Inferring Risks of Coronavirus Transmission from Community Household Data

Thomas House[1,2,3,*], Lorenzo Pellis[1,3], Koen B. Pouwels[4,5,6], Sebastian Bacon[7], Arturas Eidukas[8], Kaveh Jahanshahi[8], Rosalind M. Eggo[9], A. Sarah Walker[4,5,10,11], and ONS CIS Team[12]

[1]Department of Mathematics, University of Manchester, Manchester, UK
[2]IBM Research, Hartree Centre, Daresbury, UK
[3]The Alan Turing Institute for Data Science and Artificial Intelligence, London, UK
[4]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[5]The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, Oxford, UK
[6]Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[7]The DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[8]Data Science Campus, Office for National Statistics (ONS)
[9]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK
[10]The National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK
[11]MRC Clinical Trials Unit at UCL, UCL, London, UK
[12]Office for National Statistics, Newport, UK
[*]Corresponding Author: thomas.house@manchester.ac.uk

Abstract

The response of many governments to the COVID-19 pandemic has involved measures to control within- and between-household transmission, providing motivation to improve understanding of the absolute and relative risks in these contexts. Here, we perform exploratory, residual-based, and transmission-dynamic household analysis of the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS) data from 26 April 2020 to 8 March 2021 in England. This provides evidence for: (i) temporally varying rates of introduction of infection into households broadly following the trajectory of the overall epidemic; (ii) Susceptible-Infectious Transmission Probabilities (SITPs) of within-household transmission in the 15-35% range; (iii) the emergence of the B.1.1.7 variant, being around 50% more infectious within households; (iv) significantly (in the range 25-300%) more risk of bringing infection into the household for workers in patient-facing roles; (v) increased risk for secondary school-age children of bringing the infection into the household when schools are open (in the range 64-235%); (vi) increased risk for primary school-age children of bringing the infection into the household when schools were open in late autumn 2020 (around 40%).

1 Introduction

1.1 Analysis of household infection data

Households have often played an important role in infectious disease epidemiology, with policies in place and under consideration in the UK to reduce both within- and between-household transmission (Scientific Advisory Group for Emergencies, 2021). This is because the close, repeated nature of contact within the household means that within-household transmission of infectious disease is common. Also, most of the population lives in relatively small, stable households (Office for National Statistics, 2019). From the point of view of scientific research, the household is a natural unit for epidemiological data collection and households are small enough to allow for explicit solution of relatively complex transmission models. Some of the earliest work in this field was carried out by Reed and Frost, whose model was first described in the literature by Abbey (1952) in a paper that analysed transmission in boarding schools and other closed populations. Frost's 1928 lecture was later published posthumously (Frost, 1976), with a re-analysis of his original household dataset from the 1918 influenza pandemic carried out using modern computational and modelling approaches by Fraser et al. (2011).

Subsequent important contributions were made in empirical studies of transmission in households, for example the highly influential study of childhood diseases by Hope Simpson (1952), and epidemic theory based on the analyses of discrete- and continuous-time Markovian epidemics presented by Bailey (1957). A key development was the solution by Ball (1986) of the final size distribution of a random epidemic in a household without requiring Markovian recovery from infection, which then enabled statistical analyses of household infection data such as that by Addy et al. (1991). Still further progress is possible due to the use of modern computational methods, particularly Monte Carlo approaches, to augment datasets (O'Neill and Roberts, 1999; Cauchemez et al., 2004; Demiris and O'Neill, 2005) or to avoid likelihood calculations (Neal, 2012).

Continued methodological developments and data availability have enabled increasingly sophisticated inferences to be drawn from household studies of respiratory pathogens, dealing with for example interactions between adults and children (van Boven et al., 2010), case ascertainment (House et al., 2012), interactions between strains (Kombe et al., 2019), and details of family structure (Endo et al., 2019). During the current pandemic, there have been numerous household studies (Madewell et al., 2020), with three recently published studies being notable for combining fitting of a transmission model with significant differentiation of risks being those of Dattner et al. (2021), Li et al. (2021) and Reukers et al. (2021).

1.2 Context for this study

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in the human population in late 2019 and the WHO declared a pandemic in March 2020 (World Health Organization, 2020). Early in the pandemic, it became clear that risks of transmission, mortality and morbidity from the associated coronavirus disease (COVID-19) were highly heterogeneous with age (Davies et al., 2020), and also that work in patient-facing roles was associated with increased risk of positivity in the community (Pouwels et al., 2021) as would be expected given the risks of healthcare-associated transmission (Bhattacharya et al., 2021).

In the UK, a coronavirus variant of concern (VOC) 202012/01, the B.1.1.7 lineage, has largely replaced the original variant. This variant was relatively easy to track due to failure of the S gene target in commonly-used polymerase chain reaction (PCR) tests, with more details on these provided in §2.1 below. There is growing evidence for both increased transmissibility of this new variant, and increased mortality amongst infected cases (Rambaut et al., 2020; Davies et al., 2021; Challen et al., 2021; Grint et al., 2021).

Here, we apply a combination of methods, including a regression that explicitly accounts for transmission, to the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS) data from 26 April 2020 to 8 March 2021 (Pouwels et al., 2021). We particularly consider the absolute magnitude of transmission within and between households, as well as the associations between these and household size, age, infection with the VOC (inferred via S gene target failure) and work in patient-facing roles.

2 Methods

2.1 Description of data

ONS CIS[1] has a design based on variable levels of recruitment by region and time as required by policy, but otherwise uniformly random selection of households from address lists and previous ONS studies on an ongoing basis. If verbal agreement to participate is obtained, a study worker visits each household to take written informed consent, which is obtained from parents/carers for those aged 2-15 years. Participants aged 10-15 years provide written assent and those under 2 years old are not eligible.

Participants are asked questions on issues including work and age[2] as well as being tested for SARS-CoV-2 infection via reverse transcription PCR (RT-PCR). To reduce transmission risks, participants aged 12 years and over self-collect nose and throat swabs following study worker instructions, and parents/carers take swabs from children aged under 12 years. At the first visit, participants are asked for optional consent for follow-up visits every week for the next month, then monthly for 12 months from enrolment. The first few weeks of a hypothetical household participating in this study are shown schematically in Figure 1.

Swabs were analysed at the UK's national Lighthouse Laboratories at Milton Keynes and Glasgow using identical methodology. RT-PCR for three SARS-CoV-2 genes (N protein, S protein and ORF1ab) used the Thermo Fisher TaqPath RT-PCR COVID-19 Kit, and analysed using UgenTec FastFinder 3.300.5, with an assay-specific algorithm and decision mechanism that allows conversion of amplification assay raw data from the ABI 7500 Fast into test results with minimal manual intervention. Samples are called positive if at least a single N-gene and/or ORF1ab are detected. Although S gene cycle threshold (Ct) values are determined, S gene detection alone is not considered sufficient to call a sample positive.

This analysis includes all SARS-CoV-2 RT-PCR tests of nose and throat swabs from 26 April 2020 to 15 February 2021 for English households in the ONS CIS. We restrict our analysis to households of size 6 and under, partly for computational reasons that we will discuss below, and partly because this captures the overwhelming majority of households, with larger households being atypical in various ways (Office for National Statistics, 2019). In contrast to other studies, the households we select constitute an approximately representative sample from the population when stratified by date and region. The restriction to England was chosen because we split the data into four time periods, corresponding to changing situations about policies that are devolved (i.e. policies are different in

[1] https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/protocol-and-information-sheets; ISRCTN number ISRCTN21086382; The study received ethical approval from the South Central Berkshire B Research Ethics Committee (20/SC/0195).

[2] https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/case-record-forms

Scotland, Wales and Northern Ireland). These time periods split the data into the following tranches, with associated time periods and notable events (described broadly).

- **Tranche 1:** 26 April 2020 to 1 September 2020; low prevalence; schools closed; B.1.1.7 variant not emerged yet.
- **Tranche 2:** 1 September 2020 to 15 November 2020; high prevalence; schools open; negligible B.1.1.7 variant.
- **Tranche 3:** 15 November 2020 to 1 January 2021; high prevalence; schools open; B.1.1.7 variant emerged.
- **Tranche 4:** 1 January 2021 to 15 February 2021; high prevalence; schools closed (except for pre-school); B.1.1.7 variant dominant.

These properties are summarised again in Table 1. The properties of the data allocated to these tranches are shown in Table 2. Note that, while we have taken the survey data to 8 March 2021, we do not include new primary infections in households after 15 February 2021, but do include secondary infections in households where the primary infection happened before 15 February 2021. This is done to reduce problems with censoring.

2.2 Mathematical representation of data

Suppose we have a set of n individuals (participants), indexed $i, j, \ldots \in [n]$, where we use the notation $[k]$ to stand for the set of integers from 1 to k inclusive. These individuals are members of m households, and we represent the a-th household using a set of individual indices H_a. These are specified such that each individual is in exactly one household, so formally,

$$H_a \subseteq [n], \forall a \in [m] \ , \quad H_a \cap H_b = \varnothing, \forall a \in [m], b \in [m] \setminus \{a\} \ , \quad \bigcup_{a=1}^{m} H_a = [n] \ .$$

The size of the a-th household is then $n_a = |H_a|$. The a-th household is visited at a set of times $\mathcal{T}_a$, and for each $t \in \mathcal{T}_a$ we let $\mathbf{x}_{i,t}$ be the length-p feature vector (also called covariates) associated with the i-th individual at time t, and $y_{i,t}$ be the test result so that $y_{i,t} = 1$ if the swab is positive and $y_{i,t} = 0$ if not. Note that not all $i \in H_a$ will register a valid observation for features and swab results for each $t \in \mathcal{T}_a$.

We let a tranche be defined by a time interval $T = [t_1, t_2)$, and the household H_a will appear in the analysis associated with the tranche T if $\mathcal{T}_a \cap T \neq \varnothing$. For the analysis that we will perform, we require a method for associating a unique positivity and feature vector with each individual for the duration of the tranche. Under our modelling assumptions, the following definition of tranche positivity is most natural. For each household H_a associated with tranche T,

$$\forall i \in H_a \ , \quad y_i = \begin{cases} 1 & \text{if } \exists t, y_{i,t} = 1 \ \& \ \min\{\tau | \exists j \in H_a, y_{j,\tau} = 1\} \in T \ , \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This means that we associate every positive in the household with the tranche in which the first positive appears in that household. Such an approach would need revision for a situation where individuals were infected a large number of times (i.e. common reinfection) or if incidence were so high that a significant number of households would be expected to have multiple introductions, but we do not see these scenarios in our data. For features, the appropriate rule will depend on the feature. An example such rule for the case where there is only one feature $x_{i,t} \in \{0, 1\}, \forall i, t$ would be

$$x_i = \max\{x_{i,t} | t \in \mathcal{T}_a \cap T\} \ ,$$

i.e. we take this feature to be 1 if it is measured as 1 at any point during the tranche in question.

2.3 Exploratory analysis of density

An important part of our analysis will be consideration of counts / proportions of households with a given composition of cases displayed as histograms as in Figure 2, and density plots as in Figure 3.

The heights of the histogram bars are given by

$$Z_{k,\ell} = \sum_{a=1}^{m} \mathbb{1}_{\{n_a=\ell\}} \mathbb{1}_{\{\sum_{i\in H_a} y_i = k\}} \ , \quad k \in \{2,3,4,5,6\} \ , \quad \ell \in \{0,\ldots,k\} \ ,$$

where $\mathbb{1}$ stands for the indicator function. Verbally, $Z_{k,\ell}$ is the count of households of size ℓ with k participants testing positive.

The density plots are obtained by considering some feature (in this case, age) that takes values 0 or 1. We then form a point $\mathbf{r}_a \in [0,1]^2$ for each household H_a such that

$$\sum_{i\in H_a} \mathbb{1}_{\{y_i=1\}} > 0 \ , \quad \sum_{i\in H_a} \mathbb{1}_{\{x_i=1\}} > 0 \ , \quad \sum_{i\in H_a} \mathbb{1}_{\{x_i=0\}} > 0 \ ,$$

through the definition

$$\mathbf{r}_a = \left(\frac{\sum_{i\in H_a} \mathbb{1}_{\{y_i=1 \& x_i=1\}}}{\sum_{i\in H_a} \mathbb{1}_{\{x_i=1\}}}, \frac{\sum_{i\in H_a} \mathbb{1}_{\{y_i=1 \& x_i=0\}}}{\sum_{i\in H_a} \mathbb{1}_{\{x_i=0\}}} \right) \ .$$

Then we can construct a kernel density estimate in the usual way by summing then normalising kernel functions around the points, in particular the width-w square kernel function

$$\mathcal{K}(\mathbf{r}, \mathbf{r}_a) = \mathbb{1}_{\{||\mathbf{r}-\mathbf{r}_a||_\infty < w\}} \ .$$

We use age (16 years old and under versus over 16 years old) as the feature in making the density plots in Figure 3.

2.4 Residual analysis

We are also interested in tabulation of features and positives in households in a manner that allows their clustering to be assessed. In particular, this involves calculation of Pearson residuals for the within-household pairs of features and positives. Let x_i be the feature for individual i that takes values with generic labels $A, B, \ldots$ (here mainly patterns of PCR target failure indicative of viral strain). We are then interested in the table of pairs of individuals in households in the set $\mathcal{H} \subseteq [m]$ with certain properties,

$$Y_{AB} = \sum_{a\in\mathcal{H}, i\in H_a, j\in H_a\setminus\{i\}} \mathbb{1}_{\{x_i=A\}} \mathbb{1}_{\{x_j=B\}} \ .$$

Verbally, Y_{AB} is the count in the sample of A-B pairs in the set of households. On its own, this does not indicate whether A and B are more strongly associated with each other in households than would be expected from their overall prevalence, and so we want to think about what the expected value for such a quantity is under the null hypothesis that each individual picks its state with independent probabilities given by the vector $\boldsymbol{\pi} = (\pi_A)$. The standard maximum likelihood estimator for such probabilities is

$$\hat{\pi}_A = \frac{1}{n} \sum_{i\in[n]} \mathbb{1}_{\{x_i=A\}} \ .$$

The expected table under the null hypothesis of independence will then be

$$M_{AB} = \pi_A \pi_B \sum_{a \in \mathcal{H}} n_a(n_a - 1) \ .$$

The Pearson residual associated with the (A, B)-th table entry is then

$$R_{AB} = \frac{Y_{AB} - M_{AB}}{\sqrt{M_{AB}}} \ .$$

Such residuals are typically asymptotically standard normal under the null hypothesis (Bishop et al., 1975), and so values larger than 2 are indicative of significant positive correlation of the feature combination tabulated within households, and values smaller than -2 are indicative of significant negative correlation in households of the feature combination tabulated.

Here we will use pattern of PCR target failure as a feature and the restriction of households to those in which there is at least one infection (to avoid domination of the tables by all-negative households), i.e.

$$\mathcal{H} = \left\{ a \in [m] \,\middle|\, \sum_{i \in H_a} y_i > 0 \right\} \ ,$$

to produce the plots in Figure 4. Where an individual is positive on multiple visits with varying PCR gene positivity patterns, here and throughout we consider the *maximal* pattern, i.e. that containing the least target failures.

2.5 Probability model

While the more exploratory methods above are useful for formulating hypotheses, the main part of our analysis will be household regression, using time, household size and individual features to predict positivity. We start by defining a vector and matrix for each household $H_a, a \in [m]$,

$$\mathbf{y}_a := (y_i)_{i \in H_a} \ , \qquad \boldsymbol{X}_a := [(\mathbf{x}_i)_k]_{i \in H_a, k \in [p]} \ . \tag{2}$$

Note that the outcomes of swab positivity are not independent of each other due to transmission within households, but otherwise the households are selected as uniformly as possible from the population. This means that an independent-households assumption is appropriate, in which we write the likelihood function as

$$L(\boldsymbol{\theta}) = \prod_{a \in [m]} P_{\mathbf{y}_a}(\boldsymbol{X}_a, \boldsymbol{\theta}) \ . \tag{3}$$

Here, $\boldsymbol{\theta}$ is a vector of model parameters, and $P_{\mathbf{y}}$ is a function mapping a household feature matrix and set of model parameters onto a probability of a given set of household positivity outcomes. We can derive a set of equations for such probabilities from equation (4) of Addy et al. (1991) as in Kinyanjui and House (2019).

We will now write down the relevant equations for a household H of size n with outcome vector $\mathbf{y}$ and feature matrix $\boldsymbol{X}$ (i.e. suppressing the household index a to simplify notation). In particular, given a map $\iota : \{0, 1\}^n \to \{1, \ldots, 2^n\}$, we will be able to form the vector $\mathbf{P} = (P_{\iota(\mathbf{y})})_{\mathbf{y} \in \{0,1\}^n}$ of probabilities of different outcomes in the household. This will be a solution to the set of linear equations

$$\boldsymbol{B}(\boldsymbol{\theta})\mathbf{P} = \mathbf{1} \ , \tag{4}$$

where $\mathbf{1}$ is a length-2^n vector of all ones, and $\boldsymbol{B} = [B_{\iota(\boldsymbol{\nu}),\iota(\boldsymbol{\omega})}]_{\boldsymbol{\nu},\boldsymbol{\omega}\in\{0,1\}^n}$, which has

$$B_{\iota(\boldsymbol{\nu}),\iota(\boldsymbol{\omega})} = \mathcal{B}_{\boldsymbol{\nu},\boldsymbol{\omega}} = \frac{1}{\prod_{j\in H} \Phi\left(\sum_{i\in H}(1-\nu_i)\lambda_{ij}\right)^{\omega_j} Q_j^{1-\nu_j}} \ , \quad \boldsymbol{\nu} \leq \boldsymbol{\omega} \in \{0,1\}^n \ , \tag{5}$$

and other elements equal to zero, where we write $\leq$ between vectors to stand for the statement that each element on the left-hand side is less than or equal to the corresponding element on the right-hand side. The associated condition imposes that each $\boldsymbol{\nu}$ above will correspond to a sub-epidemic of $\boldsymbol{\omega}$ meaning that the equation (4) can be solved iteratively. There are then three main ingredients of the transmission model that we will enumerate below and in doing so define the terms in Equation (5).

The first model component is the probability of avoiding infection from outside; for the i-th individual this is

$$Q_i = \mathrm{e}^{-\Lambda_i} \ , \qquad \Lambda_i = \Lambda \mathrm{e}^{\boldsymbol{\alpha}\cdot\mathbf{x}_i} = \mathrm{e}^{\alpha_0 + \boldsymbol{\alpha}\cdot\mathbf{x}_i} \ . \tag{6}$$

In the language of infectious disease modelling, Λ_i is the cumulative force of infection experienced by the i-th individual. Then $\exp(\alpha_k)$ is the relative external exposure associated with the k-th feature / covariate, meaning that it is the multiplier in front of the baseline force of infection, which is that for an individual whose feature vector is all zeros, $\mathbf{0}$. This baseline probability of avoiding infection from outside is then

$$q = \exp(-\Lambda) = \exp(-\exp(\alpha_0)) \ , \tag{7}$$

and we will report $(1-q)$ in tables, alongside the relative external exposures that are elements of the vector $\boldsymbol{\alpha}$, although it would also be possible to use (7) to relate this to the baseline force of infection Λ or intercept of the linear predictor, α_0. Note that some care must be taken in interpretation of this variable when the data are split into time periods as in this work, since to appear as a household with at least one positive in one tranche, it is necessary to appear as a household with no positives in the previous tranches for which the household was in the study. Values of $1-q$ will typically be low enough here that this conditional dependence is not strong, but this might not be true at higher levels of incidence for the same design.

The second component of the model is variability in the infectiousness at the individual level, usually interpreted as arising from the distribution of infectious periods. We assume that each individual picks from a unit-mean Gamma distribution since this provides a natural one-parameter distribution with appropriate support. The Laplace transform of this is used in (5) and is

$$\Phi(s) = (1 + \vartheta s)^{-1/\vartheta} \ . \tag{8}$$

The parameter ϑ is the variance of the Gamma distribution, i.e. it is larger for more individual variability.

The third component of the model is the infection rate from individual j to individual i,

$$\lambda_{ij} = n^\eta \lambda \sigma_i \tau_j = n^\eta \lambda \mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \mathrm{e}^{\boldsymbol{\gamma}\cdot\mathbf{x}_j} = \mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{x}_i} \mathrm{e}^{\gamma_0 + \eta\log(n) + \boldsymbol{\gamma}\cdot\mathbf{x}_j} \ . \tag{9}$$

In this equation: λ is the baseline rate of infection; $\sigma_i = \mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{x}_i}$ is the relative susceptibility of the i-th participant, and $\exp(\beta_k)$ is the relative susceptibility associated with the k-th feature; $\tau_j = \mathrm{e}^{\boldsymbol{\gamma}\cdot\mathbf{x}_j}$ is the relative transmissibility of the j-th participant, and $\exp(\gamma_k)$ is the relative transmissibility associated with the k-th feature / covariate. As can be seen from (9), we can interpret $\log(\lambda)$ as the intercept of the linear predictor for transmissibility. The term n^η is a modelling approach to the effect of household size usually attributed to Cauchemez et al. (2004); as can be seen from (9), this is equivalent to taking $\log(n)$ as a covariate for transmissibility. Experience with fitting these models (Kinyanjui et al., 2018) suggests that it is a good idea to impose hard bounds on the Cauchemez parameter, i.e. insist that $\eta \in [\eta_{\min}, \eta_{\max}]$, meaning that here we will treat η separately from other parameters.

2.6 Model variables and fitting

We now enumerate all of the model parameters, distinguishing between the 'natural' representations of parameters that sit in $\mathbb{R}$ and transforms of natural parameter space $\mathbb{R}^\kappa$ that are most epidemiologically interpretable and therefore suitable for reporting. Since Λ, λ and ϑ have positive support, we can use logarithmic and exponential functions to transform between epidemiological and natural parameters. As noted above, we want η to have compact support, and so note that the function $\tan : [-\pi/2, \pi/2] \to \mathbb{R}$ and its inverse can be used. We choose $\eta_{\mathrm{min}} = -2$ and $\eta_{\mathrm{max}} = 2$, meaning that our natural parameter vector is $\boldsymbol{\theta} = (\log(\Lambda), \log(\lambda), \log(\vartheta), \tan(\pi\eta/4), \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^\kappa$.

The first part of this parameter vector is the external force of infection, with natural representation $\alpha_0 = \log(\Lambda)$. Here we will quote the baseline probability of infection from outside, which is $1 - q$ for q as in (7).

The second part of the parameter space relates to baseline within-household transmission with natural representation $\gamma_0 = \log(\lambda)$, $\log(\vartheta)$, and $\tan(\pi\eta/4)$, where we use this transform for η to make a hard constraint of epidemiologically meaningful values. For interpretability, we work with probabilities of infection by household size, which from (5) are

$$p_n = 1 - \Phi(n^\eta \lambda) \ . \tag{10}$$

Such quantities have been called Susceptible-Infectious Transmission Probabilities (SITP) by e.g. Fraser et al. (2011), who estimated values close to 20% from historical data on the 1918 influenza pandemic.

The third part are features, where we consider:

- Three age groups: 2-11 years old; 12-16 years old; and older.
- Working in a patient-facing role or not.
- Pattern of PCR gene target positivity: OR+N+S, which is inconsistent with the B.1.1.7 variant; OR+N, which is consistent with the B.1.1.7 variant; or other, which is usually indicative of too low a viral load to be confident in strain.

We assume that age and working in a patient-facing role have an association with external risk, leading to natural parameters $\alpha_{\text{2-11}}$, $\alpha_{\text{12-16}}$ and α_{PF}; that age has an association with susceptibility, leading to natural parameters $\beta_{\text{2-11}}$ and $\beta_{\text{12-16}}$; and that age and gene positivity in PCR have an association with transmissibility, leading to natural parameters $\gamma_{\text{2-11}}$, $\gamma_{\text{12-16}}$, $\gamma_{\mathrm{OR+N}}$ and γ_{oth}. For any natural parameter r, we will report the multiplicative effect $\exp(r)$.

Model fitting was performed in an approximate Bayesian framework using the Laplace approximation. As noted above, households of size 7 and larger were excluded from the analysis partly because these are often very different in composition from smallerhouseholds, and partly because of the numerical cost of solving a linear system of size 2^{2n} We combine the likelihood (3) with a standard normal prior on natural parameters, $\boldsymbol{\theta} \sim \mathcal{N}_\kappa(\mathbf{0}, \boldsymbol{I})$. Sensitivity of results to this prior was considered by hand and revealed essentially no impact on the highly identifiable parameters such as Λ and λ, and that while a higher variance could slightly reduce the shrinkage of effect sizes towards zero, it could also lead to instability in fitting, meaning that this prior achieves regularisation of the inference problem without excessive bias. The maximum a posteriori estimate was obtained using multiple restarts of a Quasi-Newton optimiser. The Hessian was calculated numerically for the natural parameters and used in the Laplace approximation to the posterior on the natural parameters. The credible intervals (CIs) are then transformed from natural to epidemiologically interpretable parameters.

2.7 Data processing and software implementation

The analysis was carried out on the ONS Secure Research Server in the Python 3 language. To illustrate issues with data processing, note that the ‘flat’ form for the data extracted from the database after cleaning takes a form like:

```
    HID   PID   Visit Date   Age   Test Result   Work PF   Pattern
    ...
    123   456   2020-10-02   8     Negative      No        NA
    123   457   2020-10-02   38    Negative      No        NA
    123   456   2020-10-10   8     Negative      No        NA
    123   457   2020-10-10   38    Positive      No        OR+N+S
    123   456   2020-10-17   9     Positive      No        OR+N+A
    123   457   2020-10-17   38    Negative      No        NA
    ...
    124   458   2021-02-15   53    Negative      Yes       NA
    ...
```

In particular, there is a hierarchical structure to the data. Households, each with a unique household ID in the `HID` column, have a number of study participants with a unique participant ID in the `PID` column, and each participant being visited on a number of dates as in the `Visit Date` column. Each visit will have associated participant features (e.g. as in the `Age` column above) and a `Test Result`.

The large size of this flat file (slightly under three million rows) means that it is advantageous to use specialist libraries, in this case *pandas* (The pandas development team, 2020; McKinney, 2010) together with *NumPy* (Harris et al., 2020). To deal with the nested structure of the data, we used the ‘split-apply-combine’ paradigm that this library encourages by analogy with SQL operations. In the example above, this would involve first associating each participant with an age using `pandas.groupby('PID')` and `pandas.DataFrame.apply(numpy.min)`, and then producing an array of ages for each household using `pandas.groupby('HID')` and `pandas.DataFrame.apply(numpy.array)`. A similar approach is possible for test results and multiple features.

Apart from data processing, the main computational cost of the analysis is the linear algebra associated with solving (4), particularly for larger households. Due to portability, this was carried out in NumPy on the ONS system, however we found that implementation in *Numba* (Lam et al., 2015) can generate significant speed-ups, as might use of GPU hardware through use of e.g. *PyTorch* (Paszke et al., 2019).

Code is available on request from the corresponding author, and is being prepared for release on a public repository.

3 Results and Discussion

3.1 Exploratory analysis

Figure 2 shows the distribution of positives in households; comparison with Table 2 shows that the numbers of households with two or more positives are much greater than would be expected under the assumption of independence. In fact, some histograms even take a bimodal ‘U’ shape. This multi-modality is even more apparent in the kernel plots in Figure 3, which also demonstrate that it is common to see households with only child positives. While this could be due to failure of ascertainment of an adult case in the household, such a pattern is suggestive that in most such households, the introduction of infection to the household would have been due to a child.

3.2 Residual analysis

The Pearson residual analysis, together with pair counts, applied to the maximal PCR target gene positivity pattern being OR+N+S, OR+N, other positive, or negative, is shown in Figure 4. The pair counts show the replacement of the OR+N+S pattern as the main source of positive pairs in households with the OR+N pattern. We also see that while there is clustering of (OR+N+S)-(OR+N+S) and (OR+N)-(OR+N) pairs, as well as negative-negative pairs, there is a negative correlation associated with (OR+N+S)-(OR+N) pairs and also between pairs involving any other positive pattern. While this analysis is not causal, we expect that the main mechanism generating clustering of positives in households is transmission. As such, the results are consistent with our understanding of the B.1.1.7 variant as a more transmissible strain that replaced the original variant in England.

3.3 Regression analysis

The regression analysis has its outputs shown in Table 3, Figure 5, and Figure 6. The results can be roughly split into those that change with time and those that do not.

Conclusions that are relatively stable over time include: the approximate size of the transmission probabilities p_n (subject to a rising and falling pattern over time that may merit further investigation); that patient-facing roles are associated with increased risk of external exposure; that there is little evidence for a difference between within-household transmissibility and susceptibility for adults and children (although this needs further analysis to look at e.g. viral load as a mediating factor); and that maximal PCR gene positivity patterns other than OR+N+S or OR+N are associated with lower transmissibility, which would be expected given that such patterns are associated with lower viral load (Walker et al., 2021).

Conclusions that do change over time include: the overall risk of external introduction from outside the household, as would be expected from the overall changes in prevalence and incidence for the epidemic; the transmissibility of the OR+N maximal PCR gene positivity pattern moving from less than OR+N+S to more, as would be expected from S gene target failure changing from indicative of low viral load to indicative of the B.1.1.7 variant; and the risk of external introduction in school-age children. On this last point, both primary and secondary school-aged children are associated with greater external infection risk when schools are open. The exception to this pattern is primary school children before the emergence of the new variant, who exhibit a relative effect that is not significantly different from zero. Whether this change in association is due to some causal factor not accounted for here, or is related to the new variant spreading more efficiently amongst young children than wildtype, requires further investigation.

3.4 Limitations and directions for future work

While we have taken many steps to ensure that the results presented here are as robust as possible, there are key limitations to the analysis that need to be borne in mind. The main one of these is failures in ascertainment of positives and other missingness in the longitudinal design in question. The most likely consequence of this will be to depress susceptible-infectious transmission probability estimates. One theoretical approach to deal with this would be imputation of the transmission tree as suggested by Demiris and O'Neill (2005), but this is likely to be too computationally intensive to be practical in the current context. Another would be analytical work to include failure of ascertainment into the likelihood function as in House et al. (2012), however it is unclear how to model ascertainment probabilistically in a tractable manner. A data-driven approach would be to try to include positives from other sources such as Test and Trace case data or self-reported episodes of illness. The next most important limitation is the possibility that inclusion of other features, for example the geographical region that households are in, more detailed information about viral load and symptoms, or information about the physical structure of the household, which might play an important explanatory role in the associations observed. Finally, there are possible refinements of the work: trends in external infection over time could be modelled as a flexible functional form (e.g. a spline as in Pouwels et al. (2021)); features could be selected using formal criteria, including relaxing of the Cauchemez assumption to allow transmission probabilities to depend in a general manner on household size, and explicit correction to attack rates due to shrinking and growing epidemics could be made as proposed by Ball and Shaw (Ball and Shaw, 2015; Shaw, 2016); model parameters – e.g. baseline transmission probabilities – could be shared across tranches; the work coule be extended to Wales, Scotland and Northern Ireland; more formal analysis of causal pathways could be performed; and improvements could be made in implementation data processing, model evaluation through improved linear algebra, and fitting algorithm. These and other directions should be the subject of future studies.

Acknowledgements

The ONS CIS is funded by the Department of Health and Social Care with in-kind support from the Welsh Government, the Department of Health on behalf of the Northern Ireland Government and the Scottish Government. TH is supported by the Royal Society (grant number INF/R2/180067). LP is supported by the Wellcome Trust and the Royal Society (grant number 202562/Z/16/Z). TH and LP are also supported by the UK Research and Innovation COVID-19 rolling scheme (grant numbers EP/V027468/1, MR/V028618/1 and MR/V038613/1) as well as the Alan Turing Institute for Data Science and Artificial Intelligence. SB is supported by the Wellcome Trust, the Medical Research Council and UK Research and Innovation. KBP and ASW are supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) (NIHR200916). KBP is also supported by the Huo Family Foundation. ASW is also supported by the NIHR Oxford Biomedical Research Centre, by core support from the Medical Research Council UK to the MRC Clinical Trials Unit (MC_UU_12023/22), and is an NIHR Senior Investigator. RE is funded by HDR UK (grant number MR/S003975/1), the MRC (grant number MC_PC 19065) and the NIHR (grant number NIHR200908). TH, LP and the ONS CIS team would like to thank Frank Ball for extremely valuable comments on the work. The views expressed are those of the authors and not necessarily those of their employers, funders, the National Health Service, NIHR, Department of Health, or PHE.

References

H. Abbey. An examination of the Reed-Frost theory of epidemics. *Human Biology*, 24(3):201–233, 1952.

C. L. Addy, I. M. Longini, and M. Haber. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47(3):961–974, 1991.

N. T. J. Bailey. *The Mathematical Theory of Epidemics*. Griffin, London, 1957.

F. Ball and L. Shaw. Estimating the within-household infection rate in emerging SIR epidemics among a community of households. *Journal of Mathematical Biology*, 71(6-7):1705–1735, 2015.

F. G. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemics models. *Advances in Applied Probability*, 18(2):289–310, 1986.

A. Bhattacharya, S. M. Collin, J. Stimson, S. Thelwall, O. Nsonwu, S. Gerver, J. Robotham, M. Wilcox, S. Hopkins, and R. Hope. Healthcare-associated COVID-19 in England: a national data linkage study. *medRxiv*, 2021. DOI: 10.1101/2021.02.16.21251625.

Y. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Massachusetts Institute of Technology Press, Cambridge, 1975.

S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boëlle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23(22):3469–3487, 2004.

R. Challen, E. Brooks-Pollock, J. M. Read, L. Dyson, K. Tsaneva-Atanasova, and L. Danon. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*, 372:n579, 2021.

I. Dattner, Y. Goldberg, G. Katriel, R. Yaari, N. Gal, Y. Miron, A. Ziv, R. Sheffer, Y. Hamo, and A. Huppert. The role of children in the spread of COVID-19: Using household data from Bnei Brak, Israel, to estimate the relative susceptibility and infectivity of children. *PLOS Computational Biology*, 17(2):1–19, 2021.

N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, C. A. B. Pearson, B. J. Quilty, A. J. Kucharski, H. Gibbs, S. Clifford, A. Gimma, K. van Zandvoort, J. D. Munday, C. Diamond, W. J. Edmunds, R. M. G. J. Houben, J. Hellewell, T. W. Russell, S. Abbott, S. Funk, N. I. Bosse, Y. F. Sun, S. Flasche, A. Rosello, C. I. Jarvis, R. M. Eggo, and CMMID COVID-19 working group. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature Medicine*, 26(8): 1205–1211, 2020.

N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, CMMID COVID-19 working group, K. COVID-19 Genomics UK (COG-UK) Consortium Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, and W. J. Edmunds. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, page eabg3055, 2021. DOI: 10.1126/science.abg3055.

N. Demiris and P. D. O'Neill. Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):731–745, 2005.

A. Endo, M. Uchida, A. J. Kucharski, and S. Funk. Fine-scale family structure shapes influenza transmission risk in households: Insights from primary schools in Matsumoto city, 2014/15. *PLOS Computational Biology*, 15(12):e1007589, 2019.

C. Fraser, D. A. T. Cummings, D. Klinkenberg, D. S. Burke, and N. M. Ferguson. Influenza transmission in households during the 1918 pandemic. *American Journal of Epidemiology*, 174(5):505–514, 2011.

W. H. Frost. Some conceptions of epidemics in general. *American Journal of Epidemiology*, 103(2): 141–151, 1976.

D. J. Grint, K. Wing, E. Williamson, H. I. McDonald, K. Bhaskaran, D. Evans, S. J. Evans, A. J. Walker, G. Hickman, E. Nightingale, A. Schultze, C. T. Rentsch, C. Bates, J. Cockburn, H. J. Curtis, C. E. Morton, S. Bacon, S. Davy, A. Y. Wong, A. Mehrkar, L. Tomlinson, I. J. Douglas, R. Mathur, P. Blomquist, B. MacKenna, P. Ingelsby, R. Croker, J. Parry, F. Hester, S. Harper, N. J. DeVito, W. Hulme, J. Tazare, B. Goldacre, L. Smeeth, and R. M. Eggo. Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. *Eurosurveillance*, 26(11):2100256, 2021.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

R. E. Hope Simpson. Infectiousness of communicable diseases in the household: (measles, chickenpox, and mumps). *The Lancet*, 260(6734):549–554, 1952. Originally published as Volume 2, Issue 6734.

T. House, N. Inglis, J. V. Ross, F. Wilson, S. Suleman, O. Edeghere, G. Smith, B. Olowokure, and M. J. Keeling. Estimation of outbreak severity and transmissibility: Influenza A(H1N1)pdm09 in households. *BMC Medicine*, 10(117):https://doi.org/10.1186/1741–7015–10–117, 2012.

T. Kinyanjui and T. House. Generalised linear models for dependent binary outcomes with applications to household stratified pandemic influenza data, 2019. [arXiv:1911.12115].

T. Kinyanjui, J. Middleton, S. Güttel, J. Cassell, J. Ross, and T. House. Scabies in residential care homes: Modelling, inference and interventions for well-connected population sub-units. *PLOS Computational Biology*, 14(3):e1006046, 2018.

I. K. Kombe, P. K. Munywoki, M. Baguelin, D. J. Nokes, and G. F. Medley. Model-based estimates of transmission of respiratory syncytial virus within households. *Epidemics*, 27:1–11, 2019.

S. K. Lam, A. Pitrou, and S. Seibert. Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, number 7 in LLVM '15, 2015. DOI: 10.1145/2833157.2833162.

F. Li, Y.-Y. Li, M.-J. Liu, L.-Q. Fang, N. E. Dean, G. W. K. Wong, X.-B. Yang, I. Longini, M. E. Halloran, H.-J. Wang, P.-L. Liu, Y.-H. Pang, Y.-Q. Yan, S. Liu, W. Xia, X.-X. Lu, Q. Liu, Y. Yang, and S.-Q. Xu. Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: a retrospective observational study. *The Lancet Infectious Diseases*, 2021. https://doi.org/10.1016/S1473-3099(20)30981-6.

Z. J. Madewell, Y. Yang, J. Longini, Ira M., M. E. Halloran, and N. E. Dean. Household transmission of SARS-CoV-2: A systematic review and meta-analysis. *JAMA Network Open*, 3(12):e2031756, 2020.

W. McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. DOI: 10.25080/Majora-92bf1922-00a.

P. Neal. Efficient likelihood-free Bayesian computation for household epidemics. *Statistics and Computing*, 22(6):1239–1256, 2012.

Office for National Statistics. Families and households, Edition: 15 November, 2019. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds.

P. D. O'Neill and G. O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

K. B. Pouwels, T. House, E. Pritchard, J. V. Robotham, P. J. Birrell, A. Gelman, K.-D. Vihta, N. Bowers, I. Boreham, H. Thomas, J. Lewis, I. Bell, J. I. Bell, J. N. Newton, J. Farrar, I. Diamond, P. Benton, A. S. Walker, D. Crook, P. C. Matthews, T. Peto, N. Stoesser, A. Howarth, G. Doherty, J. Kavanagh, K. K. Chau, S. B. Hatch, D. Ebner, L. Martins Ferreira, T. Christott, B. D. Marsden, W. Dejnirattisai, J. Mongkolsapaya, S. Hoosdally, R. Cornall, D. I. Stuart, G. Screaton, D. Eyre, J. Bell, S. Cox, K. Paddon, T. James, J. N. Newton, J. V. Robotham, P. Birrell, H. Jordan, T. Sheppard, G. Athey, D. Moody, L. Curry, P. Brereton, J. Hay, H. Vansteenhouse, A. Lambert, E. Rourke, S. Hawkes, S. Henry, J. Scruton, P. Stokes, T. Thomas, J. Allen, R. Black, H. Bovill, D. Braunholtz, D. Brown, S. Collyer, M. Crees, C. Daglish, B. Davies, H. Donnarumma, J. Douglas-Mann, A. Felton, H. Finselbach, E. Fordham, A. Ipser, J. Jenkins, J. Jones, K. Kent, G. Kerai, L. Lloyd, V. Masding, E. Osborn, A. Patel, E. Pereira, T. Pett, M. Randall, D. Reeve, P. Shah, R. Snook, R. Studley, E. Sutherland, E. Swinn, A. Tudor, J. Weston, S. Leib, J. Tierney, G. Farkas, R. Cobb, F. Van Galen, L. Compton, J. Irving, J. Clarke, R. Mullis, L. Ireland, D. Airimitoaie, C. Nash, D. Cox, S. Fisher, Z. Moore, J. McLean, and M. Kerby. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *The Lancet Public Health*, 6(1):e30–e38, 2021.

A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, A. Carabelli, T. Connor, T. Peacock, D. L. Robertson, E. Volz, and COVID-19 Genomics Consortium UK (CoG-UK). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

D. F. M. Reukers, M. van Boven, A. Meijer, N. Rots, C. Reusken, I. Roof, A. B. van Gageldonk-Lafeber, W. van der Hoek, and S. van den Hof. High infection secondary attack rates of SARS-CoV-2 in Dutch households revealed by dense sampling. *Clinical Infectious Diseases*, page ciab237, 2021.

Scientific Advisory Group for Emergencies. Reducing within- and between-household transmission in light of new variant SARS-CoV-2, 14 January, 2021. Paper prepared by the Environmental Modelling Group (EMG), the Scientific Pandemic Insights Group on Behaviours (SPI-B) and the Scientific Pandemic Influenza Group on Modelling (SPI-M).

L. Shaw. *SIR epidemics in a population of households.* PhD thesis, The University of Nottingham, 2016. http://eprints.nottingham.ac.uk/38606/1/LaurenceShawThesis4185911.pdf.

The pandas development team. Pandas, 2020. https://doi.org/10.5281/zenodo.3509134.

M. van Boven, T. Donker, M. van der Lubben, R. B. van Gageldonk-Lafeber, D. E. te Beest, M. Koopmans, A. Meijer, A. Timen, C. Swaan, A. Dalhuijsen, S. Hahné, A. van den Hoek, P. Teunis, M. A. B. van der Sande, and J. Wallinga. Transmission of novel influenza A(H1N1) in households with post-exposure antiviral prophylaxis. *PLOS ONE*, 5(7):e11442, 2010.

A. S. Walker, E. Pritchard, T. House, J. V. Robotham, P. J. Birrell, I. Bell, J. I. Bell, J. N. Newton, J. Farrar, I. Diamond, R. Studley, J. Hay, K.-D. Vihta, T. E. A. Peto, N. Stoesser, P. C. Matthews, D. W. Eyre, K. Pouwels, and the COVID-19 Infection Survey team. Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. *medRxiv*, 2021. DOI: 10.1101/2020.10.25.20219048.

World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020, 2020.

Tranche	Start date	End date	Prevalence	Schools	B.1.1.7 variant
1	26-Apr-20	1-Sep-20	Low	Closed	Not emerged
2	1-Sep-20	15-Nov-20	High	Open	Negligible
3	15-Nov-20	1-Jan-21	High	Open	Emerged
4	1-Jan-21	15-Feb-21	High	Mainly closed	Dominant

Table 1: Summary of properties of the time periods (tranches) that the data are split into for analysis.

	Tranche 1	Tranche 2	Tranche 3	Tranche 4	Overall
Number of participants	89624	293570	315187	329532	371420
Number of households	43300	144904	157432	165238	181710
Number of positive individuals	242	5625	6078	6925	19548
Households with 1+ positive	206	4074	4433	5123	14345
OR+N+S positives	124	4051	2263	695	7151
OR+N positives	17	614	2690	4533	8299
Children <12	7483	23257	24045	24686	29793
Children 12–16	4815	15503	16790	18098	20091
Patient-facing participants	3335	9464	10046	10069	13412

Table 2: Features of the dataset and different tranches.

	Tranche 1	Tranche 2	Tranche 3	Tranche 4
$1-q$	0.237 (0.205,0.274) %	1.35 (1.31,1.4) %	1.27 (1.22,1.31) %	1.56 (1.51,1.61) %
p_2	18.4 (12.1,28.6) %	34.5 (32.1,36.9) %	30.2 (27.6,33.2) %	29.0 (25.2,33.4) %
p_3	16.2 (11.4,23.3) %	27.4 (25.7,29.2) %	23.8 (21.8,26.1) %	23.3 (20.1,26.9) %
p_4	14.8 (9.98,22.3) %	23.0 (21.3,25.1) %	19.8 (17.9,22.7) %	19.7 (16.8,23.4) %
p_5	13.7 (8.84,22.8) %	20.0 (18.1,22.6) %	17.1 (15.0,20.3) %	17.3 (14.5,21.2) %
p_6	12.9 (7.89,24.0) %	17.7 (15.7,20.7) %	15.1 (13.0,18.6) %	15.5 (12.7,19.3) %
$\exp(\alpha_{\text{2-11}})$	0.883 (0.525,1.49)	0.845 (0.723,0.987)	1.39 (1.23,1.56)	0.742 (0.64,0.86)
$\exp(\alpha_{\text{12-16}})$	0.546 (0.26,1.15)	1.64 (1.44,1.87)	2.35 (2.1,2.63)	0.938 (0.807,1.09)
$\exp(\alpha_{\text{PF}})$	2.93 (1.91,4.49)	1.26 (1.06,1.49)	1.61 (1.38,1.87)	1.88 (1.66,2.13)
$\exp(\beta_{\text{2-11}})$	0.984 (0.393,2.46)	0.824 (0.636,1.07)	0.865 (0.7,1.07)	0.956 (0.787,1.16)
$\exp(\beta_{\text{12-16}})$	0.786 (0.298,2.07)	0.778 (0.578,1.05)	0.872 (0.68,1.12)	0.741 (0.583,0.943)
$\exp(\gamma_{\text{2-11}})$	0.922 (0.266,3.2)	0.715 (0.476,1.07)	0.824 (0.593,1.15)	0.919 (0.652,1.29)
$\exp(\gamma_{\text{12-16}})$	0.815 (0.237,2.8)	0.771 (0.542,1.1)	0.662 (0.488,0.899)	1.11 (0.815,1.52)
$\exp(\gamma_{\text{OR+N}})$	0.576 (0.199,1.67)	0.572 (0.447,0.731)	1.52 (1.33,1.75)	1.46 (1.2,1.77)
$\exp(\gamma_{\text{oth}})$	0.157 (0.062,0.398)	0.097 (0.0626,0.15)	0.0926 (0.0607,0.141)	0.0826 (0.055,0.124)

Table 3: The parameter point estimates and CIs.

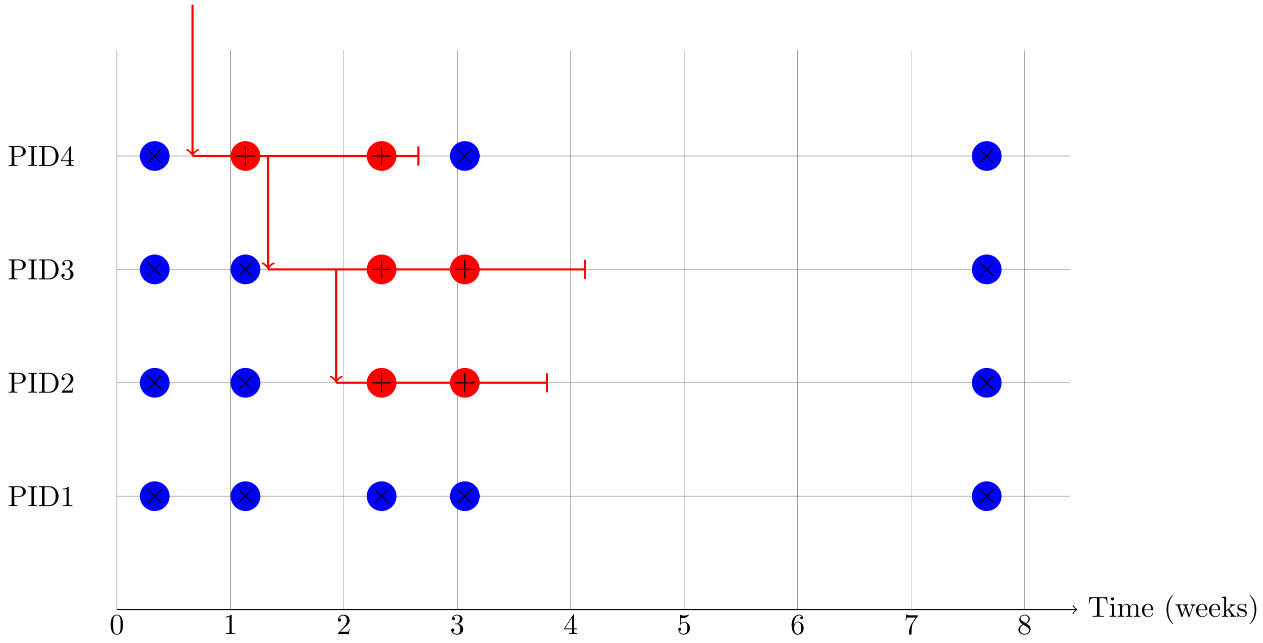
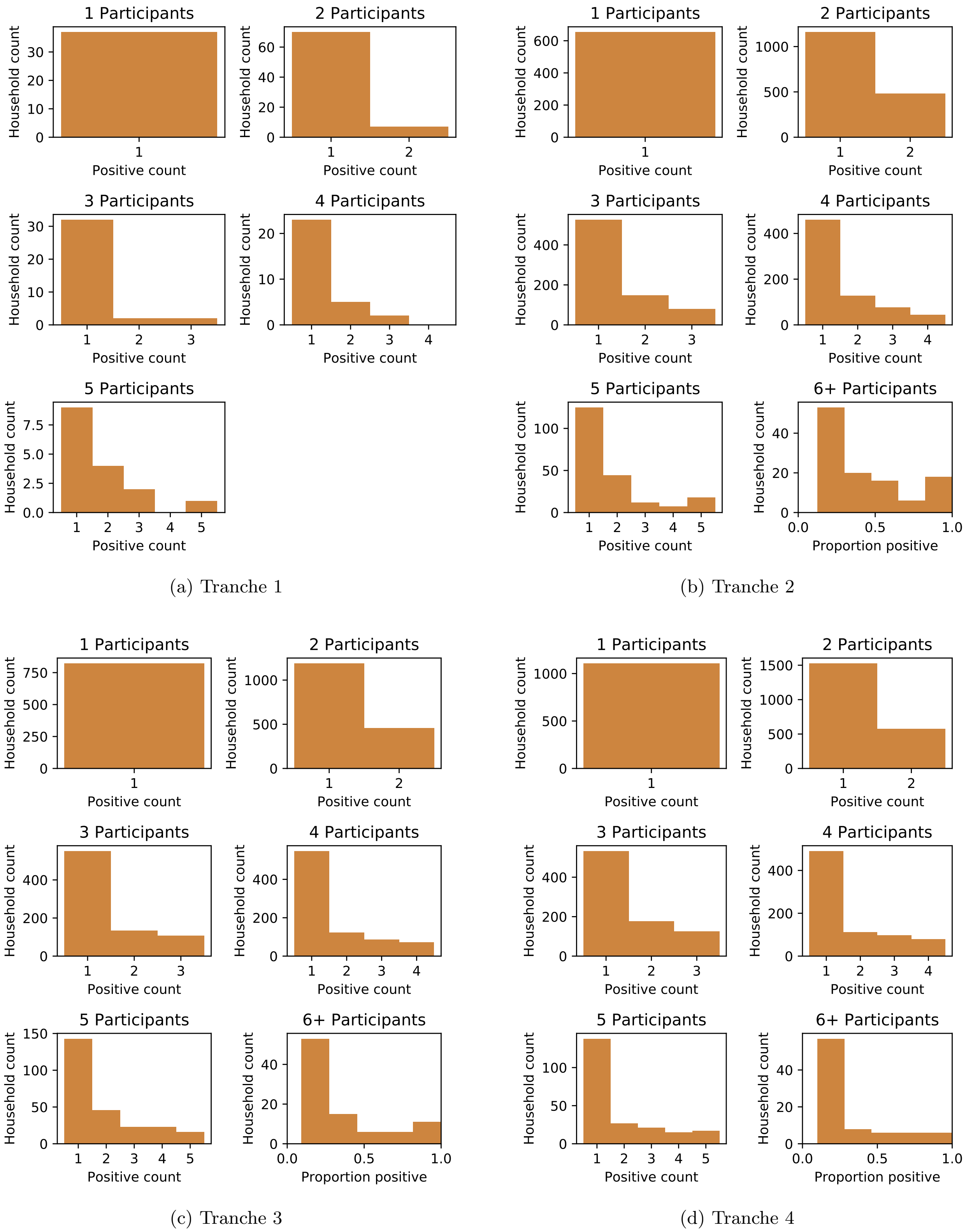


Figure 1: Schematic of a hypothetical but realistic data pattern for a four-person household in the first two months after recruitment. Each negative test is shown as a blue circle containing × and each positive test is shown as a red circle containing +. One potential route for infection coming into and transmitting within the household is shown as through a series of red arrows. This is not directly observed in the study design, and in fact other transmission trees (for example one in which PID2 is infected before PID3) are consistent with the data that would be obtained from this household.



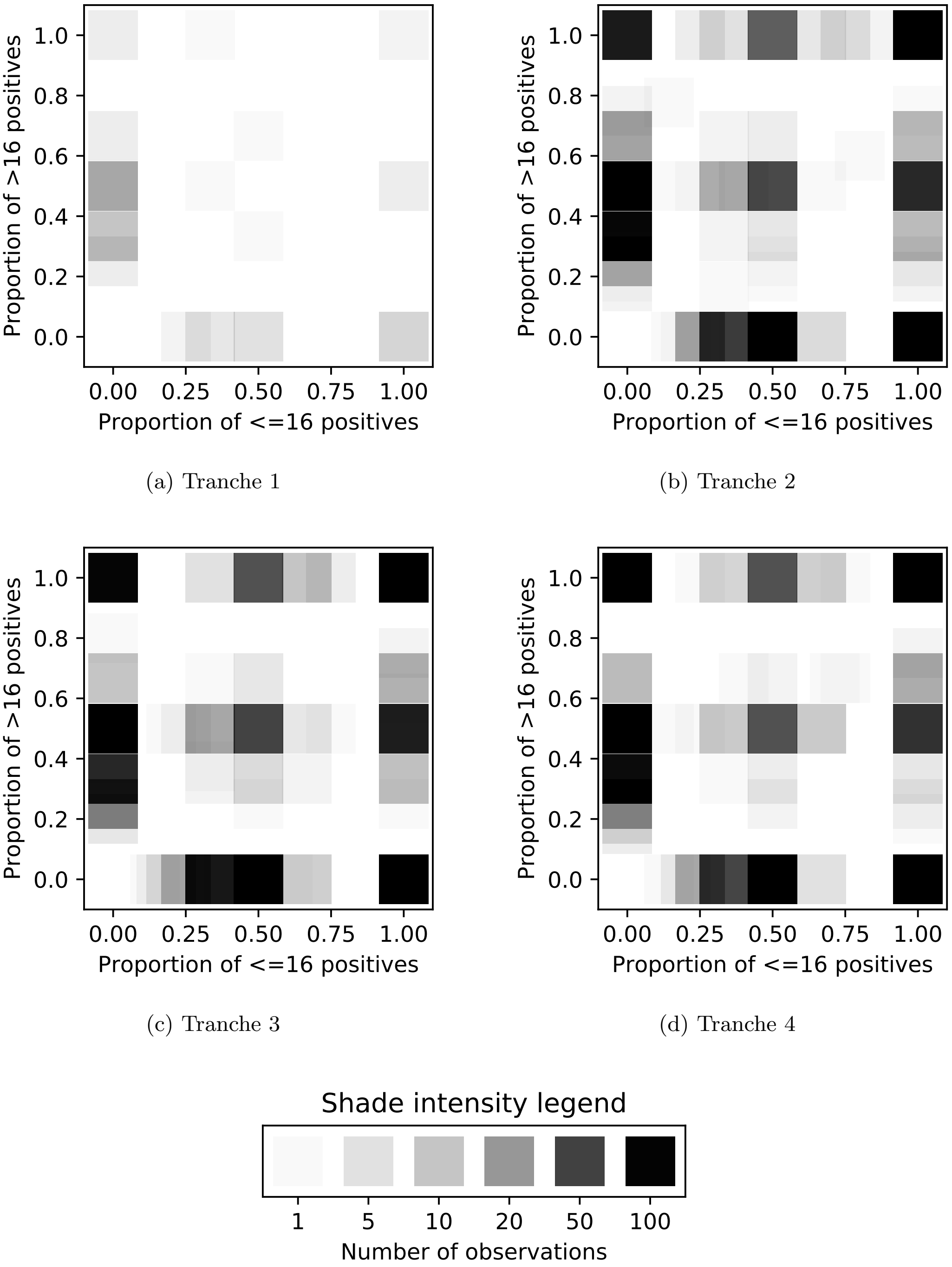
(a) Tranche 1

(b) Tranche 2

(c) Tranche 3

(d) Tranche 4

Figure 2: Histograms of household attack rates



(a) Tranche 1

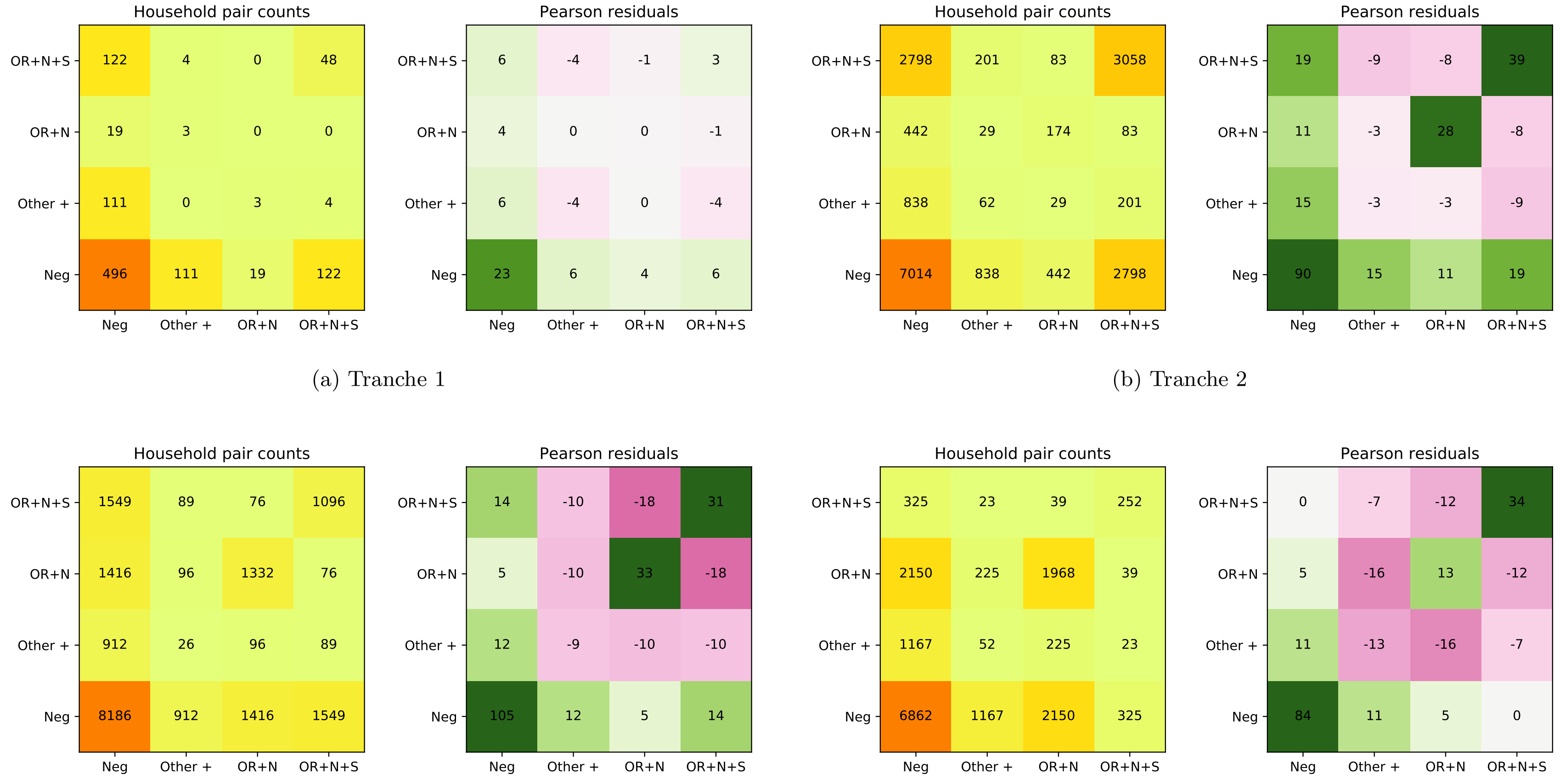
(b) Tranche 2

(c) Tranche 3

(d) Tranche 4

(e) Legend

Figure 3: Kernel density plots showing proportion of positives in different age classes in households.



(a) Tranche 1

(b) Tranche 2

(c) Tranche 3

(d) Tranche 4

Figure 4: Pair counts and residual plots for PCR gene positivity patterns.

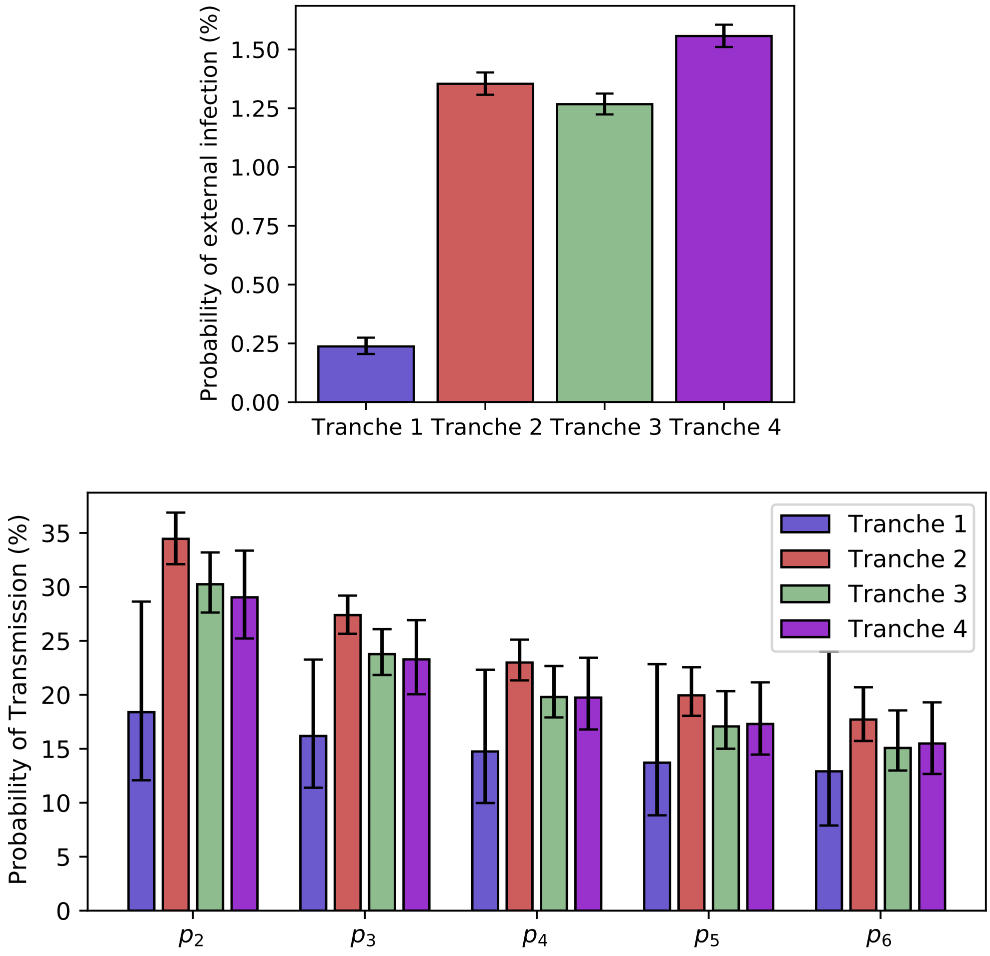


Figure 5: Visualisation of the fitted model. Top: Baseline probability of infection from outside. Bottom: Per-pair baseline probabilities of secondary transmission within the household, not including tertiary transmission effects.

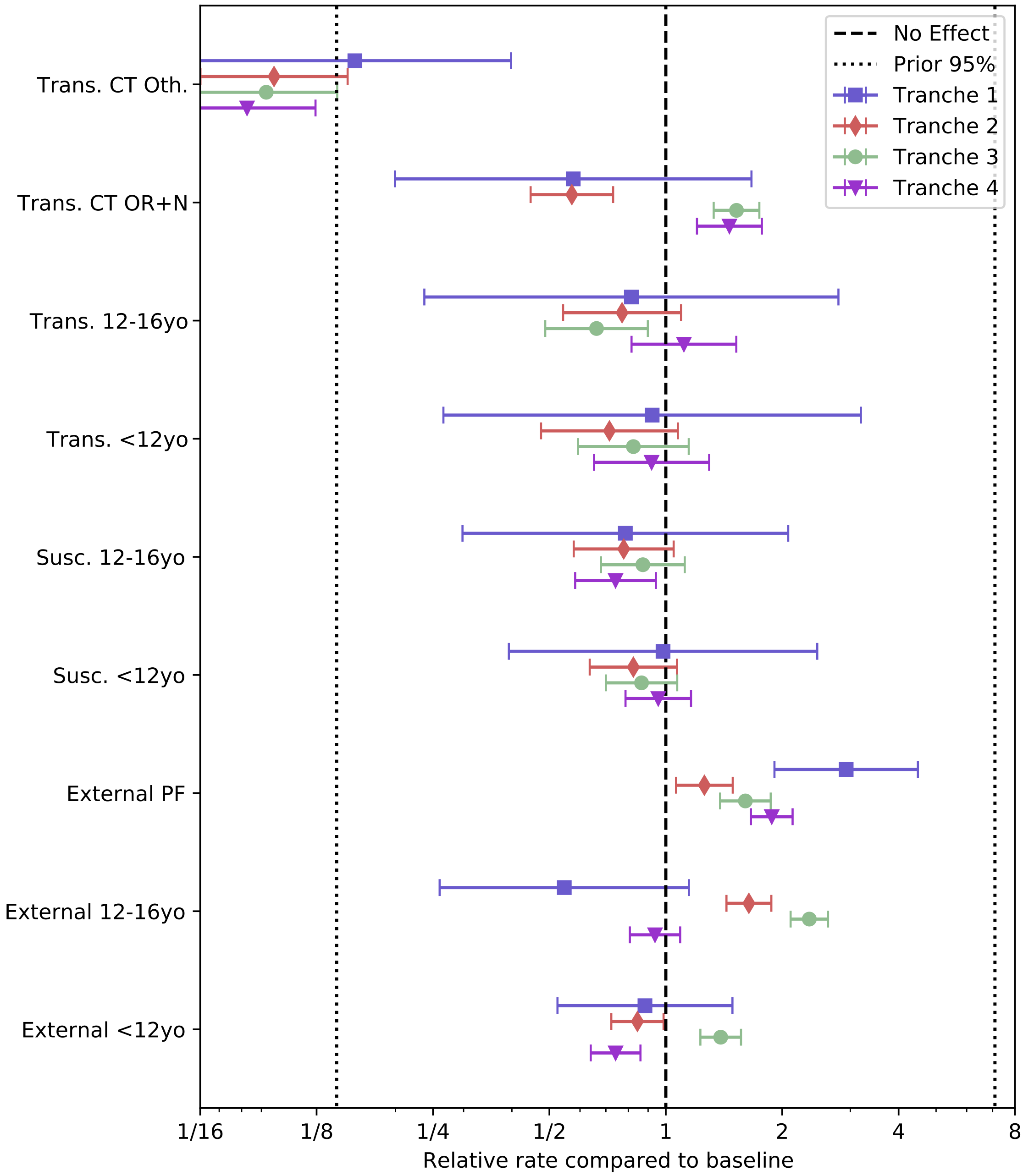


Figure 6: Visualisation of the fitted model. Relative effects on transmission, susceptibility and external exposure compared to baseline of an adult not working in a patient-facing role with OR+N+S maximal PCR gene positivity pattern if positive. 'Trans.' stands for relative transmissibility, 'Susc.' for relative susceptibility, and 'External' for relative external exposure.