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], R. Wightman[2], and R.R. Kao[*1,3]

[1]Roslin Institute, University of Edinburgh

[2]Edinburgh Medical School, University of Edinburgh

[3]Royal (Dick) School of Veterinary Studies, University of Edinburgh

September 8, 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 (census areas of order 500–1,000 residents) and reporting date of cases 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 testing frequency.

Despite differences in virus lineage, time of year, and interventions in place, we find the risk factors remained broadly consistent between the two waves. Many of the observed smaller differences could be reasonably explained by changes in control measures.

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

1 Introduction

A key challenge in the long term control of an infectious disease is to identify predictable patterns of incidence. The emergence and spread of the SARS-CoV-2 virus saw restrictions imposed globally on everyday life 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 in Scotland, as in the rest of the UK, randomised testing [1] and hospital admissions [2] indicate continued widespread transmission. The winding down of community testing and other surveillance is making it more difficult to track the transmission patterns of COVID-19 in detail.

Typically, identifying risk factors for infection rely on disease surveillance studies. While these studies can be powerful and provide important insights [3, 4, 5, 6], 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 data. The availability of COVID-19 case data at fine spatial scales with detailed metadata enables us to identify important health-related risks, with the data collected during the pandemic being made available to researchers in close to real-time.

In this work we aim to identify risk factors for COVID-19 cases in Scotland, and their change over time, to serve as an indicator for how the longer-term profile of infection may evolve. We fit the case distributions of two different waves of COVID-19, with a machine learning model informed by a range of explanatory variables relating to geography and demographics.

The first COVID-19 case in Scotland was identified on 1st March 2020 [7]. The Scottish Government imposed strict "lockdown" non-pharmaceutical intervenions (NPIs) on 23rd March 2020 [8]. While initially applied at the national level, following the initial lockdown period NPIs were adjusted by local authority (administrative areas with populations ranging between 22,540–635,130) through a "levels"-based system [9]. The seeding and rapid spread of the B.1.1.7 lineage (termed the "Alpha" variant) in December 2020 led to a tightening of NPIs and a second lockdown [10, 11]. A mass vaccination programme began in December 2020 [12, 13], prioritising the elderly and healthare workers, with all adults eventually eligible.

We focus on case data gathered between May 2021 and January 2022, a period that saw the steady relaxation of nearly all NPIs [14]. This period had two major waves of infection: the first from May 2021 triggered by the B.1.617.2 lineage ("Delta"), and a second wave from November 2021 by the B.1.1.529 B.A.1 lineage ("Omicron"). The deletion of two specific amino acids in the Omicron sub-variant distinguished it from most co-circulating variants including Delta, in PCR

tests that have an accompanying "S-gene" test result [15]. 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.

Earlier work has exploited finely-grained case data to highlight risk factors for cases and severe outcomes including (but not limited to) sex [16, 17, 18], population density [19, 20, 21], deprivation [22, 23, 24, 25], occupation [26, 27, 28], and age [29, 30, 31]. Similar studies have incorporated movement data [32] to demonstrate the protective impact of NPIs that restrict mobility [21, 33, 34, 35, 36, 37]. Many of these studies focus on "first wave" of infection, during which strict NPIs were imposed and no population immunity had been established. This study focuses on a more advanced period moving away from NPIs, and the conditions for disease spread comparatively less "exceptional". This is especially the case for the Omicron wave. A unique feature of our model is the inclusion of lateral flow test taking *frequency*. The proportion of infectons that end up reported is likely to depend on testing propensity, and we consider how that may lead to distortions in the case distribution.

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

The period November 15th 2021 – January 6th 2022 covers the first outbreak and peak of the B.1.1.529 lineage (BA.1 sublineage, hereafter referred to as the Omicron variant) (S-gene "dropout" test signature). Prior to this, the B.1.617.2 lineage (Delta variant) (S-gene positive test signature) 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 nationwide sequence data [38].

2.1 Time evolution and early patterns of spread

We identified 385,558 cases between November 15th 2021 and January 6th 2022, of which 227,286 were likely Omicron. From 1st May 2021 to 7th September 2021 we identified 269,838 cases, of which 229,073 were likely Delta. The remaining cases in these periods (those with no S-gene result, or a different result) are excluded. 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.

Omicron cases had a doubling time (the time taken for *newly reported daily* cases to double) of 2.9 days over the first 28 days, compared to 6.2 days for Delta (Supplementary Material, Fig. S1). Over half of all DZs had reported an Omicron case in the wave within 29 days, whereas for Delta this took 39 days (Supplementary Material, Fig. S2).

The reproduction number R_t consistently rose for Omicron, peaking at above 2 for nearly all local authorities 28 days in to the outbreak, and only consistently falling below 1 after 50 days (Supplementary Material, Fig. S3). 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.

In the intermediate period during which Omicron became dominant and Delta declined, 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). A Student's t test shows this difference to be statistically significant ($t = -52.2$, $p < 0.001$). This was the case from relatively early on when Omicron accounted for at least 5% of cases (Supplementary Material, Fig. S4A). 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 (Supplementary Material, Fig. S4B).

2.2 Case distribution and model fit

Fig. 1, shows the distribution of COVID-19 cases for the Omicron and Delta waves broken down 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. Fig. 2 shows case rates per DZ. Geographically, case rates fall with increasing rurality, most notably in Orkney, Shetland and the Western Isles (all island communities). The trend with respect to multiple deprivation decile is bimodal, with higher rates towards the highest and lowest deciles.

The fit case rates from our random forest regression models are overlaid onto Fig. 1. We achieve a good fit to these larger-scale trends. The model slightly under-fits the age ranges 15–24, where case rates were the highest overall. Variable importance outputs are presented in Supplementary

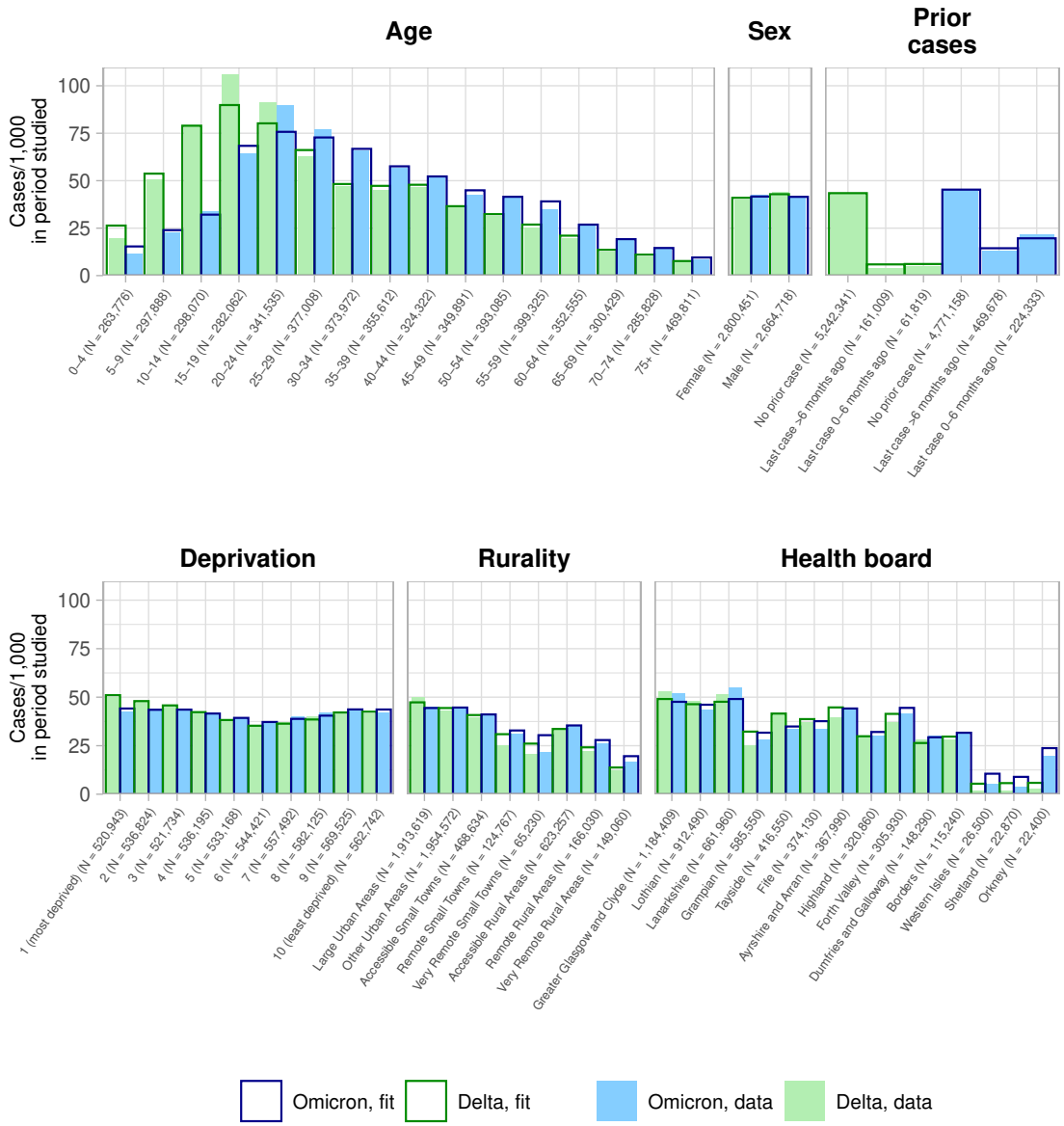

Figure 1: Summary of 227,286 Omicron COVID-19 cases in Scotland between November 15th 2021 and January 6th 2022 (blue, filled), and 229,073 Delta cases from 1st May 2021 to 7th September 2021 (green, filled). The full population ($N = 5,465,169$) is broken down by *age range*, *prior case status* (whether a person had previously reported a COVID-19 case prior to that specific wave, and when), *deprivation* (of place of residence, per the SIMD decile, with 1 the most deprived), *rurality* (of place of residence, per the census Urban/Rural Classification) and *location* (at the level of Scottish health board). Cases are given per 1,000 people in that group (with subpopulation N recorded on the axis labels). The corresponding case rates as fit by our models are superimposed. Note that the subpopulations in the *prior case status* plot change across waves, due to being at different points in time.

Material Fig. S7, with node purity and accuracy loss.

Fig. 3 (top) shows model performance at DZ level, comparing observed cases to fit cases.

Beginning with Omicron cases, 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), with a poorer fit for cohorts with very high case counts. A "reduced" random forest 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. 3 shows further deviation of the data-fit slopes away from the diagonal for these "reduced" models.

Considering now earlier Delta cases from 1st May to 7th September 2021, the geographical distribution (Fig. 2) is visually similar, with a concentration of high case rates in the denser "central belt". Cases skewed slightly younger (Fig. 1), with the highest rates within ages 15–19. The distributon with respect to deprivation decile remains bimodal, with higher rates in both the most and least deprived DZs. Model performance was similar, explaining 72% (fit: 73%, test: 61%) of DZ-level variation.

Fig. 3 (bottom) shows for both the Delta and Omicron models, autocorrelation of residuals (as measured by the Moran's I statistic, Section 4.5) within 1km is 0.35, falling to 0.15 at 5km, and 0.05 at 50km. The reduced models exhibit much higher residual autocorrelation, with the density-only model performing best, but persisting over larger distances (see Supplementary Material, Fig. S6 for a map view of residuals).

2.3 Accumulated local effects

Fig. 4 shows the accumulated local effects (ALEs) of all explanatory variables in the model (see Section 4.4 for definition).

Population, age, sex, and prior case status have ALEs that follow the empirical distributions observed in Fig. 1; ALEs are strongly positive for ages between 15–40, and those that had never reported a case before.

Beyond these variables, Fig. 4 shows that features such as low population density, high vaccination uptake, a low mean household size, and a low rate of negative LFD test reporting are protective. We note that for vaccination uptake, the protective value at zero is likely an artefact arising from cohorts with ages 0–9 that were not eligible.

The effects for many variables associated with social deprivation such as the ratio of working age people with no qualifications and the rate of income deprivation (see Supplementary Material, Section B.2 for full descriptions) are weaker. This is consistent with the small degree of deprivation-

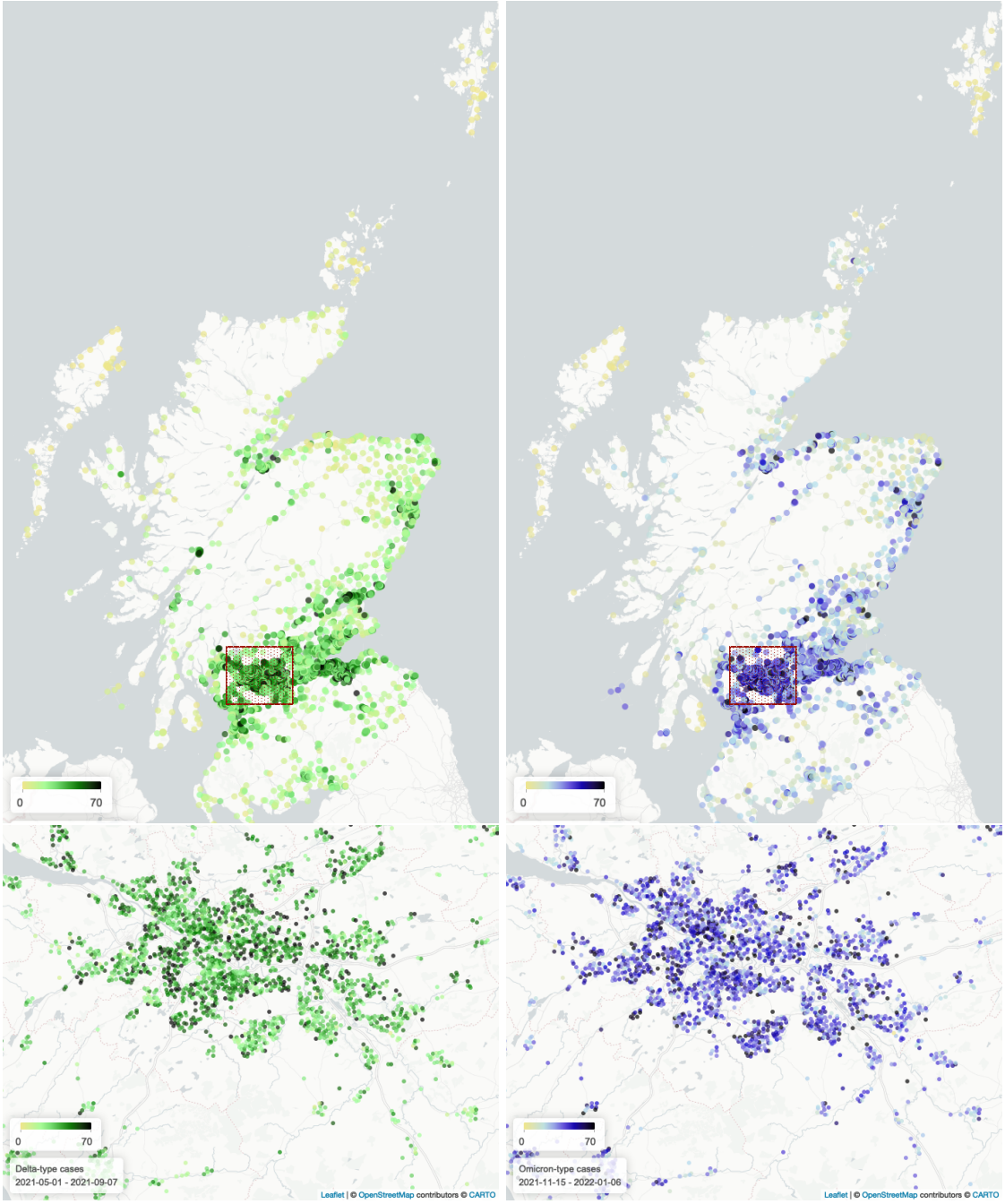

Figure 2: COVID-19 cases in Scotland (top) over the Omicron period (right, blue) as compared to Delta cases (left, green), with focus on the Greater Glasgow region (boxed, bottom). Each point indicates the population *centroid* of a DZ, with the colour representing the number of cases reported.

level variation seen in Fig. 1.

The directionality of the ALEs remain broadly consistent across both waves. Some risk factors were more pronounced in the Delta model, including in mean hosehold size, population density

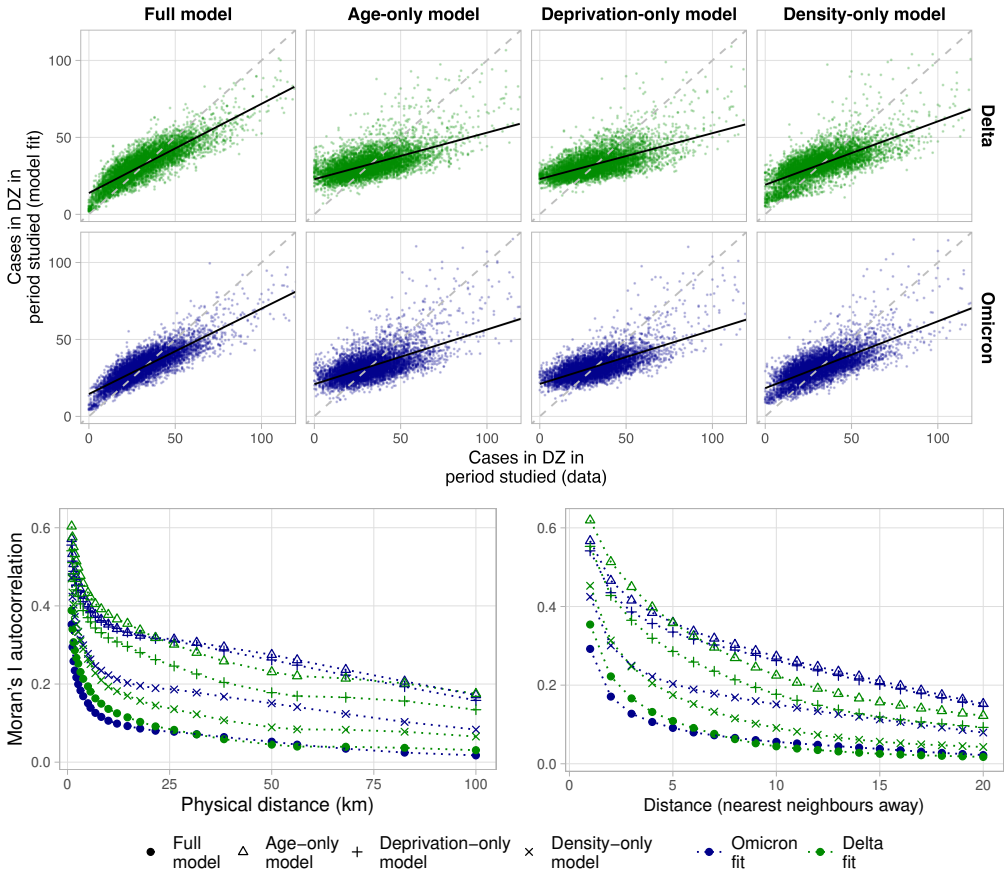

Figure 3: Top: performance of different models, comparing observed cases to fit cases at DZ level. Each point represents a DZ. Points deviating from the diagonal indicate DZs with less accurate fits. The full model is compared with performance of reduced models informed with only population, and one of either age, overall deprivation rank, or population density. Bottom: 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 model residuals, when comparing DZs sitting within a given locus. DZs are defined as nearest neighbours of one another if they share a boundary.

and the proportion of individuals belonging to a black or minority ethnicity. Conversely, cohorts with very high student populations were associated more strongly with high case rates in the Omicron fit.

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 ending of detailed surveillance since, it is more difficult to monitor the precise patterns of infection amongst the population and how that will

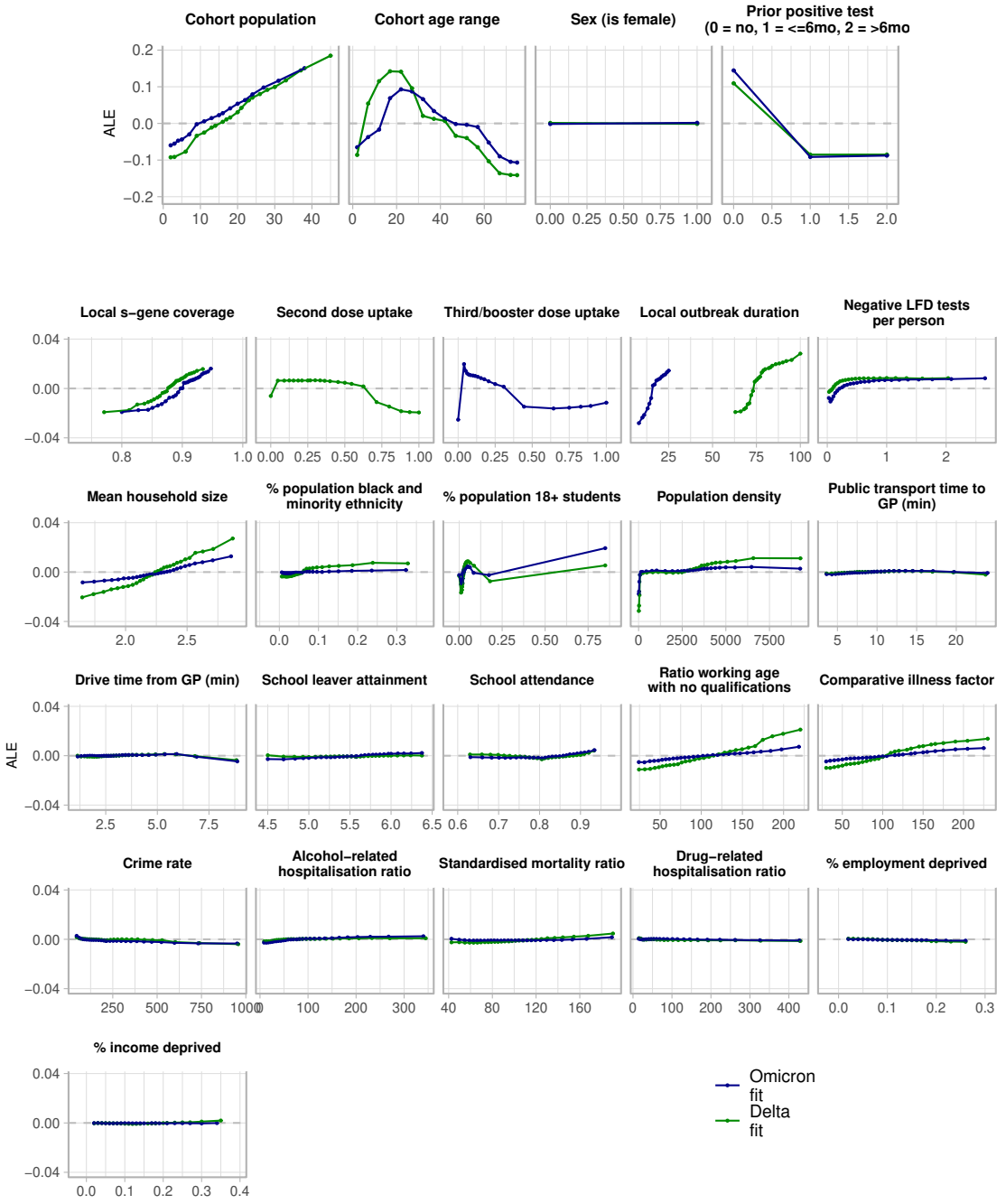

Figure 4: Accumulated local effects across all explanatory variables. For each variable, the x-axis represents the range of values of that variable in the data, and the y-axis (note scale differences for *population*, *age*, *sex* and *prior case status*) is the ALE for that variable value. The overall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

evolve over time, especially with respect to different variants.

The aim of this study was to compare the pattern of cases across two waves of COVID-19 in Scotland in 2021, during which non pharmaceutical interventions (NPIs) were being relaxed but testing remained mandatory and a mass vaccination rollout was in progress. 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. We have shown that case heterogeneity was 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. Despite differences in the severity of interventions in place, time of year, vaccination uptake and virus phenotype, these risk factors remain broadly consistent across both waves.

Our models accurately capture the case distributions (Fig. 1). However, not all variation is explained, and residual autocorrelation persists at <5km scales (Fig. 3). A reason for this may be that our model is not informed by mobility, thus explicit links between communities are not known to the model. We also do not include meteorological data (such as in e.g. [33]). This could have explained further variation as our waves occur in different seasons, where the characteristic routes of transmission may have differed. Last, the fit cases are also time-aggregated, and therefore do not account for changes in risk factors *during* each wave.

The inclusion of the *local outbreak duration* for each DZ (the time the first case was detected in the DZs wider intermediate zone, typically containing 4-6 DZs) accounts in part for local interactions between neighbouring communities, in the absence of explicit mobility data. A weakness of this is that the local outbreak duration correlates with the total number of cases, given the relatively short periods studied. We suspect this is less influential in the Omicron model where geographical spread was more rapid. The regression models applied here may be better suited to scenarios where an infectious disease is already well established in the population. For future analyses on cases at the very beginning of an outbreak with fewer cases, this approach may be adapted to instead fit case rates per day, from when the first case was identified locally.

Risk factors

We presented the accumulated local effects (Fig. 4), revealing broad indicators for higher or lower case rates, and how they changed between waves. It is difficult to fully disentangle whether a difference was caused by a change in control measures, or a change in virus strain. Nonetheless, our analyses provide some important insights.

To begin, high mean household size emerges as a risk factor, consistent with the high secondary attack rates for SARS-CoV-2 [39, 40], and increased risk of inter-household transmission relative to contacts outside of the home [41]. That this, and high population density are both stronger risk factors for Delta may reflect the stronger NPIs at this tme increasing the proportion of within-DZ or within-household transmissions.

High vaccine uptake (amongst those eligible) is also protective, more so with Delta, consistent with higher rates of immune breakthrough with the Omicron variant as compared to Delta [42, 43, 44]. We do not know the specific vaccination status of those in the test data, however, and linked data may show a stronger protective effect.

For Delta, a high proportion of individuals of black and minority ethnicity is a stronger risk factor. In the UK, this is also a risk factor for severe COVID-19 outcomes [45, 46, 47] 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 [48] and occupation [49] (thus ability to work from home or effectively physical distance), and the relative impacts of those factors changing across the two waves.

Finally, living in a deprived community was suggested from early on [50] and has since also emerged as a risk factor for *severe* COVID-19 disease [51, 52, 53, 54, 55, 56]. However, the corresponding ALEs for the variables associated with deprivation are small. Deprivation effects may be captured by proxy with other variables that correlate with deprivation such as age [57] and vaccine uptake [58, 59].

Testing frequency

The low case rate variation with deprivation (Fig. 1) contrasts with observed inequalities over severe outcomes [60, 22, 23, 24, 25], suggesting that those living in more deprived communities experience a higher inherent case-hospitalisation rate. We suspect that a lower proportion of case *ascertainment*, however, may also be a factor.

An important and unique variable in our model is the rate at which *negative* LFD tests were reported throughout the period. We found high rates of *negative* test reporting to be a *risk* factor. This suggests a variation in case ascertainment across different demographics, which may in turn lead to skews in the observed case distribution [61, 62, 22].

Further work (Supplementary Material Table S7, Fig. S8) shows that up to February 2023, the rate of LFD testing and positivity varied substantially across deprivation (quintile 1: 3.6

tests/person, 4.61% positive; quintile 5: 6.7 tests/person, 3.57% positive) as well as sex (M: 3.7 tests/person, 4.82% positive; F: 7.0 tests/person 3.30% positive). If demographic differences in testing behaviour correspond to differences in case ascertainment, the profile of all infections may then be biased from reported cases, and testing rates may be obscuring the true patterns of infection over sex and deprivation.

In addition, the magnitude of the risk factor (as seen in the ALE, Fig. 4) plateaus beyond a certain rate ($>\sim 1$ test/person in each period). This hints at a deeper relationship between true incidence, the frequency of testing (and whom amongst the population is taking those tests), and the proportion of infections that are ascertained.

Our model is unique in including negative test reporting, and has revealed strong differences between different demographics that may bias the profile of cases. Beyond the work presented here, further analysess of reported cases need to be considered with these strong skews in testing behaviour in mind.

Conclusion

The COVID-19 data studied here 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 known trends in severe outcomes suggest that the distribution of infections may have been very different to that of reported cases. By incorporating trends on cases, testing behaviour, and 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 Preparation of case data

We use COVID-19 testing data from Public Health Scotland's *electronic Data Research and Innovation Service* (eDRIS) system, dated from July 14th 2022. 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 (positive, dropout, inconclusive), 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 metadata — in particular the DZ, specifying location to within an area as small as 0.1km^2 in densely populated areas — therefore identify cases at a fine spatio-temporal scale. Data on vaccine administrations are also provided by eDRIS.

This analysis considers the BA.1 sub-variant of the Omicron lineage only. The sub-variant BA.2/B.1.1.529.2 later replaced BA.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 [63], 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 [64], after which cases with S-gene results fell to about 50% by February 2022 (per eDRIS data).

For Omicron cases, we gather from the data S-gene dropout cases between 15th November 2021 and 6th January 2022, and for the Delta outbreak, S-gene positive cases between 1st May and 7th September 2021 (choosing this end date to have a similar number of cases in each set). We exclude cases that have a different, or no S-gene result.

Using the linked historical tests, we label cases based on whether the individual had either: never tested positive before; had tested positive in the last six months prior to the start of that wave, or; last tested positive over six months prior to the start of that wave. We denote this the *prior case status*, as a proxy for infection-based immunity.

Finally to prepare the cases data to be fit, we group individuals that have the same age range,

sex, residing datazone, and *prior case status*, terming these subsets of individuals *cohorts*. As an illustrative example, a cohort may be a population of 38 males aged between 50–54 residing in a given datazone "X", that have never tested positive for COVID-19 before, among whom 9 Omicron COVID-19 cases were identified. This is the highest practical resolution we can acheive using the eDRIS case data, and our model (Section 4.3) fits case counts at this resolution.

4.2 Time series analysis

Time-dependent reproduction number

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 δ_{ij} 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 [65]:

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

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 [56]).

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.

In this work the reproductive number is measured at local authority level, the level at which

the Scottish Government monitored and adjusted NPIs.

Case doubling time

At the start of each wave we assume exponential growth of cases:

$$\text{new cases} \propto e^{rt}$$

where the gradient of a linear regression on $\log(\text{new cases})$ against t returns the growth rate r. The evolution of new cases an also be rewritten in terms of of a doubling time t_D:

$$\text{new cases} \propto 2^{t/t_D}$$

where $t_D = \frac{\log 2}{r}$.

4.3 Model

Our statistical model is designed to explain variation in COVID-19 case numbers as prepared in Section 4.1, and identify risk factors amongst a broad range of variables, using random forest regression. We fit models to the distribution of Delta and Omicron cases respectively, allowing for comparison of risk factors across the two waves.

Explanatory variables

We include demographic factors (population, age, sex, ethnicity, student population), COVID-19 related factors (testing volume, prior case status, vaccination uptake), 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) [66]. 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. To account for local spread of infection between neighbourhoods that are geographically close to one another, we include an *local outbreak duration* parameter, which specifies the date at which the *first* case of the variant was identified at the *intermediate zone* (IZ, an administrative area containing of order 4–6 DZs).

A comprehensive description of all individual variables used in given in Supplementary Material, Section B.2.

Random forest model

We use random forest regression [67] on the distribution of COVID-19 cases, as it allows us to fit the distribution without specifying any prior analytical relation between the outcome variable (cases) and any of the explanatory variables, which may themselves be correlated. We fit the time-aggregated case distribution in *R* (version 4.1.0) [68], using the *randomForest* package [69] (version 4.6-14).

We fit the outcome variable $\sqrt{\text{cases} + 1}$ at cohort level (with a *cohort* defined in Section 4.1). The fit number of cases at other scales (such as DZ level) is then an aggregation of cases from their constituent cohorts.

We extract 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 effective removal of that variable from the model.

Model hyperparameters were chosen manually so as to maximise the variance explained by a subset of the data not used to fit the model. Full hyperparameter specification is included in Supplementary Material, Section B.1. The model specifications for fitting the Omicron and Delta waves are identical with one exception: for the Omicron model, third/booster dose uptake is used, whereas for Delta, second dose uptake is used (third/booster doses were only administered later; see Supplementary Material, Section B.3 for further details).

In addition to the full model, we fit for each of Omicron and Delta three "reduced" models, under equivalent hyperparameters to the full model and the same cohort structure, but informed only by population, and one of: age; the relative deprivation of the residing DZ, as defined by the overall SIMD deprivation *rank* [70], and; population density. These outputs illustrate how effective these variables are at alone at explaining case variation, relative to our full model.

4.4 Accumulated local effects

To identify risk factors amongst the explanatory variables used to inform the model, we calculate the *accumulated local effects* (ALEs) of each variable. The 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 [71]. In this context, ALEs indicate whether a variable value is associated with fewer or more cases in general over the data. If the ALE is greater than zero, the fit cases generally increases given that variable value.

4.5 Moran's I autocorrelation statistic

To probe geographical variation in cases *not* explained by the model, we measure the Moran's I autocorrelation [72, 73] on the residuals (the difference between the data and fit value), relating to their physical location. We compare local DZ-aggregated residuals over physical distances (from 1–100km), as well as network distance (number of nearest neighbours apart). For a set of N residuals y_i, the Moran's I is a measure of autocorrelation:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

with $\bar{y}$ the mean of all residuals, and $w_{i,j}$ is an associated *weight* of the pair of observations (i, j), with $w_{i,i} = 0$. To measure the autocorrelation between residuals within a separation d (either a physical or network-based distance) of one another, we set $w_{i,j} = 1$ if $\mathrm{dist}(i, j) \leq d$, and 0 otherwise. Fully correlated residuals would have $I = 1$, whereas $I = 0$ would indicate no correlation.

This measure characterises how effective our models are at explaining geographical variation, and with different distances d shows over what length scales residual autocorrelation persists.

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. We also thank the reviewers for their feedback and suggestions, which has led to improvement of the article.

6 Author contributions

R.R.K. conceived the project. A.J.W. wrote the model code and performed the case distribution analysis. A.R.S. and P.R.B. performed the temporal analysis. R.W. performed analysis of lateral flow testing data. 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*, and can be contacted via phs.edris@phs.scot. The authors received no special privileges with respect to data access as compared to other researchers.

Data relating to deprivation are derived from the 2020 Scottish Index of Multiple Deprivation, and are publicly available (https://simd.scot). Fine-scale population statistics are drawn from publicly-available population estimates as of mid-2020 [74]. 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 *Real-time monitoring and predictive modelling of the impact of human behaviour and vaccine characteristics on COVID-19 vaccination in Scotland* (ES/W001489/1). This work has also been funded the BBSRC Institute Strategic Programme grant to the Roslin Institute (BB/J004235/1).

References

[1] Office for National Statistics. Coronavirus (COVID-19) Infection Survey: Scotland Dataset;. Available from https://www.ons.gov.uk/

peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/covid19infectionsurveyscotland (last accessed 15/08/2023).

[2] Office for National Statistics. Coronavirus (COVID-19) latest insights: Hospitals;. Available from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/hospitals (last accessed 16/08/2023).

[3] 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.

[4] 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;.

[5] 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–e598.

[6] 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.

[7] The Scottish Government. Coronavirus (COVID-19) confirmed in Scotland;. Available from https://www.gov.scot/news/coronavirus-covid-19/ (last accessed 16/08/2023).

[8] The Scottish Government. Effective 'lockdown' to be introduced;. Available from https://www.gov.scot/news/effective-lockdown-to-be-introduced/ (last accessed 16/08/2023).

[9] The Scottish Government. Coronavirus (COVID-19): protection levels — reviews and evidence;. Available from https://www.gov.scot/collections/coronavirus-covid-19-protection-levels-reviews-and-evidence/ (last accessed 16/08/2023).

[10] The Scottish Government. New guidance issued for the festive period;. Available from https://www.gov.scot/news/new-guidance-issued-for-the-festive-period/ (last accessed 16/08/2023).

[11] The Scottish Government. Scotland in Lockdown;. Available from https://www.gov.scot/news/scotland-in-lockdown/ (last accessed 16/08/2023).

[12] The Scottish Government. First COVID-19 vaccinations in Scotland take place;. Available from https://www.gov.scot/news/first-covid-19-vaccinations-in-scotland-take-place/ (last accessed 16/08/2023).

[13] The Scottish Government. Coronavirus (COVID-19): vaccine deployment plan 2021;. Available from https://www.gov.scot/publications/coronavirus-covid-19-vaccine-deployment-plan-2021/ (last accessed 15/08/2023).

[14] Hale T, Angrist N, Kira B, Petherick A, Phillips T, Webster S. Variation in government responses to COVID-19. 2020;.

[15] McMillen T, Jani K, Robilotti EV, Kamboj M, Babady NE. The spike gene target failure (SGTF) genomic signature is highly accurate for the identification of Alpha and Omicron SARS-CoV-2 variants. Scientific reports. 2022;12(1):18968.

[16] Gebhard C, Regitz-Zagrosek V, Neuhauser HK, Morgan R, Klein SL. Impact of sex and gender on COVID-19 outcomes in Europe. Biology of sex differences. 2020;11:1–13.

[17] Galbadage T, Peterson BM, Awada J, Buck AS, Ramirez DA, Wilson J, et al. Systematic review and meta-analysis of sex-specific COVID-19 clinical outcomes. Frontiers in medicine. 2020;7:348.

[18] Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nature communications. 2020;11(1):6317.

[19] Sartorius B, Lawson A, Pullan R. Modelling and predicting the spatio-temporal spread of COVID-19, associated deaths and impact of key risk factors in England. Scientific reports. 2021;11(1):5378.

[20] Diao Y, Kodera S, Anzai D, Gomez-Tames J, Rashed EA, Hirata A. Influence of population density, temperature, and absolute humidity on spread and decay durations of COVID-19:

A comparative study of scenarios in China, England, Germany, and Japan. One Health. 2021;12:100203.

[21] Smith TP, Flaxman S, Gallinat AS, Kinosian SP, Stemkovski M, Unwin HJT, et al. Temperature and population density influence SARS-CoV-2 transmission in the absence of nonpharmaceutical interventions. Proceedings of the National Academy of Sciences. 2021;118(25):e2019284118.

[22] 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.

[23] Meurisse M, Lajot A, Devleesschauwer B, Van Cauteren D, Van Oyen H, Van den Borre L, et al. The association between area deprivation and COVID-19 incidence: a municipality-level spatio-temporal study in Belgium, 2020–2021. Archives of Public Health. 2022;80(1):1–10.

[24] KC M, Oral E, Straif-Bourgeois S, Rung AL, Peters ES. The effect of area deprivation on COVID-19 risk in Louisiana. PLoS One. 2020;15(12):e0243028.

[25] Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases. 2020;20(11):1247–1254.

[26] Reuter M, Rigó M, Formazin M, Liebers F, Latza U, Castell S, et al. Occupation and SARS-CoV-2 infection risk among 108 960 workers during the first pandemic wave in Germany. Scandinavian Journal of Work, Environment & Health. 2022;48(6):446.

[27] Rhodes S, Wilkinson J, Pearce N, Mueller W, Cherrie M, Stocking K, et al. Occupational differences in SARS-CoV-2 infection: analysis of the UK ONS COVID-19 infection survey. J Epidemiol Community Health. 2022;76(10):841–846.

[28] Zhang M. Estimation of differential occupational risk of COVID-19 by comparing risk factors with case data by occupational group. American journal of industrial medicine. 2021;64(1):39–47.

[29] Chadeau-Hyam M, Bodinier B, Elliott J, Whitaker MD, Tzoulaki I, Vermeulen R, et al. Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK biobank data. International journal of epidemiology. 2020;49(5):1454–1467.

[30] Lau MS, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proceedings of the National Academy of Sciences. 2020;117(36):22430–22435.

[31] Working group for the surveillance, control of COVID-19 in Spain, group for the surveillance W, control of COVID-19 in Spain, Redondo-Bravo L, Sierra Moros MJ, et al. The first wave of the COVID-19 pandemic in Spain: characterisation of cases and risk factors for severe outcomes, as at 27 April 2020. Eurosurveillance. 2020;25(50):2001431.

[32] Hu T, Wang S, She B, Zhang M, Huang X, Cui Y, et al. Human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges. International Journal of Digital Earth. 2021;14(9):1126–1147.

[33] Ledebur K, Kaleta M, Chen J, Lindner SD, Matzhold C, Weidle F, et al. Meteorological factors and non-pharmaceutical interventions explain local differences in the spread of SARS-CoV-2 in Austria. PLoS computational biology. 2022;18(4):e1009973.

[34] Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA. Population flow drives spatio-temporal distribution of COVID-19 in China. Nature. 2020;582(7812):389–394.

[35] Wang H, Ghosh A, Ding J, Sarkar R, Gao J. Heterogeneous interventions reduce the spread of COVID-19 in simulations on real mobility data. Scientific reports. 2021;11(1):7809.

[36] Hou X, Gao S, Li Q, Kang Y, Chen N, Chen K, et al. Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. Proceedings of the National Academy of Sciences. 2021;118(24):e2020524118.

[37] Asem N, Ramadan A, Hassany M, Ghazy RM, Abdallah M, Ibrahim M, et al. Pattern and determinants of COVID-19 infection and mortality across countries: An ecological study. Heliyon. 2021;7(7).

[38] 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):4121–4123.

[39] 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;.

[40] 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.

[41] 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.

[42] 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;.

[43] 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):654–656.

[44] 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):1646–1657.

[45] Office for National Statistics. Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 8 December 2020 to 1 December 2021;. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/updatingethniccontrastsindeathsinvolvingthecoronaviruscovid19englandandwales/8december2020to1december2021 (last accessed 15/08/2023).

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

[47] 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.

[48] Office for National Statistics. 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;. Available from: https://www.ons.gov.

uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/coronavirusandvaccinationratesinpeopleaged18yearsandoverbysociodemographiccharacterist 8december2020to31december2021 (last accessed 15/08/2023).

[49] National Records of Scotland. Census 2011: Release 3I - Detailed characteristics on Labour Market and Education in Scotland;. Available from: https://www.nrscotland.gov.uk/news/2014/census-2011-release-3i (last accessed 15/08/2023).

[50] Khalatbari-Soltani S, Cumming RC, Delpierre C, Kelly-Irving M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. J Epidemiol Community Health. 2020;74(8):620–623.

[51] 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.

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

[53] Bambra C, Riordan R, Ford J, Matthews F. The COVID-19 pandemic and health inequalities. J Epidemiol Community Health. 2020;74(11):964–968.

[54] 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.

[55] McGurnaghan SJ, Weir A, Bishop J, Kennedy S, Blackbourn LA, McAllister DA, et al. Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland. The lancet Diabetes & endocrinology. 2021;9(2):82–93.

[56] 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.

[57] National Records of Scotland. Mid-2021 Small Area Population Estimates, Scotland (Report);. Available from https://www.nrscotland.gov.uk/files//statistics/population-estimates/sape-2021/sape-21-report.pdf (last accessed 15/08/2023).

[58] Office for National Statistics. Coronavirus (COVID-19) Infection Survey technical article: Analysis of characteristics associated with vaccination uptake;. Available from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsurveytechnicalarticleanalysisofcharacteristicsassociatedwit 2021-11-15 (last accessed 15/08/2023).

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

[60] Wood AJ, Kao RR. Empirical distributions of time intervals between COVID-19 cases and more severe outcomes in Scotland. PloS one. 2023;18(8):e0287397.

[61] 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;p. 111333. Available from: https://www.sciencedirect.com/science/article/pii/S0022519322003241.

[62] 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.

[63] The UK Health Security Agency. SARS-CoV-2 variants of concern and variants under investigation in England: technical briefing 35;. Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1050999/Technical-Briefing-35-28January2022.pdf (last accessed 15/08/2023).

[64] The Scottish Government. Self-Isolation and testing changes;. Available from https://www.gov.scot/news/self-isolation-and-testing-changes/ (last accessed 15/08/2023).

[65] 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):509–516.

[66] The Scottish Government. SIMD 2020 Technical Notes;. Available from https://www.gov.scot/publications/simd-2020-technical-notes/ (last accessed 15/08/2023).

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

[68] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2022. Available from: https://www.R-project.org/.

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

[70] The Scottish Government. Scottish Index of Multiple Deprivation 2020;. Available from https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-indicator-data/ (last accessed 15/08/2023).

[71] Apley D, Apley MD. Package 'ALEPlot'. 2018;.

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

[73] Gittleman JL, Kot M. Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology. 1990;39(3):227–241.

[74] National Records of Scotland. Mid-2020 Small Area Population Estimates for 2011 Data Zones;. Available from 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/ (last accessed 15/08/2023).

[75] The Scottish Government. Major milestone in vaccination programme;. Available from https://www.gov.scot/news/major-milestone-in-vaccination-programme/ (last accessed 15/08/2023).

[76] The Cabinet Secretary for Health, Care S. Scotland's autumn/winter vaccination strategy 2021;. Available from https://www.gov.scot/publications/scotlands-autumn-winter-vaccination-strategy-2021/ (last accessed 15/08/2023).

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. Wightman, R.R. Kao

A Supplementary plots for time evolution of cases

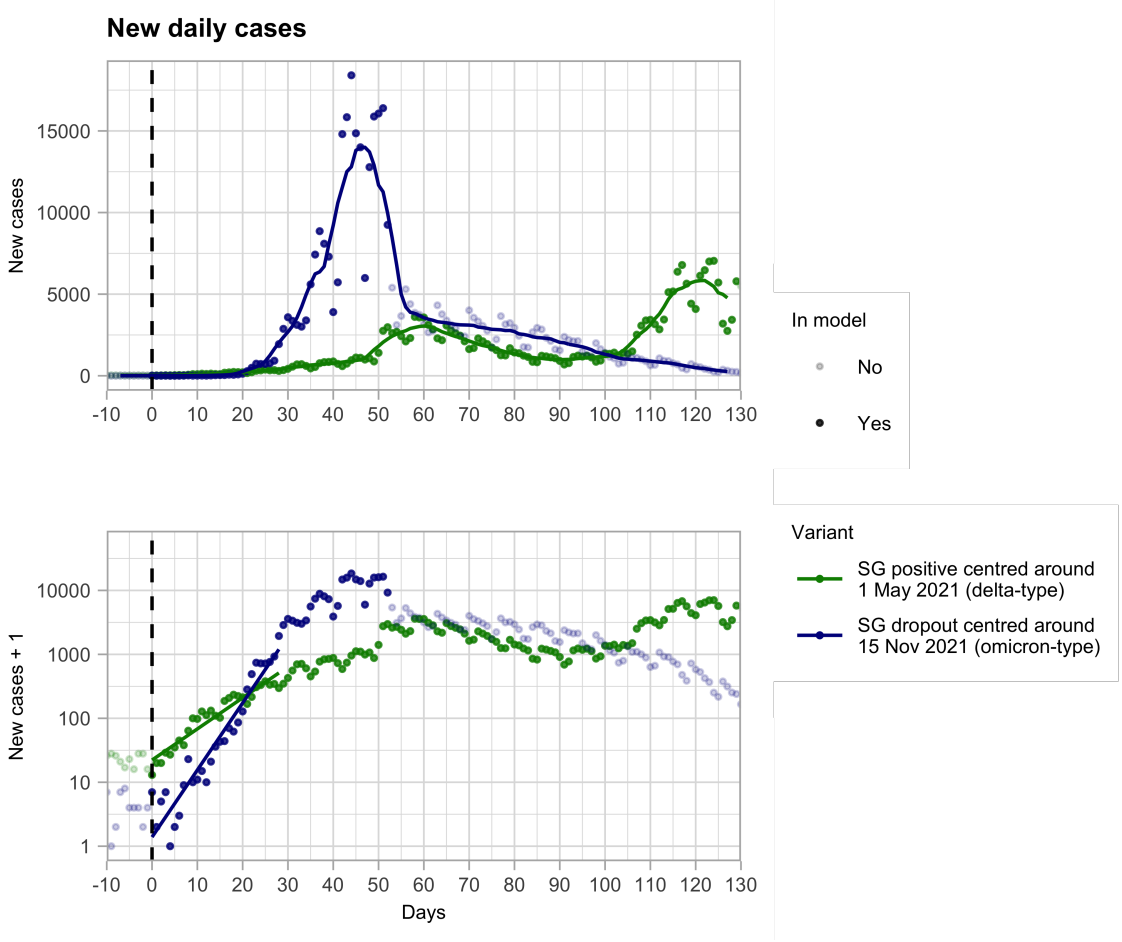

Figure S1: Timeseries of the initial outbreaks of the Delta and Omicron variants in terms of newly reported cases. The gradient of the linear regression (straight line) of the early trajectory of log(new cases + 1) is inversely proportional to the case doubling time.

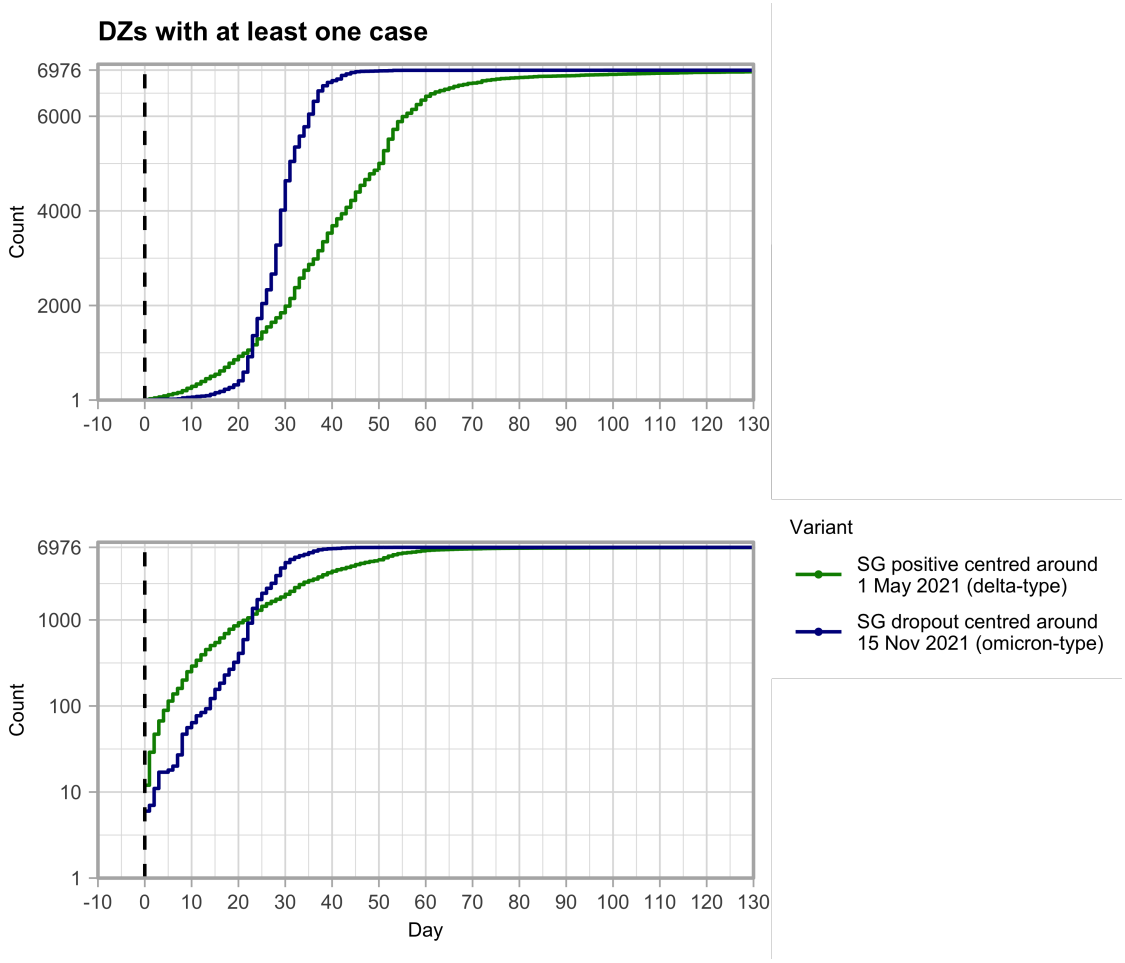

Figure S2: Timeseries of the initial outbreaks of the Delta and Omicron variants in terms of the cumulative number of DZs to have reported at least one case associated with the variant of interest.

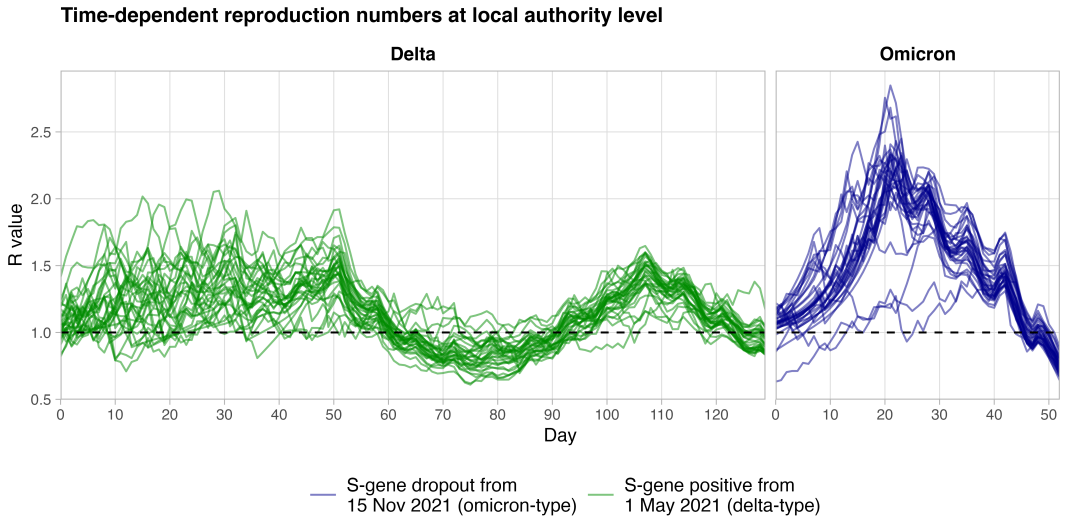

Figure S3: Time-dependent reproduction numbers for the Delta (left) and Omicron waves (right), over each of the 32 individual local authorities.

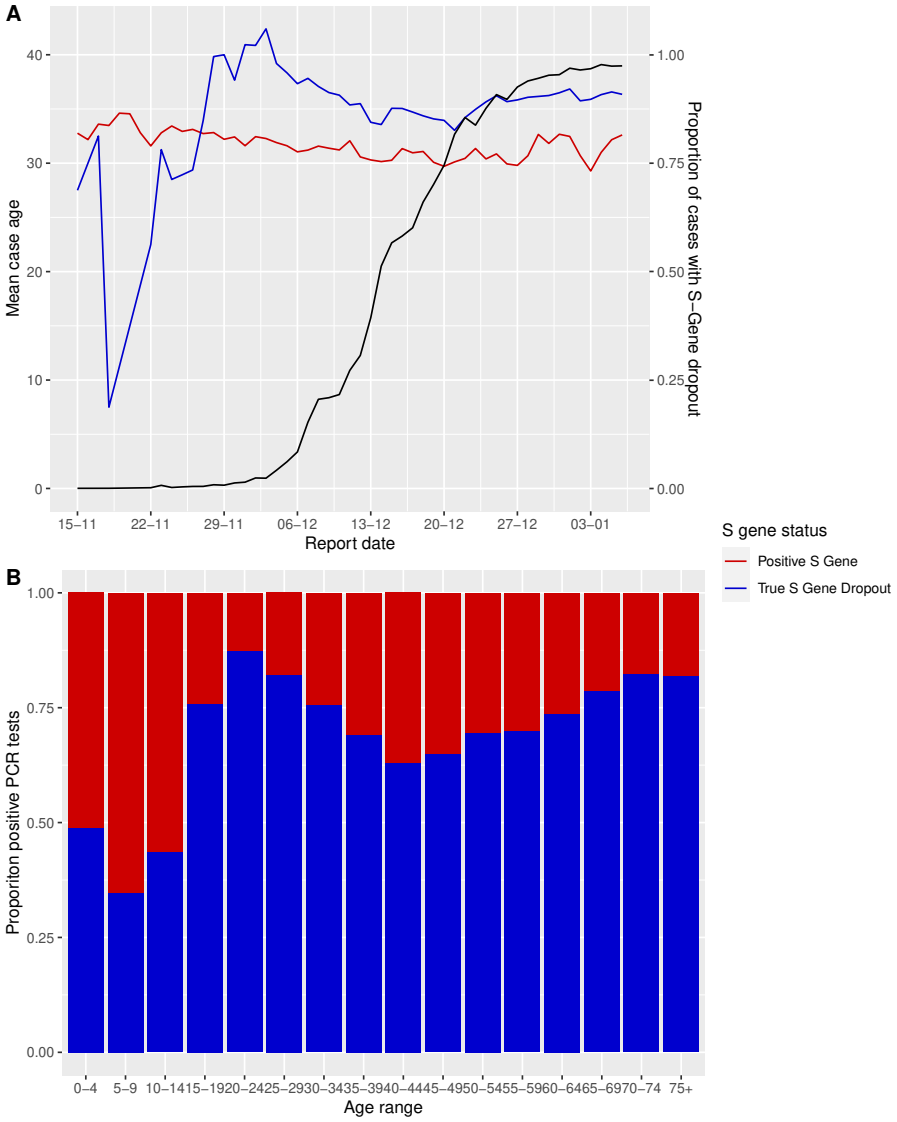

Figure S4: PCR positive cases over the period 15th November 2021 to 6th January 2022 that were S-gene dropout or true S-gene positive. (A) Daily mean case age for the two definite PCR S-gene outcomes (blue and red lines) against the proportion of the daily cases that were true S-gene dropout (presumed Omicron-type). (B) the proportion of the cases over the period by 5-year age bracket.

B Additional methodology details

B.1 Model hyperparameters

The random forest regression model is fit in *R* version 4.1.0 [68], using the *randomForest* package [69] (version 4.6-14), and ALEs analysed using the *ALEPlot* package [71] (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 used 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 building 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.

B.2 Explanatory variables used in random forest regression model

The models described in Section 4.3 are informed with the following data, first at cohort resolution:

- *Age range* (five-year windows: $[0-4]$, $[5-9]$, $\ldots$, $[70-74]$, $[75+]$), using the numeric intermediate values $2, 7, \ldots 72$, and 75 for the 75+ category;
- *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 [74], 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* [70]);
- *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 when the variant was seeded locally;
- *Student population* (% population being a full-time student aged 18 or over), also per 2011 census data;
- *Population density*, at IZ-level;
- *S-gene coverage* (the proportion of cases with an accompanying S-gene result, required to associate a likely variant) at IZ level. 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).

The measures of DZ-level deprivation included are [66]:

- *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 [64]). 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 [13]. All first doses had been offered and administered to willing adults by 18th July 2021 [75], 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 [42, 43, 44]. 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 [76]. 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 [42, 4], 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 Map views of population distribution, model residuals

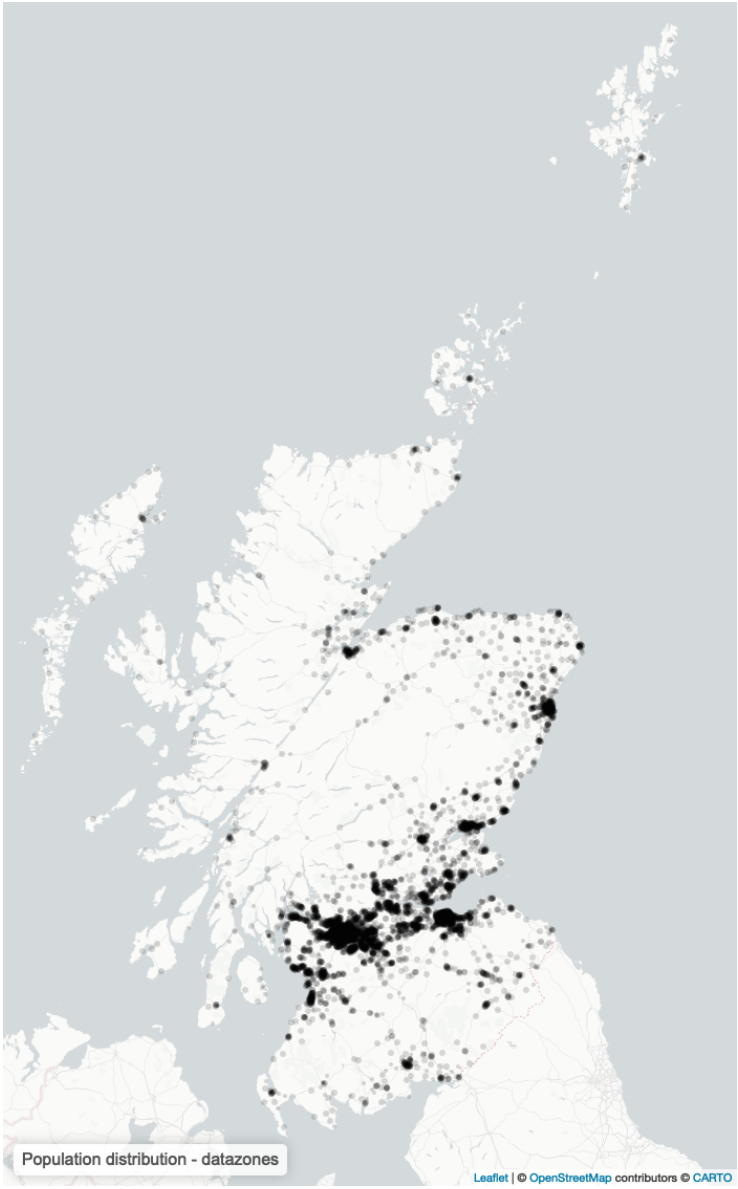

Figure S5: Distribution of population in Scotland. Each point indicates the population-weighted centroid of a datazone (DZ) of which there are 6,976 in total, with each representing a population of approximately 500-1,000 individuals.

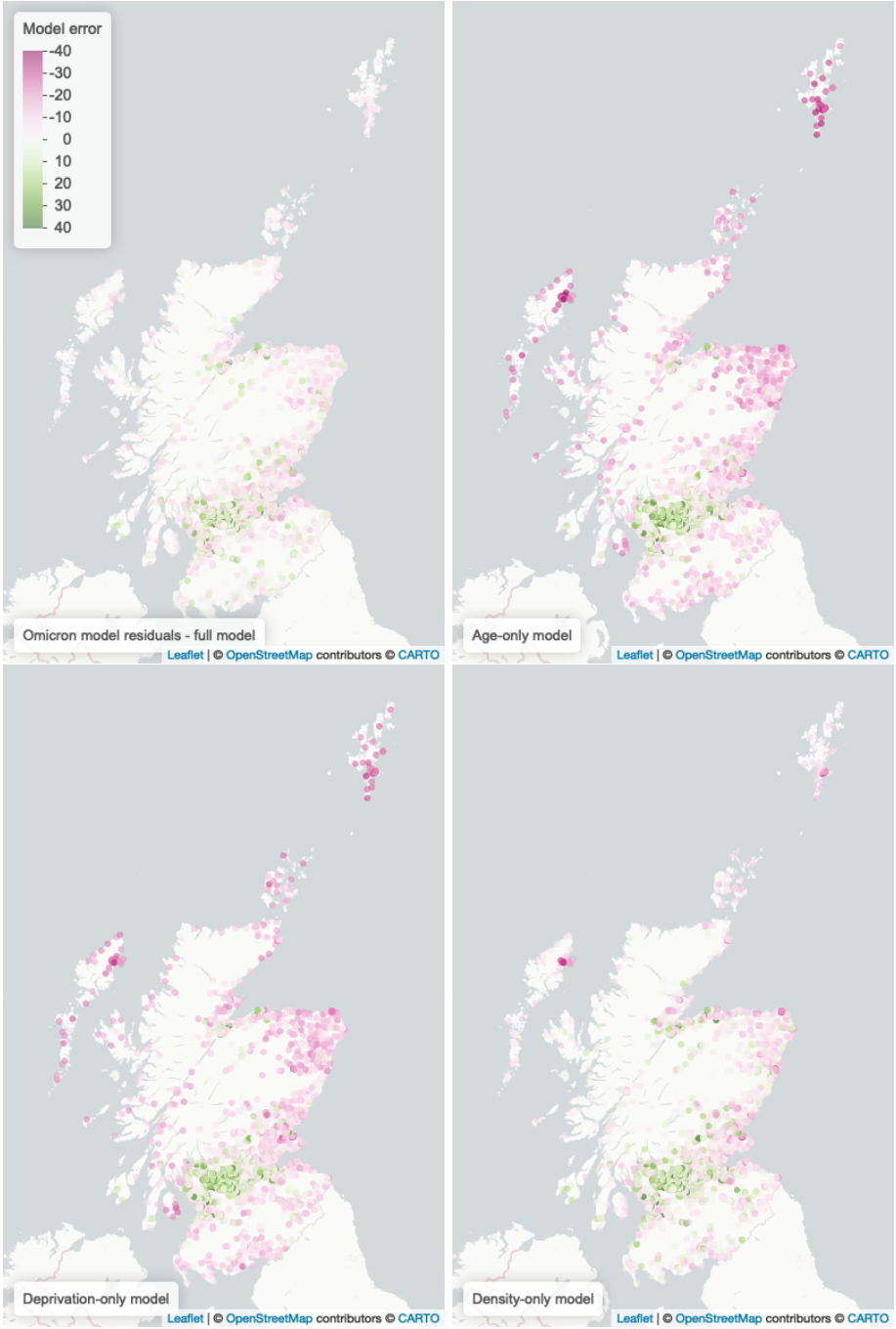

Figure S6: Comparing residuals for the distribution of Omicron cases, between the full model (top left), and the reduced models informed by population and one of age, deprivation and population density respectively. The colour scale indicates the DZ-level model error (model estimate - data), where purple points indicate DZs where the model overestimated the number of cases, and green points indicate where the model underestimated cases.

D Random forest variable importance

Fig. S7 shows variable importance measures extracted from the *RandomForest* function. Age, population and prior case status have much higher node purity (Fig. S7, top) 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 differing numbers of cases. Fig. S7, bottom, 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.

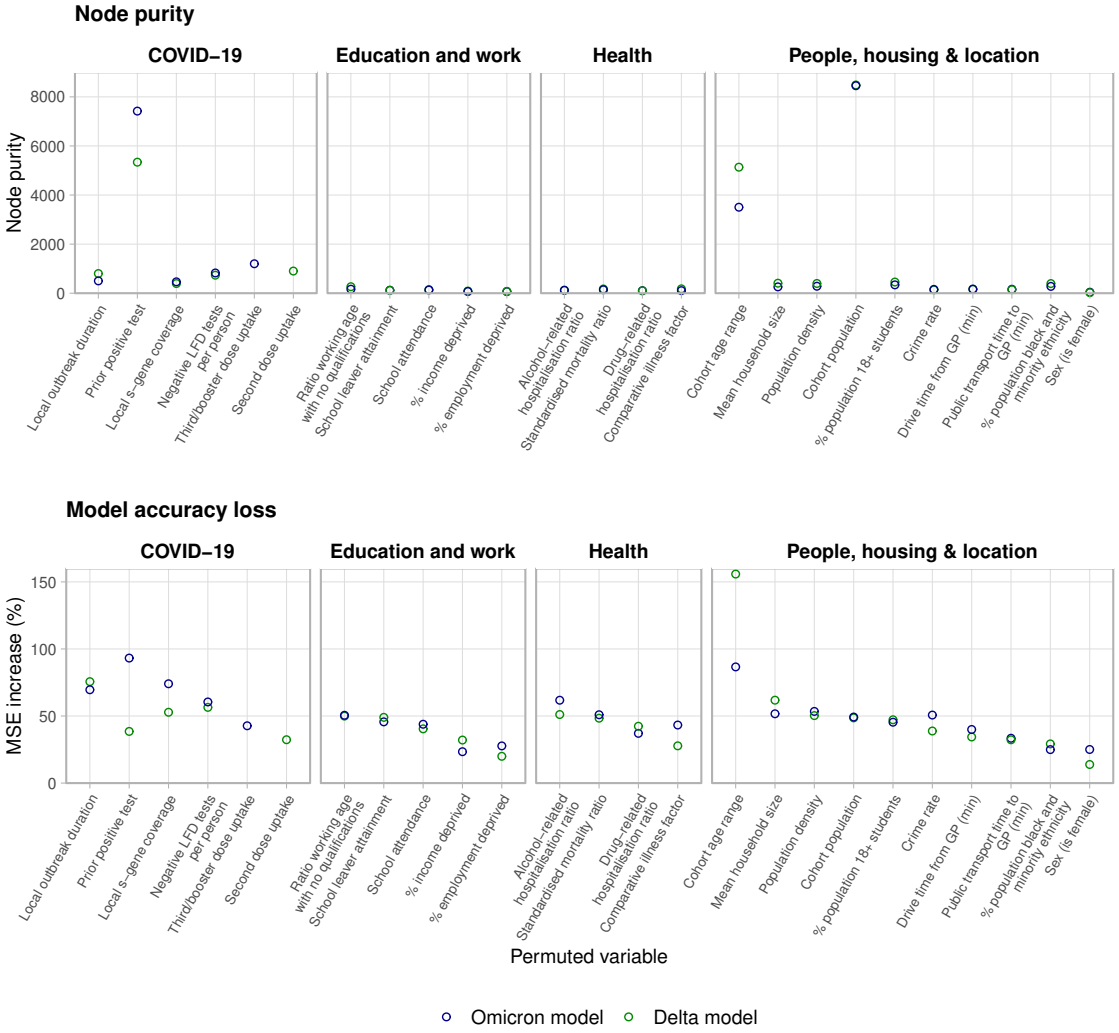

Figure S7: Feature importance outputs form the random forest regression models. Top: Node purity. Bottom: Explanatory variable mean squared error (MSE) increase on random permutation: for variable i, the increase in MSE on data not trained in each tree, if the entries of i were instead randomly permuted.

E Frequency of lateral flow testing by sex, deprivation quintile

		Population	LFD tests	Tests per person	Positive LFD tests	Positivity
Total		5,466,000	29,508,794	5.40	1,123,210	3.81%
Sex	F	2,800,788	19,639,047	7.01	647,529	3.30%
	M	2,665,212	9,869,747	3.70	475,681	4.82%
Deprivation quintile (1: most deprived)	1	1,057,767	3,827,970	3.62	176,600	4.61%
	2	1,057,929	4,992,257	4.72	201,148	4.03%
	3	1,077,589	6,023,840	5.59	220,258	3.66%
	4	1,140,448	7,134,794	6.26	256,101	3.59%
	5	1,132,267	7,529,933	6.65	269,103	3.57%

Table S7: Summary statistics of lateral flow device (LFD) tests reported in Scotland from July 2020 to February 2023, broken down by sex, and deprivation quintile of the residing datazone of individuals as ranked by the 2020 Scottish Index of Multiple Deprivation, where the most deprived datazones are in quintile 1. The test positivity is the proportion of all tests of any result that were reported as positive.

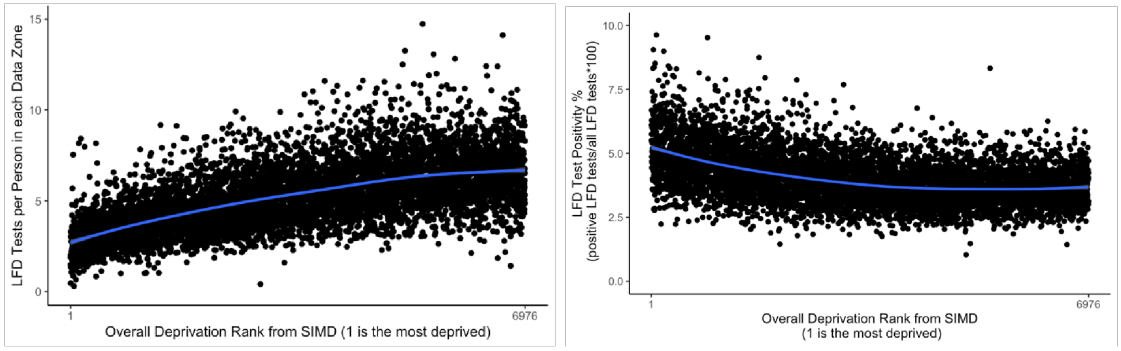

Figure S8: Lateral flow testing from July 2020 to February 2023 by datazone, ranked by deprivation per the 2020 Scottish Index of Multiple Deprivation, where the rank 1 is the datazone ranked as most deprived. Left: the number of LFD tests reported per person in each datazone. Right: the LFD test positivity, defined as the proportion of all reported LFD tests to have been positive.