PEER REVIEW HISTORY

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

ARTICLE DETAILS

TITLE (PROVISIONAL)	Staff-Pupil SARS-CoV-2 Infection Pathways in Schools in Wales: A Population Level Linked Data Approach
AUTHORS	Thompson, Daniel A Abbasizanjani, Hoda Fry, Richard Marchant, Emily Griffiths, Lucy Akbari, Ashley Hollinghurst, Joe North, Laura Lyons, Jane Torabi, Fatemeh Davies, Gareth Gravenor, Mike B Lyons, Ronan A

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Dr. Sarah Nevitt Institution and Country: University of Liverpool, United Kingdom of Great Britain and Northern Ireland Competing interests: None
REVIEW RETURNED	19-Feb-2021

GENERAL COMMENTS	I have conducted a statistical review of the manuscript "Staff-Pupil SARS-CoV-2 Infection Pathways in Schools: A Population Level Linked Data Approach" The authors describe a data linkage study, conducted using data from schools in Wales, to examine associations between testing positive for SARS-CoV-2 infection and exposures at the school and linked household level. Although a very interesting topic, I'm unsure of the suitability of some of the methods used and therefore how to interpret some of the results. Could the authors provide some clarifications? 1) I'm not completely following the linkage of data sources in Figure 1 and the accompanying text on page 5. Where does the SAIL Databank at Swansea University feature? (i.e. is the census data held within the databank?) Also, where does the 'Welsh COVID-19 ecohort' (reference 14) feature? Or is this the name given to the resulting cohort used in this analysis? Related to this, what is the C20 cohort towards the bottom of Figure 1? I suggest some additional labels may be needed on Figure 1 and perhaps some arrows rather than just lines so it is clear what is linked to what and at which stage links happen. 2) Table 1 and Table 2: I suggest that the denominator for the % positive tests in Table 1 should be the number tested rather than the total population (as those who have not had a test cannot test positive).

The number of positive tests in Table 1 and Table 2 don't add up for either staff or pupils, and I think there might be some digits missing from Table 1 for students (over 7000 positive tests for pupils listed in Table 2 and around 2000 in Table 1)?

Please check the numbers and once verified, clarify any differences in the numbers between the tables (e.g. if data on exposures are missing).

3) Unit of analysis: For the individuals within the dataset who were tested, if applicable, how were multiple tests handled? For example, would an individual be classified as receiving a positive test if any of their tests come back positive (even if they'd previously received negative test results)? And each individual would only be included as receiving a positive test once? i.e. no multiple counting of any individuals within analysis if they did receive more than one test?

4) Logistic regression analyses: A few queries about these analyses:
a) I'm not sure I understand the exposure measures of count of cases / total number of cases in the households or school.
Do these variables imply that more cases equal more exposure? i.e. number of cases is being treated as a continuous exposure measure? If so, I'm unsure about whether this is appropriate as this will be influenced by and bounded by the size of the household and size of the school; i.e. a pupil who lives only with a parent / carer could only have been exposed to one other person at home (maximum 1 case), but would it be appropriate to consider this pupil to be 'less exposed' than a pupil who lives in a much bigger household of say >6 people where there may be two cases? Similar for schools, if a small school or a small year group has one case, this could actually reflect a lot more exposure for the rest of the school than a large school or large year group with 10 cases.

The results are currently interpreted as associations between the 'total number of cases' and a positive test but really if the exposure variable is being analysed as a continuous variable then the interpretation is an increase / decrease in the odds of a positive test as the number of cases in each context increases. So currently some of the results suggest that more cases (more exposure?) actually results in significantly reduced odds of a positive test which seems counterintuitive.

I actually don't think that the authors are trying to measure the association between the 'amount of exposure' and the probability of a positive test but rather the location of exposure (household / school / both).

Therefore, I suggest that it would be more appropriate for the exposure variable included in analysis to be exposure in each setting; i.e. exposure at home (yes or no), exposure in school year group (yes or no), exposure in linked household (yes or no) etc, potentially with interaction terms to reflect exposure in multiple settings (e.g. at home and at school). The number of positive cases within each setting or the size of the household / school etc. could then be an adjustment variable to adjust for small / large households or schools, rather than part of the association.

b) This may be addressed if the authors adopt my suggestion in comment 4a, but from looking at Table 3, I wasn't sure whether this table reflected separate analyses for each exposure variable, or whether all of these variables were included in the same model (for each group; staff and pupils, staff only and pupils only). Some of these exposure variables seem to be overlapping; i.e. the number of cases within a year group will be a subset of the number of cases within the school, and are therefore not independent variables for the purposes of a regression model.

	Please add some further labelling to Table 3 regarding whether these analyses are univariable or multivariable analyses, and consider the definitions of the exposure variables (also see comment a). c) Again, this may be addressed in response to comment 4a, but I'm not sure which results some of the results text are currently referring to. Such as: "Staff members in primary and special schools had a higher odds of a SARS-CoV-2 positive test compared with middle and secondary schools, and staff had higher odds of a positive outcome compared to the reference level of pupils (OR 2.99, 95%CI 1.67-5.37, p value <0 .001)." I cannot see this result in any tables? "When stratifying by pupils, and adjusting for covariates (including household cases), the total number of cases in the school was not associated with increased risk of test positivity (Table 3)." I'm not sure which result in Table 3 this is referring to? Please also check when interpreting results of logistic regression, expressed as odds ratios, that results are interpreted as odds, rather than as 'risk.' Risk of a positive test is referred to several times in the results. 5) Limitations: The authors have performed an adjusted analysis and have highlighted some weaknesses in the available data including changes to the school environment. What I don't think it really mentioned within the interpretation or limitations is the likely scope for a lot of unmeasured confounding at the level of the individual rather than the school. In other words, behaviours and activities of individuals with or without exposures or with or without positive tests (e.g. physical distancing from others, frequency of day to day tasks such as shopping) and also length of any exposures, all of which are likely to impact on the infection transmission pathway but will not be captured within this data. I suggest adding some discussion of this and any other measured or unmeasured factors which may influence the transmission pathway into the limitations section. 6) Table S3: I was unsure whether these variables were missing at the level of the school or the level of the individual? i.e. does number of staff within the school mean that for 842 schools, the number of staff was missing? Also, please define the abbreviation RALF

REVIEWER	Reviewer name: Chris Taylor Institution and Country: Cardiff University, United Kingdom of Great Britain and Northern Ireland Competing interests: None
REVIEW RETURNED	11-Feb-2021

GENERAL COMMENTS	The results of this analysis are very important. The research utilises a novel dataset in order to answer a very sophisticated research question (although the analysis is relatively straightforward, using logistic regression). In general the paper would benefit from greater clarity in the use of terms and descriptions (see examples below) (these are minor revisions). I also highlight some questions that may require new analysis relating to BAME and SEN pupils, staff and families (unless a justification for not doing this can be provided). (NB this is the distinction between major revision and minor revisions for my recommendation).

Also I would recommend the authors consider the value of using multi-level modelling of this data - this could incorporate other geographical factors (such as local authority, regional consortia or health board) and look for school-level differences, thereby differentiating between national policies/guidance to reduce transmission and school-level mitigations to reduce transmission (however, given the urgency to publish these national results this may just be a recommendation rather than a requirement for publication, but could be acknowledged as a next step in the research).

Other comments:

A clearer description of what is meant by potential exposure in settings (e.g. Table 2) would be useful. Presumably this refers to potential exposure to someone else in those settings with a positive test? (that could be explained more clearly).

During the discussion of results it would be helpful to at least acknowledge the limitations of the analysis - e.g. potential exposure is linked only to positive test results, not necessarily all cases (particular non-symptomatic cases).

What is meant by testing positive across all outcomes (p.9 first paragraph). What is meant by outcomes here?

Explain why it is important to adjust for some characteristics (age, sex, rurality etc) but other variables are not adjusted for (e.g. ethnicity).

Given the prevalence of BAME groups amongst positive test results it is not clear why this has not been considered in the analysis. There may be justification for this but otherwise I would recommend that the logistic regression models incorporate ethnicity.

Similarly, SEN should also be considered - some SEN groups are necessarily going to require greater contact with staff and are likely to have greater contact with other pupils. Could this be considered in the analysis and results? (I note special schools seem to be included in the analysis - for similar reasons it might be useful to compare results with and without special schools included in the statistical models).

What is meant by the wider bubble of cases (p10 first paragraph)

The conclusions associated with 'within the same year group' do make theoretical sense, but this could be emphasised further by acknowledging that most classroom interactions are likely to be within the same year group, which contribute to that finding and conclusion.

Given the relatively high levels of non attendance during this time period is it worth clarifying whether attendance in school was a necessary condition for the analyses? (ie. does the analysis distinguish between pupils in school and pupils in home?)

VERSION 1 – AUTHOR RESPONSE

Staff-Pupil SARS-CoV-2 Infection Pathways in Schools: A Population Level Linked Data Approach

Editor in Chief Comments to Author

Title add "in Wales" after "schools"

Response: Done

Abstract needs to mention Wales in Methods

Response: Done

What this study adds delete "First UK" the journal style is NOT to describe the research as the first, as this up to others to decide after publication (see instructions to authors) Answer the points raised by both reviewers

Response: Done

Reviewer:	**1**
Reviewer name:	**Chris Taylor**
Institution and Country:	**Cardiff University, UK**

- **Comments to the Author**:

The results of this analysis are very important. The research utilises a novel dataset in order to answer a very sophisticated research question (although the analysis is relatively straightforward, using logistic regression).

In general the paper would benefit from greater clarity in the use of terms and descriptions (see examples below) (these are minor revisions).

I also highlight some questions that may require new analysis relating to BAME and SEN pupils, staff and families (unless a justification for not doing this can be provided). (NB this is the distinction between major revision and minor revisions for my recommendation).

Also I would recommend the authors consider the value of using multi-level modelling of this data - this could incorporate other geographical factors (such as local authority, regional consortia or health board) and look for school-level differences, thereby differentiating between national policies/guidance to reduce transmission and school-level mitigations to reduce transmission (however, given the urgency to publish these national results this may just be a recommendation rather than a requirement for publication, but could be acknowledged as a next step in the research).

Response: Many thanks for the suggestion. Indeed, including the multi-level aspect in the models is planned for future analysis. This initial analysis presented in the paper provides a generalised overview of transmission within school settings in Wales, hence assuming everything is occurring at the lowest individual-level with no cluster effects. We have drawn attention to the planned use of multi-level models in paper as follows:

(Implication section)

Also required is further work on specific subgroups of the school populations for example, pupils with Special Educational Needs and those from different ethnic minorities. As part of these future developments in the work, considerations to multi-level modelling and cluster effects within school settings will be included.

- **Other comments:**

A clearer description of what is meant by potential exposure in settings (e.g. Table 2) would be useful. Presumably this refers to potential exposure to someone else in those settings with a positive test? (that could be explained more clearly).

Response: Correct, thank you for your suggestion. Explanation has been added to Table 2 as follows:

Table 2 caption:

Table 2: Distribution of known potential exposure to infection by setting for staff and pupils (excluding staff contracted to multiple schools, and pupils aged 11 or 18+.

Table 2 heading:

Exposure to a known SARS-CoV-2 positive case for staff and pupils in the 14-day preceding window of their first SARS-CoV-2 positive test from 2020-08-01 to 2020-12-25

During the discussion of results it would be helpful to at least acknowledge the limitations of the analysis - e.g. potential exposure is linked only to positive test results, not necessarily all cases (particular non-symptomatic cases).

Response: We agree this is a limitation in the approach used. We have included a "Study strengths and limitations" section where this limitation is detailed. In addition, the following has been added to the section in response to your comment:

Hence, potential exposure is linked only to positive test results and not necessarily all cases (particularly non-symptomatic cases).

What is meant by testing positive across all outcomes (p.9 first paragraph). What is meant by outcomes here?

Response: We have adjusted the sentence in question for clarity, it now reads "we found significantly increased risk of testing positive across all settings"

Explain why it is important to adjust for some characteristics (age, sex, rurality etc) but other variables are not adjusted for (e.g. ethnicity).

Response: There were some limitations in the data available to us in this study, we have included the main risk factors where the data allowed. For example, our analysis did not adjust for ethnicity due to incomplete coding of this information in our available data; we have now included this as a limitation to our study in the appropriate section as follows:

Strength and Limitations section:

We were unable to account for ethnicity of pupils and staff in the study due to incomplete coding of this information in the available data.

Given the prevalence of BAME groups amongst positive test results it is not clear why this has not been considered in the analysis. There may be justification for this but otherwise I would recommend that the logistic regression models incorporate ethnicity.

Response: We agree, we were unable to incorporate BAME groups as described above and therefore we are unable to include ethnicity in our regression models.

Similarly, SEN should also be considered - some SEN groups are necessarily going to require greater contact with staff and are likely to have greater contact with other pupils. Could this be considered in the analysis

and results? (I note special schools seem to be included in the analysis - for similar reasons it might be useful to compare results with and without special schools included in the statistical models).

Response: We would like to thank the reviewer for this suggestion. We agree that this is an important issue however the purpose of this study was to provide an overview of transmission within school settings. Future analyses will be undertaken to investigate at-risk groups of pupils – we have added a sentence to the implications to acknowledge the need for these analyses.

What is meant by the wider bubble of cases (p10 first paragraph)

Response: Thank you, this did need clarification. We have revised the text to "linked cases in a household".

The conclusions associated with 'within the same year group' do make theoretical sense, but this could be emphasised further by acknowledging that most classroom interactions are likely to be within the same year group, which contribute to that finding and conclusion.

Response: Yes, thank you, we have now added this to the paper.

Given the relatively high levels of non attendance during this time period is it worth clarifying whether attendance in school was a necessary condition for the analyses? (ie. does the analysis distinguish between pupils in school and pupils in home?)

Response: The education attendance data for the 2020-2021 academic year is not available yet. As mentioned in the limitations section, we were unable to account for those days when pupils may not have been present in the school on an individual basis (which may have resulted in different exposures for a small number of cases) or as a result of class level covid isolation events. Our analysis does account for exposures at home, so if pupil or staff were at home during a household outbreak they would be assigned a 'home transmission' pathway.

Reviewer:	**2**
Reviewer name:	**Sarah Nevitt**
Institution and Country:	**University of Liverpool, UK**

- **Comments to the Author:**

I have conducted a statistical review of the manuscript "Staff-Pupil SARS-CoV-2 Infection Pathways in Schools: A Population Level Linked Data Approach"

The authors describe a data linkage study, conducted using data from schools in Wales, to examine associations between testing positive for SARS-CoV-2 infection and exposures at the school and linked household level.

Although a very interesting topic, I'm unsure of the suitability of some of the methods used and therefore how to interpret some of the results. Could the authors provide some clarifications?
1) I'm not completely following the linkage of data sources in Figure 1 and the accompanying text on page 5. Where does the SAIL Databank at Swansea University feature? (i.e. is the census data held within the databank?) Also, where does the 'Welsh COVID-19 ecohort' (reference 14) feature? Or is this the name given to the resulting cohort used in this analysis? Related to this, what is the C20 cohort towards the bottom of Figure 1?

Response: All the data sources used to create our e-cohort are held within the SAIL Databank at Swansea University. SAIL operates using two key anonymised linkage fields; information at the individual level can be linked together from different datasets in SAIL using an Anonymised Linking Field (ALF). Individual ALFs can also be grouped at the household level using Residential Anonymised Linking Fields (RALF) – an address based linkage system which can be enhanced using administrative and environmental data. In this study we used the ALF and RALF linkage fields to create an e-cohort of linked health and administrative education data linked.

The 'Welsh COVID-19 e-cohort' from (14) (also called the ` C20 cohort') consists of all people alive and known to the NHS in Wales on or after the 1st January 2020 (and is held within the SAIL Databank). The `C20 cohort' is also replaced with `Welsh COVID-19 e-cohort' in Figure 1 for consistency.

To the Welsh COVID-19 e-cohort we linked two administrative educational datasets, the School Workforce Annual Census (SWAC) and the Pupil Level Annual School Census (PLASC), using ALF. We also linked staff and pupils via educational settings using a School Anonymised Linking Field (SALF). Furthermore, we linked staff and pupils to their household members using RALF from the Welsh COVID-19 e-cohort.

I suggest some additional labels may be needed on Figure 1 and perhaps some arrows rather than just lines so it is clear what is linked to what and at which stage links happen.

Response: We have updated the figure (added arrows and updated data sources' name), and also the caption in the new version as follows:

Figure 1. Health and administrative education data linkages. Four data sources are used to create our e-cohort: the Welsh COVID-19 e-cohort, SWAC, PLASC and COVID-19 antigen testing data. We linked SWAC and PLASC to the Welsh COVID-19 e-cohort. We also linked staff and pupils via educational settings using a School Anonymised Linking Field (SALF). Furthermore, we linked staff and pupils to their household members using the Welsh COVID-19 e-cohort. Missing variables of staff and pupils (in the Welsh COVID-19 e-cohort) before being confirmed eligible are reported in Table S3.

2) Table 1 and Table 2: I suggest that the denominator for the % positive tests in Table 1 should be the number tested rather than the total population (as those who have not had a test cannot test positive).

Response: Thank you – we believe that what you have suggested is already included in the table. Table 1 includes the count and percentages of all individuals, those tested and those with positive results.

The number of positive tests in Table 1 and Table 2 don't add up for either staff or pupils, and I think there might be some digits missing from Table 1 for students (over 7000 positive tests for pupils listed in Table 2 and around 2000 in Table 1)?
Please check the numbers and once verified, clarify any differences in the numbers between the tables (e.g. if data on exposures are missing).

Response: Thank you for highlighting this. Table 1 was an old version of the table which was mistakenly used in the submitted version. The table contained pupils aged 18+ and test results for a shorter time period. Furthermore, we excluded staff contracted to multiple schools from all the counts (previously they were only excluded in the statistical models), and have updated Tables 1 and 2, and Figure 1 (an additional step added to the figure), accordingly.

3) **Unit of analysis: For the individuals within the dataset who were tested, if applicable, how were multiple tests handled? For example, would an individual be classified as receiving a positive test if any of their tests come back positive (even if they'd previously received negative test results)? And each individual would only be included as receiving a positive test once? i.e. no multiple counting of any individuals within analysis if they did receive more than one test?**

Response: Multiple tests were handled using the following conditions: if an individual has multiple tests and any return positive, the individual's outcome is positive and date of the positive test taken as the date-of-interest; if all tests return negative, the individual's outcome is negative, and date of the most recent negative test taken as the date-of-interest. We have clarified this in the paper as follows:

Statistical Modelling section

Our outcome was the probability of testing positive, following a pillar 1 or pillar 2 test. When an individual has multiple test results: if any return positive, the individual's outcome is positive and date of the positive test taken as the date-of-interest; if all tests return negative, the individual's outcome is negative, and date of the most recent negative test taken as the date-of-interest.

4) **Logistic regression analyses: A few queries about these analyses:**
a) I'm not sure I understand the exposure measures of count of cases / total number of cases in the households or school.
Do these variables imply that more cases equal more exposure? i.e. number of cases is being treated as a continuous exposure measure? If so, I'm unsure about whether this is appropriate as this will be influenced by and bounded by the size of the household and size of the school; i.e. a pupil who lives only with a parent / carer could only have been exposed to one other person at home (maximum 1 case), but would it be appropriate to consider this pupil to be 'less exposed' than a pupil who lives in a much bigger household of say >6 people where there may be two cases? Similar for schools, if a small school or a small year group has one case, this could actually reflect a lot more exposure for the rest of the school than a large school or large year group with 10 cases.
The results are currently interpreted as associations between the 'total number of cases' and a positive test but really if the exposure variable is being analysed as a continuous variable then the interpretation is an increase / decrease in the odds of a positive test as the number of cases in each context increases. So currently some of the results suggest that more cases (more exposure?) actually results in significantly reduced odds of a positive test which seems counterintuitive.

I actually don't think that the authors are trying to measure the association between the 'amount of exposure' and the probability of a positive test but rather the location of exposure (household / school / both). Therefore, I suggest that it would be more appropriate for the exposure variable included in analysis

to be exposure in each setting; i.e. exposure at home (yes or no), exposure in school year group (yes or no), exposure in linked household (yes or no) etc, potentially with interaction terms to reflect exposure in multiple settings (e.g. at home and at school). The number of positive cases within each setting or the size of the household / school etc. could then be an adjustment variable to adjust for small / large households or schools, rather than part of the association.

Response: Thank you for the thoughtful comments. Case number is indeed treated as a continuous exposure measure. And we agree that this will be influenced by the household/school size. In our analyses we adjusted, as suggested, for household/school size. Our aim was to model a situation where the probability of testing positive can potentially increase with case number, and also with size. We stopped short of exploring the further details of the functional relationship. We had tried the interaction between case numbers and household/school size but found we could not obtain a stable solution for the more detailed model fit. We would therefore prefer to retain our simple approach, with the following clarification in Study Strengths and Limitations: "In our analysis we could test only for additive effects (log odds scale) of the case numbers that individuals were exposed to, combined with the size of the population in which the cases were identified (household or school). As more data becomes available, the interaction, or other functional relationships between the effect of exposure to a certain number of cases and the background population size (or density) could be explored in more detail".

b) **This may be addressed if the authors adopt my suggestion in comment 4a, but from looking at Table 3, I wasn't sure whether this table reflected separate analyses for each exposure variable, or whether all of these variables were included in the same model (for each group; staff and pupils, staff only and pupils only). Some of these exposure variables seem to be overlapping; i.e. the number of cases within a year group will be a subset of the number of cases within the school, and are therefore not independent variables for the purposes of a regression model.**
Please add some further labelling to Table 3 regarding whether these analyses are univariable or multivariable analyses, and consider the definitions of the exposure variables (also see comment a).

Response: Thank you for the comment and apologies for the confusion. Further detailed labelling has been added to table to help clarify and a new row created to highlight the difference in M1, M2 and M3.

Table 3 caption:

Table 3: Fully adjusted multivariable Logistic Regression Results (M1 Staff and Pupils; M2 Stratified by Staff; M3 Stratified by Pupils). Adjustments for age, sex, residential settlement type, number of pupils and staff within the linked school, and number of people within linked household are included in the models, odds ratios of the fully adjusted covariates can be found in Table S2. Odds ratios are calculated per individual case of known exposure.

Within Table 3, count of pupil cases exposure variable now reads:

M1 and M2: Count of pupil cases within the linked school

And a new row for M3 to separate the exposure variables between the stratifications clearly:

M3: Count of non-year group pupil cases within the linked school

c) Again, this may be addressed in response to comment 4a, but I'm not sure which results some of the results text are currently referring to. Such as:

“Staff members in primary and special schools had a higher odds of a SARS-CoV-2 positive test compared with middle and secondary schools, and staff had higher odds of a positive outcome compared to the reference level of pupils (OR 2.99, 95%CI 1.67-5.37, p value <0 .001).” I cannot see this result in any tables?

Response: Thank you for highlighting the missing data and apologies for the omission. A supplementary Table (Table S2) has been added which includes the fully adjusted odds ratios of all covariates for the models presented in Table 3, reference to this table has been added to Table 3 caption.

“When stratifying by pupils, and adjusting for covariates (including household cases), the total number of cases in the school was not associated with increased risk of test positivity (Table 3).” I’m not sure which result in Table 3 this is referring to?

Thank you for highlight this, the text has been updated as follows:

When stratifying by pupils, and adjusting for covariates (including household cases), the total number of staff and non-year group cases in the school was not associated with increased risk of test positivity (Table 3). However, in contrast, the number of cases in pupils within the same year group was significantly associated with testing positive (OR 1.12, 95%CI 1.08-1.15).

Please also check when interpreting results of logistic regression, expressed as odds ratios, that results are interpreted as odds, rather than as ‘risk.’ Risk of a positive test is referred to several times in the results.

Response: We agree and have changed the manuscript throughout to reflect this when describing our results.

5) Limitations: The authors have performed an adjusted analysis and have highlighted some weaknesses in the available data including changes to the school environment.
What I don’t think it really mentioned within the interpretation or limitations is the likely scope for a lot of unmeasured confounding at the level of the individual rather than the school. In other words, behaviours and activities of individuals with or without exposures or with or without positive tests (e.g. physical distancing from others, frequency of day to day tasks such as shopping) and also length of any exposures, all of which are likely to impact on the infection transmission pathway but will not be captured within this data.
I suggest adding some discussion of this and any other measured or unmeasured factors which may influence the transmission pathway into the limitations section.

Response: Thank you for the suggestion. Indeed, it has not been possible to adjust for all behaviours and activities of individuals using the routine data available. We have acknowledged unmeasured possible confounding factors in the Study strengths and limitations section. To clarify some of the points you have raised we have added further text into the Strengths and Limitations section:

Measures to reduce transmission in the school environment, although advised at a national government level, will likely have varied subtly across schools in Wales dependent on setting, numbers of staff available and personal behaviours and activities of children, staff and parents (e.g. mask wearing, and congregating at school, opening and closing times, duration of exposures). We are unable to capture these variations in routine data which may explain some of the differences observed and we have also not examined new variants of SARS-CoV-2.

6) Table S3: I was unsure whether these variables were missing at the level of the school or the level of the individual? i.e. does number of staff within the school mean that for 842 schools, the number of staff was missing?

Response: Please note this is now Table S4. This is at the level of the individual. We have updated the caption to clarify this, which now reads:

Table S4 caption:

Table S4 – Number of individuals with missing variables at the individual-level before being confirmed eligible for the modelling cohort. Note this is not a count of distinct individuals, multiple persons may have multiple missing variables.

Also, please define the abbreviation RALF

Response: This has been added to Table S3.

VERSION 2 – REVIEW

REVIEWER	Reviewer name: Dr. Peter Flom Institution and Country: Peter Flom Consulting, New York, 10024, United States Competing interests: None
REVIEW RETURNED	07-Apr-2021

GENERAL COMMENTS	I confine my remarks to statistical aspects of this paper. These were well done and I recommend publication.

VERSION 2 – AUTHOR RESPONSE