Assessing the importance of demographic risk factors across two waves of SARS-CoV-2 using fine-scale case data

A.J. Wood[1], A.R. Sanchez[1], P.R. Bessell[1], and R.R. Kao*[1,2]

[1]Roslin Institute, University of Edinburgh
[2]Royal (Dick) School of Veterinary Studies, University of Edinburgh

March 31, 2023

Abstract

For the long term control of an infectious disease such as COVID-19, it is crucial to identify the most likely individuals to become infected and the role that differences in demographic characteristics play in the observed patterns of infection. As high-volume surveillance winds down, testing data from earlier periods are invaluable for studying risk factors for infection in detail. Observed changes in time during these periods may then inform how stable the pattern will be in the long term.

To this end we analyse the distribution of cases of COVID-19 across Scotland in 2021, where the location (geographical datazones of approx. 1000 residents) and reporting date of the case are known. We consider over 450,000 individually recorded cases, in two infection waves triggered by different lineages: B.1.1.529 ("Omicron") and B.1.617.2 ("Delta"). We use random forests, informed by measures of geography, demography, testing and vaccination. We show that the distributions are only adequately explained when considering multiple explanatory variables, implying that case heterogeneity arose from a combination of individual behaviour, immunity, and willingness to test.

Despite differences in virus lineage, time of year, and interventions in place, we find the risk factors remained broadly consistent between the two waves. Any observed differences could be explained by differences in the epidemiological conditions at the time, rather than any inherent differences in virus phenotype.

1 Introduction

A key challenge in the long term control of an infectious disease is to identify predictable patterns of incidence. The emergence and widespread circulation of the SARS-CoV-2 virus saw restrictions imposed globally on everyday life, in order to control the spread of COVID-19 infection, and to protect individuals at highest risk of severe disease. While as of March 2023 few to no restrictions remain in place, the virus continues to circulate at high volume. The winding down of community testing and other surveillance is making it more difficult to track the circulation of COVID-19 in detail, and in turn to understand what these longer-term patterns of incidence may be.

Throughout the COVID-19 pandemic, incidence has been highly heterogeneous [1, 2, 3]. However, as widespread community testing winds down globally, it is becoming increasingly unclear what the present profile of infection is, and whether it is evolving towards protecting or exposing more vulnerable individuals in the population. Typically, identifying risk factors for infection rely on disease surveillance studies. While these studies can be powerful and provide important insights [4, 5, 6, 7], they are often expensive, laborious and time consuming. "Big Data" in the health sciences offers an opportunity to gain some of the same insights using routinely collected

*Corresponding author: rowland.kao@ed.ac.uk

data. The availability of finely-grained data on COVID-19 cases allows us to investigate the utility of extremely dense spatio-temporal data to identify important health-related risks, using data that was collected over the course of the pandemic, and were made available to researchers in close to real-time.

This study uses case data collected in Scotland during 2021. This period is important in the wider context of the pandemic as it saw the steady relaxation of "lockdown"-like restrictions, to nearly all legally-enforced restrictions being lifted by the end of the year. A high-capacity testing programme was in place throughout, with free-of-charge lateral flow testing strongly encouraged, and PCR testing mandated for those with symptoms, or a lateral flow positive.

We compare two major waves of COVID-19 infection during this year, each with over 220,000 cases. The first wave from May 2021 was triggered by the B.1.617.2 lineage ("Delta"), when "lockdown"-like restrictions imposed in early 2021 were in the process of being lifted. The second wave from November 2021 was triggered by the B.1.1.529 lineage ("Omicron"). With different virus variants, time of year, and intervention severity, we have infection waves under two highly contrasted sets of conditions to compare.

To compare the two waves, we fit a machine learning model to the distribution of cases across all of Scotland. The case data are specified at high resolution, and allow us to associate to each case several explanatory variables. As well as overall population structure (sex, age, ethnicity), we inform our model with measures for vaccination coverage at the time, immunity from prior cases, rates of testing, and measures derived primarily from census data relating to levels of community deprivation. We show that the distributions of cases are only adequately explained when using a combination of all these variables. From the full model, we finally identify factors that were broadly indicative of a subpopulation having more or fewer cases in each wave. Our main finding is that the risk factors for cases remained broadly consistent across both waves. Differences between the two waves either offer relatively small scale changes in demographic risk or are consistent with the impact of changes in approaches to control.

2 Results

2.1 Time evolution and early patterns of spread

We identified 385,558 cases between November 15th 2021 and January 6th from the eDRIS data, of which 227,286 were S-gene dropout, therefore likely Omicron. From 1st May to 7th September 2021 we identified 269,838 cases, of which 229,073 were likely Delta (S-gene positive).

The start date for each of these periods is the first date from which there are consistent rises in cases that are likely the new variant. Per regression on $\log_{10}$(cases) (illustrated in Supplementary Material A.1, Fig. A.1), Omicron cases had a doubling time of 2.9 days over the first 28 days, compared to 6.2 days for Delta. Omicron also outpaced Delta by spatial spread, measured by new DZs having reported at least one case over time (also Fig. A.1). At the scale of Scotland's 32 local authorities (LAs), the time-dependent reproduction number R_t (Fig. A.2, methodology detailed in Supplementary Material A.2) consistently rises for Omicron, peaking at above 2 for nearly all LAs 28 days in to the outbreak, and only consistently falling below 1 after 50 days. Reproduction numbers for Delta are less consistent between LAs; while the number generally remains above 1 for most LAs in the period, there is no coherent peak at the start of the wave.

At the time of introduction of Omicron, the Delta variant was still circulating in high volume, though the subsequent spread of Omicron coincided with Delta declining. In the intermediate period with both variants circulating in high volume, the *age* distributions by variant differed. Taking the mid-points of the five-year age brackets, the mean ages of the Delta-type cases was 3.9 years lower than the Omicron-type cases (31.8 years compared to 35.7 years) (Student's $t = -52.2$, $p < 0.001$). This was the case from relatively early on when Omicron accounted for at least 5% of cases (Fig. A.3A). However, the median ages are equal (both 32.5 years), as in the Omicron-type cases there is a trough in those aged 0–14, with fewer than 50% of cases in this age group Omicron, but then a peak in the 20–29 age group (Fig. A.3B).

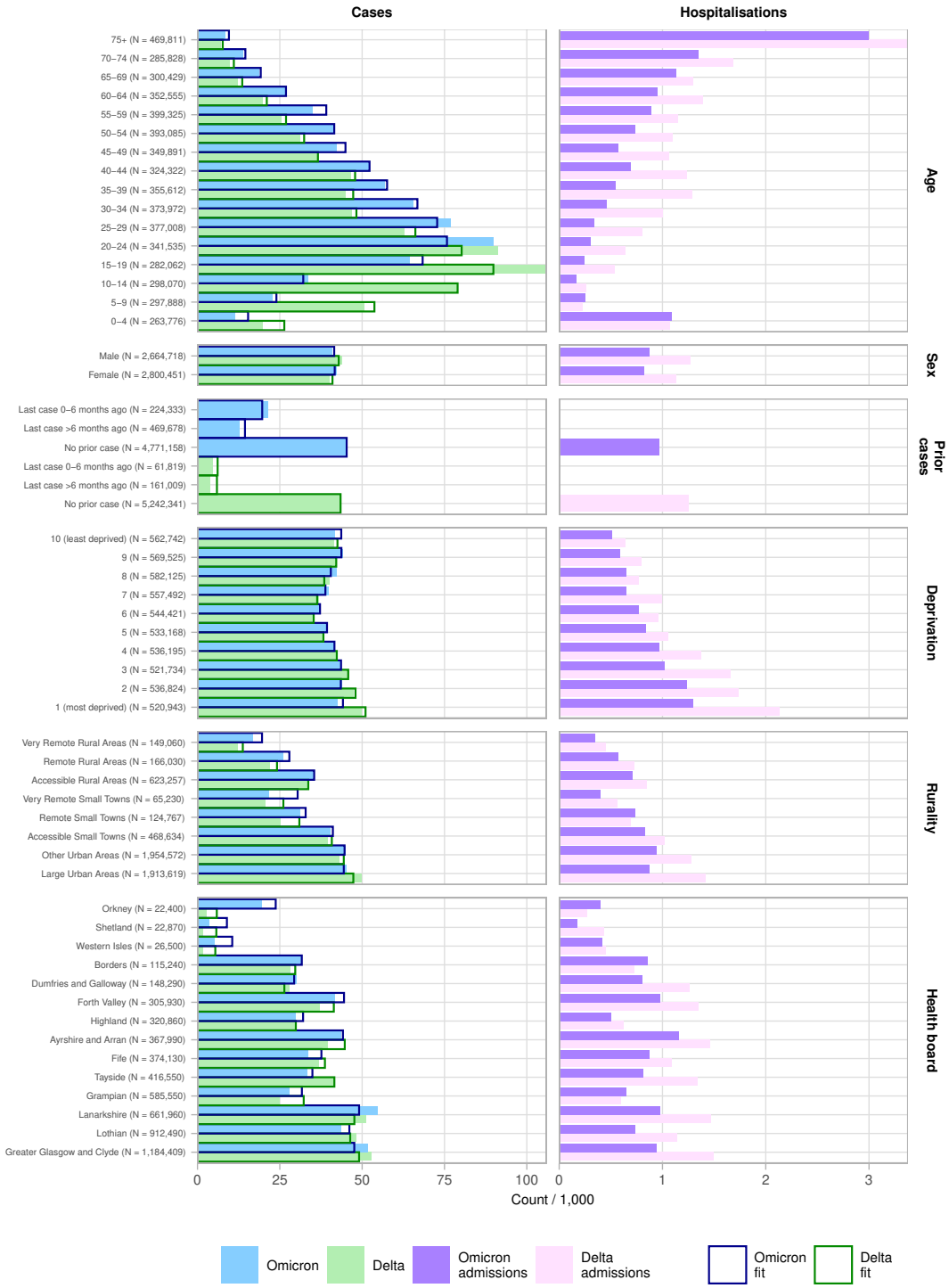

Figure 1: COVID-19 cases (left) and hospital admissions (right) by *age*, *prior case status*, *deprivation*, *rurality* and *location* (health board). Note that hospital admissions count all COVID-19 admissions in that period, and may be related to cases excluded from these analyses due to the lack of an S-gene result, or may not have an associated case at all.

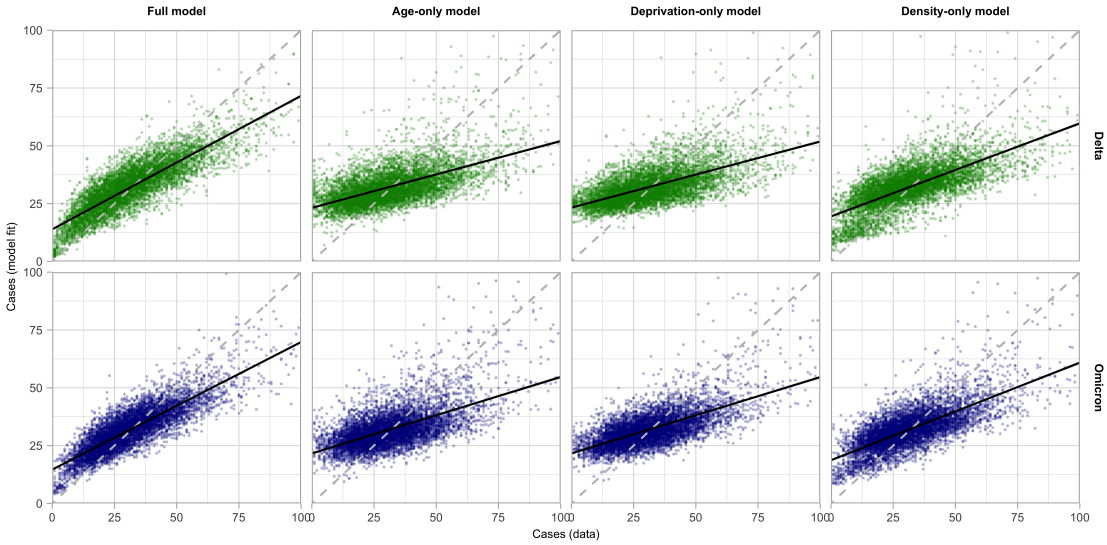

Figure 2: Performance of different models in terms of modelled cases (y-axis) in comparison with the observed cases (x-axis), for both Omicron and Delta. Each point represents a DZ. Also included are comparisons to models informed with only population, and one of either age, overall deprivation rank, or population density.

2.2 Detailed case distribution analysis

Fig. 1 shows the case distributions by age, sex, prior cases (serving as a proxy for prior immunity from infection), deprivation and health board. Omicron case rates were highest in younger adults, peaking at 90 cases/1,000 in ages 20–24. There was only a small difference in rates between men and women. Case rates were much lower amongst those that had tested positive for COVID-19 previously.

Fitting our models, Fig. 2 shows performance at DZ level, comparing observed cases to fit cases. Points deviating from the diagonal indicate DZs with less accurate fits. Our full model explains 70% (fit: 71%, test: 62%) of local variation in the case distribution (R-squared for case numbers, aggregated at a DZ level). We also simultaneously fit three "univariate" models, under equivalent hyperparameters to the full model, but informed with fewer variables. An equivalent model informed by population and population density alone explained 59% (fit: 60%, test: 55%) of variation. A model informed by only population/deprivation rank explained 53% (fit: 53%, test: 51%), and one informed by only population/age explained 48% (fit: 48%, test: 51%). Fig. 2 shows further deviation of the data-fit slopes away from the diagonal for these models. For the full model, variable importance outputs are presented in Supplementary Material C, with node purity and accuracy loss (Fig. C.6).

Returning to the data, we also see *spatial* variation. The majority of Scotland's population is densely concentrated in a narrow *central belt* (see Figure A.4), and case rates fall with increasing rurality (see Fig. A.5), most notably in Orkney, Shetland and the Western Isles (all island communities).

To probe spatial variation in cases *not* explained by the model, we measure the Moran's I autocorrelation [8] on the residuals, relating to their physical location (see also Fig. A.5). We compare local DZ-aggregated residuals over physical distances (from 1–100km), as well as network distance (number of nearest neighbours apart). Fig. 3 shows for the full model, autocorrelation of residuals within 1km (representative of correlation between close neighbours) is 0.35, falling to 0.15 at 5km (representative of correlation across distinct neighbourhoods within towns/cities), and 0.05 at 50km. The single-variable models exhibit much higher residual clustering, with the density-only model performing best, but persisting over larger distances. This is consistent with

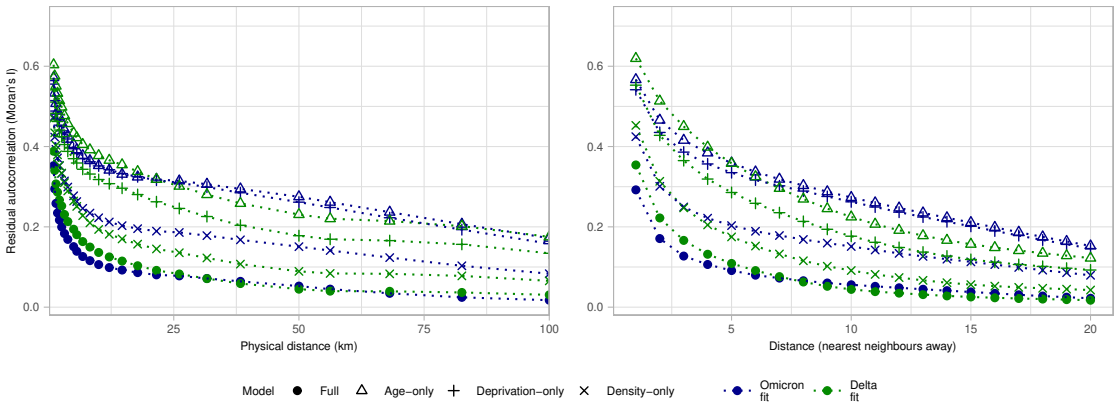

Figure 3: Residual clustering as measured by the Moran's I statistic, at different physical (left) and network-based distances (right). Higher values represent higher autocorrelation between residuals within a given locus. The point locations of each DZ are the population-weighted centroids. Nearest neighbours are defined as DZs that share a boundary.

the ability of those variables to explain spatial variation; while there is little clustering in age and deprivation beyond short neighbourhood-level distances, there is much stronger clustering in density (e.g., city centres).

The trend with respect to multiple deprivation rank is bimodal, with higher rates in both the most deprived, and least deprived communities.

Fig. 1 shows a good model fit to these larger-scale trends. It is important that the trends with deprivation, rurality and location have been captured, as the model was only informed by adjacent measures (e.g., for rurality, population density). We note that the model slightly under-fits the age ranges 15–24, where case rates were the highest overall. This under-fit is also seen in Fig. 2, where the full model is generally poorer at fitting cohorts with very high case counts. We discuss this later in Section 3.

2.2.1 Accumulated local effects

The accumulated local effects (ALEs) help identify *risk factors*; variable values that the model associates with more cases. For the range of explanatory variables in Fig. 4, the ALE is the average cohort-level "contribution" to the final fit value, over the full data set, given that variable value. If the value is greater than zero, the fit cases on average increases given that variable value.

Population has a trivial ALE, with the effect on the fit cases increasing near-linearly with cohort population. *Age*, *sex*, and *prior case status* are also known at the individual level and their ALEs in turn mirror the empirical distributions in Fig. 1. For example, very old and very young cohorts have negative ALEs, consistent with those age groups reporting fewer cases per person compared to other age groups. Conversely, Fig. 1 showed little variation by sex, and the corresponding ALE is also small.

Low *local S-gene coverage* (the proportion of cases with an accompanying S-gene result, needed to associate a likely variant) is protective. At the time S-gene coverage was 90% overall across mainland Scotland (per eDRIS data), but significantly lower in the LAs of Orkney Islands, Shetland Islands and Na h-Eileanan Siar (74%, 20% and 23% respectively). Increasing *local outbreak duration* (the time between January 6th and the first day an Omicron case was detected locally) is also indicative of higher cases.

The fit shows that the ALE for third/booster vaccination uptake is generally protective for increasing uptake. The protective value at *zero* uptake is most likely an artefact arising from differences in eligibility. During the wave, eligibility was not even across ages, with many younger groups either not yet eligible, or not eligible at all. Zero-uptake cohorts then mainly include ages

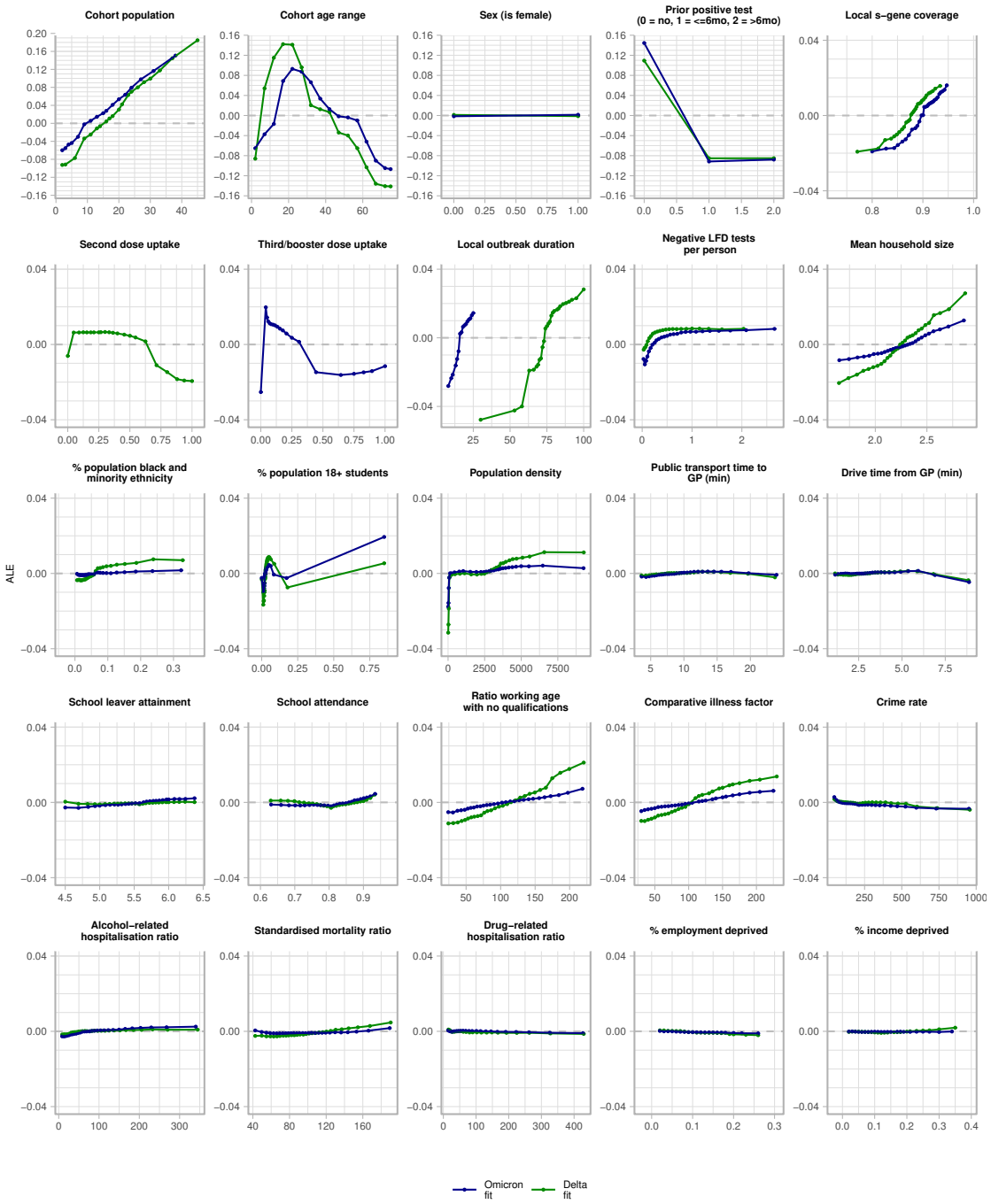

Figure 4: Accumulated local effects across all explanatory variables. For each variable, the x-axis represents the range of values of that variable in the data, and the y-axis (note scale differences for *population*, *age*, *sex* and *prior case status*) is the ALE for that variable value. The overall magnitude of the ALE represents the relative size of the effect.

0–14; individuals that reported fewer cases generally, but also may also have been protected for other reasons associated with young age.

The ALE for frequency of *negative* LFD test reporting shows a positive relationship, but exhibits a tailing-off with an increasing number of tests per person. Looking at the data explicitly, by dividing the whole population based on the number of LFD test results they reported, individuals reporting no LFD results had the lowest reporting rate of PCR positives overall (6.2%, 4.1% for Omicron and Delta respectively). Those that reported 1–2 LFD results had the highest PCR positive rate (15.1%, 10.5%), but individuals reporting *more* than 3 LFD results had a *lower* rate of reporting again (8.1%, 5.4%).

Indicators associated with deprivation appear less influential. However, modest deprivation-level variation in case rates (Fig. 1) is well captured, and deprivation indicators emerge as important in the model fit (Fig. C.6). There are correlations between different measures of deprivation, but these measures also correlate strongly with other variables in the data, such as age (with more deprived DZs skewing younger [9]), negative LFD reporting (with more deprived DZs generally reporting fewer tests [10]), and vaccine uptake (with more deprived DZs generally having lower uptake [11, 12]). It is plausible, then, that the fit is more strongly influenced with a combination of several correlated indicators of deprivation, rather than one in isolation.

Now comparing our results for the Omicron wave to earlier Delta cases from 1st May to 7th September 2021, the spatial distribution (Fig. A.5) is visually very similar to Omicron, with a concentration of high case rates in the central belt. Fig. 1 shows cases skewed slightly younger, with the highest rates within ages 15–19. Across deprivation deciles (after ranking the DZs by overall multiple deprivation [13]), as with Omicron, the distribution is bimodal, with higher rates in both the most and least deprived DZs.

The Delta model was fit using the same hyperparameters as Omicron (Supplementary Material B.1) and the same explanatory variables, with an exception that third/booster dose uptake was replaced with second dose uptake (third/booster doses were only administered later; see Supplementary Material B.3 for further details). Model performance was similar, explaining 72% (fit: 73%, test: 61%) of DZ-level variation. Fig. 4 shows the same broad risk factors for Omicron also emerged for Delta. Excluding non-eligible cohorts with zero uptake, higher second dose uptake emerged protective. A notable difference in Delta is that cohorts with higher black and minority ethnic populations were associated with higher case rates, more sharply than seen for Omicron. Conversely, communities with very high student populations were associated with high case rates for Delta, but less so compared to Omicron.

2.3 Comparison of distribution of cases to hospital admissions in the same period

Included alongside the case distribution in Fig. 1 is the distribution of COVID-19 related hospital admissions within the same periods. As our case distributions only take a subset of cases in the period (excluding those with either a likely different, or unknown variant), these should not be used to infer a raw case-to-hospitalisation rate, as not all cases in the period are accounted for.

For both periods, admissions were highly skewed towards older ages, consistent with earlier stages of the epidemic [14, 15]. By associating hospital admissions and particular COVID-19 cases with a matching DZ, age range and sex, up to 21 days prior to the admission, we find nearly all admissions had no associated case in the past – that is, very few admissions were from individuals reporting their second (or more) separate case of COVID-19.

Admissions in both periods had a strong deprivation skew; for Omicron, the lower half of DZs ranked by multiple deprivation reported 63% of admissions, and for Delta, the proportion was 64%. This pattern is not seen in cases, which we discuss further in Section 3.

3 Discussion

Scotland's programme of free community testing was an invaluable tool for tracking the spread of COVID-19 infection up to early 2022. With the winding down of detailed surveillance since, the testing data gathered in this period may prove a unique source for studying the distribution of COVID-19 cases in detail. As COVID-19 continues to circulate in high volume as of March 2023, it is increasingly uncertain who amongst the population is at the highest risk of infection, and whether that will change over time.

The aim of this study was to use data gathered in 2021 to compare how the profile of cases changed across two waves of COVID-19 under highly contrasted conditions. We analysed the distribution of cases during the B.1.617.2 "Delta" wave from May 2021, and the B.1.1.529 "Omicron" wave from November 2021. Our statistical model shows that variations in case counts in these waves of infection were associated with broad factors such as age structure and residual immunity from earlier cases, but also with factors relating to testing, vaccination, geography and demographics. However despite substantial differences in the severity of interventions in place, time of year, vaccination uptake and virus phenotype, we find the risk factors across both waves to be broadly similar. Any differences as were observed are as likely to be explained by the differences in imposed interventions as they are due to differences in the virus strains. This suggests that the profile of infection risk was not highly sensitive to demographic factors, and that similar trends may persist in future endemic-like circulation.

At the individual level, the drivers of infection are close interactions, and variation in individual susceptibility and virus transmissibility. Any existing immunity then affects the probability of infection, if exposed. Finally, the propensity of individuals to test, and report the results of those tests, affects whether an infection is identified and registered. Evidence from random community testing suggests that only a minority of infected persons did so even at the peak of testing capacity [16, 17]. The variables in our model serve as proxies for these various drivers. While it is beyond the scope of this work to fully disentangle the relative impacts of all these factors, our analyses provide some important insights.

Living in a deprived community is a known risk factor for poor COVID-19 outcomes [18, 19, 20], however the data only show modest variation in case volume. Rather, a clear deprivation-level inequality is only seen in admissions. This may arise from a higher risk of hospitalisation given *infection* with higher deprivation, compounded by potential variation in case *ascertainment* rates. With the free availability of LFD tests at the time, the reporting of negative tests may then relate to propensity for testing overall, and higher reporting of mild infections that do not result in hospital admission. Rates of LFD reporting were consistently lower in more deprived areas, and LFD positivity higher (e.g., [10], and also observed in a wider study in Liverpool, England [21]).

The influence of some variables in the models reflect known risk factors. For example, high mean household sizes are indicative of more cases, consistent with evidence of high secondary attack rates for SARS-CoV-2 [22, 23], and increased risk of inter-household transmission relative to other close contacts outside of the home [24]. Higher vaccine uptake (amongst those eligible) is also protective, to a stronger degree with Delta, consistent with higher rates of immune breakthrough seen with the Omicron variant as compared to Delta [25, 26, 27]. We do not know the specific vaccination status of those reporting cases, however, and linked data may have revealed an even stronger protective effect.

Some differences do emerge between the two waves; first, the model shows a stronger positive association for Delta, between cases and the proportion of individuals of black and minority ethnicity. In the UK, belonging to a black or minority ethnicity has been highlighted as a risk factor for poor COVID-19 outcomes overall [28, 29, 30] but without detailed, linked data, it is difficult to firmly establish drivers for a *heightened* risk during the Delta wave. Differences may emerge from known variations in vaccination uptake [31] and occupation [32] (thus ability to work from home or effectively physical distance), and the relative impacts of those factors changing across the two waves. The Delta wave also sees high population density as a stronger risk factor. This could be explained by isolated outbreaks in lower-density communities being less likely to be sustained for Delta (having a lower growth rate), as well as a reduction in imported transmissions

for Delta (with stricter travel restrictions in place).

Our model is by design not informed with data on position, or nearest neighbours. It is highly likely however that local case rates are influenced by surrounding communities, and communities connected by other means (e.g., transport links). We accounted for this in part with the *local outbreak duration* parameter, describing when cases first appeared locally, but the model may have been further improved with an explicit spatial element. Residual autocorrelations persist locally, but fall significantly past 10km. The fit cases are also time-aggregated, and therefore do not account for changes in factors during each wave. Finally, model performance is poorer in areas where cases were exceptionally high (as per the regression in Fig. 2). This may have been better captured if cohorts were instead sampled with probabilities weighted by, say, overall cases.

The COVID-19 data collated by eDRIS are remarkable in terms of volume and resolution, and has allowed us to assess a national-level epidemic at extremely fine scale. However, regardless of resolution, cases only partially represent the full underlying pattern of infection. Variations in testing frequency and differing trends in hospital admissions suggest the distribution of infections may have been very different to that of reported cases. By incorporating trends on cases, testing behaviour, and data on severe outcomes more closely linked to infection (hospitalisation, ICU admission and mortality), it may be possible to build a much more comprehensive retrospective picture of how infections were distributed amongst the population.

Importantly, while our access to such finely-grained data was exceptional, it can be expected that such data are likely to become more common in the future, and may become available in real time. As such, our demonstration of the utility of such data points the way to an important approach to improving data analysis supporting control policy response to infectious disease emergencies in the future.

4 Data and methods

4.1 Data

COVID-19 testing data are obtained from Public Health Scotland's *electronic Data Research and Innovation Service* (eDRIS) system. The data include individual tests by type (polymerase chain reaction (PCR) or rapid lateral flow device (LFD)), test result (positive, negative, void, inconclusive), test date, S-gene test result if known, age, sex, and residing data zone (*DZ*, a census area typically comprising 500–1,000 individuals). De-identified IDs link repeat tests by the same individual. We reduce the raw test data to cases by removing duplicate tests by the same individual within 60 days (taking the date of the first PCR positive as the case date, or the first LFD in the absence of any PCR). These data — in particular the DZ, specifying location to within an area as small as 0.1km^2 in densely populated areas — allow us to study cases on a fine spatio-temporal scale. Data on vaccine administrations, and hospital admissions citing COVID-19 are also provided by eDRIS. We use the version of the eDRIS data from July 14th 2022.

The period 15th November 2021 – 6th January 2022 covers the first outbreak and peak of the B.1.1.529 lineage (BA.1 sublineage, hereafter referred to as the Omicron variant). Beforehand, the B.1.617.2 lineage (Delta variant) (S-gene positive) was dominant. From 15th November 2021, S-gene dropout cases consistently rise, and all subsequent "dropout" cases are assumed Omicron. Remaining S-gene positive cases are presumed to be Delta, consistent with broader nationwide sequence data [33]. We note that the Omicron sub-variant BA.2/B.1.1.529.2 later replaced BA.1/B.1.1.529.1, becoming dominant in Scotland from around 25th February 2022. This variant, like Delta, has an S-gene positive test signature. However by the end of the period studied the BA.2 variant was only being identified in fewer than 1% of fully sequenced cases in the UK [34], and here we assume all remaining S-gene positive cases to be Delta.

Prior to January 6th 2022 in Scotland, positive LFD tests (typically taken at home) required PCR confirmation. Approximately 90% of cases in this period have a definitive S-gene result. A policy change then dropped this PCR requirement [35], after which cases with S-gene results fell to about 50% by February 2022 (per eDRIS data).

4.2 Random forest methods and choice of explanatory variables

The volume and distribution of COVID-19 cases in these waves are a result of individual interactions, immunity, and willingness to test. Taking advantage of the resolution of the data, our aim is to incorporate a large number of variables to explain variation in these distributions. These variables are not direct causative factors for infection spread, but serve to help us better understand drivers for case heterogeneity. Having a large number of variables, we fit the distribution using random forest regression. The key advantage of using a random forest regression in this context is that it allows us to fit the distribution without specifying any explicit prior relation between cases and the large number of variables involved, which may themselves be correlated.

We fit the time-aggregated case distribution in *R*, using the *randomForest* package [36, 37]. We choose model hyperparameters (in particular, the maximum number of terminal nodes per tree) so as to optimise model performance on a sample of the data it was explicitly *not* informed by in the model fit. This reduces overall model accuracy, but negates over-fitting to the data. These hyperparameters are detailed in Supplementary Material B.1.

We group individuals with a given age range, sex, DZ and when (if at all) they last tested positive, with each of these *cohorts* usually comprising 0–50 individuals. For each cohort, we count the number of associated cases from the eDRIS data. For Omicron, we seek S-gene dropout cases between 15th November 2021 and 6th January 2022, and for the earlier Delta outbreak, S-gene positive cases between 1st May and 7th September 2021 (choosing this limit so as to have a similar number of cases in each set). We then fit the case distribution using a set of explanatory variables. These variables fall in to demographic factors (cohort population, age, sex, ethnicity, student population), COVID-19 related factors (testing volume, prior case history to serve as a proxy for existing infection-based immunity, vaccination uptake and the time of first detection of the new variant locally), geography (local population density and transport time to public services to serve as proxies for connectivity and geographic remoteness), as well as deprivation. Data on deprivation are taken from the *Scottish Indices of Multiple Deprivation* (SIMD). The SIMD ranks DZs in Scotland by "multiple" deprivation, incorporating measures relating to local health, housing, geographic access, employment, income, crime, and education. In our model we use the raw measures of deprivation as explanatory variables. A full detailed description of all variables used is given in Supplementary Material B.2.

We take two metrics for variable importance from the *RandomForest* function output: the node purity (a measure of how effective variables are at partitioning cohorts with differing numbers of cases in the tree), and the loss of model accuracy on random permutation of a given variable's values. We also calculate the *accumulated local effects* (ALEs) of each variable. ALEs describe how the model fit value changes, in response to changing one variable value in isolation, averaged over many different entries in the data [38]. In this context, they attribute a "directionality" to the importance of each variable, and indicate whether a variable value is associated with fewer or more cases in general.

5 Acknowledgements

We thank Public Health Scotland's *electronic Data Research and Innovation Service* (eDRIS) for the provision of COVID-19 testing, vaccination and severe outcomes data.

6 Author contributions

R.R.K. conceived the project. A.J.W. wrote the model code and performed the spatial analysis. A.R.S. and P.R.B. performed the temporal analysis. A.J.W. and R.R.K. wrote the manuscript.

7 Competing interests

The authors declare no competing interests.

8 Code availability

Analysis code is available at
https://git.ecdf.ed.ac.uk/awood310/scotland-covid-case-distribution-random-forest-model.

9 Data availability

The COVID-19 testing, vaccination and hospitalisation data utilised in this work are not publicly available. They are provided to the authors for academic research by Public Health Scotland's *electronic Data Research and Innovation Service*, under a data sharing agreement (*Spatial and Network Analysis of SARS-CoV-2 Sequences to Inform COVID-19 Control in Scotland*).

Data relating to deprivation are derived from the 2020 Scottish Index of Multiple Deprivation, and are publicly available [13]. Fine-scale population statistics are drawn from publicly-available population estimates as of mid-2020 [39]. The population of students and those belonging to a minority ethnicity are drawn from public Scottish census data (tables KS501SC, LC2101SC respectively).

10 Funding statement

This work has been funded by the ESRC grant ES/W001489/1: *Real-time monitoring and predictive modelling of the impact of human behaviour and vaccine characteristics on COVID-19 vaccination in Scotland*.

References

[1] Lee D, Robertson C, Marques D. Quantifying the small-area spatio-temporal dynamics of the Covid-19 pandemic in Scotland during a period with limited testing capacity. Spatial Statistics. 2021:100508.

[2] Banks CJ, Colman E, Doherty T, Tearne O, Arnold M, Atkins KE, et al. SCoVMod–a spatially explicit mobility and deprivation adjusted model of first wave COVID-19 transmission dynamics. Wellcome Open Research. 2022;7(161):161.

[3] Lone NI, McPeake J, Stewart NI, Blayney MC, Seem RC, Donaldson L, et al. Influence of socioeconomic deprivation on interventions and outcomes for patients admitted with COVID-19 to critical care units in Scotland: a national cohort study. The Lancet Regional Health-Europe. 2021;1:100005.

[4] Simpson CR, Robertson C, Vasileiou E, McMenamin J, Gunson R, Ritchie LD, et al. Early pandemic evaluation and enhanced surveillance of COVID-19 (EAVE II): protocol for an observational study using linked Scottish national data. BMJ open. 2020;10(6):e039097.

[5] Sheikh A, Kerr S, Woolhouse M, McMenamin J, Robertson C. Severity of Omicron variant of concern and vaccine effectiveness against symptomatic disease: national cohort with nested test negative design study in Scotland. 2021.

[6] Canas LS, Sudre CH, Pujol JC, Polidori L, Murray B, Molteni E, et al. Early detection of COVID-19 in the UK using self-reported symptoms: a large-scale, prospective, epidemiological surveillance study. The Lancet Digital Health. 2021;3(9):e587-98.

[7] Antonelli M, Penfold RS, Merino J, Sudre CH, Molteni E, Berry S, et al. Risk factors and disease profile of post-vaccination SARS-CoV-2 infection in UK users of the COVID Symptom Study app: a prospective, community-based, nested, case-control study. The Lancet Infectious Diseases. 2022;22(1):43-55.

[8] Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1/2):17-23.

[9] Mid-2021 Small Area Population Estimates, Scotland (Report);.

[10] Coronavirus (COVID-19): modelling the epidemic (issue no.93);. The Scottish Government. url: https://www.gov.scot/publications/coronavirus-covid-19-modelling-epidemic-issue-no-93/documents/.

[11] Coronavirus (COVID-19) Infection Survey technical article: Analysis of characteristics associated with vaccination uptake;. "Office for National Statistics. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticleanalysisofcharacteristicsassociatedwit 2021-11-15.

[12] Wood AJ, MacKintosh AM, Stead M, Kao RR. Predicting future spatial patterns in COVID-19 booster vaccine uptake. medRxiv. 2022:2022-08.

[13] Scottish Index of Multiple Deprivation 2020;. The Scottish Government. url: https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-indicator-data/.

[14] COVID-19 statistical report, as at 15 November 2021;. Public Health Scotland. url: https://www.publichealthscotland.scot/publications/covid-19-statistical-report/covid-19-statistical-report-17-november-2021/.

[15] Coronavirus (COVID-19) weekly insights: latest health indicators in England, 12 February 2021;. Office for National Statistics. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19weeklyinsights/latesthealthindicatorsinengland12february2021.

[16] Colman E, Puspitarani GA, Enright J, Kao RR. Ascertainment rate of SARS-CoV-2 infections from healthcare and community testing in the UK. Journal of Theoretical Biology. 2022:111333. Available from: https://www.sciencedirect.com/science/article/pii/S0022519322003241.

[17] Nightingale ES, Abbott S, Russell TW. The local burden of disease during the first wave of the COVID-19 epidemic in England: estimation using different data sources from changing surveillance practices. BMC public health. 2022;22(1):1-14.

[18] Blundell R, Costa Dias M, Joyce R, Xu X. COVID-19 and Inequalities. Fiscal studies. 2020;41(2):291-319.

[19] Bambra C, Riordan R, Ford J, Matthews F. The COVID-19 pandemic and health inequalities. J Epidemiol Community Health. 2020;74(11):964-8.

[20] Baena-Díez JM, Barroso M, Cordeiro-Coelho SI, Díaz JL, Grau M. Impact of COVID-19 outbreak by income: hitting hardest the most deprived. Journal of Public Health. 2020;42(4):698-703.

[21] Green MA, García-Fiñana M, Barr B, Burnside G, Cheyne CP, Hughes D, et al. Evaluating social and spatial inequalities of large scale rapid lateral flow SARS-CoV-2 antigen testing in COVID-19 management: An observational study of Liverpool, UK (November 2020 to January 2021). The Lancet Regional Health-Europe. 2021;6:100107.

[22] Jalali N, Brustad HK, Frigessi A, MacDonald EA, Meijerink H, Feruglio SL, et al. Increased household transmission and immune escape of the SARS-CoV-2 Omicron variant compared to the Delta variant: evidence from Norwegian contact tracing and vaccination data. medRxiv. 2022.

[23] Fonager J, Bennedbæk M, Bager P, Wohlfahrt J, Ellegaard KM, Ingham AC, et al. Molecular epidemiology of the SARS-CoV-2 variant Omicron BA. 2 sub-lineage in Denmark, 29 November 2021 to 2 January 2022. Eurosurveillance. 2022;27(10):2200181.

[24] Dupraz J, Butty A, Duperrex O, Estoppey S, Faivre V, Thabard J, et al. Prevalence of SARS-CoV-2 in household members and other close contacts of COVID-19 cases: a serologic study in canton of Vaud, Switzerland. In: Open forum infectious diseases. vol. 8. Oxford University Press US; 2021. p. ofab149.

[25] Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 vaccine effectiveness against the omicron (B. 1.1. 529) variant. New England Journal of Medicine. 2022.

[26] Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature. 2022;602(7898):654-6.

[27] Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. The Lancet. 2021;397(10285):1646-57.

[28] Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 8 December 2020 to 1 December 2021;. Office for National Statistics. url: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/updatingethniccontrastsindeathsinvolvingthecoronaviruscovid19englandandwales/8december2020to1december2021.

[29] Platt L, Warwick R. Are some ethnic groups more vulnerable to COVID-19 than others. Institute for fiscal studies. 2020;1(05):2020.

[30] Lo CH, Nguyen LH, Drew DA, Warner ET, Joshi AD, Graham MS, et al. Race, ethnicity, community-level socioeconomic factors, and risk of COVID-19 in the United States and the United Kingdom. EClinicalMedicine. 2021;38.

[31] for National Statistics O. Coronavirus and vaccination rates in people aged 18 years and over by socio-demographic characteristic and occupation, England: 8 December 2020 to 31 December 2021;. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/coronavirusandvaccinationratesinpeopleaged18yearsandoverbysociodemographiccharacterist 8december2020to31december2021.

[32] Census 2011: Release 3I - Detailed characteristics on Labour Market and Education in Scotland;. National Records of Scotland. url: https://www.nrscotland.gov.uk/news/2014/census-2011-release-3i.

[33] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.

[34] SARS-CoV-2 variants of concern and variants under investigation in England: technical briefing 35;. The UK Health Security Agency. url: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1050999/Technical-Briefing-35-28January2022.pdf.

[35] Self-Isolation and testing changes;. The Scottish Government. url: https://www.gov.scot/news/self-isolation-and-testing-changes/.

[36] Liaw A, Wiener M, et al. Classification and regression by randomForest. R news. 2002;2(3):18-22.

[37] Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

[38] Apley D, Apley MD. Package 'ALEPlot'. 2018.

[39] Mid-2020 Small Area Population Estimates for 2011 Data Zones;. National Records of Scotland. url: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/small-area-population-estimates-2011-data-zone-based/mid-2020/.

[40] Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of epidemiology. 2004;160(6):509-16.

[41] SIMND 2020 Technical Notes;. The Scottish Government. url: https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2020/09/simd-2020-technical-notes/documents/simd-2020-technical-notes/simd-2020-technical-notes/govscot%3Adocument/SIMD%2B2020%2Btechnical%2Bnotes.pdf.

[42] Coronavirus (COVID-19): vaccine deployment plan 2021;. The Scottish Government. url: https://www.gov.scot/publications/coronavirus-covid-19-vaccine-deployment-plan-2021/.

[43] Major milestone in vaccination programme;. The Scottish Government. url: https://www.gov.scot/news/major-milestone-in-vaccination-programme/.

[44] Scotland's autumn/ winter vaccination strategy 2021;. The Cabinet Secretary for Health and Social Care. url: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2021/09/scotlands-autumn-winter-vaccination-strategy-20212/documents/scotlands-autumn-winter-vaccination-strategy/scotlands-autumn-winter-vaccination-strategy/govscot%3Adocument/scotlands-autumn-winter-vaccination-strategy.pdf.

Supplementary material

Assessing the importance of demographic risk factors across two waves of SARS-CoV-2 using fine-scale case data

A.J. Wood, A.R. Sanchez, P.R. Bessell, R.R. Kao

A Temporal analysis of Omicron and Delta waves

A.1 Supplementary plots of time evolution

Supplementary plots of the time evolution and spatial distribution of Omicron and Delta cases are presented in Figures A.1–A.5.

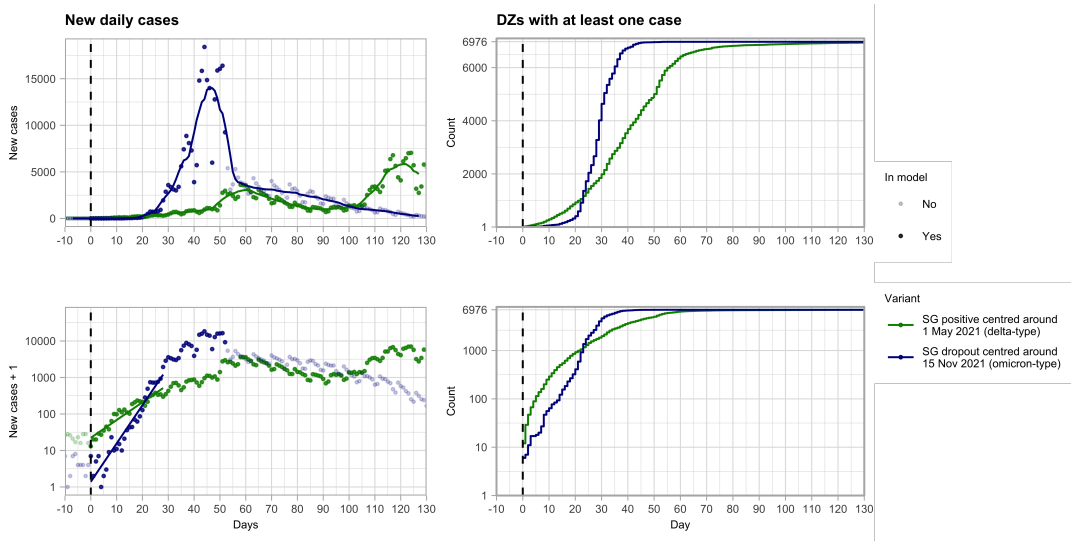

Figure A.1: Timeseries of the initial outbreaks of the Delta and Omicron variants. The gradient of the linear regression (straight line) of the early trajectory of log(new cases + 1) is inversely proportional to the case doubling time.

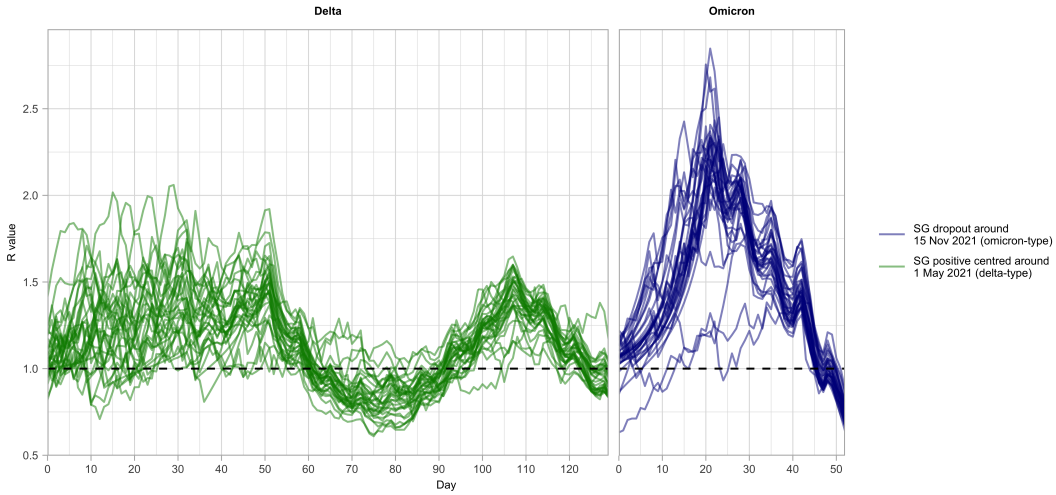

Figure A.2: Time-dependent reproduction numbers for the Delta (left) and Omicron waves (right), over each of the 32 individual local authorities.

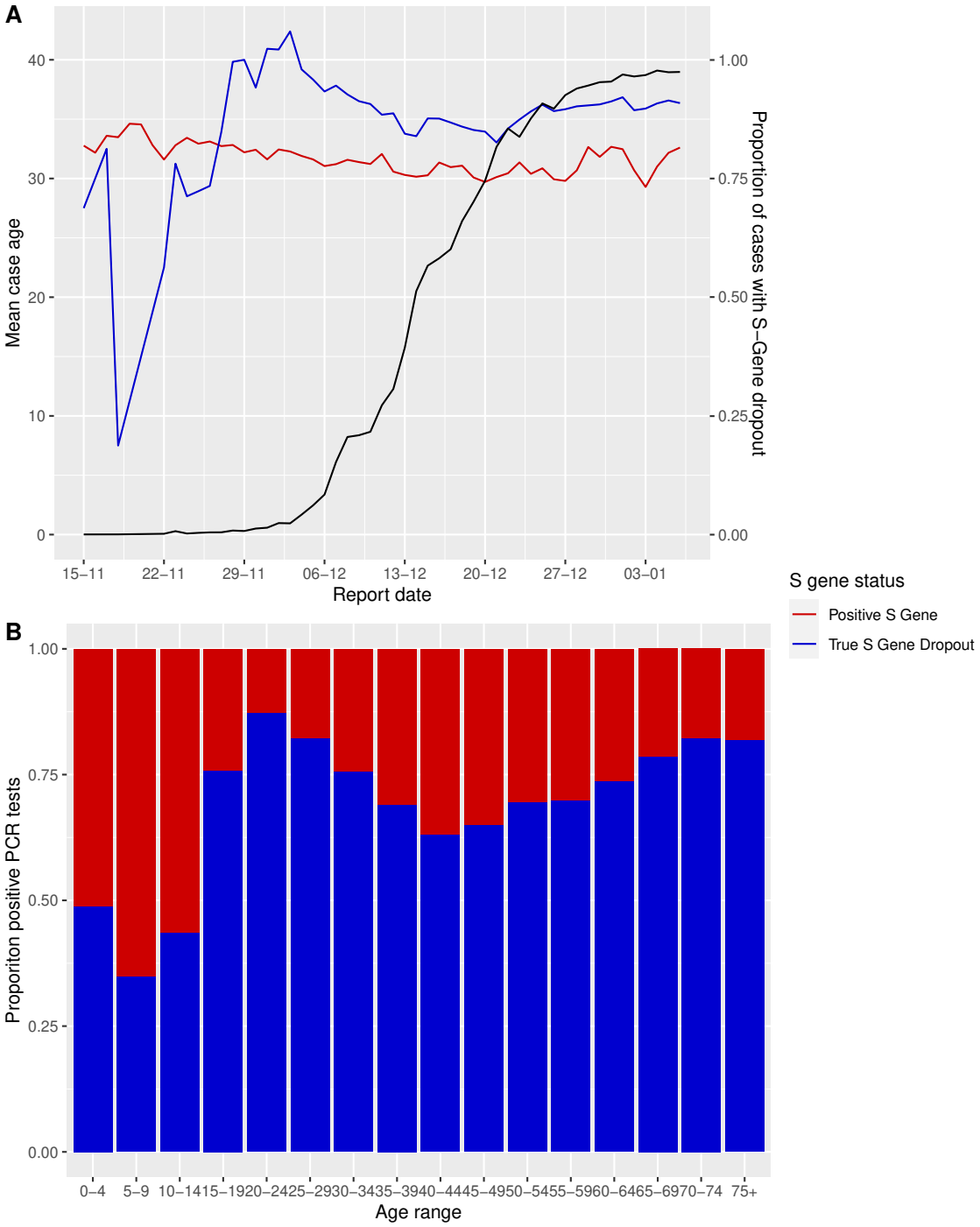

Figure A.3: PCR positive cases over the period 15th November 2021 to 6th January 2022 that were S-gene dropout or true S-gene positive. (A) Daily mean case age for the two definite PCR S-gene outcomes (blue and red lines) against the proportion of the daily cases that were true S-gene dropout (presumed Omicron-type). (B) the proportion of the cases over the period by 5-year age bracket.

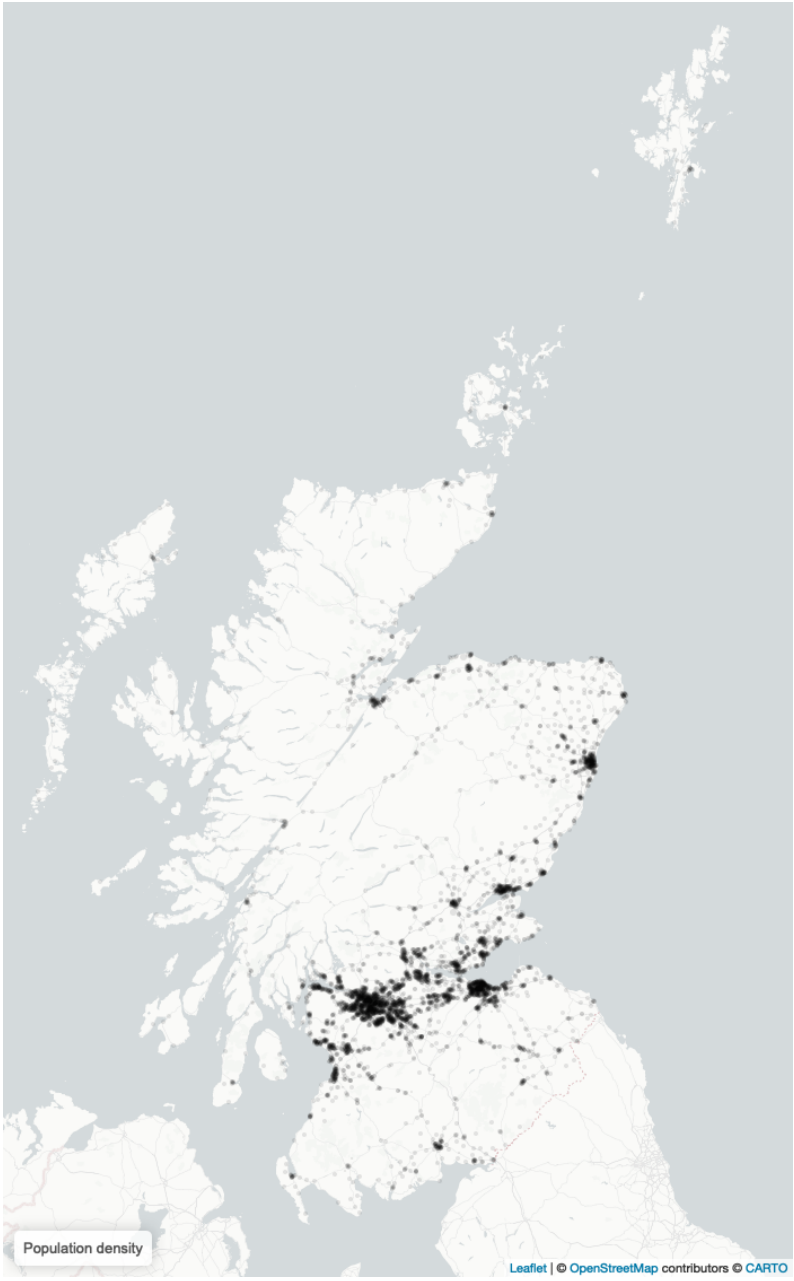

Figure A.4: Distribution of population in Scotland. Each point indicates the population-weighted centroid of a datazone (DZ) of which there are 6,976 in total, with each representing a population of order 500-1,000 individuals.

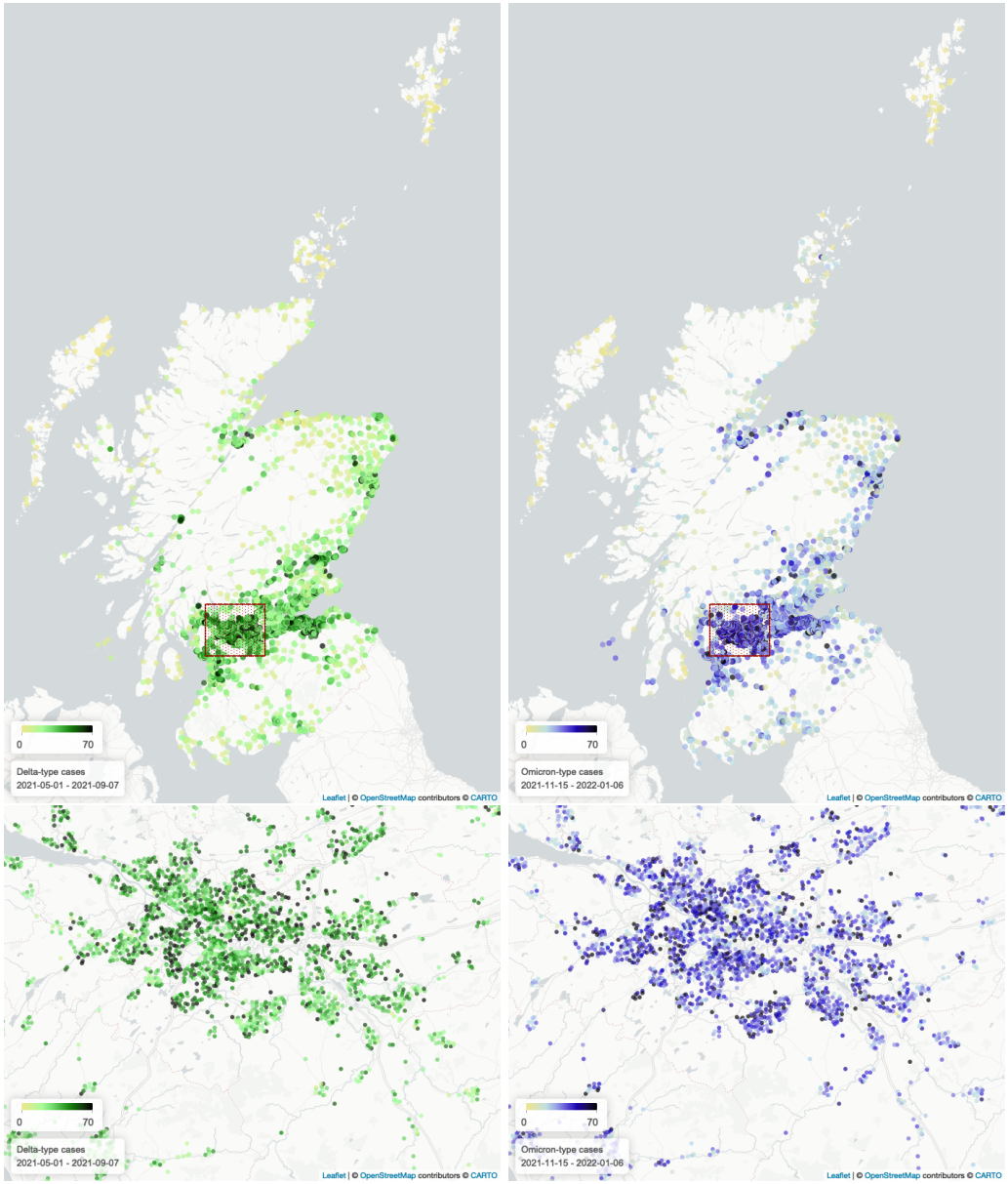

Figure A.5: Omicron COVID-19 cases in Scotland (top) between 15th November 2021 and 6th January 2022 (right), compared to Delta cases between 1st May and 7th September 2021 (left), as well as focus on the Greater Glasgow region (boxed, bottom). Each point indicates the population *centroid* of a DZ, with the colour representing the number of cases reported. The dense clustering of DZs in the south is the *central belt*.

A.2 Estimation of reproduction numbers

The time-dependent reproduction number R_i is the average number of forward infections caused by a person infected on day t_i. Define n_j as the number of new infections on day t_j. These new infections came from individuals infected on days on, or prior to t_j. Define A_{ij} as the number of new infections on day t_j *specifically* from those infected on day $t_i \leq t_j$:

$$A_{ij} = \frac{(n_i - \delta_{ij})P(t_j - t_i)}{\sum_{i' \leq j}(n_{i'} - \delta_{i'j})P(t_j - t_{i'})}\, n_j \ .$$

$P(\Delta t)$ is the probability of an individual passing on the infection, Δt days after being infected. The presence of the Kronecker delta excludes the possibility of infected individuals infecting themselves. The reproduction number R_i is then the average total of infections generated over all subsequent days [40]:

$$\begin{aligned} R_i &= \frac{1}{n_i}\sum_{j \geq i} A_{ij} \\ &= \frac{1}{n_i}\sum_{j \geq i} \frac{n_j(n_i - \delta_{ij})P(t_j - t_i)}{\sum_{i' \leq j}(n_{i'} - \delta_{i'j})P(t_j - t_{i'})} \ . \end{aligned}$$

We take $P(\Delta t)$ to be

$$P(\Delta t) \sim \mathrm{e}^{-\lambda \Delta t}$$

with λ^{-1} the mean infectious period. Individuals are equally infectious throughout the entire infection. In our calculations we estimate $1/\lambda = 6.26$ days, using the posterior mean duration of infectiousness obtained from the *SCoVMod* compartmental model (for more detail see Reference [2]).

As we estimate the infection reproduction number using the cases data, we implicitly assume that case ascertainment does not change over time, and does not account for the delay between infection, and registering a case.

B Methodology details

B.1 Model hyperparameters

The random forest regression model is fit in *R* (version 4.1.0), using the *randomForest* package [36] (version 4.6-14), and ALEs analysed using the *ALEPlot* package [38] (version 1.1).

From 6,976 DZs, 2 sexes, 16 age ranges, and 3 prior case states, there were a total of 669,696 cohorts (of which a fraction will have population zero and are excluded). Cohorts from 90% of DZs were randomly sampled for the fit, with 10% reserved to test model performance against data it explicitly did not fit. The fit was made to $\sqrt{\text{cases} + 1}$. The RF comprised 500 trees, with cohorts sampled for each tree weighted by population. 5 variables were tested at each split, and each tree had a maximum of 30,000 terminal nodes, with a minimum node size of 300. This choice of hyperparameters maximises accuracy at a whole-DZ level across areas with both high and low case numbers, both on the data the model were informed by, and the "unseen" data.

B.2 Explanatory variables used in random forest regression model

The models described in Section 4.2 are informed with the following data, first at cohort resolution:

- *Age range* (five-year windows, up to 75+);
- *Sex*;
- *Prior case status*: the time of the last reported case, broken into three categories: never tested positive before, last tested positive in the 6 months prior to the first day of the outbreak, last tested positive over 6 months prior;
- *Cohort population* (derived using historical testing data for those testing positive before, and estimated populations as of mid-2020 collated by the National Records of Scotland [39], for the remainder that had not tested positive before).

At age/sex/DZ resolution, we then include:

- COVID-19 *vaccination uptake* (eDRIS) (see also Supplementary material B.3);
- *Ethnicity* (% population belonging to a minority ethnicity), as per the most recent Scottish census data (2011);
- The per-population, time-aggregated number of *negative LFD tests* reported in that period.

Finally included are the following at DZ resolution or broader:

- Measures of DZ-level deprivation (obtained from Scottish census data, and the 2020 *Scottish Index of Multiple Deprivation* [13]);
- *Local outbreak duration*: the difference between the final date of the period studied, and the date the variant was first detected in that cohort's corresponding *intermediate zone* (*IZ*). An IZ typically contains 4–6 DZs, and 3,000–5,000 individuals, with this granularity chosen to give a reasonable proxy for how many days during the period studied that variant was circulating locally;
- *Student population* (% population being a full-time student aged 18 or over), also per 2011 census data;
- *S-gene coverage* (the proportion of cases with an accompanying S-gene result, required to associate a likely variant) at IZ level;
- *Population density*, at IZ-level.

The measures of DZ-level deprivation included are [41]:

- *Drive time from GP*: Average drive time to a GP surgery in minutes;
- *Public transport time to GP*: Public transport travel time to a GP surgery in minutes;
- *% Income deprived*: Proportion of individuals in receipt of income support payments, such as Job Seekers Allowance;
- *% Employment deprived*: Proportion of working age population claiming employment-related payments, such as Incapacity Benefit;
- *Standardised mortality ratio*: Age/sex-standardised mortality rate as compared to the overall population;
- *Comparative illness factor*: Proportion of individuals claiming from a variety of illness and disability-related payments as compared to the overall population;
- *Drug-related hospitalisation ratio*: Rate of hospitalisations relating to drug use, as compared to the overall population;
- *Alcohol-related hospitalisation ratio*: Rate of hospitalisations relating to alcohol use, as compared to the overall population;
- *Crime rate*: Rate of recorded crimes per population;
- *Attendance*: Percentage of pupils with school attendance of over 90%;
- *Attainment*: Measure for average attainment of school leavers from 2015–2018;
- *Ratio working age with no qualifications*: Proportion of working age people with no qualifications, as compared to the overall population.

We do not use data on *PCR* negative tests. In the Omicron wave PCR positivity peaked at 30% (per eDRIS data), with testing capacity being reached (resulting in a policy change on 5th January 2022 removing the need for a confirmatory PCR after an LFD positive [35]). Thus with this "ceiling" capacity being reached, we exclude negative PCR tests as a poorer proxy for propensity to test as compared to LFD negatives, and being too closely related to overall cases (requiring an S-gene sequenced positive PCR test).

B.3 Vaccination uptake as an explanatory variable

Scotland's COVID-19 vaccination programme began on December 8th 2020, with initial priority given to healthcare workers, the elderly and those otherwise especially vulnerable to COVID-19, then generally by decreasing age [42]. All first doses had been offered and administered to willing adults by 18th July 2021 [43], with rates of first dose administration declining thereafter. By 15th November 2021, then, the first dose date may have differed between two individuals by up to 11 months. This likely led to substantial variation in protection offered by the first dose at the time of the Omicron wave, given both evidence of efficacy waning over timescales of six months, and high rates of breakthrough for Omicron against vaccines originally designed against earlier "wild-type" SARS-CoV-2 lineages, particularly for non-mRNA vaccines [25, 26, 27]. This, combined with high uncertainty in the cohort-level population denominator used to determine uptake, leads us to exclude first and second dose uptake (being highly correlated with first dose uptake) as an explanatory variable for Omicron cases. We do, however, include third/booster dose uptake, as the proportion receiving a first dose to have *returned* for a third/booster dose by 15th November 2021 (and zero if nobody in the cohort had yet received a first dose). This definition eliminates uncertainty in the underlying population. Prior to the detection of Omicron, those aged 50+ or otherwise vulnerable to COVID-19 were due to be offered a third or booster dose, twelve weeks after their second [44]. The booster programme began on September 20th 2021, and a snapshot on 15th November 2021 shows substantial variation between different cohorts, particularly by age.

With these doses being delivered more recently, as well as evidence of this dose proving more protective against Omicron [25, 5], we include this definition of third/booster dose uptake as a reasonable proxy for vaccine-induced protection against Omicron at the time.

The initial Delta wave occurred while the bulk of first and second doses were still being administered, thus we include second dose uptake on 1st May 2021 as an explanatory variable, as the proportion of individuals that had returned for a second dose, having received a first (and zero, if nobody in the cohort had yet received their first dose).

C Random forest variable importance

Figure C.6 shows variable importance measures extracted from the *RandomForest* function. Age, population and prior case status have much higher node purity than the other variables, indicating that splits in individual trees using values of these variables in particular are characteristically more "effective" at separating cohorts with fewer cases from those with more cases. These are consistent with explicit variation in Fig. 1 with respect to age and prior case status (and population trivially). Fig. C.6 then shows random permutation of each of the variables results in appreciable increase in fit error, confirming that this larger collection of variables are important to explain finer patterns in the data.

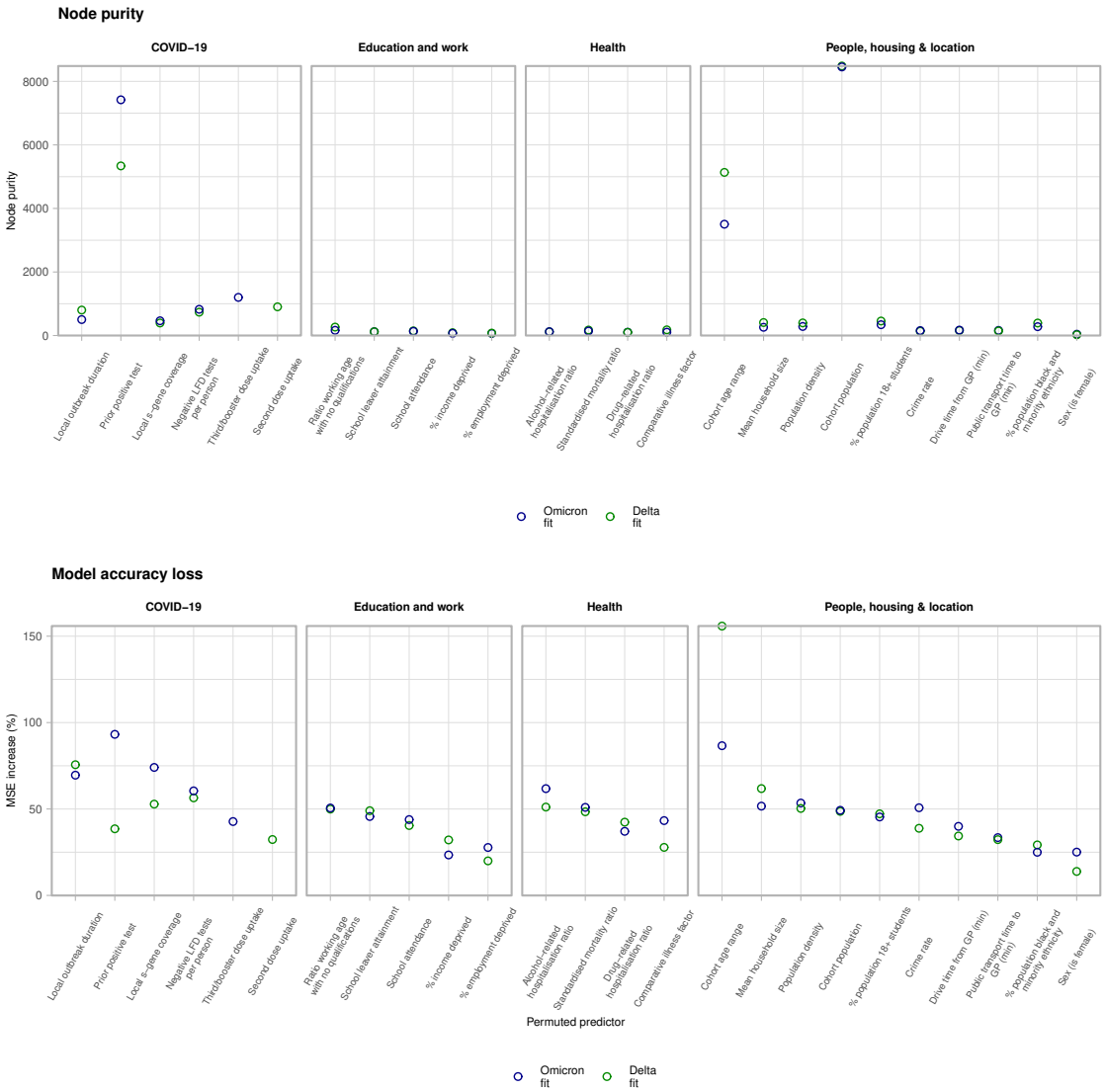

Figure C.6: Feature importance outputs form the random forest regression models. Top: Node purity. Bottom: Explanatory variable mean squared error (MSE) increase on random permutation: for variable i, the increase in MSE on data not trained in each tree, if the entries of i were instead randomly permuted.