**METHODS**

**Competitive advantage and increased transmissibility of the SARS-CoV2 VOC-202012/01**

To infer the competitive advantage of the VOC-202012/01 over other circulating SARS-CoV2 strains we use the COG-UK sequencing data to calculate the rate by which the strain is displacing other variants and increases in relative abundance *p*. Formally, this is quantified based on the selection rate coefficient *s*, which for a newly invading variant is defined as

(1)

This coefficient measures the rate at which any new variant would displace the resident variant in terms of the increase in the log(odds) to encounter the new variant. A great advantage of the selection rate coefficient is that it can readily be calculated from a logistic regression model as the slope of the proportion of the new variant on a logit (log-odds) link scale. We can further observe that since the ratio of relative frequencies is equal to the ratio of the absolute representation of the new variant *V* and the wild-type *W* that

(2)

Hence, if selection is density independent and there are no interactions between genotypes, the selection rate is also equal to the difference in Malthusian growth rates between the new variant (*rV*) and wild-type (*rW*) :

(3)

If we further multiply the selection rate by mean generation time *T* then we obtain the dimensionless selection coefficient

(4)

Selection coefficients and represent the most direct measures possible of the fitness advantage enjoyed by any new variant, and are the best possible predictors of whether or not it it expected to increase in frequency during an outbreak. However, assuming that the generation time of the competing variants remain unaltered (e.g. that the non-infectious period after exposure remains the same), it is also possible to relate the selection coefficient to the expected multiplicative increase in the infectiousness of the virus, as measured by the ratio of the basic reproduction number *Rt* of the new variant relative to that of the wild type. Specifically, if generation time is gamma distributed with mean *T* and *SD* *σ*, and if we set , it is the case that the basic reproduction number *Rt*

(5)

Furthermore, for small *k* (small *SD* of the generation time *σ* relative to the mean *T*), the following approximation holds

(6)

From this, it follows that the ratio of the effective reproduction number of the invading new variant *RV* relative to that of the wild type *RW*, i.e. the expected multiplicative increase in the *Rt* value *M*, assuming no change in generation time *T* between the variants, equals approximately

(7)

Although this formula is strictly speaking only exact for the limit of *k* → 0, in practice with our parameter estimates, the error made is extremely small even for larger *k*. E.g. with 1, , *T* = 5.5 days and *σ* = 1.8, *k* = 0.33 and application of the exact formula (5) would yield *M*=1.71, whilst the approximate formula (7) would yield *M*=1.73, which would amount to an error on *M* of only 1.6%. The exact formula (5) could only be used if we would be able to estimate the variant-specific intrinsic growth rates *rV* and*rW* separately, e.g. using the raw counts, to which one could fit a spline-based Poisson GLM, to yield intrinsic growth rates as the first derivative of the fitted curve on the log link scale. Such a fit, however, would show very large fluctuations due to the implementation of various non-pharmaceutical interventions, and would also require accurate corrections for changes in testing and sequencing intensity over time. Hence, such a calculation would carry a much larger error. Instead, it is much more accurate to estimate the expected multiplicative effect on from the rate of change in the log(odds) of the relative abundance of any new variant *p*, .

To estimate pairwise differences in growth rates between the VOC variant and other sets of lineages, i.e. pairwise selection rate coefficients, we used both binomial GLMMs (generalized linear mixed models), using data on the representation of pairs of lineages in the COG-UK sequencing data at time of invasion, as well as a multinomial spline regression, where we could simultaneously consider the competition for representation among all the major SARS-CoV2 variants and lineages in different regions across the UK. In both sets of analyses, we considered both the of the VOC 202012/01 (defined as lineage B.1.1.7 and carrying defining mutation N501Y and deletion ∆69/∆70 in the spike protein) as well of the earlier dominant lineage B.1.177, in relation to either all other circulating variants or a set of 440 minority variants, which never reached >15% in the aggregated UK counts in any week. For lineage B.1.177, we included any later descendent lineages into the same group.

Binomial GLMMs fit to the UK data included a fixed factor for NHS England region, a continuous covariate for sampling date, the interaction between both if this yielded a more parsimonious fit (based on the Bayesian Information Criterion) or if we were specifically interested to test for differences in rates of spread across regions, as well as random effects for the local-tier local authority (LTLA) and an observation-level random effect to take into account overdispersion. These GLMMs were fit using R’s *glmer* function in the *lme4* package version 1.1.23. For these binomial GLMMs, we used the part of the data where either variant VOC 202012/01 or lineage B.1.177 were initially invading, and for which there was good linearity on a logit scale (Fig. S2). For VOC 202012/01, we therefore used the subset of the data from August 1 2020 onwards, while for lineage B.1.177 we used data for the period between July 1st 2020 and September 30 2020, before it starting to be displaced by VOC 202012/01. From these binomial GLMMs, we subsequently estimated the selection rate Δr from the slope in the log(odds) to encounter the focal variant. Both this slope as well as it’s 95% confidence intervals were estimated using *emtrends* function in the *emmeans* R package version 1.5.1. Model predictions or marginal mean model predictions and 95% confidence intervals as well as Tukey posthoc tests to test for differences in slopes (rates of displacement of other strains) across regions were also calculated using this same package. In the calculation of marginal means, we used a bias correction for the presence of the random effects. Under the assumption of unaltered generation times, we also made two estimates of the expected multiplicative effect on the *Rt* value, *M1* and *M2*, based on eqn. (7) above, , using estimated SARS-CoV2 mean generation times *T* of either 5.5 days (Ferretti et al. 2020) or 3.6 days (Abbott et al. 2020, Ganyani et al. 2020)*.* Both the mean and confidence intervals on were exponentiated, in this way resulting in the estimated geometric mean multiplicative effect on *Rt*.

To be able to make another independent baseline estimate of outside the UK, we also used a binomial GLMM to estimate the rate of spread by which VOC 202012/01 is displacing other variants in Denmark, where SARS-CoV2 sequencing is carried out approximately randomly with respect to sample variant identity, and for which data on the incidence of the VOC 202012/01 (lineage B.1.1.7) aggregated by week and by region are openly available. These analyses either used the Danish data alone (using data from week 39 of 2020 until week 1 of 2021), or used a combined analysis of the Danish and UK data (also aggregated by region and week, to match the Danish data, and including data from August 1 2020 onwards). These analyses included region and sample date as fixed factors and continuous covariate as well as country and country x sample date in the combined DK+UK analysis, as well as an observation-level random effect to take into account overdispersion.

Finally, we also fitted a multinomial spline model to the UK data, considering the frequencies of 9 major SARS-CoV2 lineages (all reaching at least 15% in some week) as separate variant outcome levels, and subsuming the remaining 440 variants in a category of “minor variants”, thereby allowing us to simultaneously model the competition for representation among all the major variants. This model included a fixed factor region plus a natural cubic spline in function of sample date to allow for slight variation in the selection rate in function of time, plus the interaction between both to allow for different selection rates across regions. A two-degree of freedom natural cubic spline was chosen, as this model both resulted in a visually realistic fit and in a stable and realistic extrapolation (which was no longer the case for any natural cubic splines with more knots). In this multinomial model, pairwise values between variants VOC 202012/01, B.1.177 and the category of minority variants were calculated using the *emmeans emtrends* function as contrasts in the above-average growth rates of each variant (using argument mode=”latent”). Since the growth differences () in this model were time-dependent, we calculated the average growth difference for the VOC vs. minority variants and for the VOC vs. B.1.177 variant contrasts for the period from November 1 2020 onwards and from July 1st 2020 until the 30th of September 2020, respectively, when each of these variants were actively invading in the population. The predictions of this model were also used to produce Muller plots, to display the change in relative frequencies of the major SARS-CoV2 lineages over time in the UK (Fig. 1).

**Analysis of differential age susceptibility for the VOC based on secondary attack rates**

To determine if there was any difference across age cohorts in susceptibility for the new VOC-202012/01, we analysed the age-stratified aggregated data of secondary attack rates reported by Public Health England using a binomial GLM. These data comprise secondary attack rates among contact tracing data (from NHS Test and Trace) for the variant of concern (VOC 202012/01), with the identity of strain carried by the index patients (VOC or not) called based on either genomic sequence or S-gene target failure (SGTF) data, and with data split by age bracket of the person that was infected. In total, the dataset contains 17,701 and 456,086 secondary contact records of known age with index patients for which either sequence or SGTF data were available, for the period between 30 November 2020 and 20 December 2020. Out of these secondary contacts, 2,455 and 64,325 became cases, which translates into overall secondary attack rates of 13.87% and 14.10%. To determine the odds ratios for people to be infected by index patients carrying the VOC vs. by those carrying other variants, we fitted a binomial GLM with factors data type (sequence data or SGTF data), age group, variant (VOC or other strains) plus all first order interaction effects. Overdispersion was tested for by fitting an equivalent quasibinomial GLM, but was found to be absent. The R package *emmeans* was used to make effect plots of marginal and predicted means and carry out Sidak posthoc tests to test if the odds for people to be infected by index patients carrying the VOC was higher than that for those carrying other strains across the different age categories as well as overall. Possible differential age susceptibility was tested for by comparing the log(odds ratios) for people of different age to be infected by the VOC against the average log(odds ratio) for people to be infected by the VOC overall. These age group x variant interaction contrasts were again calculated using the *emmeans* package, employing a Sidak *p* value correction for multiple testing. Type III Anova tests were carried out using the *Anova* function in R’s *MASS* package.