3.2 of the supplementary material for the final (best fitting) models for each dataset (S25, S28: S37).
- Added validation figures to section 3.2 of the supplementary material for the all-cases models that include splines for time (a restricted cubic spline (using 4 knots) and a smooth spline) as an example of the poorer fit that these approaches provide, relative to the AR1 term (S26, S27).

Reviewer 2

I found the figure elegant and the effort to tease apart the effects on SARS-CoV-2 incidence commendable. Some of my major issues with the manuscript can easily be addressed and have to do with the lack of details (especially for a non-UK audience) and with the interpretation of some of the results. A third concern is more generic and has to do with the choice to restrict the study to a single area in 2020. Finally, I am unsure about the soundness of some statistical decisions (but this is not my strong suit).

1. *Study focus*

   The authors decided to focus on a community of 550,000 inhabitants from January to December 2020.

   1.1. *They justify the geographical scope by the lack of studies performed at a local scale. This might be the case but then, why only include a single community? Is it for technical reasons (e.g. if there is a lot of manual curation to perform)? Otherwise, it seems that including more communities could have brought more power to disentangle factors that would be not identifiable at such a local scale.*

   [Response]

   We agree that we failed to properly articulate our study aims or justify our use of Teesside in the previous version of the paper. The reason that we looked at local and national interventions in one specific spatial context, is because we wanted to compare the effectiveness of the interventions that were applied based on the current situation in the present context (cases in Teesside), to those that were applied based on a different context (cases in England). Using a relatively small geographic area allows us to match the scale of the local interventions to the local cases, and using a relatively isolated sub-region like Teesside means that we can be confident that neighbouring effects are minimised. While Teesside may be a single sub-region, it contains many separate communities. We have now re-written parts of the introduction and added a new subsection in the methodology, which should alleviate these concerns.

   In addition to the these more abstract considerations, there were also some practical factors that would have made it extremely time consuming to extend our study area. The reviewer is correct in that there was a great deal of data sourcing and manual curation required, across many different datasets (e.g. IMD calculation, weather data sourcing and processing, sourcing the school/college/university term dates, collecting all the specific local scale tier levels and timings, tidying the postcode shapefiles etc.). Unfortunately, we did not have sufficient funding to extend the timeframe of the project to accommodate any further work in this area. It would also have been difficult for us to identify and acquire the PCR sequence data from another suitable sub-region, as we are less familiar with other areas of England and would lack the necessary contacts that could provide the required data (several of the study authors were involved in sequencing the PCR tests from Teesside).

   [Changes]

   - Re-wrote the second paragraph of the introduction, to more clearly link the discussion of relevant literature to our study aims, including this new text from line 82: "The UK government introduced a heterogenous "tier" system of restrictions in England during the second wave, which were applied on local scales and were intended to be more responsive and appropriate to the local disease context (UK Government, 2020c). While analysis of the tier system across the entirety of England has revealed the more stringent tiers to be more effective than the lower ones (Davies et al., 2021; Laydon et al., 2021), there has been very

little consideration of these tiers, or comparison with the national-level restrictions, within their specific geographic and community contexts."
- Re-wrote start of the final paragraph of the introduction, to state our study aims more clearly: "Our aim was to investigate how a range of variables that can influence COVID-19 cases affected positive tests in the Teesside sub-region of North East England during 2020. We wanted to understand how the national and local government interventions introduced over the course of this year affected the population of this sub-region, which has high levels of deprivation and is reasonably geographically isolated. Examining cases in his context also allows us make policy recommendations aiming to improve outcomes in future epidemics, for this and other vulnerable populations in the UK."
- Added a new subsection to the methods section of the paper (Location and timeframe) that gives some background information about Teesside and helps to justify our focus on this area, from line 142: "The wider Teesside area is a sub-region of the North East region of England, centred around the mouth of the river Tees. While Teesside has reasonably good local and national transport links, it is relatively isolated geographically as it borders the North Sea to the north east, The North York Moors national park to the south east, and extensive farmland to the west. Teesside has a distinct cultural identity due to its industrial history, which has also left its population with a greater burden of diseases and socio-economic deprivation that stigmatises and further culturally isolates the area (Bush et al., 2001). Teesside contains a mixture of urban, suburban, and rural environments and is formed from a collection of separate communities including Middlesborough, Redcar, Thornaby-on-Tees, Billingham, Hartlepool, and Stockton-on-Tees, each with their own identities, facilities, schools, etc. All of these characteristics make Teesside an interesting case study that deserves research focus."

1.2. *The authors also restrict their approach to the year 2020 and, unless I missed it, do not justify this focus. I thought that Pillar 2 had continued at least until 2022 and it seems strange not to include this data because by doing so the study has the same limitation as the myriad of studies from 2021 that were hampered by only being able to follow a single season.*

[Response]

The reviewer is right in that we neglected to explain our reasons for using data from 2020. We have now corrected this oversight. One of the main factors relates to our study aim regarding comparing national vs local restrictions; the UK government stopped using local restrictions from early January 2021, and only used national-level ones after that point. There are also issues of confounding for data in 2021 that relate to information that may be difficult to ascertain at the necessary levels of detail (such as vaccination uptake).

[Changes]

- Added a new paragraph explaining our use of data from 2020 only, in the new 'Location and timeframe' subsection of the methods section of the paper, from line 152: "We chose to examine positive tests within Teesside during only 2020 for several reasons. Firstly, the local tiered restrictions were abandoned when the country entered a new national lockdown in early January 2021, and were not reinstated (UK Government, 2021), which means there were no further locally applied restrictions that could be examined. Secondly, the UK's vaccination programmed begun in December 2020, which means that analysing data that included cases into 2021 would have required additional data to properly adjust for this confounder. Inclusion of vaccination uptake information would have been needed, rather than numbers/proportion of eligible people in the population, as uptake is lower amongst deprived people (Mounier-Jack et al., 2023), but this would have been difficult to ascertain at the necessary spatial and temporal scales. And finally, including data collected over a longer time period would open up the analysis to other potential sources of confounding that cannot easily be controlled for, such as 'pandemic fatigue'."

1.3. *At any rate, given the focus of the article, the introduction may not need to discuss too much the unfolding of the pandemic.*
[Response]
We have trimmed some of the content from the introduction.
[Changes]
- Moved the description of the epidemic progress and management in the UK from the second paragraph of the introduction to subsection 1.1 of the supplementary material file.

2. *Results interpretation*

2.1. *I found that the authors put a lot of emphasis on demographic factors but these appear rather limited in Figure 4. For example, the estimate for the population size is very low compared to that of the NPIs (but perhaps it needs to be rescaled?). Furthermore, the one for social deprivation has a confidence interval that spans from 0 to 1 when analysing all the cases (I am surprised that it comes out as significant), while it is non-significant for nearly all the lineage-specific effects.*
[Response]
After reading the reviews, we reanalysed our data using a different model specification that was a better fit for the data (please see our response to comment 4 below). The results for these new models are slightly different: the lower bound of the confidence intervals for deprivation for the all-cases and B.1.1.119 models are further from 0, and the positive relationship is now significant for two more of the lineages (B.1.1.1, B.1.1.315). Regarding population, we neglected to explain in the previous submission of the paper that this variable had been rescaled, so we have now amended this. This effect may be very small, but it is consistent. However, we recognise that we should be more conservative with our interpretation of the population and deprivation results, and so we have changed the wording of the relevant part of the discussion.
[Changes]
- Added some text to the methods section, from line 256: "We controlled for population size by including district total population (rescaled to measure population in thousands rather than single people) as a fixed effect."
- Altered the wording of the first sentence of the second paragraph of the discussion, from line 441 now reads: "The spatial variation in total positive SARS-CoV-2 tests across the Teesside area was influenced by demographic factors…", instead of: "The spatial variation in total positive SARS-CoV-2 tests across the Teesside area was caused (at least in part) by demographic factors…".
- Reanalysed the data with new models, which show clearer effects of deprivation.

2.2. *About the NPI effects, the opposite trend between lockdown 1 and the other NPIs perhaps warrants some more discussion. In particular, I was puzzled because the methods indicate that the model tests "the time since imposition" of an effect whereas the discussion (line 411) interprets it as an effect of the lockdown. Perhaps this is the case (meaning that the* *figure and Discussion were written after correcting for the effect) but at least it is potentially misleading.*

[Response]

Upon re-reading the methods and results sections we agree that the wording here was confusing. The government interventions and subsidy were coded numerically: 0 before they were imposed, 1 for the first week of imposition, increasing by 1 for each further week of duration, and reverting to 0 after the end of the intervention (though these then had the time-lag applied). This means that we were looking at the effect of the interventions (imposed/not imposed), but their duration was also incorporated into this. We have now simplified the language used regarding these effects in the paper and described the coding in more detail in the supplementary material.

However, we also agree that further discussion of the contrasting NPI effects is needed, and so we have added a bit more of this as well.

[Changes]

- Changed the wording in the mixed effects modelling subsection of the results section of the paper to remove "time since imposition" in reference to the NPIs.
- Added a description of the NPI and eat-out subsidy coding in a new section (2.1 Model Specifications) of the supplementary material: "The interventions and subsidy were coded numerically according to the procedure of Hunter et al., 2021 (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.28.2001401): 0 before they were imposed, 1 for the first week of imposition, increasing by 1 for each further week of duration, and reverting to 0 after the end of the intervention. This was to allow the effect to incorporate duration as well as presence/absence."
- Added some new text to the end of the relevant discussion paragraph discussing NPIs, from line 491: "Other research has also found no effect or positive effects of non-pharmaceutical interventions on case counts, and it has been suggested that this could represent an association with increases in testing capacity or with changes in testing policy (Giudici et al., 2023; Lison et al., 2023). However, this is effect will be minimised in our dataset (total sequenced PCR tests rather than total positive PCR and/or antigen tests) by the sequencing capacity of the COG-UK consortium, which did not increase at the same rate as testing capacity in 2020. It is also possible that the difference in effectiveness could reflect differences in exposure between lockdowns due to changes in behaviour or routine, e.g. increased use of public transport during the second wave. People experiencing greater socio-economic deprivation in the UK have been shown to experience increased exposure to high infection risk activities permitted during the lockdowns, and that this varied slightly over time between different lockdowns/restrictions (Beale et al., 2022). Further research is needed to understand the factors affecting lockdown success in different communities, particularly ones with high levels of deprivation, such as Teesside."

2.3. *The authors mention urban vs. rural post-codes (line 383) but it was unclear how they tested this because this co-variate does not appear in the GLMM.*

[Response]

We did not formally investigate this in our models. The comment in question was in reference to the maps produced by the models (disease mapping and the GLMM with a gradient for week for each postcode), which we interpreted using the demographic data and our knowledge of the region. We appreciate that this paragraph is confusing and unclear on this point, and so we have added some clarifying text and shortened it.

[Changes]

- The sentence starting on line 446 now reads: “Our model output maps, when viewed in combination with the demographic information and regional knowledge, did not demonstrate a clear pattern of risk of, or rate of increase in, positive tests in relation to how urban or rural the postcode districts are.”
- We have also added a sentence to line 450: “Because we did not formally investigate land use in any of our models, we do not know what the true effect was, though we can speculate as to why we did not detect any sort of clear signal.”
- Removed the sentences relating the studies conducted in North America (Huang et al., 2021; Rifat & Liu, 2022), to make the paragraph shorter and more relevant to the context of our study.

2.4. *At a spatial scale, from Figure 3, it looks like the first wave (B.1.1.119) hit the center (more urban?) and the second wave (B.1.177 and B.1.17) hit the north of the area (more rural?). The authors mention that immunity can be ruled out because of a lack of correlation between the positive tests (lines 458-459) but it is unclear how this was tested for.*

[Response]

This sentence refers to the GLMM that examined the number of cases of second-wave lineages and included the number of first-wave lineage cases as a fixed effect. We appreciate this was not clear though, and so we have expanded on this.

We also stated in this paragraph that we believe our failure to detect a relationship does not rule out the effect of immunity, rather it is likely due to the low numbers of positive tests of each lineage in our dataset.

[Changes]

The sentence starting on line 540 now reads: “However, when we included the number of cases of first-wave lineages as a fixed effect in a GLMM modelling the number of cases of second-wave lineages, there was no apparent relationship.”

3. *Correlated variables*

Many of the variables the authors investigate appear to be very correlated. This is the case for instance regarding the virus lineage and the day of the year since the vast majority only peaks one (B.1.177 being the main exception). I think this might also be the case for weather-related variables.

To address this issue in the GLMM, the authors use a VIF test to remove one of the correlated variables and reperform the test. This is fine but, unless I am mistaken the choice of the correlated variable removed is arbitrary. For example, in Figure 4 the week is removed but couldn't another variable correlated with it be removed instead? Perhaps showing which of the variables were highly correlated with the ones removed (at least for the global model) would help interpret the results.

[Response]

The process we followed (outlined in Zuur et al. (2010): http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00001.x), is effectively an arbitrary method. We could have selected other variables to remove, but that would have perhaps opened up the results to bias via the choices that we made. The method we employed was a clearer and more prescriptive one that can be replicated more easily. However, we realise that this aspect of the analysis is unsatisfactory, and so we have added additional methods for assessing collinearity, which are outlined in section 3.2 of the supplementary material. We have plotted the correlations from hierarchical variable clustering analyses and conducted redundancy analyses on each dataset and included the outputs in the supplementary material. In most cases, variables that were identified as collinear via VIF values were also identified as redundant via the redundancy analyses.

We also recognise that we need to be clearer about the implications of our choice to drop any variables due to collinearity on model interpretation, and the need to add some detail regarding the nature of the collinearity in the full models.

[Changes]

- Added a new section (3.2) to the supplementary materials file that includes: some text explaining the collinearity/redundancy analysis process, a table (S10) of the outputs of the redundancy analyses for each dataset (the R2 with which each variable can be predicted from all other variables), and variable clustering plots (S23 and S25) for each dataset (hierarchical cluster analysis using both squared Pearson and Spearman correlation).
- Added this text to the methods section, from line 285: “Predictor variables were only removed if they were found to be collinear via variance inflation factor (VIF) (calculated using the “performance” R package (Lüdecke et al., 2021)), whereupon we followed the procedure defined by Zuur et al. (2010) of removing variables sequentially until none of the recalculated VIF values are above 3. We also confirmed variable removal via redundancy analysis and variable clustering (conducted using the “Hmisc” R package (Harrell & Dupont, 2021)) (see section 3.2 of supplementary materials for further details, redundancy analysis outputs, and variable clustering plots).”
- Added this text to the first paragraph of the results section covering the GLMM models, from line 358: “See Supplementary Materials for model output from both the full and final (simplified) versions (the coefficients from the full models are accurate in terms of the magnitude and direction of any effects, but the errors are unreliable). This collinearity was between the temporal variables and was most often strongest for the weather variables, though the collinearity was also very strong for the interventions or subsidy where lineages had a very narrow temporal window (e.g. B.1.1.7).”
- Added an extra point to the first paragraph of the discussion (covering study limitations), from line 430: “Thirdly, collinearity between temporal variables was present in most of the GLMM models. While we were able to assess and correct for this using a systematic approach, our results must be viewed in the context of the variables that were dropped. It is possible that some of the effects we see in our simplified models are actually being driven by those variables that were not included.”

4. *Spatial random effects*

This is outside my area of expertise but I was unsure how the spatial and temporal auto-correlations were handled in the GLMM. Based on my reading, it looks like a random intercept is used to capture both the postcode and the time effect (line 226). I was expecting two random effects, one for the spatial aspect (e.g. in 2D with the spatial coordinates) and another for the temporal effect (e.g. with a spline) but perhaps I am missing something here.

[Response]

Regarding the spatial autocorrelation, unfortunately this cannot be accounted for in a frequentist GLMM. We did apply such a method in the Bayesian disease mapping models (conditional autoregression using the INLA R package), but unfortunately our datasets are too small (especially the separate lineage ones) to be able to incorporate both spatial and temporal autoregression and covariates. We instead included postcode as a random intercept in our frequentist GLMMs to account for repeated sampling and unmeasured geographical variation. Regarding the temporal auto-correlation, after receiving the reviewer comments we went back and reanalysed all of the datasets using several additional methods in the GLMMs: a restricted cubic spline for time (using both 3 and 4 knots) (implemented via the “splines” R package); a smooth spline for time (implemented via the “mgcv” R package); an AR1 term (autoregressive order 1) for time (without any grouping); and an AR1 term for time grouped by postcode. Each of these (and the original models) were compared using validation figures created with the DHARMa R package (QQ plot, residuals vs predicted, outlier check, dispersion check, residuals vs time, and ACF), alongside a plot of the observed values vs fitted values from the model. For each dataset, the AR1 model that was grouped by postcode proved to be the best fit. This model structure accounts for the temporal autocorrelation present in the data, across time, for each postcode and fits the data well (see the observed vs fitted values across week and postcode for the all-cases model in figure S3). The validation figures for all of the different model specifications for each dataset can be found in the analysis repository (https://doi.org/10.25405/data.ncl.23815077). We have included some example validation figures for these alternative models in the supplementary material (section 3.2).

[Changes]

- Reanalysed the datasets using several alternative specifications for time in the models, including smooth and restricted cubic splines, and an autoregressive term of order 1.
- Changed the description of the GLMMs in the methods section of the paper to include the new AR1 term, from line 251: “We included an autoregressive term of order 1 (AR1) for week of year for each postcode district to account for temporal auto-correlation (non-independence) of cases over time within each separate district.”
- Added a description to the methods section of the new alternative model specifications that were tried, from line 278: “Because it would be reasonable to assume that the number of cases over time would follow a non-linear trajectory, we also fit several alternative model specifications that included week of year as a fixed effect with either a smooth or restricted cubic spline instead of an AR1 term, however, these models were a poorer fit for all datasets than the AR1 models (see section 3.2 of the supplementary material).”
- Added validation figures to section 3.2 of the supplementary material for the final (best fitting) models for each dataset (S25, S28: S37).
- Added validation figures to section 3.2 of the supplementary material for the all-cases models that include splines for time (a restricted cubic spline (using 4 knots) and a smooth spline) as an example of the poorer fit that these approaches provide, relative to the AR1 term (S26, S27).

Detailed comments

5. *line 75: I was a bit surprised by this statement because there have been loads of articles trying to identify the effect of a variety of interventions. The problem is that many of them only analysed the first wave, which means they have very little power. But perhaps few of these models have considered UK data (although it would seem surprising given the quality and availability of this data).*

   [Response]

   We agree that this sentence is misleading, as there have been studies examining the second

wave of the epidemic in the UK. We were trying to make a point about the lack of studies at local scales that looked across waves. We have re-written this sentence to make our point clearer and link it more explicitly to our study aims.
[Changes]
- Rewritten the final sentence of the first paragraph of the introduction, from line 73: "While there is evidence that strict national restrictions reduced cases and mortality on a national-scale in the UK (Hunter et al., 2021; Sharma et al., 2021), there has been very little evaluation at local or regional scales or in socio-economically deprived populations."

6. *line 151: is the testing in Pillar 2 done at random? Put differently, is the testing a good proxy of the incidence? If so, it should be clarified. If not, it is obviously a huge issue.*
[Response]
None of the UK government mass testing programmes that utilised PCR testing were conducted at random, they targeted patients with high clinical need, symptomatic individuals, or asymptomatic individuals who were at a high risk due to exposure. The pillar 2 PCR testing during 2020 was predominantly of symptomatic individuals, though it began to also be used for asymptomatic testing from the autumn in 2020 (https://www.gov.uk/government/publications/pcr-testing-for-sars-cov-2-during-the-covid-19-pandemic). We appreciate this was not explained in the paper and so have added some details.
[Changes]
- The first part of the 'Data Collation' section, from line 164, now reads: "The UK government introduced two 'pillars' of COVID-19 testing, which utilised polymerase chain reaction (PCR) to identify positive cases. Pillar 1 testing was of staff and patients in hospitals and care homes, mainly focussing on symptomatic individuals, but also included asymptomatic staff (e.g. contacts of confirmed cases). Pillar 2 was community testing of symptomatic individuals, which also began to include asymptomatic testing of suspected cases and high-risk situations (e.g. confirming lateral flow results, elective care settings, care homes, and contacts of confirmed cases) from autumn 2020 (Dept. Health & Social Care, 2021; UK Health Security Agency, 2023)."

7. *line 153-154: Were all the positive tests sequenced? If so, please clarify it otherwise it seems that two databases are used (one for the tests and another for the sequencing).*
[Response]
The reviewer is right that this section is confusing, we thank them for highlighting this oversight. There are two separate databases for PCR data in the UK: one containing all positive and negative PCR test results; and another (the COG-UK genomic dataset) that contains the genome sequencing data from a randomly selected sample of the positive PCR test samples. However, we only used the genomic sequence dataset, which represents a random subsample of all positive tests. We have now re-written this section to clarify this.
[Changes]
- The text from line 170 now reads: "Genomic sequencing was conducted on a random sample of the PCR samples that tested positive for SARS-CoV-2, which permitted characterisation of genetic lineages in individual cases, and the information was stored in the Covid-19 Genomics UK (COG-UK) dataset (COVID-19 Genomics UK (COG-UK) Consortium, 2020; Smallman-Raynor et al., 2022). We collated records of PCR tests positive for SARS-CoV-2 in the Teesside area during 2020, from the COG-UK genomic dataset, which provided the dates, postcodes, and viral lineages of all sequenced positive tests from January 2020 to January 2021."

8. *line 167: Were the weather variables assumed to be homogeneous throughout the area studied?*

[Response]
Yes, we assumed the weather would be relatively homogenous across the area. This is because we used mean weekly values, which should even out any differences in localised weather effects that move across the area (e.g. heavy rain showers that are patchy over space and time).
[Changes]
- Added some extra text to the methods section, from line 189: "(we assumed weather to be homogenous across the area at this temporal scale)".

9. *line 215: please explain the DIC acronym.*
[Response]
We have changed our model validation/fit methods and no longer use DIC. We have made sure to expand any new acronyms.

10. *line 218-245: writing a formal version of the model in the Appendix would help understand the methodological steps.*
[Response]
All of the model-specific information (fixed effects, random effects, family, link function etc.) can be found in the data analysis scripts, which have been uploaded to a repository and given a DOI (listed in the Data/Code section at the end of the paper: https://doi.org/10.25405/data.ncl.23815077). However, we appreciate that this information should also be more widely accessible, and so these details have been added to the supplementary materials.
[Changes]
- Added this text to line 292 of the paper: "Full details of all the model specifications (disease mapping and GLMMs) can be seen in the analysis scripts (https://doi.org/10.25405/data.ncl.23815077) and the supplementary materials".
- Added a new section to the supplementary material file, "Model Specifications", which gives details of all model parameters.

11. *line 235-237: Why did you choose a two-weeks delay? This seems appropriate for Pillar 1 (hospital admissions occur 10 to 14 days post-infection) but it seems exaggerated for Pillar 2 (95% of the testing positive would be between day 2 and day 11). Furthermore, as the sentence is written, it looks like the individual test-seeking behaviour was at play, which seems inconsistent with the fact that Pillar 2 was testing at random and not based on symptoms.*
[Response]
As we explained in response to comment 6 from this reviewer, the pillar 2 PCR testing was mainly of symptomatic individuals, though it began to include asymptomatic testing from the autumn. It is also worth noting that pillar 1 PCR testing was not just of symptomatic hospital and care staff and patients, but also included some asymptomatic frontline workers in hospital and care settings. We believe that the inclusion of both symptomatic and asymptomatic people in both pillars will minimise differences in the timing of testing in relation to the timing of infection. The availability and promotion of regional and local testing centres, mobile testing units, and at-home postal tests (all of which were free) should also minimise differences between the two pillars.
We chose to use a 14-day delay period for two reasons:
i) One of the studies we referenced (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.28.2001401) included a sensitivity analysis using different length delays. They found that the impact of NPIs did not become apparent until after 14 days:

"For time periods 1–7 and 8–14 days, the IRR values were above 1, indicating a positive association between cases/death and the intervention variable. For periods starting 15 days onwards the IRR was generally below 1 suggesting a negative association between the outcome and the intervention. This pattern probably reflects the time lag between exposure, latency and disease detection, so that the impact of interventions only kicks in after what is effectively a lag period of 14 days."

ii) We aggregated cases in our dataset by 7-day periods. If we had used a 7-day delay, this could potentially allow situations where a case at the beginning of a week is considered to be separated from the initiation of an NPI by 7 days, when in fact it may be considerably fewer (and too close in time to demonstrate a relationship).

We have now clarified this reasoning.

[Changes]

Added this text to line 267 of the paper: "Because we aggregated case numbers and summarised temporal variables by week, a one-week delay could have artificially assumed a greater separation between cases and temporal events than actually happened. Additionally, the sensitivity analyses in Hunter et al. (2021) demonstrated that non-pharmaceutical interventions in the UK did not show an impact on numbers of cases until after 14 days."

12. *line 262: When testing the correlation between the tests in Pillar 1 and Pillar 2, were the time series treated as a set of independent (paired) observations? Because the temporal correlation could amplify the signal.*

[Response]

This correlation did not include a temporal component; it was conducted on data that had been summed across all time points. We have now clarified the wording.

[Changes]

The description of the correlation, from line 177, now reads: "In addition, we also summarised the total number of cases of each lineage recorded by each pillar during the period when sampling was run contemporaneously for both pillars (from week of year 19 to 53), creating a dataset of total cases of each lineage summed across time for each pillar (each lineage had one value per pillar). We ran a Spearman's correlation on this dataset, to check for a sampling bias for different lineages between the two pillars, as only pillar 1 tested the most severely ill patients, and illness severity can vary between lineages (Sievers et al., 2022; Goethem et al., 2022)."

13. *Figure 3 is beautiful, but it suggests that the pattern is driven by lineages whereas it could also be temporal. Perhaps the latter representation would be more neutral.*

[Response]

We thank the reviewer for this compliment.

Figure 3 is created from the output of the disease mapping models (each map within the figure came from a separate model). These were spatial models that did not include a temporal component. At present, some temporal interpretation is possible by comparing the lineages that clustered together temporally (e.g. first wave vs second wave), but we cannot combine the output from separate models into 1 map.

To create a temporal disease mapping figure would require creating a different model, which would generate a separate risk estimate for each temporal window (week-of-year) in each postcode. We would only be able to include a relatively small number of maps of these risk estimates in a figure, which would mean arbitrarily selecting a subset that may not demonstrate the true underlying pattern. The full output from such models is better viewed as an animated video or gif, which is unfortunately not easy to include in publications.

We instead attempted to show the broader temporal patterns in positive tests via the raw data shown in Figure 2, and from the modelling output from the GLMM with a random gradient for week for each postcode (Figure 5).

14. *Figure 4 is nice but having different scales for the different panels can be misleading. Perhaps use a single scale, at least for the lineage-specific figures.*
    [Response]
    We created an alternative version of this plot with fixed scales, but unfortunately the magnitude of the effects and errors are too different between the different models to be able to plot them with a single shared scale, and still have the details visible on all panels. We have instead highlighted the use of different scales.
    [Changes]
    - Added this text to the legend of figure 4: "Note that each panel has a different y axis scale."

15. *lines 427-429: This seems to be a bit redundant with the part in lines 414-416.*
    [Response]
    We thank the reviewer for highlighting this repetition. We have attempted to make the difference in the two points being made more obvious.
    [Changes]
    - The first part of this paragraph now reads: "Our findings suggest that the local tier system of interventions was less effective at reducing cases than a long and strict national lockdown, which has also been found in other studies (Davies et al., 2021; Torres et al., 2022). We also found that the tier restrictions were equally ineffective as the second national lockdown, all of which were applied during the second wave. This suggests that if there are any benefits to applying local-scale interventions in response to local-scale cases (rather than cases on the national scale, which in this context were determined by more populous and distant regions), they are masked by the effects of other factors, such as stringency and duration of restrictions, introduction events, and transmissibility of present lineages."

Reviewer 3

I would like to thank the editor for the opportunity to review this interesting paper and the authors for their nice work. First and foremost, the quality of figures should be denoted, notably the maps.

In its current state, I found that the study suffers from reporting loopholes and sometimes inappropriate or not sufficiently supported statistical choices. In light of this, I found that some authors' statements were too clear-cut.

1. *Introduction:*
    1.1. *I found the introduction interesting but maybe a bit too long. I would suggest shrinking the conclusion and maybe including some of the § in the supplementary files for readers, interested in getting additional background.*
    [Response]
    We have trimmed some of the content from the introduction.
    [Changes]

- Moved the description of the epidemic progress and management in the UK from the second paragraph of the introduction to subsection 1.1 of the supplementary material file.
- Re-written the final paragraph of the introduction to be more concise and clear.

1.2. *At this step, it is not clear why the authors focused only on 2020 and did not consider subsequent years. Similarly, I did not get why they restricted the analysis to the Teesside. What is preventing authors from a larger spatial field?*

[Response]

We agree that we failed to justify our choice of 2020 data or justify our use of Teesside in the previous version of the paper. Our decisions here relate to our study aims, but we did not properly articulate these either.

The reason that we looked at cases in one specific spatial context, is because we wanted to compare the effectiveness of the government interventions that were applied based on the current situation in the present context (local restrictions in response to cases in Teesside), to those that were applied based on a different context (national restrictions in response to cases in England). Using a relatively small geographic area allows us to match the scale of the local interventions to the local cases, and using a relatively isolated sub-region like Teesside means that we can be confident that neighbouring effects are minimised. While Teesside may be a single sub-region, it contains many separate communities. We have now re-written parts of the introduction and added a new subsection in the methodology to properly explain this.

Regarding use of data from 2020, one of the main reasons relates to our study aim regarding comparing national vs local restrictions; the UK government stopped using local restrictions from early January 2021, and only used national-level ones after that point. There are also issues of confounding for data in 2021 that relate to information that may be difficult to ascertain at the necessary levels of detail (such as vaccination uptake). We have also added an explanation for this decision.

In addition to the these more abstract considerations, there were also some practical factors that would have made it extremely time consuming to extend our study area and/or timeframe. There was a great deal of data sourcing and manual curation required across many different datasets (e.g. IMD calculation, weather data sourcing and processing, sourcing the school/college/university term dates, collecting all the specific local scale tier levels and timings, tidying the postcode shapefiles etc.). Unfortunately, we did not have sufficient funding to extend the timeframe of the project to accommodate any further work in this area. It would also have been difficult for us to identify and acquire the PCR sequence data from another suitable sub-region, as we are less familiar with other areas of England and would lack the necessary contacts that could provide the required data (several of the study authors were involved in sequencing the PCR tests from Teesside).

[Changes]

- Re-wrote the second paragraph of the introduction, to more clearly link the discussion of relevant literature to our study aims, including this new text from line 82: “The UK government introduced a heterogenous “tier” system of restrictions in England during the second wave, which were applied on local scales and were intended to be more responsive and appropriate to the local disease context (UK Government, 2020c). While analysis of the tier system across the entirety of England has revealed the more stringent tiers to be more effective than the lower ones (Davies et al., 2021; Laydon et al., 2021), there has been very little consideration of these tiers, or comparison with the national-level restrictions, within their specific geographic and community contexts.”

- Re-wrote the start of the final paragraph of the introduction, to state our study aims more clearly: "Our aim was to investigate how a range of variables that can influence COVID-19 cases affected positive tests in the Teesside sub-region of North East England during 2020. We wanted to understand how the national and local government interventions introduced over the course of this year affected the population of this sub-region, which has high levels of deprivation and is reasonably geographically isolated. Examining cases in his context also allows us make policy recommendations aiming to improve outcomes in future epidemics, for this and other vulnerable populations in the UK."
- Added a new subsection to the methods section of the paper (Location and timeframe) that gives some background information about Teesside and helps to justify our focus on this area, from line 142: "The wider Teesside area is a sub-region of the North East region of England, centred around the mouth of the river Tees. While Teesside has reasonably good local and national transport links, it is relatively isolated geographically as it borders the North Sea to the north east, The North York Moors national park to the south east, and extensive farmland to the west. Teesside has a distinct cultural identity due to its industrial history, which has also left its population with a greater burden of diseases and socio-economic deprivation that stigmatises and further culturally isolates the area (Bush et al., 2001). Teesside contains a mixture of urban, suburban, and rural environments and is formed from a collection of separate communities including Middlesborough, Redcar, Thornaby-on-Tees, Billingham, Hartlepool, and Stockton-on-Tees, each with their own identities, facilities, schools, etc. All of these characteristics make Teesside an interesting case study that deserves research focus."
- Added a new paragraph explaining our use of data from 2020 only, in the new 'Location and timeframe' subsection of the methods section of the paper, from line 152: "We chose to examine positive tests within Teesside during only 2020 for several reasons. Firstly, the local tiered restrictions were abandoned when the country entered a new national lockdown in early January 2021, and were not reinstated (UK Government, 2021), which means there were no further locally applied restrictions that could be examined. Secondly, the UK's vaccination programmed begun in December 2020, which means that analysing data that included cases into 2021 would have required additional data to properly adjust for this confounder. Inclusion of vaccination uptake information would have been needed, rather than numbers/proportion of eligible people in the population, as uptake is lower amongst deprived people (Mounier-Jack et al., 2023), but this would have been difficult to ascertain at the necessary spatial and temporal scales. And finally, including data collected over a longer time period would open up the analysis to other potential sources of confounding that cannot easily be controlled for, such as 'pandemic fatigue'."

2. *Comments related to the 'Data collation' part:*

   2.1 *I found the description of the data sources clear and informative.*

   [Response]

   We thank the reviewer for this feedback.

   2.2 *Major concerns - Line 160: 'contemporaneously for both pillars (from week of year 19 to 53) and ran a correlation on the two pillars.' What do authors mean by 'ran a correlation on the two pillars'? Did they simply compute a correlation coefficient between the two time series? I assume both series are non-stationary both in mean and variance? In addition, there is probably a high level of temporal correlation between the data points. In this context simply* *computing a correlation coefficient is not appropriate and basic principles of time series analysis should be considered.*

[Response]

This correlation did not include a temporal component; it was conducted on data that had been summed across all time points. We have now clarified the wording.

[Changes]

The description of the correlation, from line 177, now reads: “In addition, we also summarised the total number of cases of each lineage recorded by each pillar during the period when sampling was run contemporaneously for both pillars (from week of year 19 to 53), creating a dataset of total cases of each lineage summed across time for each pillar (each lineage had one value per pillar). We ran a Spearman’s correlation on this dataset, to check for a sampling bias for different lineages between the two pillars, as only pillar 1 tested the most severely ill patients, and illness severity can vary between lineages (Sievers et al., 2022; Goethem et al., 2022).”

3. *About model selection in the ‘Disease mapping’ part:*

In the ‘disease mapping’ part, authors compared a Poisson and Negative binomial regression within a Bayesian framework using the R-INLA package. Model selection was performed using the DIC.

The modern Bayesian approach instead recommends relying on Cross-validation. Cross-validation is a family of techniques that try to estimate how well a model would predict unseen data. This is especially appropriate in a Bayesian framework in which the goal is to model the full data-generating process. A must-read for a better understanding of this topic is: https://mc-stan.org/loo/articles/online-only/faq.html.

A classical assessment metric in this context is the CPO (conditional predictive ordinates) or the LOO-PD (leave-one-out log pointwise predictive density). This metric is defined for all observations in the dataset and measures the leave-one-out model predictive capabilities of the model. It is a special case of cross-validation in which only one observation is left out (leave-one-out cross-validation). These CPO/LOO-PD are then usually used along a logarithmic scoring rule to get a summary of the model's quality in terms of probabilistic forecasts, see https://doi.org/10.1198/016214506000001437.

The LOO and LGOO-CV can be routinely computed in R-INLA by specifying in the inla() call:

```
control.compute = list(
dic = TRUE,
cpo = TRUE,
waic = TRUE,
control.gcpo = list(enable = TRUE),
config = T
)
```

Because this is more consistent with the current recommended approach and as easy to implement as the DIC-based model selection, I recommend using the CPO along with a log-score approach, instead of the current one, if they really do want to stick with their current model selection.

I however don't think such a selection is relevant here. In the trade-off between Poisson and Negative binomial, two key elements are at stake:

1) *The Poisson distribution constraints the mean to be equal to the variance, which might be a strong assumption and prevent to specify of a good generative model;*
2) *The Negative binomial distribution allows to relaxation this assumption by introducing potential overdispersion (Var(.)>E(.)) but at the cost of additional parameters which might however lead to overfitting.*

It is known that the Negative binomial can naturally arise from a Poisson-Gamma model. In this model, the standard Poisson distribution appears as a limiting case.

R-INLA provides a class of priors (Penalised-Complexity prior, PC-prior) aiming to tame the risk of overfitting for nested models having a natural base model. In the case of the Negative binomial: it penalises the departure from a Poisson distribution. For the BYM2 model: first departure from the case in which there is no heterogeneity and then departure from the case where the whole heterogeneity is not spatially structured.

Key references include: DOI: 10.1214/16-STS576. For the Poisson-Negative binomial case, the full derivation of the PC-prior and additional details are provided on https://dansblog.netlify.app/2022-08-29-priors4/2022-08-29-priors4.html by Dan Simpson, also provided in DOI: 10.1214/17-STS576REJ

I would therefore recommend to just stick with the Negative binomial distribution along with a PC-prior (which is the R-INLA default). If authors really do want to strongly penalise the deviation from the Poisson case, priors parameters might be changed accordingly (see https://inla.r-inla-download.org/r-inla.org/doc/likelihood/nbinomial.pdf). Note that doi: 10.2807/1560-7917.ES.2021.26.28.2001401, which you are referring to multiple times only used the Negative binomial distribution.

[Response]
We thank the reviewer for this detailed comment (and the detailed review more generally). The authors agree with both of the main points made here: that there are better methods to compare INLA models than DIC (cross-validation), and that it makes perfect sense to simply use a negative binomial model in this context. One of the authors was concerned that we may have to justify our choice of negative binomial, given that much of the relevant literature appears to use Poisson models (or at least use them as the initial model). We also fell back on using DIC here as it is again very commonly used in the literature. We have therefore swapped the disease mapping models to negative binomial ones.
[Changes]
- Changed the final models to negative binomial. All output and figures have been changed to reflect this, though it has not made any difference to the results.
- From line 233, the relevant sentence in the paper now reads: "As the response variables in these models are aggregated counts or rates, we fit the models with negative binomial error distributions (using log link functions) and penalised complexity priors, which reduce the chance of overfitting (Riebler et al., 2016)."

4. *Critical reporting issues:*

[Response]

All of the model-specific information (priors, adjacency matrix, package versions, link functions, approximation and integrations methods etc.) can be found in the data analysis scripts, which have been uploaded to a repository and given a DOI (listed in the Data/Code section at the end of the paper: https://doi.org/10.25405/data.ncl.23815077). However, we appreciate that this information should also be more widely accessible, and so these details have been added to the paper or supplementary materials where relevant. See the specific changes for each point below.

[Changes]

- Added this text to line 292 of the paper: “Full details of all the model specifications (disease mapping and GLMMs) can be seen in the analysis scripts (https://doi.org/10.25405/data.ncl.23815077) and the supplementary materials”.
- Added a new section to the supplementary material file, “Model Specifications”, which gives details of all model parameters.

4.1. *About priors in the ‘Disease mapping’ part: Please specify in the main text or supplementary files the priors (distribution and values) for all parameters, notably the Negative binomial hyperparameter and the BYM2 latent effect.*

[Changes]

Added a new section to the supplementary material file, “Model Specifications”, which gives details of all model parameters, including hyperparameter values.

4.2. *About the adjacency matrix in the ‘Disease mapping’ part: Please, specify in main text or supplementary files which adjacency matrix was used for the BYM2 latent effect. Did you build it using the Queen contiguity definition? (i.e., what definition of ‘adjacent spatial units’ did you use?).*

[Response]

Yes, we did previously build the adjacency matrix using the queen contiguity definition. Adjacent polygons in our dataset are always connected by multiple points, so it would make no difference whether or not we specified queen or rook contiguity. However, we appreciate that rook contiguity could be argued to be more appropriate in the context of our research, so we have explicitly changed it to rook. We have added this detail and explanation to the supplementary materials.

[Changes]

- The analysis script now references the fact that queen and rook create the same adjacency matrix, and explicitly uses rook contiguity.
- The supplementary material file now includes explanation of the adjacency matrix method: “The adjacency matrix for the Teesside postcode districts was constructed using rook contiguity: shared borders between adjacent areas must be edges (multiple points) rather than single vertices (single points). This is appropriate in the context of disease transmission between UK postcodes districts, as a single point would represent a tiny area of linear contact that is not representative of a true geographical boundary. However, the arrangement and irregular shape of the postcode polygons in the shapefile means that it would make no difference whether or not we used queen or rook contiguity, as all polygons that connect do so by more than one point.”

4.3. *About R-INLA in the ‘Disease mapping’ part:*

4.3.1. *Please specify in the main text or supplementary files the version of R-INLA that you used. This is especially important because R-INLA is a package with very active development.*

[Changes]

- Added INLA package version (22.05.07) to line 233.

4.3.2. *Please specify in the main text or supplementary files how the results were approximated. Especially, to increase the accuracy, please use the 'Laplace' strategy.*

[Response]

The approximation method that we used was 'simplified Laplace'. We have now changed it as requested.

[Changes]

- Changed the approximation method of the INLA models to 'Laplace'.

4.3.3. *Please, specify in the main text or supplementary files which integration strategy was used. Especially, note that the 'eb' strategy severely underestimates uncertainty and is therefore not recommended.*

[Response]

We used the default integration strategy (control.inla = list(int.strategy = "auto")), which applies the grid method when the number of hyperparameters is ≤2 and the CCD (complete composite design) method when it is not. There were 3 hyperparameters in our models, so the CCD method was used. This has now been clarified in the supplementary material.

[Changes]

- Added this text to the supplementary material file: "Integration strategy: CCD (complete composite design). This was the default for our models as the number of hyperparameters was greater than 2 ("size for the nbinomial observations (1/overdispersion)", "Precision for PC_num", "Phi for PC_num"). e.g. control.inla = list(int.strategy = "auto")".

4.4. *About the link used in the model: The link used for the Poisson and Negative binomial distribution should be stated in the main text to avoid any ambiguity. I got from Figures and Supplementary that authors used the log link but it should be included in the methods part.*

[Changes]

- Added text to line 235 stating that we used the 'log' link function.

4.5. *About the posterior distribution for the hyperparameters of the 'Disease mapping' part: Please report in supplementary files the posterior distribution of all hyperparameters for the model (notably the two hyperparameters from the BYM2 model, which have an intuitive interpretation, as denoted in https://doi.org/10.1177/0962280216660421).*

[Changes]

- Added a table (Table S2) to section 1.2 of the supplementary material that contains the summary information of the posterior distributions for all fixed and random effects for each disease mapping model.

- The distributions can also be viewed in the figures in the new "Model Fit" section of the supplementary material file.

4.6. *About posterior predictive check in the 'Disease mapping part': In line with the view that the aim of a Bayesian model is to specify a generative model, a key step in the Bayesian analysis workflow is to compare the posterior predicted values against the observed one ('posterior predictive check, see https://doi.org/10.48550/arXiv.2011.01808 page 30). Please provide the posterior predictive check in the supplementary files. Note that with R-INLA, you must use inla.posterior.sample and inla.posterior.sample.eval to draw from the joint posterior distribution.*

This has now been conducted and the output has been included in the supplementary materials file as a series of figures. We have also included some other plots used to assess model fit. These demonstrated that the models fit the data well for total cases (all-cases) and for most of the lineages, but not very well for 3 of them, which have now been dropped from the paper.

[Changes]

- Added figures to the supplementary materials file in a new section (3) "Model validation" (figures S6, S8, S10, S12, S14, S16, S18, S20, S22), which include a comparison of the observed values to a sample generated from the posterior distribution (density plot), and a comparison of the observed values per postcode area to a sample generated from the posterior distribution (violin plot).
- Added figures to the supplementary materials file in the "Model validation" section (figures S5, S7, S9, S11, S13, S15, S17, S19, S21), which include observed vs fitted, CPO, PIT, and the posterior density of the intercept and hyperparameters.
- Added this text to the methods section of the paper, from line 238: "Model validation, via PIT (probability integral transforms) and plotting observed values against a sample generated from the posterior distribution, demonstrated a poor fit to the data for three of our lineages (B.1.1.309, B.1.1.315, and B.1.1.37), which have been dropped from the results section of the paper (see section 3.1 in the supplementary material for model validation and hyperparameter sensitivity analyses)."

4.7. *Authors should be praised for using the BYM2 latent effect.*

[Response]

We thank the reviewer for this compliment.

5. *Minor elements in the 'Disease mapping part':*

5.1. *Line 201: 'To help us define the spatial dynamics of COVID-19 in Teesside' this does not seem appropriate as your model does not include any interaction between space and time. As such, it cannot help you assess any 'spatial dynamic'. Please, consider changing 'dynamic' by 'heterogeneity'.*

[Changes]

- We have made the recommended change to this sentence (line 221).
- We have also removed use of the word 'dynamic' throughout the paper.

5.2. *Line 208-209: 'We included the expected number of cases for each postcode district […]', am I right to assume that by 'included' you meant 'included as offset'?*

[Response]

Yes, this has now been clarified.

[Changes]

- Line 230 now includes: "…as an offset…"

5.3. *I would recommend a mathematical description of the full model (including all priors) in supplementary files to avoid any misunderstanding.*
[Response]
As stated under "4. Critical reporting errors", this information was and is available in the analysis code in the analysis repository and full model specifications are now also in the supplementary materials. We believe that the extra details added in the paper and supplementary material, in combination with access to the analysis code, should provide sufficient explanation of the analysis method.
[Changes]
- Added a new section to the supplementary material file, "Model Specifications", which gives details of all model parameters.

6. *Other major concerns: What about sensitivity analyses (adjacency matrix+prior)?*
[Response]
As stated in the response to comment 4.2, all Teesside postcode polygons that connect do so by more than one point. Therefore, we did not do any sensitivity analyses comparing queen vs rook contiguity. Creating more complex adjacency matrices would have required extra data and resources that were beyond the scope of this research.
The models were run using a variety of different precision and phi hyperparameter values (larger precision, larger phi, and a combination of the two – see section 2.2 of the supplementary materials file or the analysis script for the exact values), none of which showed any tangible difference or improvement over the default PC prior. This code was originally omitted from the analysis script in the DOI repository, but it has now been added back in. We have also added some text to the paper and included the output of the prior comparisons to the supplementary material.
[Changes]
- The INLA analysis script now includes a section trialling different prior specifications.
- This text has been added to line 236 of the paper: "We used the default penalised complexity priors for the BYM2 model (Riebler et al., 2016) throughout, as a sensitivity analysis using different hyperparameter values showed no improvement to model fit."
- Added figures to the supplementary materials file in the section 3.1 (figures S6, S8, S10, S12, S14, S16, S18, S20, S22) demonstrating the lack of sensitivity of the models to different hyperparameter values. These figures compare 4 models, each fit with different hyperparameter values (see model code in the analysis repository (https://doi.org/10.25405/data.ncl.23815077), including: density of the posterior marginals of the intercept and phi and precision hyperparameters, mean and 97.5% quantiles for the intercept, comparison of the observed values to a sample generated from the posterior distribution, and comparison of the observed values per postcode area to a sample generated from the posterior distribution.

7. *About the 'Mixed-effect modeling' part:*

   7.1. *Please, explain why you did not consider any spatially-structured prior for the postcode district. Especially, why did you not consider a BYM2 latent effect, which might reduce toward the I.I.D. case that you imposed on your model?*

[Response]
The mixed-effect models were not Bayesian, and as such could not include the suggested BYM2 component. We see now that the description was ambiguous, and so we have clarified this.
[Changes]
- Added a sentence to the mixed effect modelling section on line 249: "These models were non-Bayesian and were fitted using maximum likelihood estimation."

7.2. *Line 239, the authors wrote: "which demonstrated negative binomial to be a superior fit for the data in all". Please, explain how you came up with this conclusion. Please report in supplementary files all the diagnostics provided by Dharma that you may have looked at.*
[Response]
The text output from each of the DHARMa model validation methods for each model was already copy/pasted into the GLMM analysis scripts. However, we appreciate that some details of the validation methods would be beneficial in the paper, and that the full validation details should be present in the supplementary materials, and so this has been added.
[Changes]
- Added a new section the supplementary material file (3) "Model validation". The subsection 3.2 contains an extensive description of the validation methods for the GLMM models, including the DHARMa R package tests. This section also includes validation figures (S25, S28: S37) for the final model for each dataset, which show several subplots created with the DHARMa tests: QQ plot, residuals vs predicted, outlier check, dispersion check, residuals vs time, and ACF. They also include a plot of the VIF values, and a plot of the observed values vs fitted values from the model.
- The text from line 271 in the paper now reads: "As our response variable (positive tests) is a count, we fit our models with Poisson and negative binomial distributions (using log link functions) before validating them with the "DHARMa" R package (Hartig, 2022), which uses a simulation approach to create interpretable scaled residuals. We used DHARMa's test and plotting functions to assess deviations from the expected distribution, dispersion, heteroskedasticity, temporal autocorrelation, and zero-inflation, alongside plots of observed values against those fitted from the models, which demonstrated that Poisson was a better fit for the data in all models, except for the all-cases model fit with a random gradient (for full details and output see section 3.2 of the supplementary material)."

7.3. *Is there any reasons to assume a fixed 'effect' of time (i.e., to include time as a fixed coefficient?). Why did you not consider penalised smoothing splines (which may in addition shrink toward the case you imposed) allowing the handle time in a more flexible way?*
[Response]
After receiving the reviewer comments we went back and reanalysed all of the datasets using several additional methods: a restricted cubic spline for time (using both 3 and 4 knots) (implemented via the "splines" R package); a smooth spline for time (implemented via the "mgcv" R package); an AR1 term for time (without any grouping); and an AR1 term for time grouped by postcode. Each of these (and the original models) were compared using the validation figures mentioned in the response to the previous comment (comment 7.2). For each dataset, the AR1 model that was grouped by postcode proved to be the best fit. This model structure accounts for the temporal autocorrelation present in the data across time, for each postcode, and fits the data well (see the observed vs fitted values across

week and postcode for the all-cases model in figure S3). The validation figures for all of the different model specifications for each dataset can be found in the analysis repository (https://doi.org/10.25405/data.ncl.23815077). We have included some example validation figures for these alternative models in the supplementary material (section 3.2).

[Changes]

- Reanalysed the datasets using several alternative specifications for time in the models, including smooth and restricted cubic splines, and an autoregressive term of order 1.
- Changed the description of the GLMMs in the methods section of the paper to include the new AR1 term, from line 251: “We included an autoregressive term of order 1 (AR1) for week of year for each postcode district to account for temporal auto-correlation (non-independence) of cases over time within each separate district.”
- Added a description to the methods section of the new alternative model specifications that were tried, from line 278: “Because it would be reasonable to assume that the number of cases over time would follow a non-linear trajectory, we also fit several alternative model specifications that included week of year as a fixed effect with either a smooth or restricted cubic spline instead of an AR1 term, however, these models were a poorer fit for all datasets than the AR1 models (see section 3.2 of the supplementary material).”
- Added validation figures to section 3.2 of the supplementary material for the final (best fitting) models for each dataset (S25, S28: S37).
- Added validation figures to section 3.2 of the supplementary material for the all-cases models that include splines for time (a restricted cubic spline (using 4 knots) and a smooth spline) as an example of the poorer fit that these approaches provide, relative to the AR1 term (S26, S27).

7.4. *What do authors mean by ‘random gradient for week for each postcode district’? Is it a time-specific intercept varying by postcode district? Are these intercepts temporally correlated (e.g., AR(p) or RW(p)) process? Is that a smoothing spline? I don’t understand if there is any assumption about the space-time interaction? Please clarify the relation between space and time in your model using Knorr-Held’s classical typology (https://pubmed.ncbi.nlm.nih.gov/10960871/). To avoid any ambiguity for readers, please provide the mathematical counterpart of your model in the supplementary files.*

[Response]

Because the models that the reviewer is referring to are frequentist GLMMs (not Bayesian models), the Knorr-Held typology does not quite translate. This model includes a random intercept for postcode (just like the other GLMMs), to control for repeated sampling of postcodes, plus an additional random gradient that allows the effect of week on cases to vary by postcode. In R, this is included with the following code: “+ (Week + 1|Postcode)”. The full code for all the models is available in the analysis scripts, and as stated in response to the “Critical reporting issues”, full model specifications are now also in the supplementary materials. However, we appreciate that the model description in the paper is confusing, and so we have clarified it.

[Changes]

- Added some text to this sentence from line 260: “We fitted an additional model to the all-case data, including the same fixed effects and the random intercept for postcode, but this model did not include an AR1 term and instead used a random gradient for week for each postcode district (allowing the effect of week on cases to vary by postcode), which allowed us to map and compare the rate of change in cases reported each week across the different postcodes."

7.5. *Line 235: 'To account for variation in symptom onset and testing delay after infection, we applied a two-week time lag to all temporal variables'. In the cited reference doi: 10.2807/1560-7917.ES.2021.26.28.2001401 the baseline analysis relied on a 7-day delay. Here, the authors used a 2-week delay. The choice of a 2-week delay seems very arbitrary without any theoretical support in the cited reference.*
[Response]
While the paper we reference (https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2021.26.28.2001401) does indeed use a 7-day delay in the main analyses, they also conducted sensitivity analyses using different length delays. They found that the impact of NPIs did not become apparent until after 14 days:
"For time periods 1–7 and 8–14 days, the IRR values were above 1, indicating a positive association between cases/death and the intervention variable. For periods starting 15 days onwards the IRR was generally below 1 suggesting a negative association between the outcome and the intervention. This pattern probably reflects the time lag between exposure, latency and disease detection, so that the impact of interventions only kicks in after what is effectively a lag period of 14 days."
We also chose to use a 14-day rather than 7-day delay period because we aggregated cases in our dataset by 7-day periods. If we had used a 7-day delay, this could potentially allow situations where a case at the beginning of a week is considered to be separated from the initiation of an NPI by 7 days, when in fact it may be considerably fewer (and too close in time to demonstrate a relationship).
We have now clarified this reasoning.
[Changes]
- Added this text to line 267 of the paper: "Because we aggregated case numbers and summarised temporal variables by week, a one-week delay could have artificially assumed a greater separation between cases and temporal events than actually happened. Additionally, the sensitivity analyses in Hunter et al. (2021) demonstrated that non-pharmaceutical interventions in the UK did not show an impact on numbers of cases until after 14 days."

7.6. *Please make it clear that you used the log link for your model (this is specified in Figure 4 but should be specified in the method section).*
[Changes]
- Added text to line 272 stating that we used the 'log' link function.

7.7. *Authors should be praised for avoiding AIC-based variables selection, which would have introduced post-selection bias.*
[Response]
We thank the reviewer for this compliment and confirmation of our choice to avoid such a method.

8. *General comment about the method:*

8.1. *Mixing both Bayesian and Frequentist approaches is very weird as parameters have fundamentally not the same nature in these two framework and methodologies are different.*

[Response]
We appreciate the reviewer's opinion, but we believe that there can be good reasons to combine these different approaches within the same study. We initially approached this research from the perspective of trying to understand the relationships between the number of positive tests of different lineages in Teesside in relation to a range of covariates and we were able to use frequentist GLMMs to help us understand these associations. However, we also wanted to examine and visually represent the spatial pattern of positive tests across the different postcodes, but simple maps of positive tests or standardised incidence ratios (SIRs) would not account for spatial structure (postcode adjacency), variable population sizes, and low positive test numbers. This latter objective could only be achieved by mapping the relative risk with a Bayesian conditional autoregressive model. We believe that the maps of relative risk (and the corresponding exceedance probabilities) provide additional insight to the frequentist model outputs and complement the paper as a whole.

8.2. *Authors restricted the spatial domain to Teesside, what about possible border effects?*
[Response]
Teesside is a conurbation of towns and villages that form a small and semi-isolated sub-region, which has a strong sense of identity that is distinct and separate from nearby towns and cities (https://doi.org/10.1016/S1353-8292(00)00037-X, https://doi.org/10.1080/09595237500185411). We believe the features of this area reduce the chance for any border effects. Teesside has a high population density, the population is focussed in the centre, and it is bordered by rural landscapes (including a large national park) and the North Sea. It has strong connections to distant places via railway and several major roads.
We also observed that COVID-19 cases tended to be lower and later in the outer postcodes, while the cases in the more central postcodes tended to be higher and earlier, indicating that the virus was introduced into the centre of Teesside from outside of the area.
We appreciate that the justification for the spatial scale and use of Teesside was insufficient, and so this has been expanded upon in the paper (as described in our response to comment 1.2 from this reviewer).
[Changes]
- Added a new subsection to the methods section of the paper (Location and timeframe) that gives some background information about Teesside and helps to justify our focus on this area: "The wider Teesside area is a sub-region of the North East region of England, centred around the mouth of the river Tees. While Teesside has reasonably good local and national transport links, it is relatively isolated geographically as it borders the North Sea to the north east, The North York Moors national park to the south east, and extensive farmland to the west. Teesside has a distinct cultural identity due to its industrial history, which has also left its population with a greater burden of diseases and socio-economic deprivation that stigmatises and further culturally isolates the area (Bush et al., 2001). Teesside contains a mixture of urban, suburban, and rural environments and is formed from a collection of separate communities including Middlesborough, Redcar, Thornaby-on-Tees, Billingham, Hartlepool, and Stockton-on-Tees, each with their own identities, facilities, schools, etc. All of these characteristics make Teesside an interesting case study that deserves research focus."

9. *Results:*

   Figure 4: ', asterisks indicate significance', please add the matching between number of stars and confidence level.

   [Changes]

   The following text has been added to the legend of figure 4: "Asterisks indicate significance level: '***' indicates p ≤ 0.001, '**' indicates p ≤ 0.01, and '*' indicates p ≤ 0.05."

10. *Discussion:*

    10.1. *Line 462: 'Our study has demonstrated the effects of weather and government interventions on the spatio-temporal'. In the meantime, the authors wrote, line 362: 'These results must be interpreted with caution due to several limitations.'. 'Demonstrated' seems way too strong given the empirical approach and the data (plagued with reporting bias and other issues as denoted) on which the analysis relies on. This statement seems even more inappropriate given the lack of sensitivity analyses to priors and adjacency matrix.*

    Note that many more limits could be raised regarding the analysis, among which imperfect ascertainment (false positive and negative due to tests' imperfect clinical performances), reverse causality (e.g., endogeneity issues between Tier and number of positive tests), unplausible underlying assumption that the observed variables are uncorrelated with the unobserved variables (see Mundlak's model). All these could bias the coefficients, which, again, call for greater care in model interpretation.

    [Response]

    This sentence on line 462 referred to the results of the (frequentist) GLMM models, therefore a lack of sensitivity analyses for the (Bayesian) disease mapping models is not relevant here (and those have now been added anyway). We have also altered the wording of this sentence slightly, to make it more relevant to our study context. We also believe that we have already adequately addressed this reviewer's concerns regarding reporting of model specifications and validation.

    Regarding test results, while false positives/negatives are of course possible in PCR tests (though very unlikely in a standardised and nationalised testing programme), the dataset that we analysed was of PCR samples that had been whole genome sequenced to identify the specific viral lineage present in the samples. This means that false positives/negatives are not a problem for our data. While it is possible that there could be incorrect results due to contamination, the chance of this should be vanishingly small due to the sequencing being conducted in established laboratories with experienced staff, overseen by a national organisation (COG-UK https://webarchive.nationalarchives.gov.uk/ukgwa/20230505083137/https://www.cogconsortium.uk/), and following specific protocols (https://webarchive.nationalarchives.gov.uk/ukgwa/20230507113734/https://www.cogconsortium.uk/priority-areas/data-linkage-analysis/protocols/).

    Regarding endogeneity between tier level and positive tests, we acknowledge that tier level at a given time point (t0) will have been determined by positive test numbers in the previous 1-2 weeks (t-1 and t-2). However, because we applied a time-lag to our temporal variables, this means that we are looking at the effect of tier level during t0 on cases at t+2. This represents a separation of 3-4 weeks, which we believe should be enough to decouple the causal effects, especially as we included an AR1 temporal autocorrelation term in our models. The tier levels also did not change in Teesside once they had been put in place (there was a national lockdown in between tier2 and tier3), so they can be viewed as more rigidly fixed than fluidly changeable.

Regarding the model assumptions, we believe that we have now correctly specified the fixed and random effects in our models. Including the AR1 term for week by postcode has drastically improved the fit of the models.

[Changes]

- Changed the wording of the first sentence of the final paragraph of the discussion, from line 545: "Our study has demonstrated the effects of weather and government interventions on the number of SARS-CoV-2 positive tests at a sub-regional scale in Teesside, UK."

10.2. *After reading the whole study, I do not believe that labelling the study as 'spatio-temporal' is really appropriate. First, the disease mapping model does not include any temporal or spatiotemporal component (everything being aggregated over the entire period). Second, the GLMM model does not include any terms aiming to handle residual spatial autocorrelation or residual interaction between space and time. Spatiotemporal models can be broadly classified into 4 types, each assuming a type of interaction between space and time, see https://pubmed.ncbi.nlm.nih.gov/10960871/. Type IV models are usually the right type of spatiotemporal model when modelling infectious diseases spread in a spatio-temporal disease mapping context. The 'group' argument in the latent effect in R-INLA can help set such spatio-temporal model. Else, space-time interaction can be easily handled through the INLAspacetime package: https://cran.r-project.org/web/packages/INLAspacetime/index.html.*

[Response]

We agree with the reviewer that the analysis methods used in the submitted paper do not include any specific spatio-temporal models. Our use of the term was in reference to a more general collective approach that was used during the paper as a whole. However, we acknowledge that our interpretation was unhelpful and could be misleading, and so we have changed the wording where relevant.

[Changes]

- Changed the paper title to: "Spatial and temporal…" instead of "Spatio-temporal..."
- Changed the wording throughout paper to remove "spatio-temporal" and replace it with more appropriate wording.