MANCHESTER 1824
The University of Manchester

Why we don't have a generation time estimate from ONS data (yet)

INI Workshop on "Understanding the Generation Time for COVID-19"
10:55, 25 July 2021

Thomas House
Department of Mathematics,
University of Manchester

MANCHESTER 1824
The University of Manchester

Why bother?

- In general, we have been quite lucky that cases have followed the epidemic.
- But ultimately they are biased by factors we don't understand like testing behaviour – e.g. now (!) and last September when testing capacity was an issue.
- The ONS and REACT community surveys do not suffer from such bias, particularly important when asymptomatic transmission is so key for policy.

Why households

Household models are an integral part of the history of infectious disease epidemiology, alongside the better known whole-population models like the SIR equations. Households are important for various reasons:

- The close, repeated nature of contact within the household means that within-household transmission of infectious disease is common.
- Most of the population lives in relatively small, stable households and so these are a natural unit for data collection.
- We can design interventions at the household level – this pandemic, the emphasis has been on whole-household isolation, and school LFD testing has a strong household element, for example.

History

Personal view – there have been three ‘eras’:

1. Early-mid 20th century: Reed and Frost’s unpublished work in the 1920s on the first stochastic epidemic model (simulated using a modified roulette table). Theoretical developments by e.g. Bailey and symptom-based empirical observations by e.g. Hope Simpson.
2. Late 20th century: General final-size formula from Ball, Statistical work using this by e.g. Addy, Longini, Halloran on e.g. Tecumseh study based on viral culture.
3. 21st century: Modern computational methods (e.g. MCMC – Demiris and O’Neill) available as well as modern molecular techniques such as PCR for empirical work.

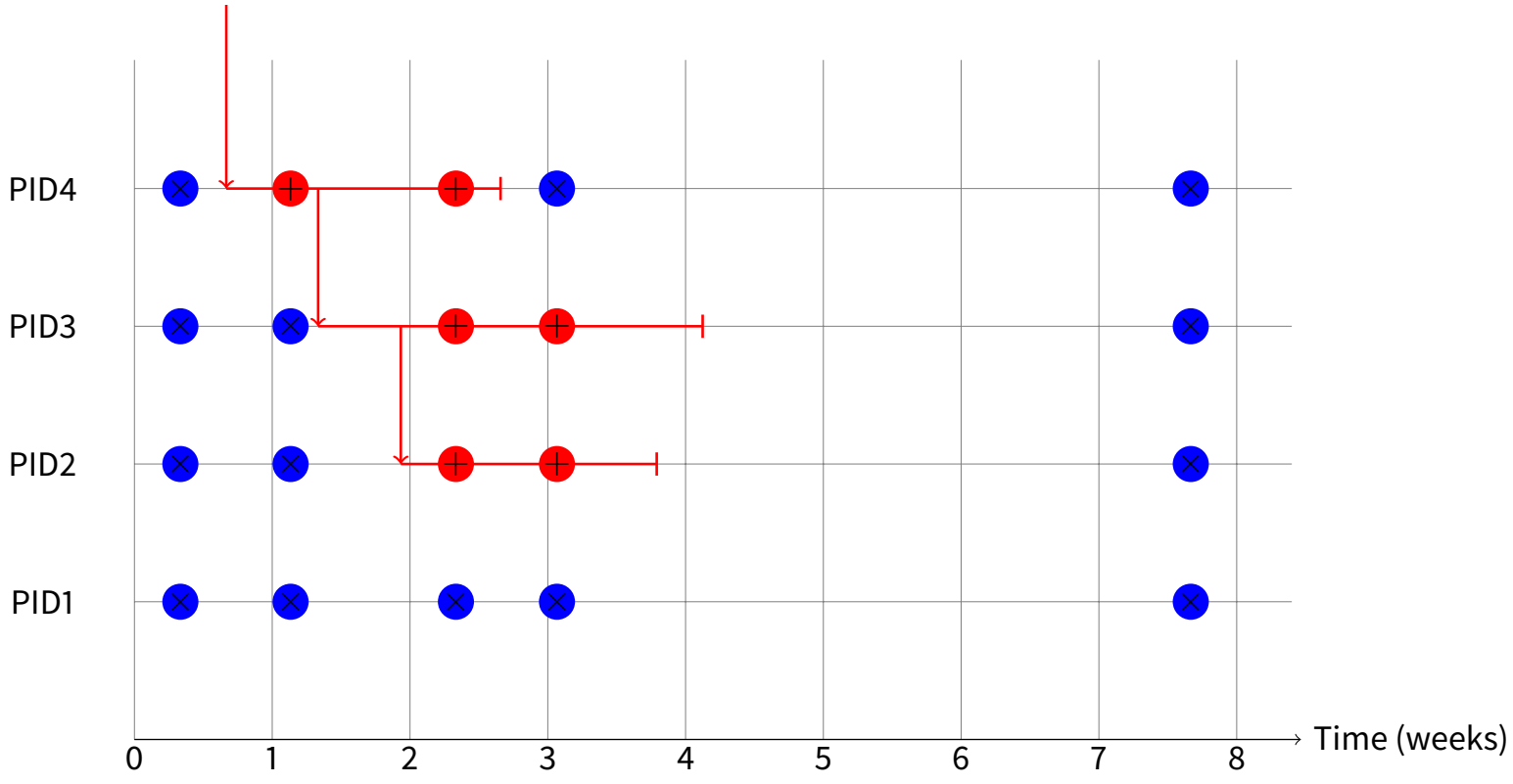
As is often the case, in an emergency, we will use the last era’s methods to get a timely answer!

This work

- Paper is on arXiv:2104.04605: **T. House** L. Pellis, K. B. Pouwels, S. Bacon, A. Eidukas, K. Jahanshahi, R. M. Eggo, A. S. Walker, "Inferring Risks of Coronavirus Transmission from Community Household Data."
- Methodology developed in arXiv:1911.12115: T. M. Kinyanjui and **T. House**, "Generalised Linear Models for Dependent Binary Outcomes with Applications to Household Stratified Pandemic Influenza Data."
- Uses data from the UK Office for National Statistics' COVID-19 Infection Survey, a large longitudinal household study based on (approximately) uniformly random sampling from the population.[†]

[†] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/previousReleases

ONS study design



The study design involves weekly then monthly household visits for a year, and uses tests for live virus. This reduces the sampling bias of data obtained through case ascertainment in the public health system, but at the cost that the actual transmission routes are not observed.

Time periods

Our approach to analysis of the data must therefore be designed to deal with the study design as detailed. We start by splitting the data into the following tranches, with associated time periods and notable events.

- **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 mainly open; B.1.1.7 variant emerged.
- **Tranche 4:** 1 January 2021 to 15 February 2021; high prevalence; schools mainly closed; B.1.1.7 variant dominant.

Assume now that the following properties are indexed by tranche.

‘Table 1’

We are also interested in the impact of the following on infection risk:

- Household size
- The B.1.1.7 variant (identified via S-gene target failure)
- Age of participant
- Work in patient-facing roles

These are distributed in the sample as below:

	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

Setup

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,

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

$$\bigcup_{a=1}^{m} H_a = [n]\,.$$

The size of the a-th household is then $n_a = |H_a|$. We let $\mathbf{x}_i$ be the length-p feature vector (also called covariates) associated with the i-th individual, and y_i be the test result so that $y_i = 1$ if the swab is positive and $y_i = 0$ if not.

Exploratory analysis – histograms

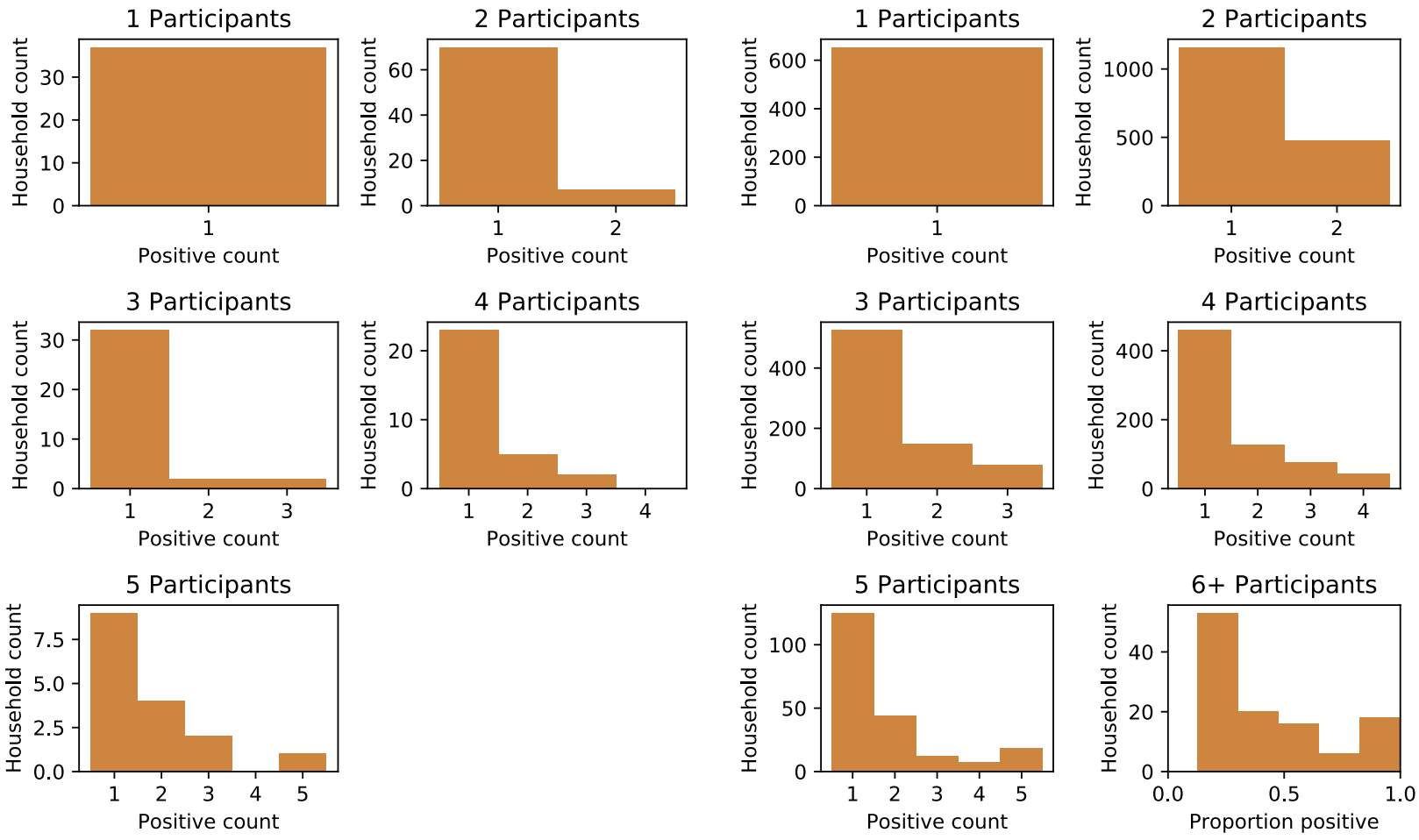
- Before jumping in to modelling (as I teach MSc students!) we should do an exploratory analysis of the data.
- The heights of the histogram bars are given by

$$Z_{k,\ell} = \sum_{a=1}^{m} \mathbb{1}_{\{n_a=\ell\}} \mathbb{1}_{\{\sum_{i\in H_a} y_i=k\}} ,$$

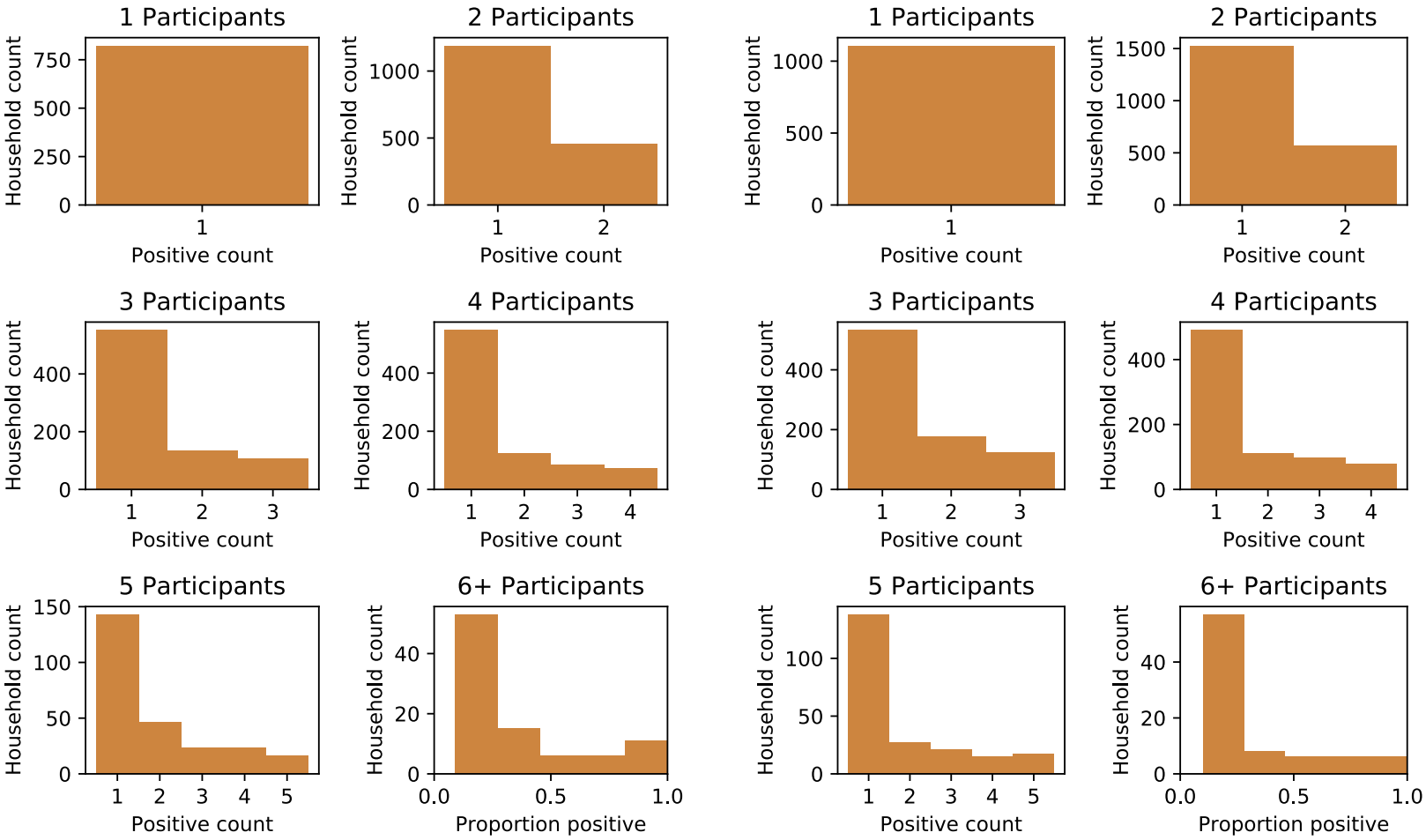
$$k \in \{2,3,4,5,6\} , \quad \ell \in \{0,\ldots,k\} ,$$

  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.

Histograms for Tranches 1 and 2



Histograms for Tranches 3 and 4



Exploratory analysis – Density plots

- 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

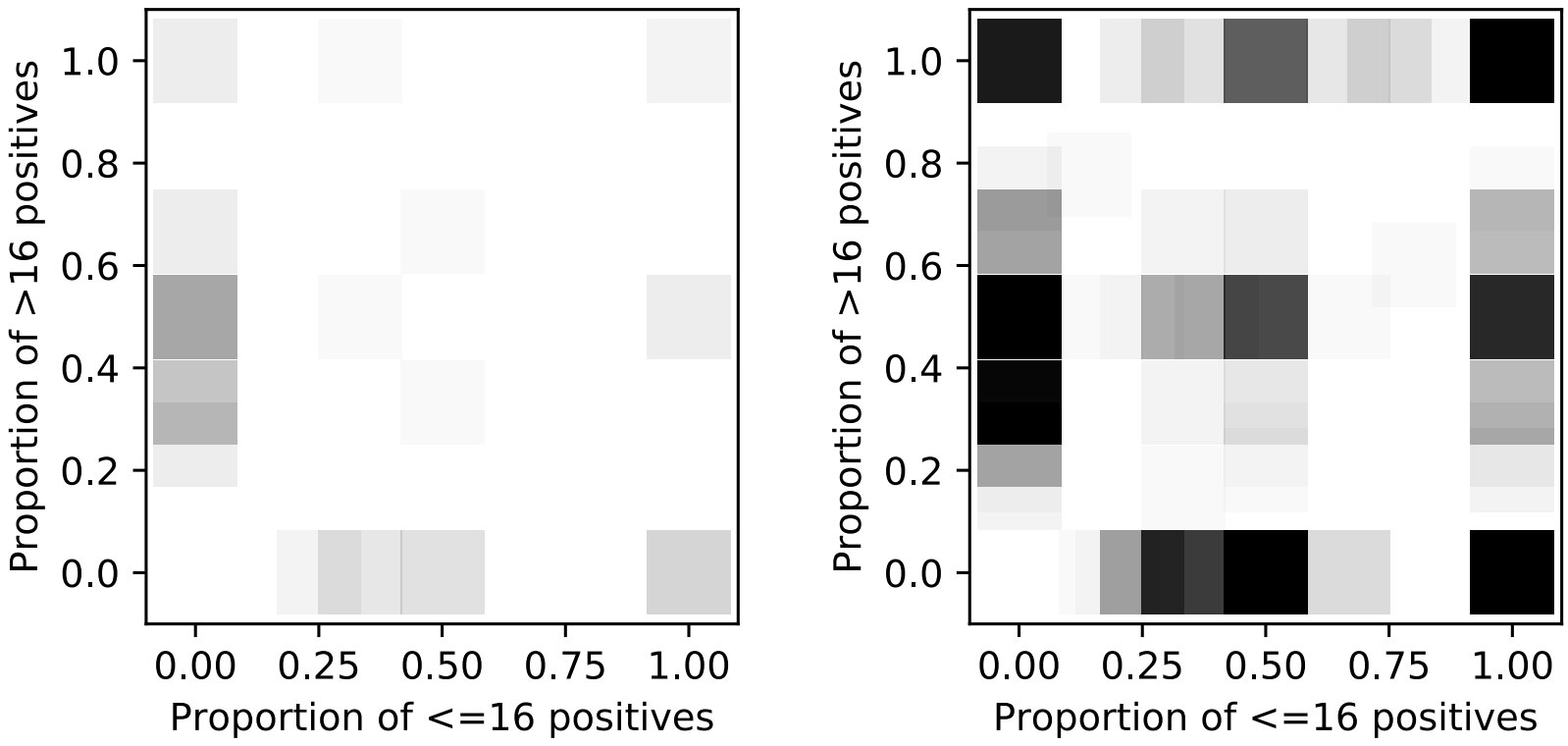
$$\mathbf{r}_a = \left(\frac{\sum_{i \in H_a} \mathbb{1}_{\{y_i=1 \& x_i=1\}}}{\sum_{i \in H_a} \mathbb{1}_{\{x_i=1\}}}, \frac{\sum_{i \in H_a} \mathbb{1}_{\{y_i=1 \& x_i=0\}}}{\sum_{i \in H_a} \mathbb{1}_{\{x_i=0\}}} \right) .$$

- 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

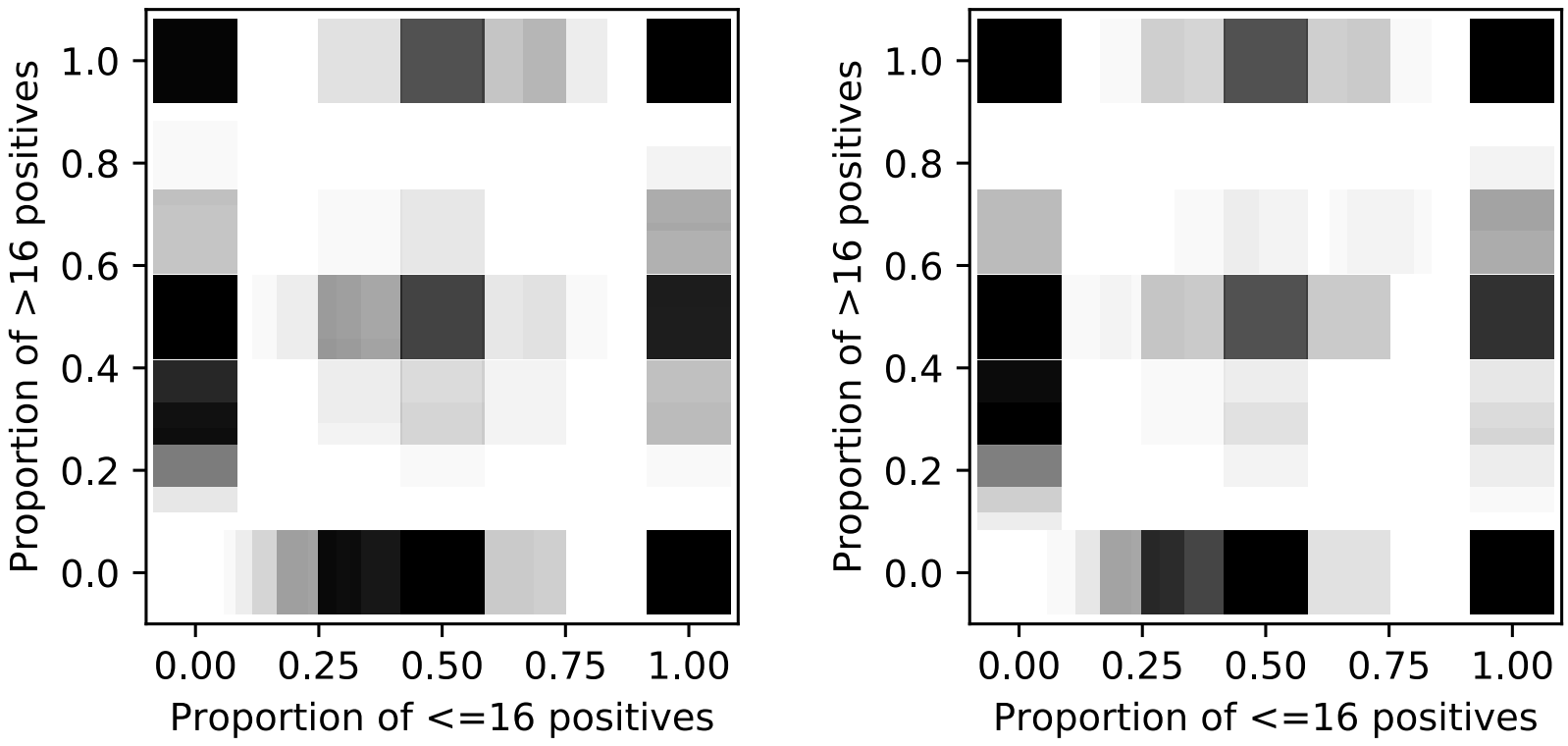
$$\mathcal{K}(\mathbf{r}, \mathbf{r}_a) = \mathbb{1}_{\{||\mathbf{r}-\mathbf{r}_a||_\infty < w\}} \,.$$

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

Density plots for Tranches 1 and 2



Density plots for Tranches 3 and 4



Residual analysis

- Pearson residuals let us tabulate features and positives in households in a manner that allows their clustering to be assessed.
- Let x_i be the feature for individual i that takes values with generic labels $A, B, \ldots$ in particular PCR gene patterns.
- We are then interested in the table of pairs of individuals in households in the set $\mathcal{H} \subseteq [m]$ with certain properties,

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

- Verbally, Y_{AB} is the count in the sample of A-B pairs in the set of households.

Residual analysis

- The null hypothesis to compare to is independent state probabilities $\boldsymbol{\pi} = (\pi_A)$ with MLE

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

- The expected table under the null is

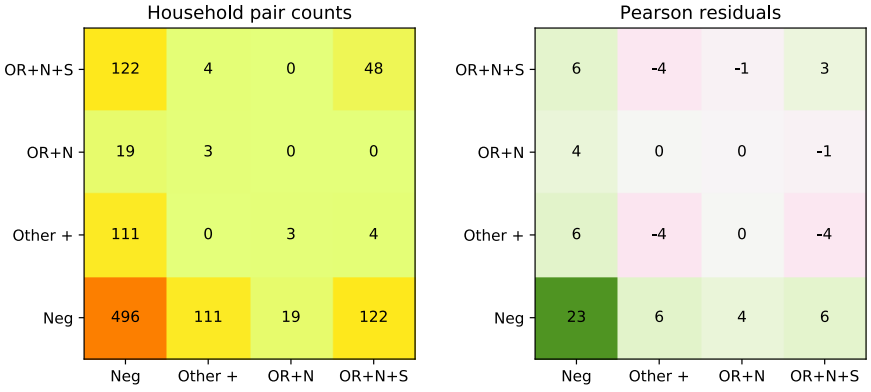
$$M_{AB} = \pi_A \pi_B \sum_{a \in \mathcal{H}} n_a (n_a - 1) \, .$$

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

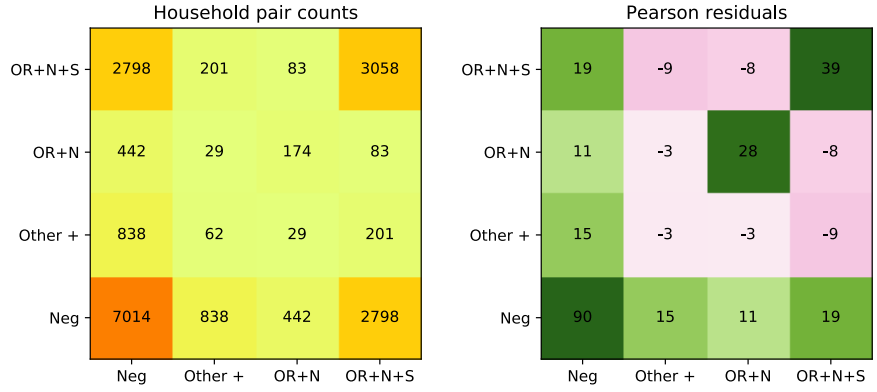
$$R_{AB} = \frac{Y_{AB} - M_{AB}}{\sqrt{M_{AB}}} \, .$$

- Such residuals are typically interpreted such that values larger than 2 are indicative of significant positive correlation, and values smaller than -2 are indicative of significant negative correlation.

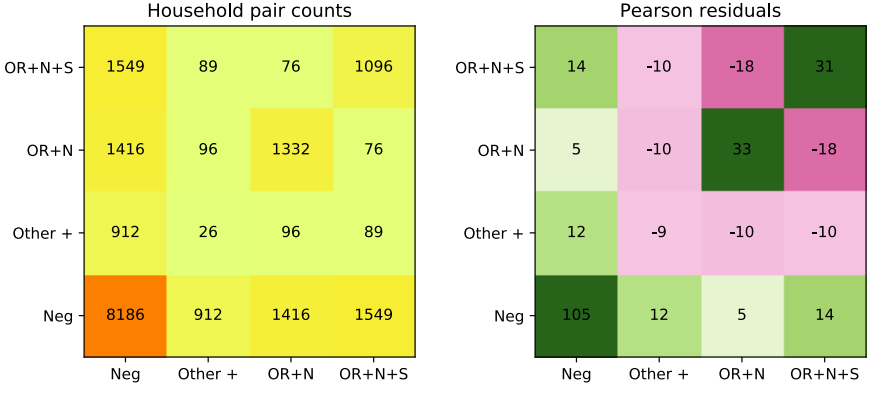
Residual plots



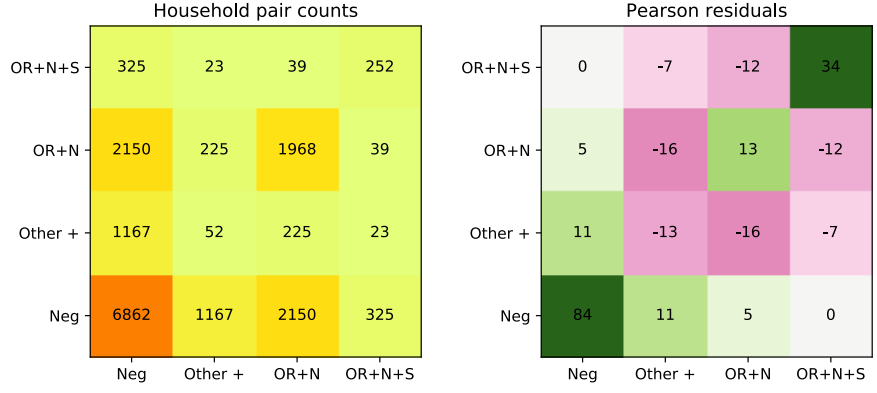
(a) Tranche 1



(b) Tranche 2



(c) Tranche 3



(d) Tranche 4

Sellke construction

We are now going to think about how to model the within-household epidemic, which starts with the Sellke construction.

- We suppose that each individual i has a stochastic variable T_i for its infectious period, picked from the infectious period distribution, and that susceptible individuals have a random threshold $\Xi_i \sim \mathrm{Exp}(1)$.
- The individual then becomes infectious when their threshold is exceeded by the total force of infection they have experienced. To see why this is equivalent to the standard definition, consider

$$\begin{aligned}\Pr(\Xi > \Lambda(t+\delta t) | \Xi > \Lambda(t)) &= \frac{\int_0^{\Lambda(t+\delta t)} \exp(-\xi)\mathrm{d}\xi}{\int_0^{\Lambda(t)} \exp(-\xi)\mathrm{d}\xi} \\ &= 1 - \Lambda(t)\delta t + o(\delta t)\,.\end{aligned}$$

Final size equations

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

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

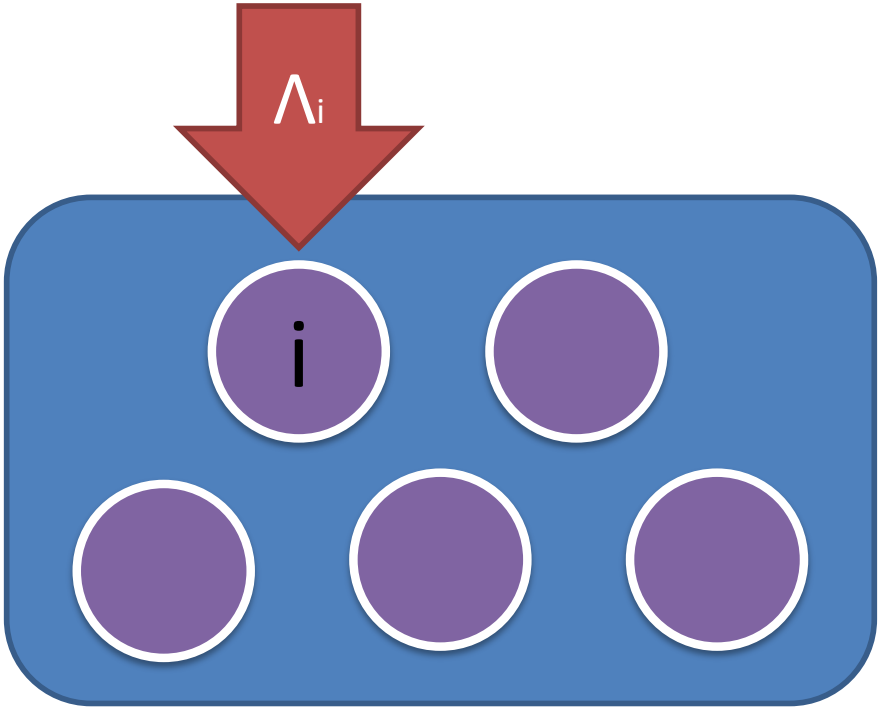
where $\mathbf{1}$ is a length-2^n vector of all ones, and $\boldsymbol{B} = [B_{\iota(\boldsymbol{\nu}),\iota(\boldsymbol{\omega})}]_{\boldsymbol{\nu},\boldsymbol{\omega}\in\{0,1\}^n}$, which has

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

$\boldsymbol{\nu} \leq \boldsymbol{\omega} \in \{0,1\}^n$, 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.

MANCHESTER 1824
The University of Manchester

Component 1: External infection



Final size equations

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

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

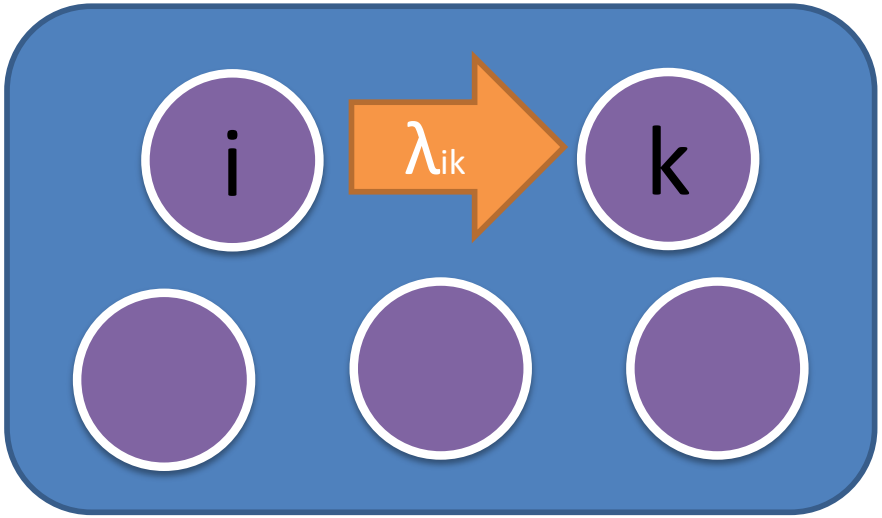
In the language of infectious disease modelling, Λ_i is the cumulative force of infection experienced by the i-th individual. Then $\exp(\alpha_k)$ is the relative external exposure associated with the k-th feature / covariate, meaning that it is the multiplier in front of the baseline force of infection,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

$$q = \exp(-\Lambda) = \exp(-\exp(\alpha_0)) \,, \tag{2}$$

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 (2) to relate this to the baseline force of infection Λ or intercept of the linear predictor, α_0.

MANCHESTER 1824
The University of Manchester

Component 2: Within-household infection



Final size equations

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 and is

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

The parameter ϑ is the variance of the Gamma distribution, i.e. it is larger for more individual variability. To see why the Laplace transformation is appropriate, consider the Sellke construction and assume a baseline rate of infection, λ, to be multiplied by infectious duration T to give total force of infection $\Lambda = \lambda T$, so

$$\Pr(\Xi > \Lambda) = \int_0^\infty F_\Xi(\lambda t) f_T(t) \mathrm{d}t = \int_0^\infty \mathrm{e}^{-\lambda t} f_T(t) \mathrm{d}t = \mathcal{L}[f_T](\lambda) .$$

Final size equations

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

$$\lambda_{ij} = n^{\eta}\lambda\sigma_i\tau_j = n^{\eta}\lambda \mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{x}_i}\mathrm{e}^{\boldsymbol{\gamma}\cdot\mathbf{x}_j} = \mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{x}_i}\mathrm{e}^{\gamma_0+\eta\log(n)+\boldsymbol{\gamma}\cdot\mathbf{x}_j}\ . \quad (3)$$

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 (3), 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; as can be seen from (3), this is equivalent to taking $\log(n)$ as a covariate for transmissibility.

Likelihood function and fitting

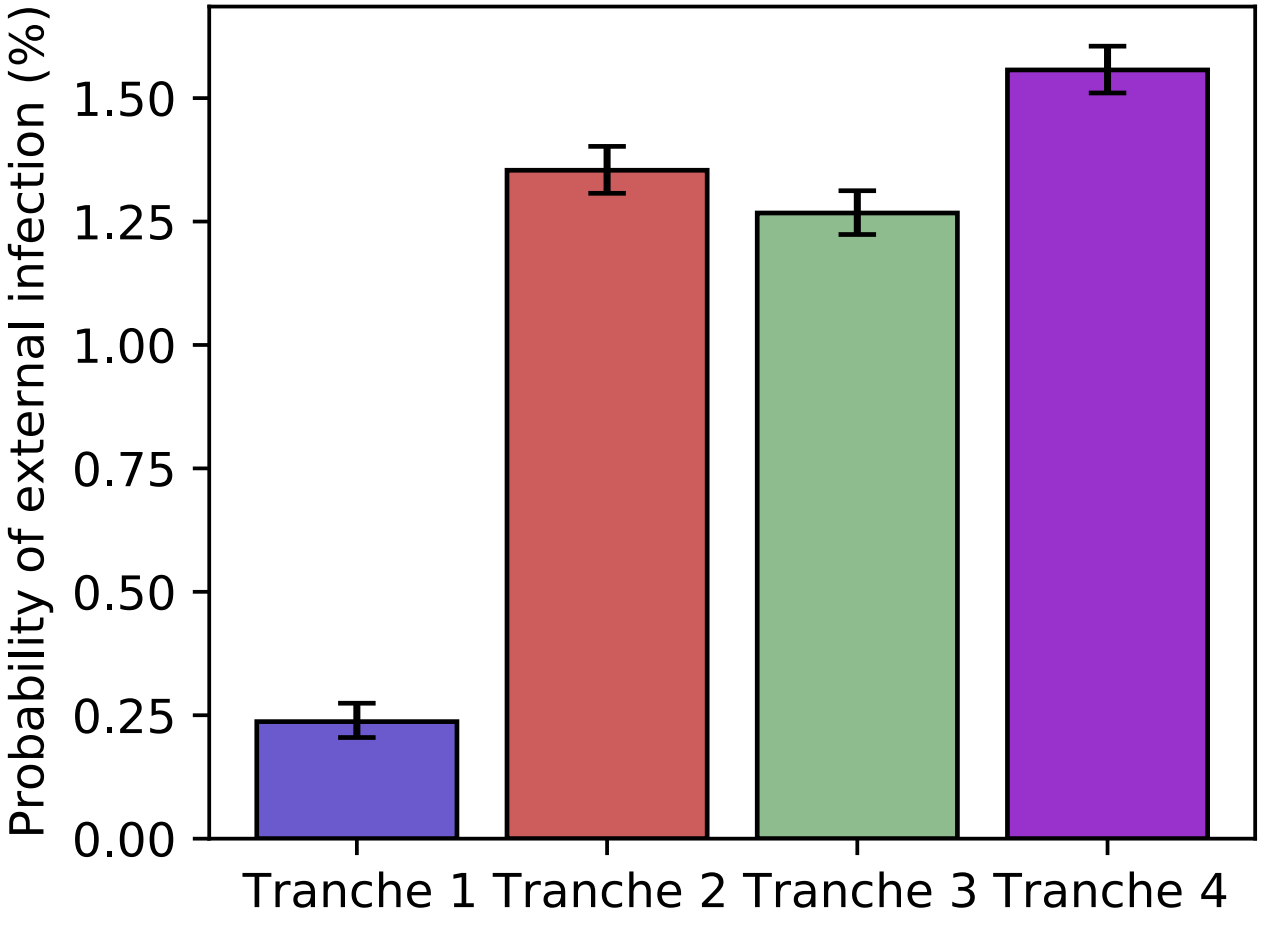
- We can then produce a likelihood for the data from the probability model.
- This will take the form of a product of probabilities derived from solving the Ball equations (1).
- Actually fitting this model to 3M observations on a secure environment is non-trivial, and involves a significant numerical linear algebra computational effort.
- For the results here, NumPy was sufficient, but we are experimenting with implementation in Numba.
- We carried out approximate Bayesian inference.
- This was done using the Laplace approximation and a standard normal prior on each parameter.
- Multi-restart numerical optimisation using a Quasi-Newton method was used.

Results – ‘Table 2’

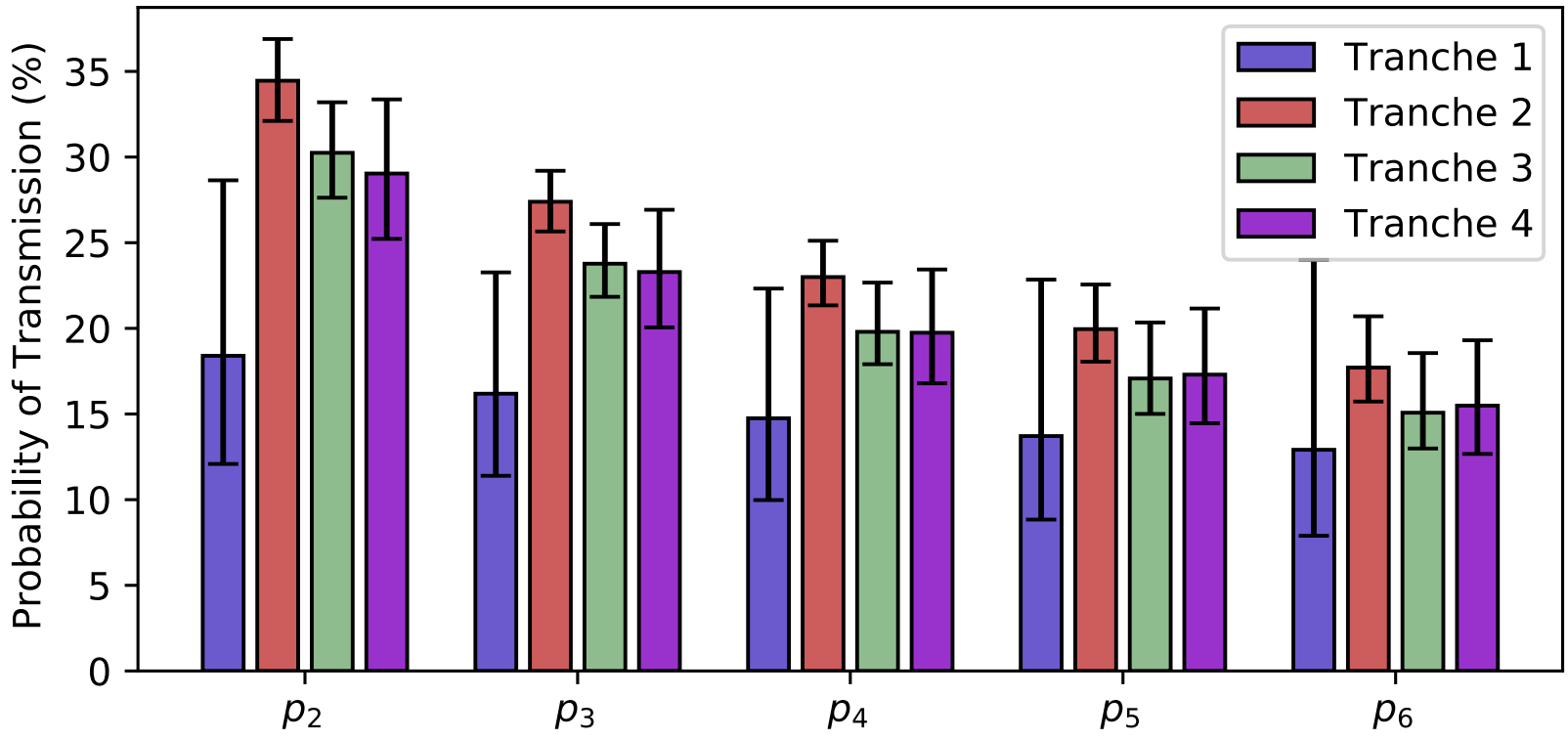
Point estimates and CrIs for each parameter are:

	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)

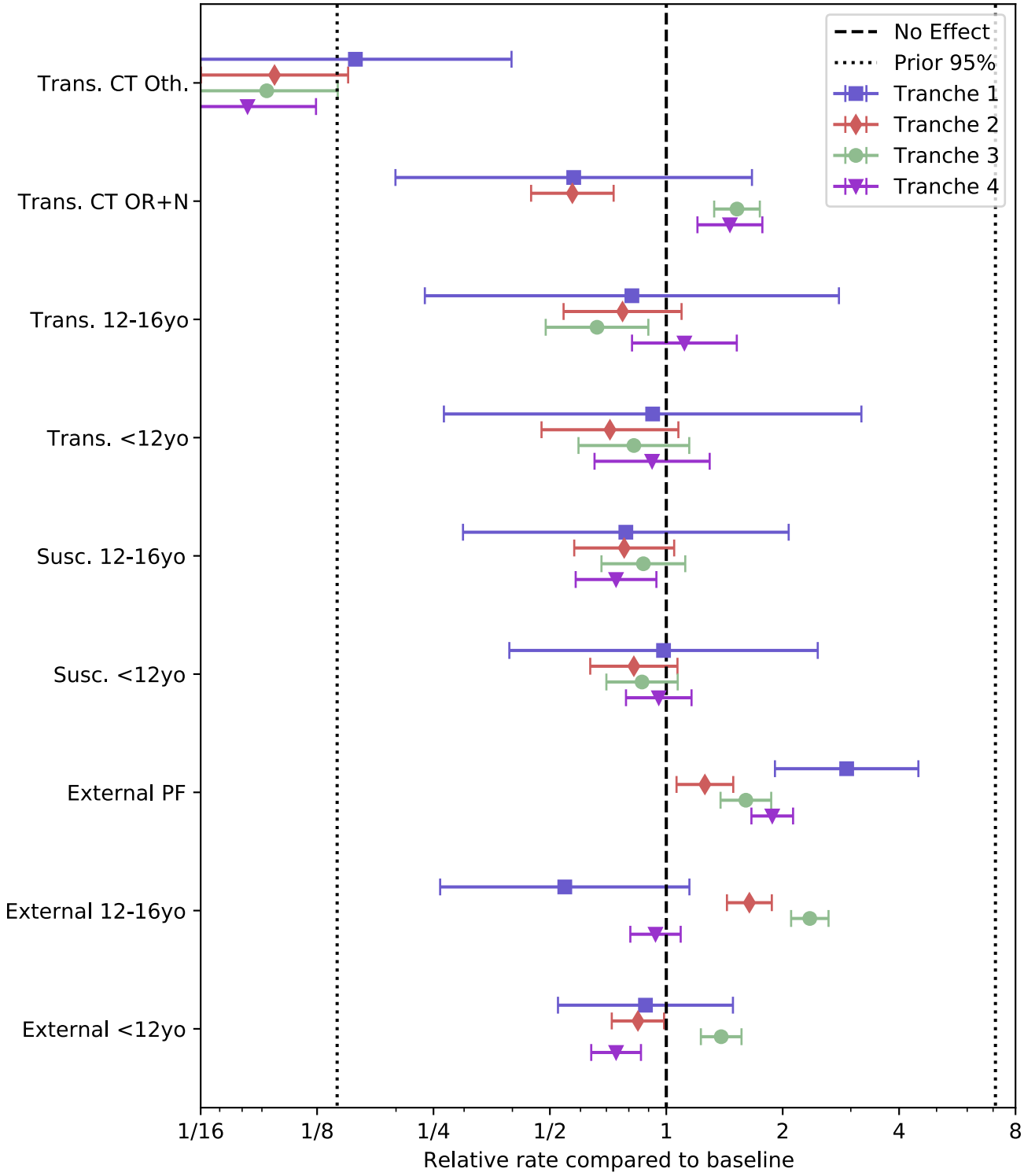
Results – Overall infection from outside



Results – Within-household infection



Results – Effect sizes

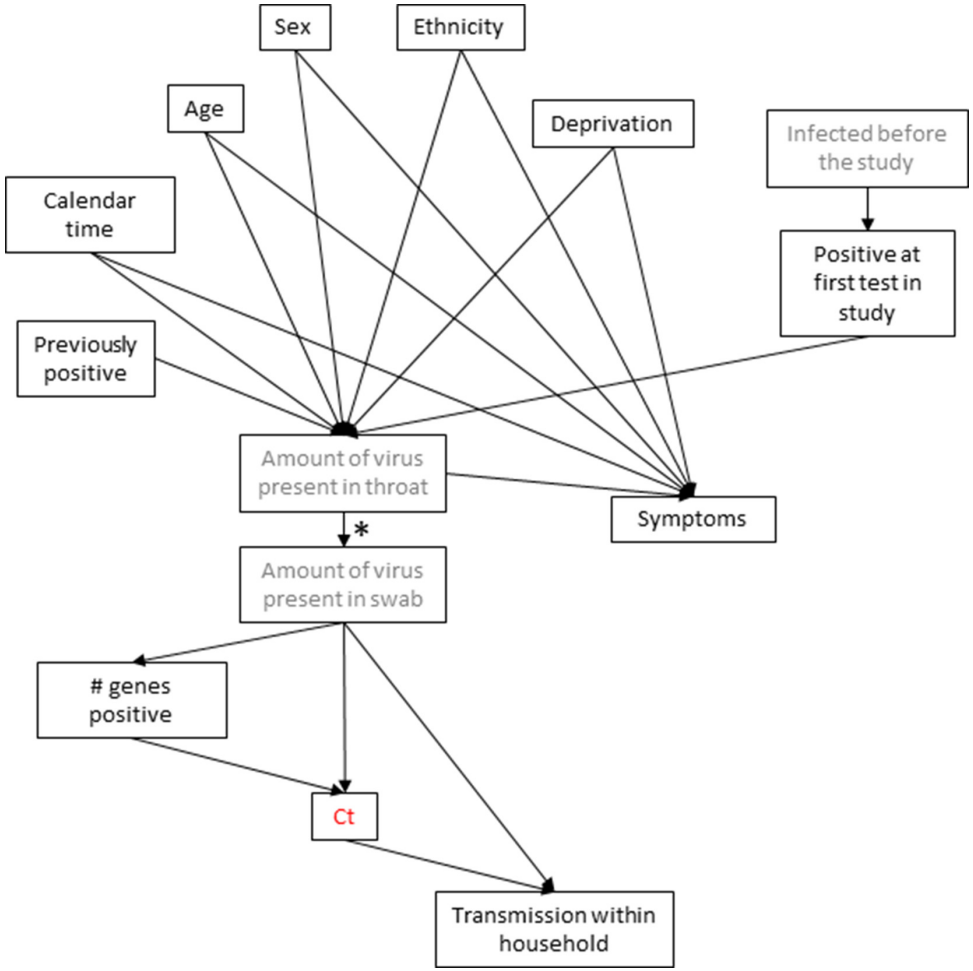


eLife
RESEARCH ARTICLE

MANCHESTER 1824
The University of Manchester

Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time

A Sarah Walker[1,2,3,4]*, Emma Pritchard[1,2], Thomas House[5,6], Julie V Robotham[2,7], Paul J Birrell[7,8], Iain Bell[9], John I Bell[10], John N Newton[11], Jeremy Farrar[12], Ian Diamond[9], Ruth Studley[9], Jodie Hay[13,14], Karina-Doris Vihta[1,2], Timothy EA Peto[1,2,3,15], Nicole Stoesser[1,2,3,15†], Philippa C Matthews[1,15†], David W Eyre[1,2,14,16†], Koen B Pouwels[1,2,17], COVID-19 Infection Survey team



Appendix 1—figure 2. Directed acyclic graph of potential relationships between factors. *May also depend on factors which effect self-swabbing efficiency, e.g., demographics.

- We have plenty of evidence that there's a lot behind actual transmission

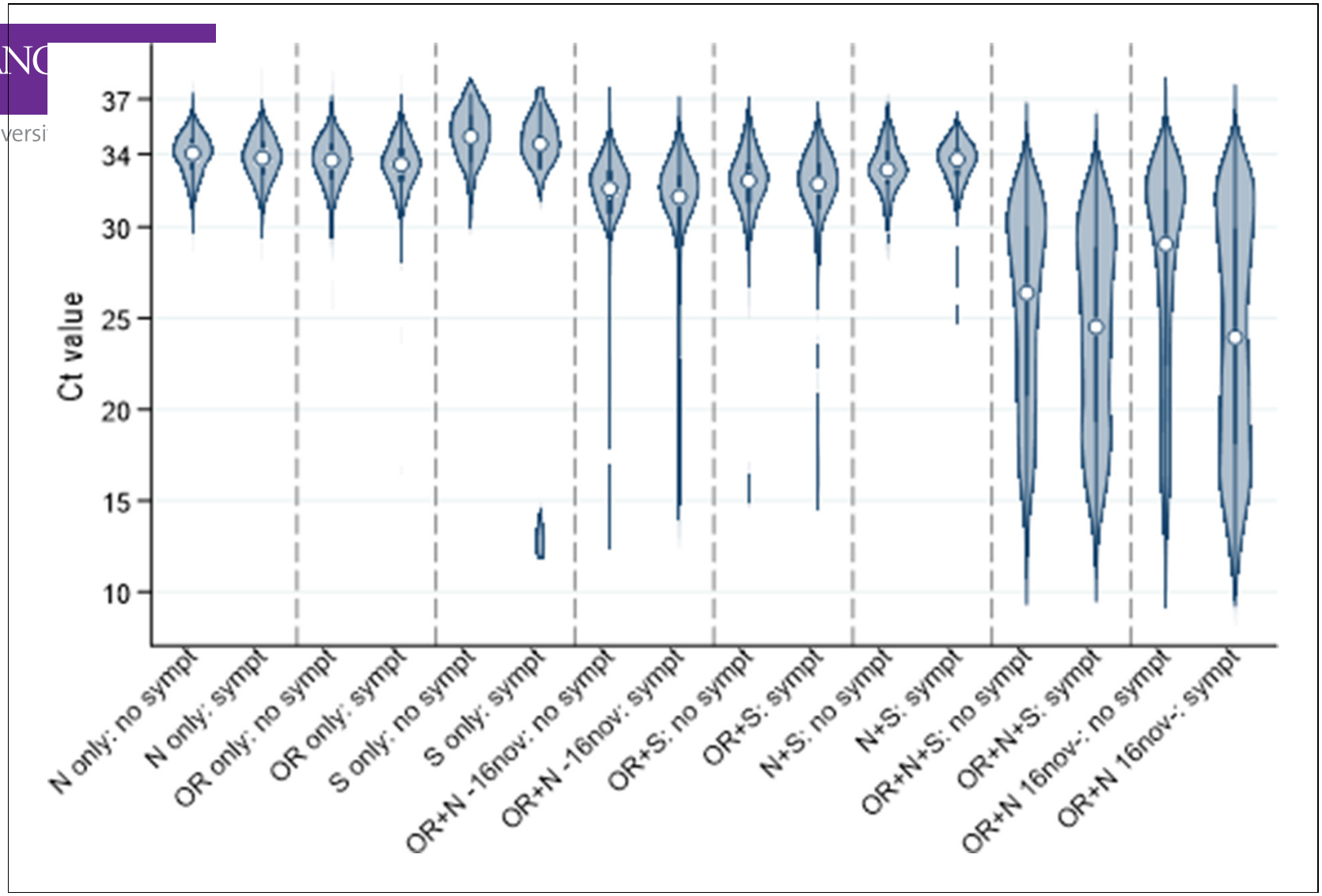


Figure 1. Distribution of Ct values at each positive test by genes detected and self-reported symptoms. Note: Points show the median and boxes the interquartile range. OR=ORF1ab. Positives where only the ORF1ab+N genes were detected are split by whether the swab was taken before or after 16 November 2020, reflecting the expansion of B.1.1.7 (which has S-gene target failure on the assay used in the survey).

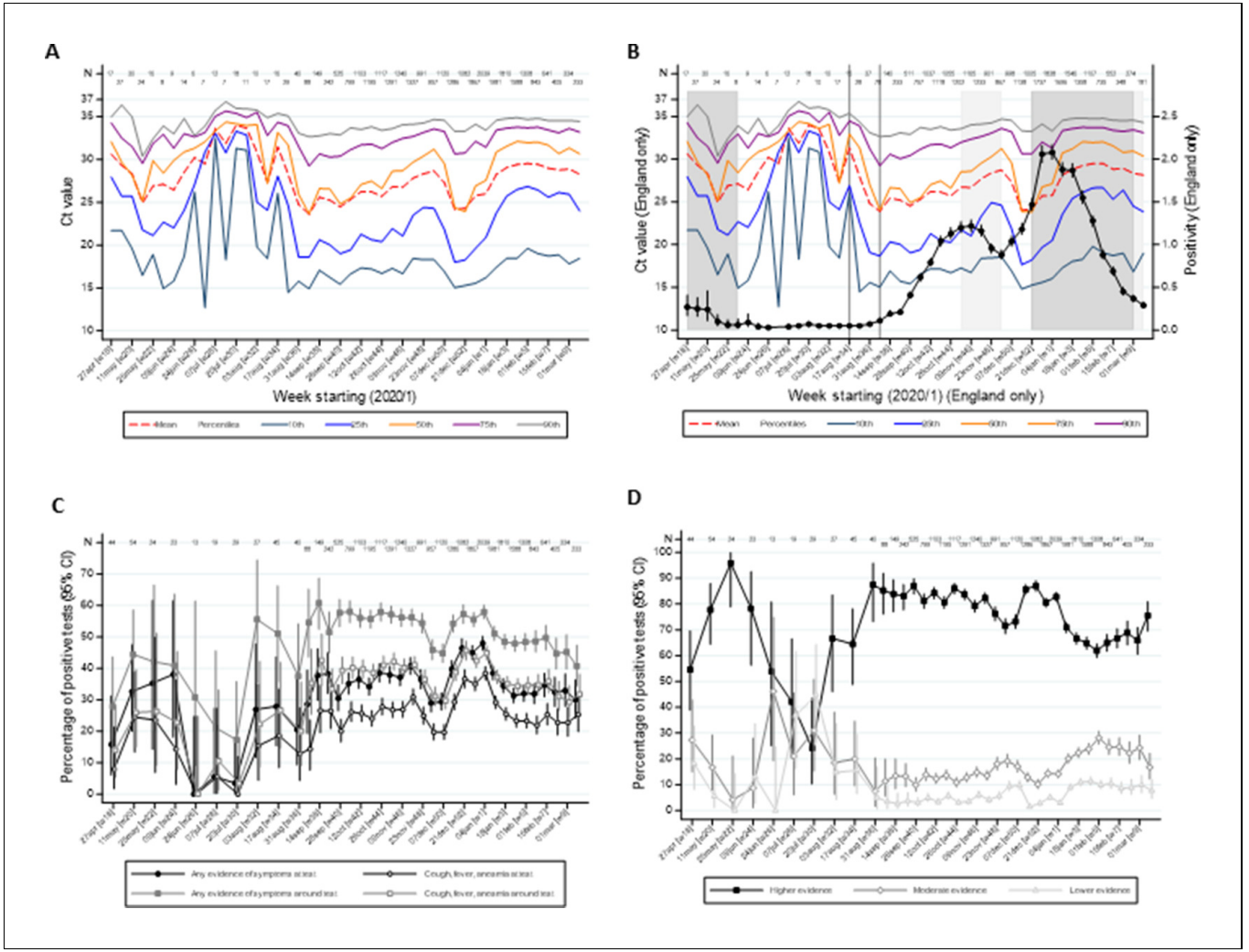


Figure 3. Variation over calendar time in the distribution of Ct values in the UK (**A**) and England (**B**) together with percentage positivity in England (**B**), and in self-reported symptoms (**C**) and evidence supporting positives (**D**). Note: Panel (**A**) shows the distribution of Ct values each week including all positives across the UK. Panel (**B**) is restricted to England shown together with the official estimates of positivity as reported by the Office for National Statistics (black line) and periods of national 'stay-at-home' restrictions (schools shut in dark grey, schools open in light gray). Panels (**C**) and (**D**) show the proportions reporting symptoms and with different levels of evidence supporting the positive test, respectively. Variation in the width of 95% CI reflects the increase in size of the survey from mid August (***Supplementary file 1***).

Keeping up with COVID-19

Rachel Thomas

Submitted by Rachel on April 9, 2021

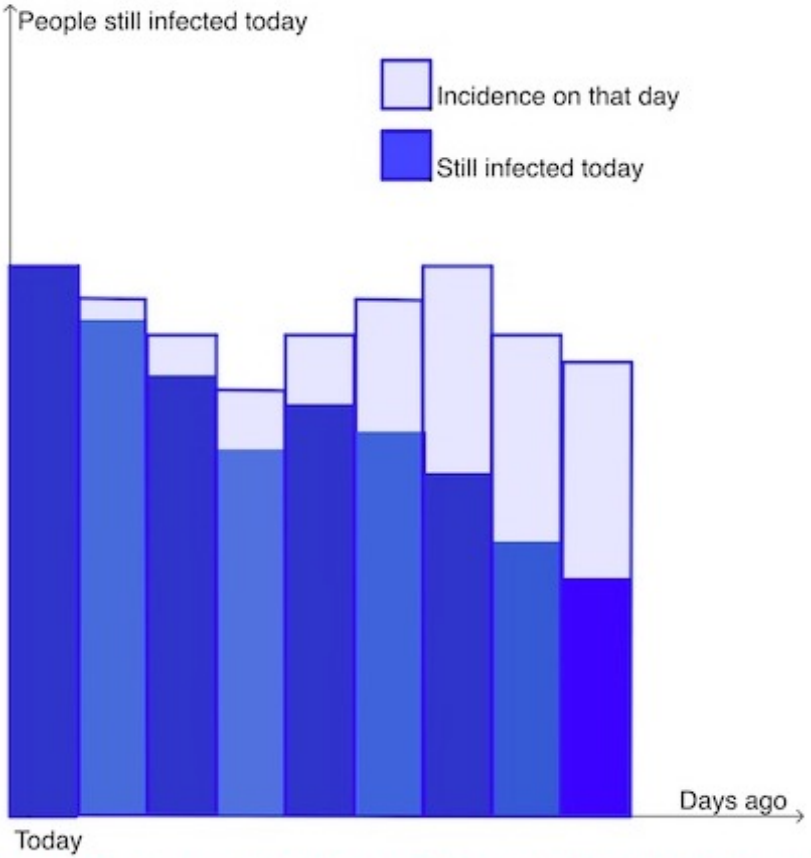
Many of us wait for the daily numbers announcing the state of the pandemic: people testing positive for the first time, hospitalisations, and sadly, deaths in the last 24 hours. But you might have wondered why two numbers are missing from the daily government statistics: how many people currently have COVID-19 and how many new COVID-19 infections there have been in the UK.

Accurately knowing these numbers, the *prevalence* and *incidence* of the disease, would seem to be vital during the pandemic, but these numbers aren't announced alongside other daily statistics. Since we can't possibly test everyone in the population all the time, these numbers are hard to come by. This is why the Office for National Statistics (ONS) started the COVID-19 Infection Survey back in April 2020. The results and analysis of the ONS survey are reported weekly to Government and the public.

See here for all our coverage of the COVID-19 pandemic.

The prevalence of the disease today also includes the people who caught the disease two days ago who are still testing positive today: given by $Inc(t-2) \times Dur(2)$. And we can continue with this train of thought to give the prevalence of the disease to be:

$$Prev(t) = Inc(t) + Inc(t-1) \times Dur(1) + Inc(t-2) \times Dur(2) + Inc(t-3) \times Dur(3) + \ldots$$



The prevalence of a disease today is the sum of the people who caught the disease today, plus those who caught it yesterday and are still infected today, plus those who caught it two days ago and are still infected today... and so on. The light blue rectangles illustrate the incidence on previous days, and the dark blue rectangles illustrate the proportion of those (given by *Inc(t-k) x Dur(k)* where *k* is the number of days ago) who are still infected today.

Acknowledgements

- The ONS CIS team
- Manchester Mathematical Epidemiology group
- Juniper consortium
- Funders: Royal Society; Alan Turing Institute; UKRI; DHSC
- SAGE and in particular SPI-M members and secretariats
- Many others!

Thanks for your time!