Coronavirus (COVID-19) related deaths by ethnic group, England and Wales methodology

Technical appendix for the updated ethnic contrasts in deaths involving coronavirus COVID-19 article.

Contact:
Chris White and Daniel Ayoubkhani
health.data@ons.gov.uk
+44 (0)1633 455865

Release date:
16 October 2020

Next release:
To be announced

Table of contents

1 . Introduction

This technical appendix provides the detail around the data and methods used in the article Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England and Wales: deaths occurring 2 March to 28 July 2020.

2 . Data

These analyses are based on a unique linked dataset that encompasses Census 2011 records, death registrations in England and Wales, and hospital episode statistics (HES) with England coverage only. It was created by:

- linking the 2011 Census to NHS Patient Register (PR) records between 2011 and 2013, where NHS number was added to those Census records identified in the Patient Register

- using NHS number and a deterministic match key linkage method where NHS number was unavailable – death registrations were linked to 2011 Census records up to 24 August 2020

- joining HES records from April 2017 onto the census-deaths linked data using a combination of date of birth and NHS number

The linked population has a very similar distribution across a range of characteristics as the full census population, and so can be considered representative of the general population of England and Wales in 2011. Examination of linkage rates for ethnic groups showed distributions at 2011 Census and the linked population were relatively consistent across all categories, although there was more significant variation in unlinked records. For all ethnic groups, linkage rates of NHS number exceeded 80% in all cases.

The study population included all usual residents coded to an ethnic group in 2011 and not known to have died before 2 March 2020 (number surveyed (N) equals 48,468,645). Those enumerated in 2011 answering the “Intention to Stay” question, because they had entered the UK in the year before the 2011 Census took place, were excluded from the analyses because of their high propensity to have left the UK before the analysis period under investigation. However, this leaves uncertainty in the extent of emigration of usual residents between 27 March 2011 and 2 March 2020, which is dealt with later in this section. Analyses using HES data were limited to usual residents thought to be alive on 2 March 2020 in England only (number surveyed (N) equals 45,842,599).

We use data from the Office for National Statistics (ONS) Longitudinal Study and the International Passenger Survey (IPS) to estimate emigration between March 2011 and March 2020 by broad age group and ethnicity. As we only have IPS data up to year-ending March 2019, we assume emigration rates observed between March 2019 and March 2020 are the same as those observed in the previous year.

These emigrations and deaths are used to ensure that the analysis refers to people still in the population of England and Wales and at risk of the coronavirus (COVID-19) from 2 March 2020, by applying out-migration adjustment factors to deplete the population sizes resulting from expected emigration since the 2011 Census.

The number of deaths occurring between 2 March 2020 and 28 July 2020 that were registered by 24 August 2020 amounted to 253,194. Of these, 229,983 were successfully linked to a 2011 Census record (90.8%). However, only 229,929 were usable because 48 were linked to non-usual residents and six to individuals over 110 years of age, which we excluded from our study population. Of these, 216,406 were resident in England and 13,523 were resident in Wales.

Causes of death were defined using the International Classification of Diseases, 10th Revision (ICD-10). Deaths involving COVID-19 include those with an underlying cause, or any mention, of ICD-10 codes U07.1 (COVID-19, virus identified) or U07.2 (COVID-19, virus not identified).

The study population is not currently refreshed with new births or immigrations. Some COVID-19 deaths will therefore have occurred to immigrants entering the country since 2011; deaths involving COVID-19 to those born since the 2011 Census and resident in England and Wales will be very small as they will be nine years old or younger.

3 . Hospital episode statistics

For this analysis, we used hospital episode statistics (HES) data from April 2017 sourced from three datasets: Accident and Emergency (AE), Outpatients (OP) and Admitted Patient Care (APC). The information within these three datasets is at episode level (each finished period of care under a consultant). We created a person-level dataset from the record-level HES data to preserve all information when linking to 2011 Census and deaths data.

The analytical variables derived from HES were:

- flags for ICD10 diagnoses codes of interest in the OP and APC datasets
- the total number of episodes per NHS number and date of birth (our method to identify an individual) for all datasets.
- the number of first admission episode flags in the APC dataset to derive the number of admissions per person.
- the number of days spent in admitted patient care from the APC dataset

These were then aggregated up to the person level by stacking and deduplicating all datasets on NHS number and date of birth, to create one row per individual. Records with blank or invalid NHS numbers and/or dates of birth were dropped, as these could not be linked to the Census. The total number of individuals in our HES data was 43,562,505. The HES data was then linked to the Census and deaths data through a simple deterministic link on NHS number and date of birth. 31,903,383 of the HES records linked to the 2011 Census (73.2%). The remaining unlinked 26.8% are likely to have not been registered on the 2011 Census, because they were born after 27 March 2011, migrated to England after that date or were not enumerated at the 2011 Census despite being a resident. In addition, some individuals in the unlinked group may not have been able to have an NHS number assigned to their Census record. This could be due to conflicting addresses, name changes or other reasons, and thus the deterministic and probabilistic linkage methods would have failed, though this is only in a small number of cases.

4 . Age-standardisation method

This Microsoft Excel template demonstrates how age-standardised rates and 95% confidence intervals are calculated.

Age-standardised rates are calculated as follows:

$$\frac{\sum_i w_i r_i}{\sum_i w_i} = \times 100,000\, study\, population\, alive\, at\, 2\, March\; 2020$$

where:

- i is the age group
- w_i is the number, or proportion, of individuals in the standard population in age group i
- r_i is the observed age-specific rate in the subject population in age group i, given by:

where:

$$r_i = d_i / n_i$$

- $d_i$is the observed number of deaths in the subject population in age group i
- n_i is the population at risk in age-group i

The age-standardised rate is a weighted sum of age-specific death rates where the age-specific weights represent the relative age distribution of the standard population (in this case the 2013 European Standard Population (ESP)). The variance is the sum of the age-specific variances and its standard error is the square root of the variance:

$$SE(ASR) = \sqrt{\frac{\sum\left(w_i^2 \frac{r_i^2}{d_i}\right)}{\left(\sum w_i\right)^2}}$$

- r_i is the crude age-specific rate in the local population in age group i
- d_i is the number of deaths in the local population in age group i

Confidence intervals

The mortality data in this release are not subject to sampling variation as they were not drawn from a sample. Nevertheless, they may be affected by random variation, particularly where the number of deaths or probability of dying is small. To help assess the variability in the rates, they have been presented alongside 95% confidence intervals.

The choice of the method used in calculating confidence intervals for rates will, in part, depend on the assumptions made about the distribution of the deaths data these rates are based on. Traditionally, a normal approximation method has been used to calculate confidence intervals on the assumption that deaths are normally distributed. However, if the number of deaths is relatively small (fewer than 100), it may be assumed to follow a Poisson probability distribution. In such cases, it is more appropriate to use the confidence limit factors from a Poisson distribution table to calculate the confidence intervals instead of a normal approximation method.

The method used in calculating confidence intervals for rates based on fewer than 100 deaths was proposed by Dobson and others (1991) as described in APHO (2008). In this method, confidence intervals are obtained by scaling and shifting (weighting) the exact interval for the Poisson distributed counts (number of deaths in each year). The weight used is the ratio of the standard error of the age-standardised rate to the standard error of the number of deaths.

The lower and upper 95% confidence intervals are denoted as ASR lower and ASR upper, respectively, and calculated as:

$$ASR_{lower} = ASR + (D_I - D) \cdot \sqrt{\frac{v\ (ASR)}{v\ (D)}}$$

$$ASR_{upper} = ASR + (D_u - D) \cdot \sqrt{\frac{v\ (ASR)}{v\ (D)}}$$

where:

- D_l and D_u are the exact lower and upper confidence limits for the number of deaths, calculated using confidence limit factors from a Poisson probability distribution table
- D is the number of deaths in each year
- v(ASR) is the variance of the age-standardised rate
- v(D) is the variance of the number of deaths

Where there are 100 or more deaths in a year, the 95% confidence intervals for age-standardised rates are calculated using the normal approximation method:

$ASR_{LL/UL}$ = ASR± 1.96*SE

where:

$ASR_{LL/U}$ represents the upper and lower 95% confidence limits, respectively, for the age-standardised rate and SE is the standard error.

5 . Modelling analysis

We use Cox proportional hazard models to assess how the risk of dying with coronavirus (COVID-19) varies among ethnic groups once we adjust for a range of geographical, demographic, socio-economic, household, occupational exposure and health-related factors. Most individual characteristics are retrieved from the 2011 Census, except for pre-existing health conditions which are derived from hospital episode statistics (HES) records from April 2017 onwards.

We model the hazard of dying with COVID-19 between 2 March 2020 and 28 July 2020. In our analytical dataset, we include all those who died of any cause during this period and a weighted random sample of those who did not (the sampling fractions are 5% for the White population and 20% among other ethnic groups combined). The regression estimates are further weighted using the probability not to have migrated between 2011 and 2020.

We estimate separate models for males and females, as the risk of death involving COVID-19 differs markedly by sex. We also estimate separate models for people in private households and in care homes according to place of residence in 2011 and the Patient Register in 2019. We present results from several models, adding different control variables step by step. This allows us to see how the differences across ethnic groups vary as we include further explanatory variables.

All our models are adjusted for age. We include age as a second-order polynomial to account for the non-linear relationship between age and the hazard of death involving COVID-19.

We then adjust for geographical factors. The probability to be infected by COVID-19 is likely to vary by region of residence. We therefore allow the baseline mortality hazard to vary by local authority district. We also adjust for population density for the Lower Super Output Area (LSOA) of residence at the time of the 2011 Census. To account for the non-linear relationship between population density and the hazard of death involving COVID-19 we include population density as a second-order polynomial, allowing for different slopes for the top 1% of the population density distribution to account for outliers.

We then account for deprivation and wider measures of socio-economic status. We adjust for neighbourhood deprivation by adding decile of the Index of Multiple Deprivation (IMD) 2015 at the time of the 2011 Census in our model. The IMD is an overall measure of deprivation based on factors such as income, employment and health.

We also adjust for the level of household deprivation, a summary measure of disadvantage based on four selected household characteristics (employment, education, health and housing). We include in our model the highest level of qualification (degree, A-level or equivalent, GCSE or equivalent, no qualification) of the individual, and the National Statistics Socio-Economic Classification (NS-SEC) of the household head (higher managerial, administrative and professional occupations, intermediate occupations, routine and manual occupations, never worked or long-term unemployed, not applicable).

We further adjust for household composition and circumstances. We include in our models the number of people in the household, the family type (not a family, couple with children, lone parent), and binary variables for living in a multigenerational household (defined as three generations living together) or with any children (aged 18 years or under). We also adjust for the tenure of the household (owned outright, owned with mortgage, social rented, private rented, other).

In addition, we adjust for a set of measures of occupational exposure. We include binary variables indicating if the individual is a key worker, and if so, what type. This data is taken from occupation as recorded on the 2011 Census. We also include a binary variable indicating if anyone in the household is a key worker. We account for exposure to diseases and contact with others using scores ranging from 0 (no exposure) to 100 (maximum exposure). Exposure to disease and physical proximity scores were originally obtained using O*NET data based on US Standard Occupational Classification (SOC) codes and were mapped to UK SOC codes. The derivation of the scores is in line with the methodology previously used by the Office for National Statistics (ONS). We include these scores for all individuals with a valid occupation and derive the maximum value amongst all household members.

Finally, we adjust for several measures of health. We include in the model self-reported health status (very good, good, fair, bad, very bad) and whether the individual has activity limitation (disability) (not limited, daily activity limited a lot, daily activity limited a little) as recorded in the 2011 Census. We also adjust for pre-existing conditions derived from HES records from April 2017 onwards, as discussed in Section 4 of the main article:

- history of cancer
- history of cardiovascular disease
- history of digestive system conditions
- history of mental health conditions
- history of metabolic conditions
- history of musculoskeletal conditions
- history of neurological conditions
- history of renal conditions
- history of respiratory conditions
- number of admitted patient care (APC) admissions (0, 1, 2 to 3, 4 to 5, 6 to 9, 10 plus)
- number of days spent in APC (0, 1, 2 to 4, 5 to 9, 10 to 19, 20 to 39, 40 to 69, 70 plus)

To allow for the effect of all these health-related factors to vary depending on the age of the individuals, we interact each of them with a binary variable indicating if the individual is aged 70 years or over.

In the article we report the hazard ratios for each ethnic minority group relative to the White population for people in private households in England, after adjusting for age, geographical factors, socio-economic factors and health-related variables. The corresponding model goodness-of-fit statistics can be found in the dataset.

We find that much of the difference in COVID-19 mortality risk across ethnic groups that is attributable to health-related factors can be explained by self-reported health and disability status in 2011. For most ethnic groups, further adjusting for hospital-based comorbidities has little impact on the hazard ratios, though there is notable attenuation for Bangladeshi and Pakistani males. For females, including hospital-based comorbidities increases the hazard ratios for several ethnic groups, most notably for those of Black African or Chinese ethnic background.

Figure 1: Rate of death involving COVID-19 by ethnic group and sex relative to the White population for people in private households, England, 2 March to 28 July 2020

Download the data

.xlsx

Notes:

1. Cox proportional hazards models adjusting for age, geography (local authority and population density), socio-economic factors (area deprivation, household composition, socio-economic position, highest qualification held, household tenure, multigenerational household flags and occupation indicators (including key workers and exposure to others), and health (self-reported health and disability status in March 2011, and hospital-based comorbidities since April 2017).

2. Figures based on death registrations up to 24 August 2020 that occurred between 2 March 2020 and 28 July 2020 and could be linked to the 2011 Census.

3. Deaths were defined using the International Classification of Diseases, 10th Revision (ICD-10). Deaths involving COVID-19 include those with an underlying cause, or any mention, of ICD-10 codes U07.1 (COVID-19, virus identified) or U07.2 (COVID-19, virus not identified).

4. Other ethnic group encompasses Asian other, Black other, Arab, and other ethnic group categories in the classification.

5. Error bars not crossing the x axis at value 1.0 denote a statistically significant difference in relative rates of death.