Analysis

The Data

ClassAct is a pilot study aiming to describe air quality in schools and test its effect on the health of students. 31 primary schools were fitted with air quality monitoring devices, and were assigned to one of three study arms: 10 schools were assigned HEPA air filters, 10 were fitted with UVC air purifiers, and a control group of 11 schools without any air quality intervention. In this study, we aim to evaluate the effect of air quality interventions on illness-related absences, by comparing the attendance between the schools from the different study arms: following delays in the installation and operation of the UVC purifiers, this comparison will consider only the HEPA intervation as it compares to the control schools.

- talk about limiting study period to Sep-Apr window - something about covid regs??
- discuss schools switching from UVC to HEPA arm?
- Data on school demographics?

Attendance Reports

The outcome of interest in this study is illness-related absence reported by each of the participating schools. School attendance records are collected and reported in accordance with DfE guidance (REFERENCE) and raw attendance records were provided by the Information Management Team at Bradford Council covering the full study period. The data provided comprise anonymised attendance "strings" (lines of text) represented as individual characters that signify one of 30+ attendance types, per student for each of 14 weekly am/pm sessions Monday-Sunday. These codes were cleaned and aggregated into weekly counts of the various attendance types. Weeks in which schools were closed were removed from the data, as they contain no information about attendance, illness or otherwise. A minority of codes were blank or did not correspond to a recognised attendance type. Pending potential future data quality improvement efforts, school weeks with a 1% or greater proportion of "junk" sessions were removed from the dataset - 2.4% of the aggregated data after removing school closures.

The aggregated attendance records were then divided into one of three categories: codes signifying an illness-related absence, codes signifying in-school attendance, and remaining codes that signify absence from school for reasons other than illness (TABLE REF). The response/outcome variable was then defined as a ratio of the illness-related absence sessions to the in-school attendance sessions.

	Code	Description
In-school Attendance Codes:	/	AM: Present in school
	\	PM: Present in school
	L	Late arrival before the register has closed
	U	Arrived in school after registration closed
Illness Attendance Codes:	**I**	Illness (not medical or dental appointments)
	I01*	Illness - "non-coronavirus (COVID-19) related illness or sickness"
	I02*	Illness Confirmed case of coronavirus (COVID-19)
	X02*	Pupil self-isolating with coronavirus (COVID-19) symptoms

Figure 1: DfE attendance codes that signify a pupil's attendance in the school building or absence from school because of illness. The remaining attendance codes are not considered in this study.

***** Added to the pre-2020 attendance codes following the DfE's updates to the attendance guidelines following the Covid-19 pandemic

Though we would have liked to study covid absences more specifically, ambiguities in the DfE guidelines on

covid-related attendance records and the significant strain on resources resulting from the pandemic lead to numerous conflicting approaches to recording covid/non-covid absences.

Air Quality Data

An important mediator of the transmission of airborne illnesses is the ventilation of the classroom environment. Ventilation rates have a potential to confound the effect of the HEPA intervention in this study. As such, measurements of indoor CO^2 concentration, a means of inferring ventilation rates, were included in this analysis. Each of the participating Class-Act schools are fitted with air quality sensors, collecting timestamped recordings of CO^2 concentration (along with other IAQ measures) each minute. These data were cleaned to remove measurements above unreasonably high thresholds values and dates with greater than 25% missing/dropped recordings. The cleaned data were then aggregated to calculate a weekly mean CO^2 concentration over the period covered in this study for each of the participating HEPA/control schools.

Covid Rates

Rates of Covid-19 in the general population also have the potential to confound the effect of the HEPA intervention. The United Kindom government provide statistics on Covid-19 related metrics as a publicly available resource (REFERENCE). Data are available at varying granularities, from National at the highest level to Middle Layer Super Output Areas (MSOA) at the lowest level. For the purposes of this study, "new cases by specimen date" was selected as the metric to be used as a potential covariate. This metric is available at different levels of specificity. At the lower tier local authority level, daily 7-day rolling rates are available broken down into demographic age groups. At the MSOA-level weekly data are available for the total population for the 7-day period ending each Saturday. It was unclear which level of specificity would be more valuable in describing illness outcomes in schools, age or geographic location and, as such, both the local authority and MSOA levels of the "new cases by specimen date" were selected as covariates for analysis (REFERENCE DEFINITIONS). To ensure parity between the data recorded weekly, rolling rates were taken on the Saturday of each week for the data available daily.

Variable Name	newCasesBySpecimenDateAgeDemographics
Lowest Geographic Level	Lower Tier Local Authority
Frequency	Daily
Description	New cases by specimen date age demographics - Age breakdown of new cases. Data are shown by the date the sample was taken from the person being tested. Data are shown for rate per 100,000 people of the number of new cases in the rolling 7-day period ending on the dates shown .

Variable Name	newCasesBySpecimenDateRollingRate
Lowest Geographic Level	Medium Super Output Area (MSOA)
Frequency	Weekly
Description	New cases rolling rate by specimen date - Rate per 100,000 people of the number of new cases in the rolling 7-day period ending on the dates shown. Data are shown by the date the sample was taken from the person being tested.

Figure 2: Details of the UK Health Security Agency (UKHSA) Covid-19 statistics used in this study as described by the documentation at coronavirus.data.gov.uk

	Control (N=285)	**HEPA** (N=260)	**Overall** (N=545)
Illness Ratio:			
Mean (SD)	0.0569 (0.0304)	0.0459 (0.0392)	0.0517 (0.0353)
Median [Min, Max]	0.0507 [0.00332, 0.177]	0.0375 [0.00498, 0.398]	0.0436 [0.00332, 0.398]
Missing	4 (1.4%)	6 (2.3%)	10 (1.8%)
Mean CO^2 Concentration:			
Mean (SD)	1220 (277)	1270 (288)	1240 (284)
Median [Min, Max]	1230 [609, 2110]	1270 [651, 1880]	1250 [609, 2110]
Missing	31 (10.9%)	28 (10.8%)	59 (10.8%)
Covid Rate - MSOA:			
Mean (SD)	522 (531)	516 (485)	519 (509)
Median [Min, Max]	319 [72.3, 3100]	344 [49.7, 2460]	332 [49.7, 3100]
Missing	1 (0.4%)	1 (0.4%)	2 (0.4%)
Covid Rate - Age 4-14:			
Mean (SD)	712 (617)	712 (616)	712 (616)
Median [Min, Max]	417 [80.4, 2360]	432 [80.4, 2360]	417 [80.4, 2360]

Table 1: Statistics for the sample. "Illness Ratio" refers to the ratio of illness related absences to in-school attendance. The missing data from the "Illness ratio" statistics relate to the entries with >1% "junk" entries as discussed.

Methods

Basic descriptive analysis of the data were performed, visualising the distribution of illness ratios across time and subdivided by the different air quality interventions - this included mean estimates using basic linear regression and LOESS (Locally Estimated Scatterplot Smoothing) plots to identify the trend in illness rates over time across the different interventions.

The rate of illness-absences was then modeled using a multi-level gamma regression model. Parameters were estimated with a bayesian sampling methodology, using the R brms implelentaion of the no-U-turn (NUTS) sampling sampling algorithm. The various co-variate options were tested by fitting a basic model with only a HEPA parameter and random intercepts for each school:

$$illness.ratio_i \sim Gamma(log(\alpha_i), \beta_i)$$
$$\alpha_i = \mu + \mu_{school_i} + \beta_1 HEPA$$

Further models then included the various permutations of the co-variates, mean CO^2 concentration and the two Covid-19 rate metrics, including interraction effects for each of the covariates with the HEPA intervention, ending with a "saturated" model with parameters for each of the covariates and interactions:

$$illness.ratio_i \sim Gamma(log(\alpha_i), \beta_i)$$
$$\alpha_i = \mu + \mu_{school_i} + \beta_1 HEPA + \beta_2 Covid.Rate + \beta_3 Mean.CO^2$$
$$+\beta_4 HEPA : Mean.CO^2 + \beta_5 HEPA : Covid.Rate$$

Priors were assigned to each of the unspecified parameters in these models by the default algorithm used in the brms package - this assigns "uninformative" or "weakly informative" prior distributions, resulting in conservative posterior parameter estimates in the resulting fitted models (EXPLANATION IN SUPPLEMENTARY MATERIAL??).

- mention leave one out cross validation using loo package

Results & Discussion

Seasonal Variation of Parameters and Relationship With Illness Absences

Covid rates

The data on positive COVID-19 tests show distinct peaks that correspond with the waves of the Delta and Omicron variants in the UK during December/January, but the magnitude of these peaks differ depending on the statistics considered. MSOA-level rates of positive tests among all age groups show only a slight increase in positive tests as schools re-opened in October and a pronounced peak in January. The Bradford-wide data for the 4-14 year-old age group show similar rates of positive tests during the winter Omicron wave, but a much higher peak in October suggesting the increase in infections at this time was driven by the younger age groups.

The peak of infections in January during the Omicron wave is reflected in the attendance data from the schools participating in this study. However, the magnitude of this peak in illness-related absences in comparison to periods of lower infections is much lower than the difference seen in the positive tests data. The attendance data do not clearly correlate with the increase in infections among 4-14 year-olds in October; instead, a similar pattern of absences can be seen in each half of the autumn terms and the spring term after February, where illness rates begin low following a return to school and increase gradually as the weeks progress toward the next scheduled school break.

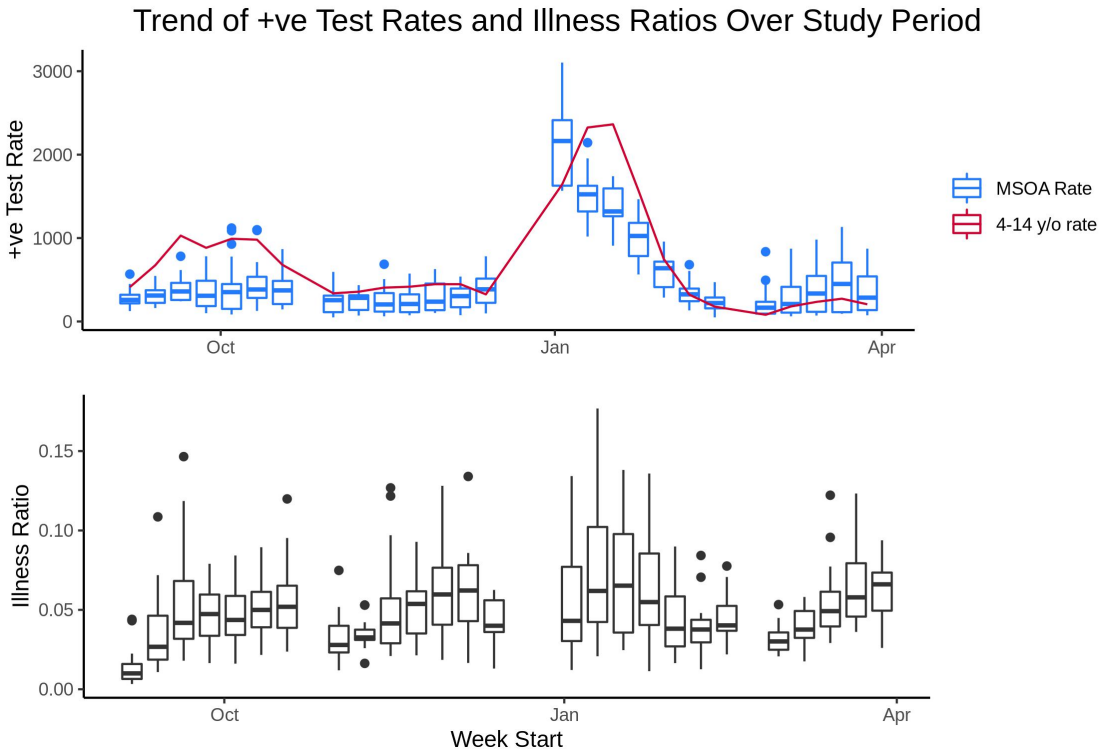


Figure 3: Top: Rates of positive tests over the study period for the MSOA-level data for all ages and the Bradford-wide data for 4-14 year olds. The MSOA-level data are represented as box plots, as there are multiple data points for each week. The Bradford-wide data is a single line, as there is only a single mesure per week.
Bottom: Box plots displaying the distribution of illness absence ratios between the participating schools over the study period. Outlying points above 0.17 have been omitted to make the trend in ratios easier to see.

A direct comparison of the positive test statistics with ratios of illness related absences demonstrates a small positive correlation, as would be expected: increases in positive tests in the general population correlate with an increase in illness-related absences. Dividing this comparison between the two arms of the study shows that this correlation differs between the HEPA and control schools - schools with no air quality interventions

show a stronger show a stronger positive slope with rates of covid in the general population. Both the Bradford-wide and MSOA-level data show a very similar relationship to illness-related absences.

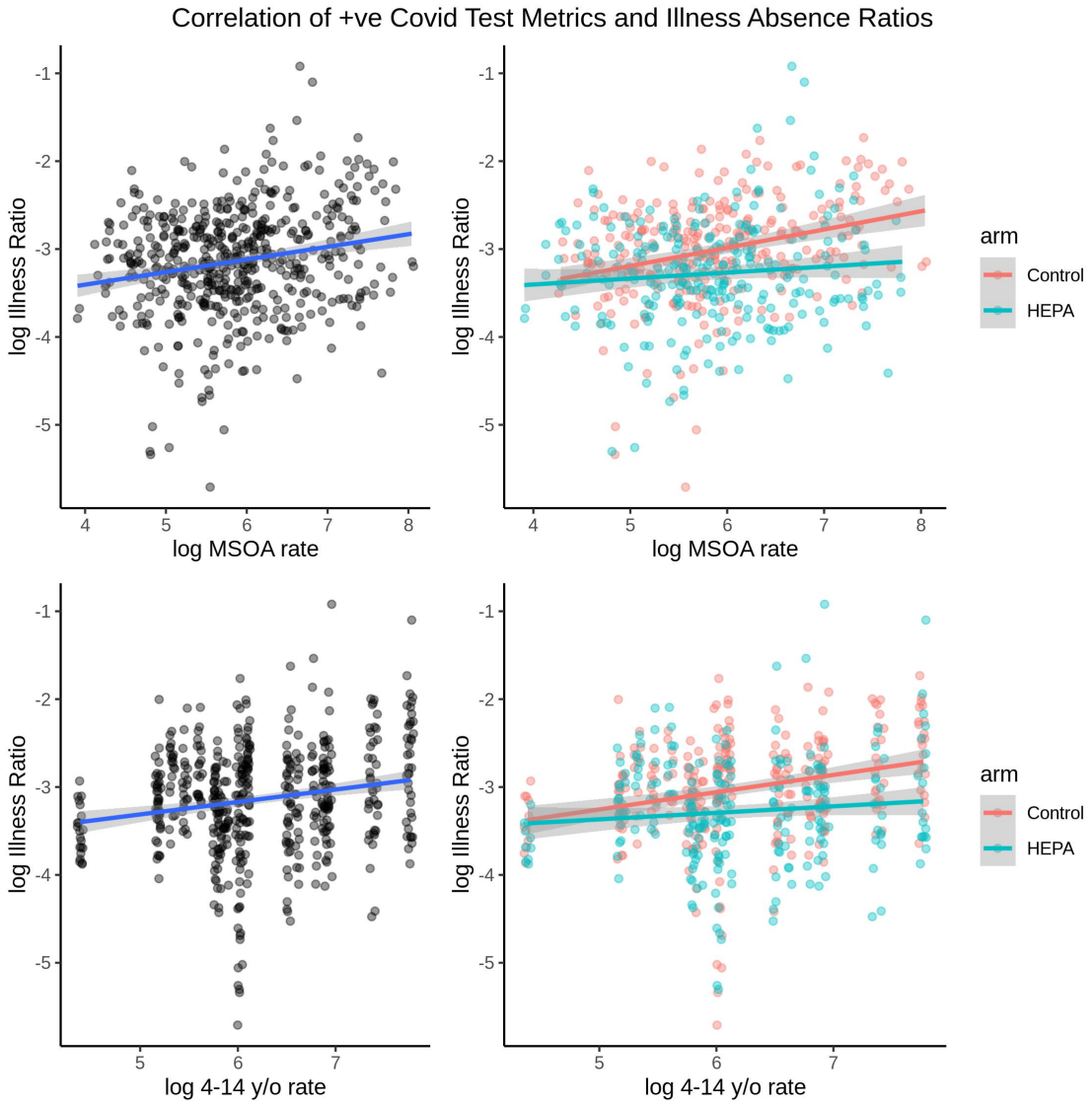


Figure 4: Scatter plots of illness ratios plotted against the two different covid prevelance statistics. Top: MSOA-level positive tests per 100,000 over a rolling 7 day period. Bottom: Bradford-level positive tests for 4-14 year old age group per 100,000 over a rolling 7 day period. Small "jitter" or random noise is added to the Bradford-level rates to aid visualisation. The plots on the left show the overall statistics with a simple regression line of best fit, on the right are the statistics broken down by the two study arms with simple regression fits. Covid prevelance is seen to have a smaller positive correlation with illness ratios among the HEPA schools when compared with the control schools. Log statistics are shown given the influence of outliers in both illness ratios and covid prevelances.

CO^2

Mean CO^2 concentration is seen to increase gradually from the schools re-opening in Autumn to Winter, and then begin to decrease again as the new year progresses. This pattern is as expected, as teachers are likely reducing ventilation (open windows/doors) during the winter months to maintain comfortable classroom temperatures as outdoor temperatures decrease. The general relationship between CO^2 levels and illness related absences appears negligible. However, a comparison of CO^2 levels and illness absences between the different study arms shows a positive correlation in the control group schools without any air quality

interventions - in schools without HEPA filtration, illness-related absences are seen to increase as CO^2 levels increase.

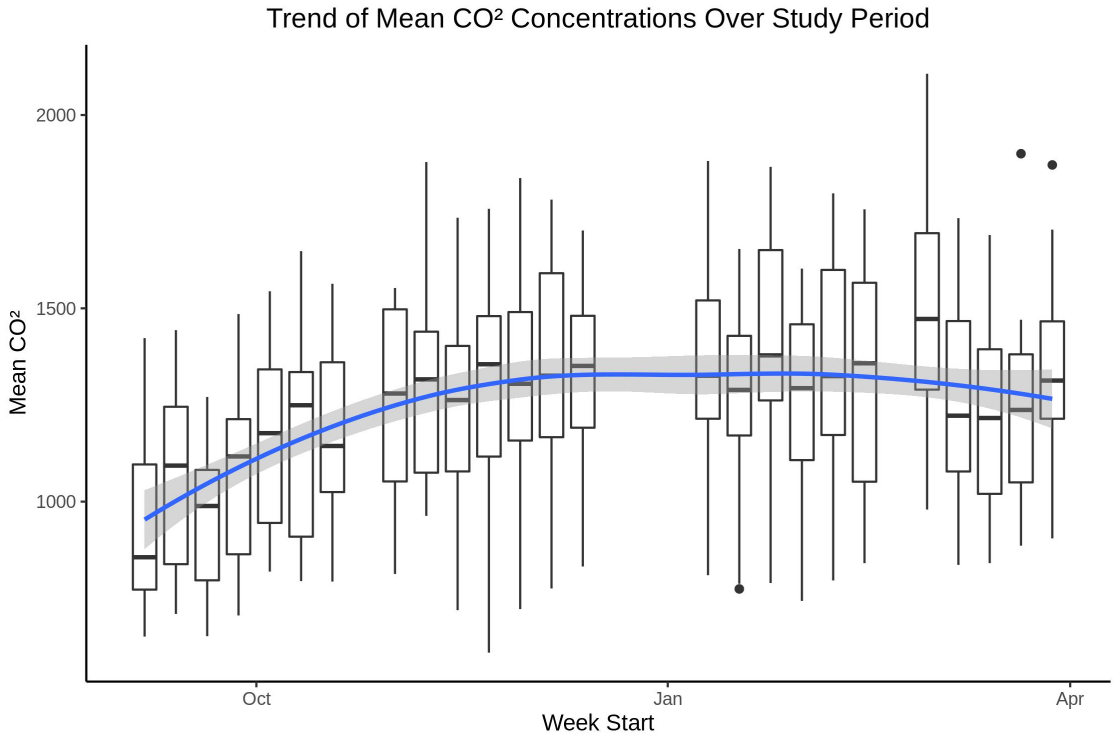


Figure 5: Box plots of mean CO^2 concentrations for each of the schools for each week of the study period. The line represents a local regression (LOESS) line of best fit, to estimate the trend in CO^2 concentrations over the duration of the study. CO^2 concentrations appear to increase gradually from September to January, and begin to decrease from March onward.

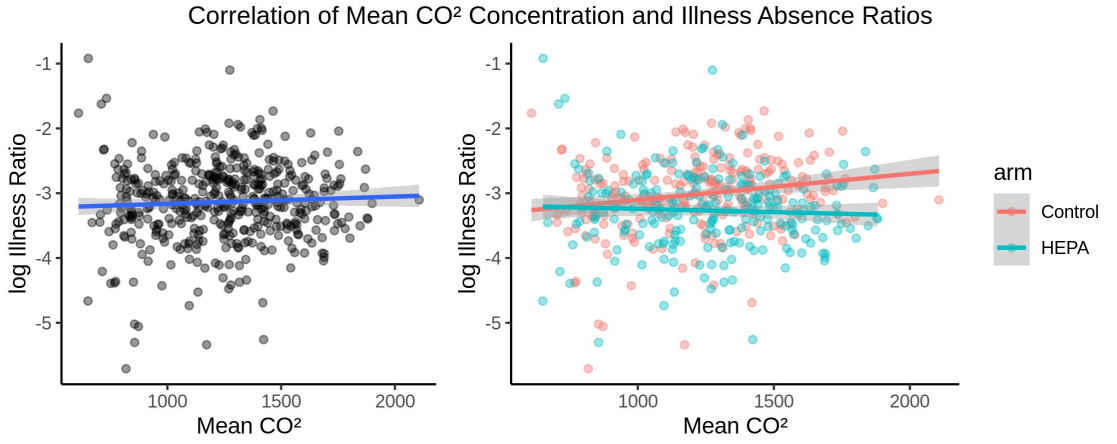


Figure 6: Scatter plots of illness ratios plotted against mean CO^2 concentrations. Left: overall statistics with a simple linear regression line of best fit. Right: statistics broken down by the two study arms with simple linear regression lines for each arm. CO^2 is seen to have a small negative correlation with illness ratios among the HEPA schools and a positive correlation in the control schools.

Study Arms

A direct comparison of illness ratios between the HEPA and control arms of the study shows a lower ratio of illness-related absences in the HEPA schools. The difference in illness patterns is seen over the winter period, with near identical ratios of illness absences at the beginning (September) and end (April) of the study period. There are also some extreme outlying values in October within the HEPA arm - it should be noted that these outliers relate to a particularly extreme outbreak in one school over a 3 week period. No other comparable outbreaks were seen in any of the other schools during the period of the study.

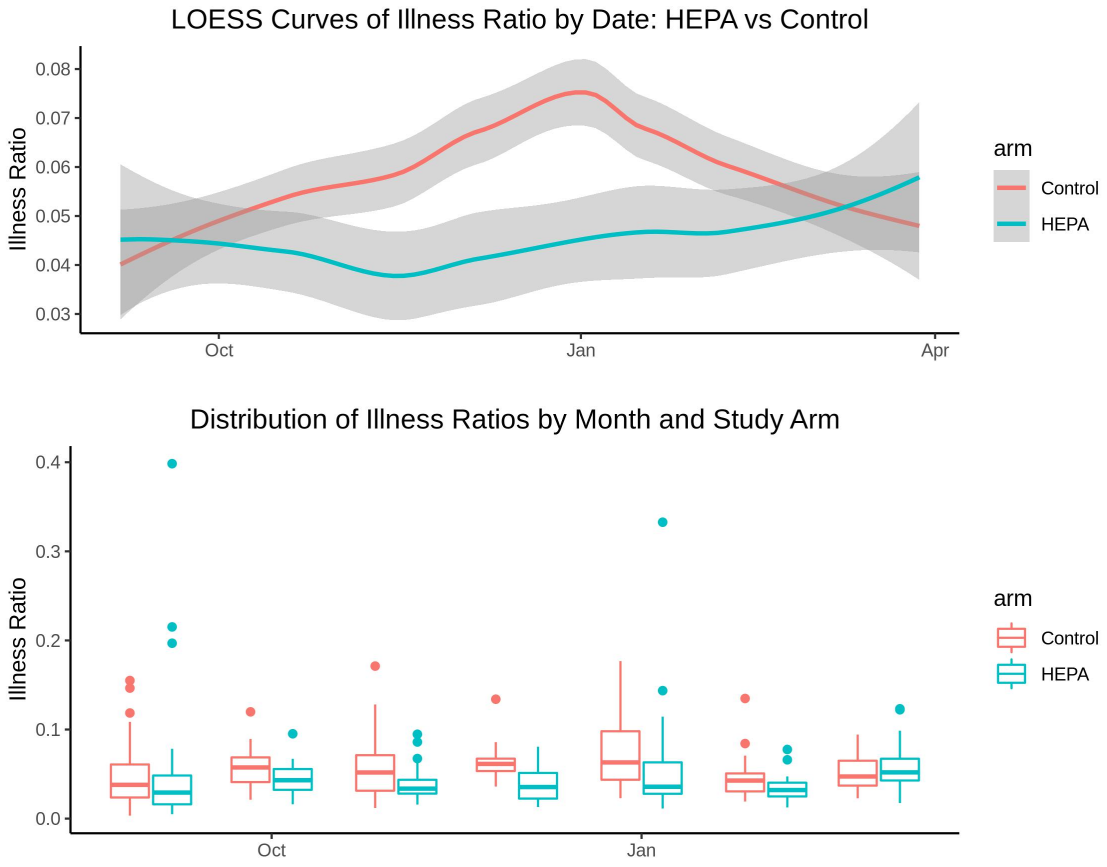


Figure 7: Trend of illness ratios over study period, broken down by study arm. Top: Local regression (LOESS) curves of illness ratio trend for HEPA and control schools. Bottom: Box plots of illness ratios grouped by month for HEPA and control Schools. Both plots show a relatively flat trend of illness ratios in HEPA schools, and a peak in illness ratios over the winter in control schools. Note also the significant outliers in September and January in the HEPA schools.

Regression Modelling

After considering all permutations of the predictors (mean CO^2, Bradford 4-14 years positive covid tests, MSOA-level positive covid tests) with the HEPA intervention, including interaction effects, the best fitting model featured parameters for the MSOA-level positive covid tests rate, mean CO^2 and an the interaction between mean CO^2 and HEPA filters:

$$y_i \sim Gamma(log(\alpha_i), \beta_i)$$
$$\alpha_i = \mu + \mu_{school_i} + \beta_1 HEPA + \beta_2 MSOA.Covid.Rate$$
$$+\beta_3 Mean.CO^2 + \beta_4 HEPA : Mean.CO^2$$

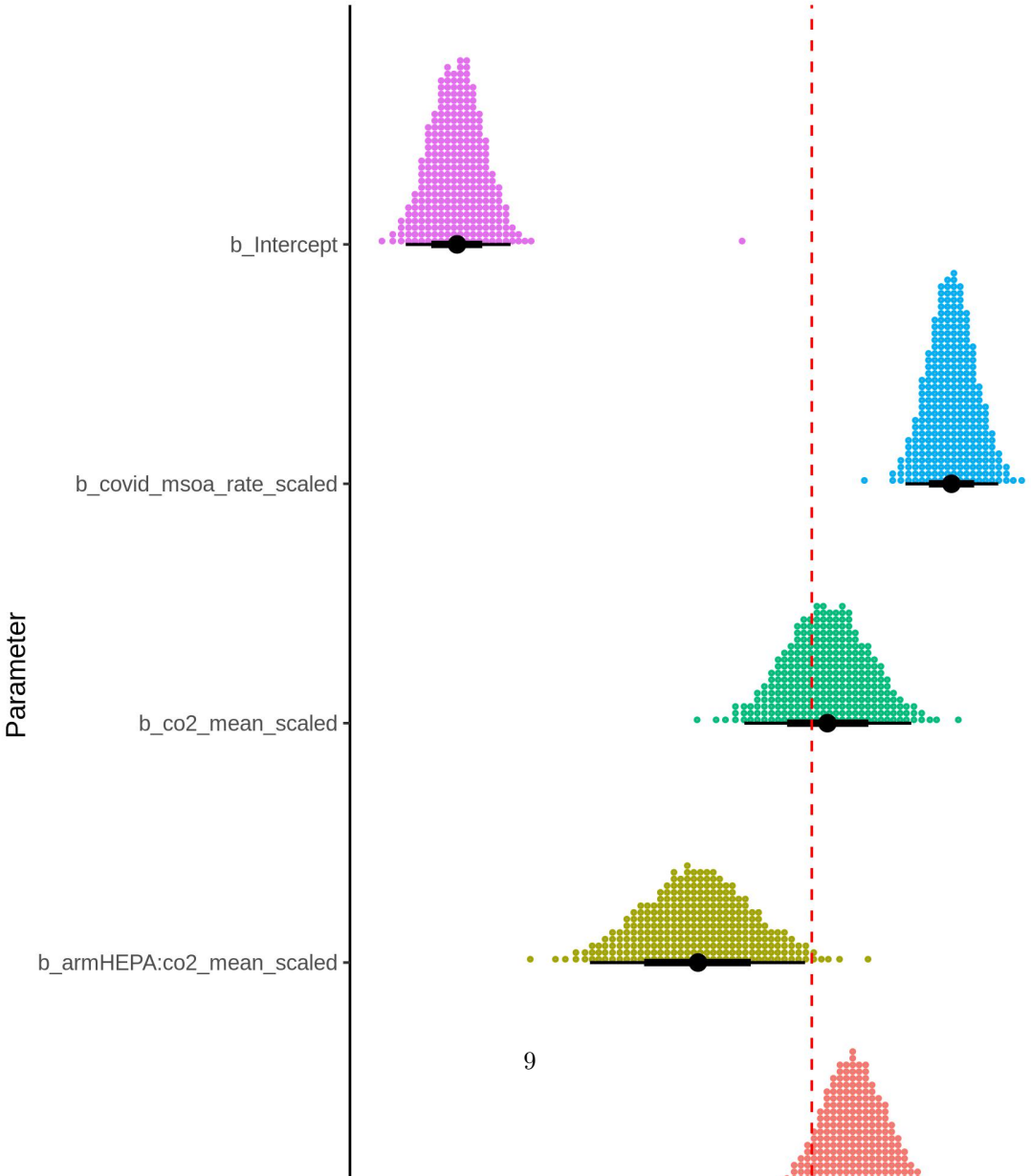
Model Parameters:					**CV Results:**				
CO^2:		**Covid Rate:**		**Rate Type:**	**ELPD**			**Params**	
Individ.	Interact.	Individ.	Interact.		Value	Diff	SE	Value	SE
-	**x**	**x**	-	MSOA	1119.73	0	25.97	26.27	4.59
x	-	**x**	-	MSOA	1119	-0.72	26.65	26.32	4.85
-	-	**x**	-	MSOA	1118.24	-1.49	27.25	24.92	4.86
x	-	-	**x**	MSOA	1118.22	-1.51	27.15	27.37	5.35
-	**x**	-	**x**	MSOA	1118.19	-1.53	26.46	28.03	5.02
-	-	**x**	-	Age	1117.73	-2	25.37	24.8	4.22
-	-	-	**x**	MSOA	1117.44	-2.29	27.69	26.03	5.17
x	-	**x**	-	Age	1117.25	-2.47	25.17	25.82	4.45
-	**x**	**x**	-	Age	1117.06	-2.66	24.65	26.58	4.3
-	-	-	**x**	Age	1116.49	-3.24	25.85	26.25	4.84
x	-	-	**x**	Age	1115.85	-3.88	25.66	27.53	5.1
-	**x**	-	**x**	Age	1115.55	-4.18	25.4	28.69	5.18
x	-	-	-	-	1101.57	-18.16	26.98	25.46	5.07
-	**x**	-	-	-	1101.51	-18.21	26.45	26.19	5

Table 2: Cross validation results for each of the models fitted in the analysis. The parameters used in each model are signified by the first 5 columns of the table: an **x** under "individ." signifies that only an individual parameter for that covariate was included in the model, an **x** under "interact." signifies both an individual and a HEPA interaction parameter for that covariate were included. Rate type signifies which of the covid prevelance statistics were used. ELPD is the leave-one-out (LOO) Expected Log Pointwise Predictive Density for each model - higher values indicate a higher probability that each datapoint could have been generated by the model proposed. Params is the difference between the leave-one-out ELPD and the non-cross-validated log psoterior predictive density, and can be though of as the effective number of parameters. Each model includes 21 group-level parameters as well as the specified population parameters, and so it appears that each has a full compliment of "effective" parameters.

Posterior parameter estimates from the best model show statistically significant effects for the MOSA-level Covid rate (1.24, [0.84, 1.65]) and the interaction parameter for HEPA filtration and CO^2 levels (-1.02, [-1.97,-0.04]); the remaining parameters are not predicted to have a statistically significant effect. Therefore, this model attributes the reduction in winter illness-related absences in the HEPA filtered schools to a contrasting of higher classroom CO^2 concentrations over this period - schools with HEPA filters see a reduction in illness related absences as classroom CO^2 levels increase.

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	BulkESS	TailESS
Group-Level Effects:							
School (Number of levels: 21)							
sd(Intercept)	0.27	0.06	0.18	0.41	1.00	1790	3048
Population-Level Effects:							
Intercept	-3.15	0.24	-3.60	-2.68	1.00	2561	3990
armHEPA	0.37	0.31	-0.23	1.00	1.00	2366	3491
covid msoa rate scaled	1.24	0.21	0.84	1.66	1.00	6788	5625
co2 mean scaled	0.14	0.37	-0.59	0.86	1.00	2718	4316
armHEPA:co2 mean scaled	-1.02	0.49	-1.97	-0.08	1.00	2449	3652
Family Specific Parameters:							
shape	4.08	0.26	3.59	4.60	1.00	7294	5833

Table 3: Summary of parameter statistics from the best performing model. "l-95% CI" and "u-95% CI" are the lower and upper bounds of the 95% credible interval respectively - these can be though of in similar terms to a confidence interval, and suggest a parameter has a statistically sinificant effect if the interval is uniformly negative or positive. The "Rhat" or Gellman-Rubin statistics are uniformly 1 suggesting good convergence and coverage of the sample space for each of the parameter samples, as do the bulk and tail effective sample sizes ("BulkESS" and "TailESS")



This predicted effect of HEPA filtration can be seen clearly from the marginal effect of CO^2 mean concentrations on posterior predictions of illness-ratios. The model reduces the predicted ratio of illness-related absences as CO^2 levels increase in HEPA schools, but predicts slightly increased rates of infection in schools without filters.

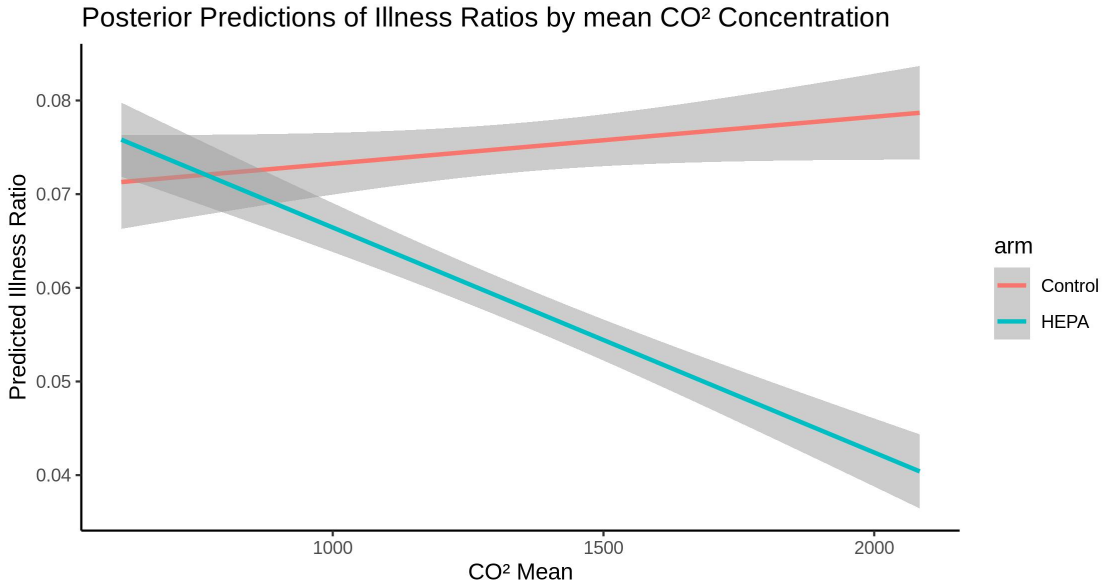


Figure 9: Mean posterior prediction estimates of illness ratios as mean CO^2 concentration is varied, between HEPA and control schools. The best model predicts that, as mean CO^2 concentration increases, the mean predicted illness ratio decreases in HEPA schools, and slightly increases in control schools. It should be noted that mean CO^2 concentration and covid prevelance are positively correlated in the data, and so higer CO^2 concentration typicaly coincides with a higher covid prevelance.

The overall marginal effect of HEPA filtration can be see in Figure XXXX. HEPA filtration is seen to reduce predicted illness-related absence ratios by 0.011 (95% HDI [-0.032, 0.01]) - that is approximately 20.1% reduction from the mean illness absence ratio of 0.052 across all schools. Posterior predictions of illness-related absence ratios in HEPA schools average 0.0446 (95% HDI [0.0366, 0.0530]) compared to an average of 0.0555 (95% HDI [0.0454, 0.0651]) in schools without HEPA filters.

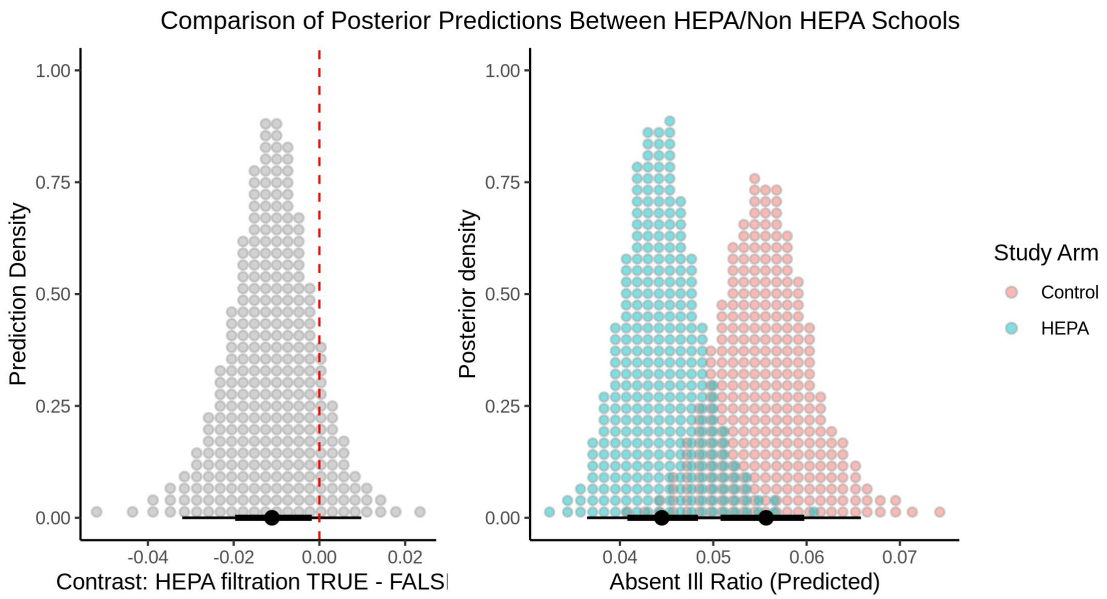


Figure 10: A comparison of posterior predicted illness ratios between HEPA and control schools. Left: density plot of the difference between posterior predictions of illness ratio across the entire distribution of MSOA-level covid rates and mean CO^2 concentrations when study arm is changed from control to HEPA. The vertical red dashed line represents no difference, or no effect of HEPA/control on predictions - points to the left of the line represent lower predicted rates with HEPA filtration, and to the right higher rates with HEPA filtration. Right: Posterior prediction density for all control and all HEPA predictions. The distribution of HEPA predictions is generally lower than predictions of schools without filters