**Household composition and risk of hospitalisation or death due to COVID-19: a UK population-based cohort using the OpenSAFELY platform**

**Authors (so far, others to be added):**

Kevin Wing, Rosalind Eggo, Daniel Grint, Rohini Mathur, James Cheshire (UCL), Fizz Williamson, Ben Goldacre, Liam Smeeth, Seb Bacon, Stephen Evans, Christopher Rentsch, Anna Schultze, Ian Douglas, Helen Curtis, Laurie Tomlinson, Emily Nightingale, Krishnan Bhaskaran

**Version:** v0.4

**Date:** [Date once finalised at v1.0 and uploaded to GitHub]

| Version history | Date | Comment |
| --- | --- | --- |
| 0.1 | 29 Oct 2020 | Initial draft created |
| 0.2 | 14 Jan 2021 | Refinements following comments received on v0.1 |
| 0.3 | 30 Mar 2021 | Edited based on comments received on version 0.2 |
| 0.4 | 02 April 2021 | Removed MSOA (stratifying on UTLA is sufficient), added DAG |
| 0.5 | 15 April 2021 | Incorporated comments received on version 0.4 |
| 1.0 | 11 July | Final version (refined censoring date re: vaccination, added power calculation) |

## Background

Household size and composition have been shown to vary by factors such as ethnic group and socioeconomic status (SES) and affect risk of SARS-CoV-2 infection and severe COVID-19 outcomes, with increased exposure/transmission between children/working age adults and older/vulnerable household members a particular concern. An investigation of how household composition (in terms of age of household occupants) influences both the absolute and relative risk of severe COVID-19 outcomes in a very large cohort of households could (1) improve understanding of the observed disparities in COVID-19 outcomes by ethnic group and socioeconomic status and (2) facilitate targeted vaccine deployment in the future.

## Hypothesis

Household composition (in terms of the ages of household occupants) is associated with severe COVID outcomes and accounts for a large proportion of the disparity in outcomes observed by ethnicity and by socioeconomic status.

## Aims and Objectives

### Objectives

1. To determine if there is an association between household composition and serious COVID outcomes (being hospitalised with COVID-19 or dying from COVID-19), and to quantify this association for different age groups
2. To assess whether the association between household composition and COVID-19 differs by ethnicity and SES
3. To report absolute measures of risk of serious COVID outcomes by category of household composition by age, by ethnicity and by SES
4. To quantify the extent to which household composition contributes to the observed disparities in serious COVID outcomes by ethnicity and by SES

## Methods

### Database Description (Copied from Google docs link)

We will use data from general practice (GP) records, obtained from the GP software provider The Phoenix Project (TPP), linked to [placeholder - whichever datasources you linked to]. The data will be accessed, linked and analysed through openSAFELY.org - a new data analytics platform created by our team on behalf of NHS England to address urgent questions relating to the epidemiology and treatment of COVID-19(REF). OpenSAFELY provides a secure software interface allowing the analysis of pseudonymised primary care patient records from England in near real-time within the EHR vendor’s highly secure data centre, avoiding the need for large volumes of potentially disclosive pseudonymised patient data to be transferred off-site. This, in addition to other technical and organisational controls, minimises any risk of re-identification. Similarly pseudonymised datasets from other data providers are securely provided to the EHR vendor and linked to the primary care data. Descriptions of OpenSAFELY have been previously published (REF), and more information can be found on<https://opensafely.org/>.

Primary care records retrieved from the TPP SystmOne electronic health record system include diagnoses (Read 3 CTV3), prescriptions (dm+d), basic sociodemographics and vital signs for 22 million individuals – approximately 40% of the English population. Data extracted by TPP SystmOne have previously been used in medical research, as part of the ResearchOne dataset (REFS).

All data will be held in a secure research environment hosted by TPP, which is a Tier 3 data centre, accredited to NHS Digital standards for centrally hosted clinical systems (ISO 27001 standard and IG Toolkit version 2). We received ethics approval to conduct the data linkage and analyses by the London - City & East Research Ethics Committee on the 2nd of April 2020 (REC reference: 20/LO/0651) and LSHTM Ethics Board (ref 21863). No further ethical or research governance approval was required by the University of Oxford but copies of the approval documents were reviewed and held on record.

*[Latest Database Description available here:*

<https://docs.google.com/document/d/1d6fw9sc80_N_UQO7qib_R8yBZGObLEzPS_xcri222rA/edit>]

### Information Governance

*[Please note: phrasing in bold needs to be retained in any consequent publication; the text below represents the gold standard. Any abstract ought to contain phrasing showing that the work is done on behalf of NHSE, i.e “Working on behalf of NHS England, the OpenSAFELY platform was used to analyse" or "On behalf of NHS England, we analysed..” ]*

**NHS England is the data controller; TPP is the data processor; and the key researchers on OpenSAFELY are acting on behalf of NHS England.** This implementation of OpenSAFELY is hosted within the TPP environment, which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant 52,53; patient data have been pseudonymized for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymized datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is through a virtual private network (VPN) connection; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; and only aggregate statistical outputs leave the platform environment following best practice for anonymization of results such as statistical disclosure control for low cell counts54. **The OpenSAFELY research platform adheres to the data protection principles of the UK Data Protection Act 2018 and the EU General Data Protection Regulation (GDPR) 2016.** In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organizations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure55. Together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic, and have been informed of the OpenSAFELY analytics platform.

### Study design and population

We will use a cohort study design with entry into the cohort defined as 1st February 2020 to assess how the incidence of severe COVID outcomes varies by household composition, and will follow people up until the earliest of:

1. The outcome of interest (admission to hospital with COVID-19 or death from COVID-19)
2. Deregistration of GP practice from TPP
3. Death from other cause
4. Latest TPP linkage

#### Analysis population

Our main analysis will focus on adults (males and females 18 years and above) registered in a TPP general practice in England on 1st February 2020. We will perform analysis over 2 separate time periods corresponding to case reporting and epidemic changes:

1. Wave 1: 1 Feb 2020 - 31 Aug 2020
2. Wave 2: 1 Sept 2020 - 31 January 2021

Although our analysis population for assessing COVID-19 outcomes will be adults (over the age of 18), the household composition variable that is our main exposure of interest will be derived from all individuals residing at the same address (regardless of age).

For the second wave, we have selected 31 January as a date that we consider far enough into Wave 2 that will allow sufficient numbers for analysis before the introduction of the COVID vaccine. We expect that as more of the 67+ age group are vaccinated, then the impact of household composition may reduce (due to people being incorrectly assigned as at risk of severe outcomes across household composition exposure groups) - we will also perform an analysis up to 31 April 2021 to check this assumption.

#### Inclusion criteria

Adults over the age of 18 alive and under follow-up as of 1st February 2020 registered with a primary care practice using TPP software, with at least one year of continuous GP registration and a valid address or postcode allowing household identification.

#### Exclusion criteria

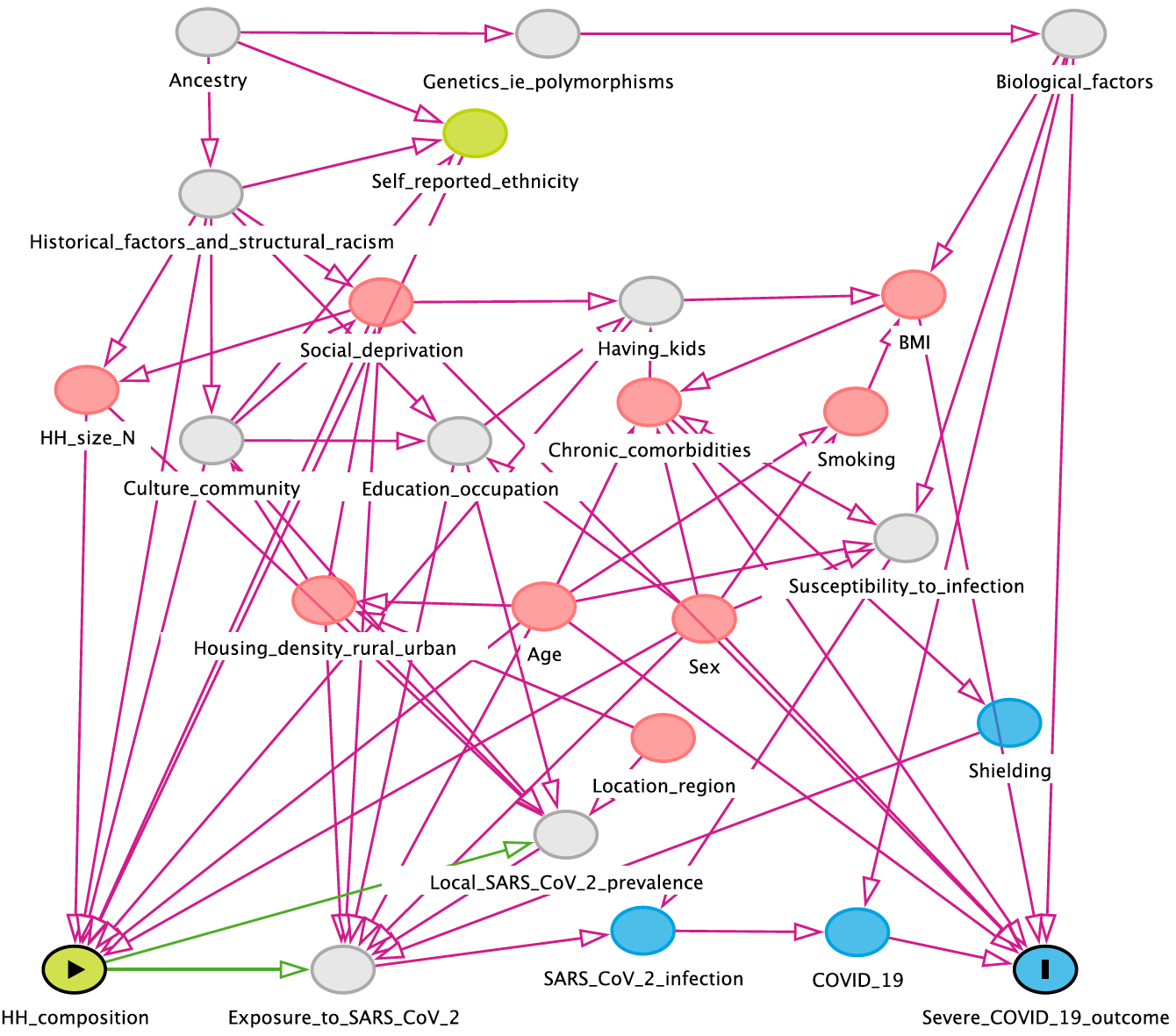
People in households greater than 10 individuals (based on TPP derived household size), in order to reduce the likelihood that we will be including institutions such as care homes in our analyses. People with missing age, sex, or index of multiple deprivation, since these are likely to indicate poor data quality. Individuals identified as living in a care home (based on TPP’s care home indicator).

**Causal framework**

We prepared a causal framework DAG in order to assist in the identification of important covariates when considering the association between household composition and severe COVID-19 outcomes. In the DAG displayed below in Figure 1 we have included variables that are not available in OpenSAFELY data (i.e. unobserved, denoted by light grey nodes) and have also included ethnicity and socioeconomic status. For ethnicity and socioeconomic status, despite their importance to the DAG, we made the a-priori decision to stratify rather than adjust for these variables, as we were not interested in trying to account for (i.e. remove) the effect of these variables in our analysis. Instead, we wanted to be able to investigate how the effect of household composition on severe COVID-19 outcomes varied by each category of socioeconomic status and ethnicity. We therefore made the a-priori decision to report results by category of (1) ethnicity and (2) socioeconomic status in order to assess and report on the impact of household composition on severe COVID-9 outcomes within individual strata of these characteristics.

The minimum sufficient adjustment set implied by a reduced version of Figure 1 that does not include social deprivation or ethnicity is: age, chronic comorbidities, housing density (rural/urban), sex.

**Figure 1 Causal framework DAG**



### Study measures

#### Primary exposure

Our main exposure of interest will be household composition (as at 1 Feb), with respect to the age of people living in the household across age categories, with particular interest in whether an individual lives within (1) within a multigenerational household and (2) whether working age adults are included as part of the household. We will therefore assign categories of 0-17 yrs (child), 18-29 yrs (younger working age adult), 30-66 yrs (older working age adult), and 67+ yrs (adult likely to have retired). Potential exposure categories will be as shown in Table 1 below.

**Table 1: household composition exposure categories**

| **Category number** | **Category description** | **0-17** | **18-29** | **30-66** | **67+** |
| --- | --- | --- | --- | --- | --- |
| 1 | Single generation 1 |  | **x** |  |  |
| 2 | Single generation 2 |  |  | **x** |  |
| 3 | Single generation 3 |  |  |  | **x** |
| 4 | Two generations 1 | **x** | **x** |  |  |
| 5 | Two generations 2 | **x** |  | **x** |  |
| 6 | Two generations 3 | **x** |  |  | **x** |
| 7 | Two generations 4 |  | **x** | **x** |  |
| 8 | Two generations 5 |  | **x** |  | **x** |
| 9 | Two generations 6 |  |  | **x** | **x** |
| 10 | Multi-generations 1 | **x** | **x** | **x** |  |
| 11 | Multi-generations 2 | **x** | **x** |  | **x** |
| 12 | Multi-generations 3 | **x** |  | **x** | **x** |
| 13 | Multi-generations 4 |  | **x** | **x** | **x** |
| 14 | Multi-generations 5 | **x** | **x** | **x** | **x** |

We will have a high-level version of the household composition variable for initial descriptive analysis analysing all people 18 years and over that includes three broad categories corresponding to households made up of people from a single generation (individual can be any of categories 1-3 from Table 1), from two generations (any of categories 4-9 from Table 1) or from more than two i.e. multiple generations (any of categories 10-14 from Table 1). For subsequent analysis steps, we will present results stratified by age group, with more granular household composition categories that are specific to the age group being analysed (e.g. for the over 67 year olds we will present granular categories specific to the over 67 year olds, namely categories 3, 6, 8, 9, 11-14 from Table 1) provided numbers of outcomes are sufficient within each category. The baseline categories for each age strata will also differ, and will be the category for that age strata containing the only a single generation.

#### Outcomes

Our outcomes of interest are:

1. Hospitalisation with COVID (defined as a COVID-19 ICD-10 code in the primary diagnosis field, ascertained from SUS data).
2. COVID-19 related death defined as a COVID-19 ICD-10 code anywhere on the death certificate (ascertained from ONS death certificate data and CPNS data, with ONS death date used if present in both)

We will also analyse non-COVID death (defined as death from any other cause on the death certificate from ONS death certificate data) in order to assess whether the results for our COVID-death analysis are specific to COVID or not.

**Covariates:**

For the association between household composition and severe COVID outcomes, we will either (1) adjust for or (2) report results stratified by the following covariates (adjusted unless specified):

* Age (spline)
* Sex
* BMI
* Smoking
* Geographic area e.g. Upper Tier Local Authority (UTLA), or other geographic level that is available
* Chronic comorbidities shown to be associated with poor COVID-19 outcomes2
* Eligibility for shielding3  (based on presence of comorbidities)
* Density of housing (5 category variable describing rural/urban)
* Ethnicity [stratified]
* SES [stratified]

We will then explore and report the (absolute and relative) effects by categories of each of the following variables:

* Age (categorical in 5 year age groups)
* Sex
* Household size\*

\*Our main exposure variable (household composition) does not capture the size of the household so we will assess this by reporting results by categories of household size, with particular interest on the impact of living in houses that contain large numbers of working-age adults. We are aware that co-linearity between household size and household composition is a possibility and these will be explored/described and alternative parameterisations investigated if needed.

### Statistical analysis

#### Main analysis (by age group and stratified by ethnicity and SES)

We will describe the proportion of individuals within each exposure and outcome category, by the covariates. We will then describe the absolute rate of outcomes according to each category of the household composition exposure variable. For the household composition variable, we will have a high-level version of the variable for initial descriptive analysis where we will include all people 18 years and over assigned to three broad household composition categories corresponding to single generation (individual can be living in a house with composition according to any of categories 1-3 from Table 1), two generations (any of categories 4-9 from Table 1) and multi-generations (any of categories 10-14 from Table 1).

For subsequent analysis steps, we will perform analysis separately for each age group, with more granular household composition categories that are specific to the age group being analysed (e.g. for the over 67 year olds we will present granular categories specific to the over 67 year olds, namely categories 3, 6, 8, 9, 11-14 from Table 1). We are performing separate analyses by age group in order to report on how the effect of household composition differs by (e.g.) those between the age of 50 and 67, and those over the age of 67 (likely to be very different, and important in order to be able to consider results in relation to recommendations for vaccine roll-out).

We will then use multivariable Cox proportional hazards models stratified by geographic area to determine hazard ratios (HRs) for each outcome using robust standard errors to account for clustering by household identifier, split over the time periods specified above (wave 1 and wave 2), stratified by ethnicity and by SES. We will present results for each category of ethnicity and by SES in order to assess how effects differ by ethnic group and by SES, and will also perform statistical tests for interaction by these variables. Violations of the proportional hazards assumption underpinning Cox regression will be explored by testing for a zero slope in the scaled Schoenfeld residuals and by graphical assessment. We will adjust for age using restricted cubic splines, demographic covariates (IMD, BMI and smoking) and fully adjusted models (+additional clinical comorbidities as a single variable coding for 0, 1 or 2+ comorbidities).2 We will report Hazard ratios with 95% confidence intervals. We will check for collinearity of variables. We will also analyse models that are adjusted for only the current JCVI priority variables (age and selected comorbidities).

All analyses will be performed by the time periods specified in the study period section above to account for changes in case reporting and epidemic intensity over time.

Analysis by category of ethnicity will initially be based upon a 5 category variable, but based upon our findings we will also assess reporting by the 16 category variable (for example, to further explore variation in associations within the south Asian group (e.g., Indian, Pakistani, Bangladeshi)).6

#### Calculation of PAF for key strata of interest, and analysis by categories of sex and household size

In order to assess how much of the effect of ethnicity or SES can be attributed to household composition, we will then calculate the Population Attributable Fraction (PAF) within key strata of interest (e.g. south Asian ethnicity), in order to be able to report how much of the effect of south Asian ethnicity severe COVID outcomes can be attributed to household composition.4,5 We will also perform a comparator analysis for illustrative purposes, where we calculate the PAF for additional characteristics within the strata of interest (such as presence of comorbidities).

Finally, we will also analyse whether the association between household composition and COVID outcomes differs by sex, based upon results of previous analysis of household composition and COVID outcomes utilising 2011 census data,7 and assess whether any observed associations differ by size of the household i.e. total number of people residing in the household, and total number of people in each age group in the household (based on the age grouping defined in Table 1).

**Missing data**

For stratifying by ethnicity, we will use multiple imputation (10 imputations) to account for missing ethnicity, with the imputation model including all covariates from the main model and an indicator for the outcome. Those with missing body mass index will be assumed to be non-obese, and those with missing smoking data will be assumed to be never-smokers; multiple imputation will not be used for these variables as they are expected to be missing not at random in UK primary care.10

#### Secondary and sensitivity analyses

We will perform the following sensitivity analyses:

1. Excluding people in households where any single member of the household meets an exclusion criteria.
2. Creating a “buffer” of 5 years below each age category in the household composition variable, within which people would be assigned to the age category above (in order to assess the impact of categorising people with similar ages that span age categories in different categories).
3. Including COVID hospitalisation where COVID is mentioned as a secondary diagnosis in hospital.
4. For the non-COVID death outcome, repeat the analysis so that for those people who died from causes recorded as non-COVID on their death certificate, if their death is within 28 days of the COVID infection date, they are considered to be COVID death.

#### Software and Reproducibility

Data management will be performed using Python 3.8 and SQL with analysis carried out using Stata 16. Code for data management and analysis as well as codelists archived online <https://github.com/opensafely/hh-classification-research>

#### Feasibility and power calculations

Assuming an alpha-level of 0.05 and 2% of patients in the baseline category (e.g. 67+ year olds living only with other 67+ year old’s) experiencing COVID19 death or hospital admission during the study period, we would require:

1. 172807 people in an exposure category to detect a hazard ratio of 1.1 or larger with 80% power for that category
2. 22805 people in an exposure category to detect a hazard ratio of 1.3 or larger with 80% power for that category
3. 9549 people in an exposure category to detect a hazard ratio of 1.5 or larger with 80% power for that category
4. 3268 people in an exposure category to detect a hazard ratio of 2.0 or larger with 80% power for that category

Feasibility analysis indicates that in the cohort, all of our main exposure categories have greater than 3268 people in them, and only two categories have less than 9549 people. Considering the cohort by ethnicity, the majority of both the white and south asian ethnicities contain over 9549 people within each of our main exposure categories.

#### Limitations

1. We are lacking the following variables that may be strongly associated with poor COVID outcomes:
   1. Household crowding
   2. Occupation - depending on availability of ONS survey data linked to OpenSAFELY primary care data, we may be able to obtain occupation data for a small subset of our cohort, in order to assess the potential impact on our results of not having full occupation data for the cohort. Based on the availability of this data, we will consider using quantitative bias analysis to assess the potential extent of confounding from high-risk occupation among working-age adults.
   3. Other neighbourhood characteristics (such as pollution)
2. It is not possible to censor people if their address changes during the study period. Although for many people this may result in deregistration from TPP (if they move GP) and therefore end of follow-up, there may be some people who move house but remain with the same GP, and their household composition may change with the move and we will not be able to track this.
3. We know household composition is not perfect because of people not updating their address etc, but data quality is expected to be better for homeowners than renters because TPP have made a correction for house sales, so imperfection is non-random and we may have more misclassification of household composition for younger people, those with lower SES etc
4. Household id/size in TPP is only generated on 1st Feb 2020, so any updates during the pandemic will not be captured

**References**

1. <https://www.gov.uk/government/publications/priority-groups-for-coronavirus-covid-19-vaccination-advice-from-the-jcvi-25-september-2020/jcvi-updated-interim-advice-on-priority-groups-for-covid-19-vaccination>
2. [Williamson EJ, Walker AJ, Bhaskaran K, *et al.* OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature* 2020; published online July 8. DOI:](http://paperpile.com/b/T9RN93/68Qy)[10.1038/s41586-020-2521-4](http://dx.doi.org/10.1038/s41586-020-2521-4)[.](http://paperpile.com/b/T9RN93/68Qy)
3. [Public Health England. Guidance on shielding and protecting people who are clinically extremely vulnerable from COVID-19. GOV.UK. Updated 14 July 2020.](http://paperpile.com/b/T9RN93/uS2w) <https://www.gov.uk/government/publications/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19/guidance-on-shielding-and-protecting-extremely-vulnerable-persons-from-covid-19> [(accessed July 16, 2020).](http://paperpile.com/b/T9RN93/uS2w)
4. Brooks-Pollock E, Danon, L. Defining the population attributable fraction for infectious diseases. Int J Epidemiology, 2017 Jun 1;46(3):976-982
5. Brooks-Pollock E *et al* The Population Attributable Fraction (PAF) of cases due to gatherings and groups with relevance to COVID-19 mitigation strategies. medRxiv preprint: doi: https://doi.org/10.1101/2020.03.20.20039537
6. Mathur R, Rentsch CT, Morton CE, *et al*. Ethnic differences in COVID-19 infection, hospitalisation, and mortality: an OpenSAFELY analysis of 17 million adults in England. *Accepted for publication in The Lancet.*
7. Nafilyan V *et al*. Ethnicity, household composition and COVID-19 mortality: a national linked data study. Journal of the Royal Society of Medicine [https://doi.org/10.1177/0141076821999973](https://doi.org/10.1177%2F0141076821999973)