Office for National Statistics

Deaths involving COVID-19 by religious group and ethnic group, England: methodology

Detailed quality and methodology information for “Deaths involving COVID-19 by religious group, England: 24 January 2020 to 28 February 2021” and “Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 24 January 2020 to 31 March 2021.

Contact:
Matt Bosworth and Chris White
health.data@ons.gov.uk
+44 (0)1633 455 865

Release date:
14 May 2021

Next release:
To be announced

Table of contents

1 . Overview

This methodology provides details of the data and methods used in the following publications:

- Deaths involving COVID-19 by religious group, England: 24 January 2020 to 28 February 2021
- Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: 24 January 2020 to 31 March 2021

2 . Data sources

The analyses in the above articles are based on a unique linked dataset that encompasses Census 2011 records, death registrations, Hospital Episode Statistics (HES) and primary care records retrieved from the the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), 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, death registrations data were linked to 2011 Census records
- joining HES records from April 2017 and GPES records from January 2015 onto the Census-deaths linked data using NHS number

The study population comprises 29.3 million respondents to the 2011 Census, aged between 30 and 100 years in 2020, that had not died before 24 January 2020 and could be linked to the 2011 to 2013 Patient Register and GDPPR dataset (which comprises active NHS patients at the start of the pandemic, and so are unlikely to have emigrated between 2011 and 2020).

The study population is not currently refreshed with immigrations. Some coronavirus (COVID-19) deaths will therefore have occurred to immigrants entering the country since 2011.

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), U07.2 (COVID-19, virus not identified) or U09.9 (post-COVID-19 condition).

3 . Hospital variables

For this analysis, we used Hospital Episode Statistics (HES) data from April 2017 sourced from Admitted Patient Care (APC) records. The information within this dataset 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:

- 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 were 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 were 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 because of conflicting addresses, name changes or other reasons, and so the deterministic and probabilistic linkage methods would have failed, though this is only in a small number of cases.

4 . Primary care variables

Primary care records were extracted from the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) dataset, which contains approximately 35,000 clinical codes (including diagnoses, measurements, and prescriptions) for active NHS patients at the start of the coronavirus (COVID-19) pandemic.

The GDPPR dataset was firstly used to identify individuals in the study population in 2020; of 43.6 million respondents to the 2011 Census in England who could be linked to the 2011 to 2013 Patient Registers and had not died before January 2020, 34.9 million could be linked to at least one GDPPR record. Secondly, as with the HES data, record-level data for relevant conditions were converted to binary (except for body mass index and kidney disease) person-level variables through grouping an individual's records by NHS number.

The GDPPR dataset was used to identify individuals who had primary care contact over the past five years for a range of conditions. These comorbidities were chosen because they were previously implicated in raising risk of death from the coronavirus by the QCOVID algorithm for predicting hospital admission and mortality from COVID-19 in adults and were derived using the same code lists.

We were unable to include some health variables from the QCOVID algorithm either because of an insufficient number of cases for analysis (bone marrow transplant, cerebral palsy, congenital heart disease, and sickle cell disease) or because we do not have permission to use these data (chemotherapy or radiotherapy treatment). The full list of health variables that were included comprises:

- body mass index
- ever having a solid organ transplant
- history of asthma
- history of atrial fibrillation
- history of blood cancer
- history of chronic obstructive pulmonary disease
- history of cirrhosis of the liver
- history of congestive cardiac failure
- history of coronary heart disease
- history of dementia
- history of diabetes
- history of epilepsy
- history of kidney disease
- history of learning disability
- history of mental illness
- history of osteoporotic fracture
- history of Parkinson's disease
- history of peripheral vascular disease
- history of pulmonary hypertension or pulmonary fibrosis
- history of rare neurological conditions (motor neurone disease, multiple sclerosis, myaesthenia, or Huntington's Chorea)
- history of rare pulmonary disorders (cystic fibrosis, bronchiectasis, or alveolitis)
- history of respiratory cancer
- history of rheumatoid arthritis or systemic lupus erythematosus
- history of stroke or transient ischaemic attack
- history of thrombosis or pulmonary embolus
- prescribed anti-leukotriene or long-acting beta blocker medication
- prescribed immunosuppressant medication
- prescribed prednisolone medication

5 . Age-standardisation method

Age-standardised rates (per 100,000 person-years at-risk) are calculated as follows:

$$\frac{\sum_i w_i r_i}{\sum_i w_i} \times 100,000$$

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:

$$r_i = d_i / n_i$$

where:

- d_i is the observed number of deaths in the subject population in age group i
- n_i is the person-years 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:

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

where:

- 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 the APHO public health guide (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 as follows:

$$\mathrm{ASR_{LL/UL}} = \mathrm{ASR} \pm 1.96 * \mathrm{SE}$$

where:

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

6 . Modelling analysis

We use Cox proportional hazard models to assess how the risk of death involving the coronavirus (COVID-19) varies among groups for exposure variables of interest once we adjust for age, residence type (private household, care home, or other communal establishment) and a range of other characteristics; specifically, location, measures of disadvantage, occupation, living arrangements, and pre-pandemic health status.

We model the hazard of dying with COVID-19 during the outcome period. In our analytical dataset, we include all those who died from any cause during this period and a weighted random sample of those who did not.

The hazard function was modelled as follows:

$$h(t) = h_0(t) \times \exp(b_1 x_1 + b_2 x_2 \ldots + b_i x_i)$$

where:

- t is the survival time
- h(t) is the hazard function at time t
- $h_0(t)$ is the baseline hazard at time t
- b_i is the estimated coefficient for the i_{th} covariate
- x_i is the value for the i_{th} covariate

The hazard ratio for the i_{th} term is calculated as:

exp (b_i)

We estimate separate models for males and females, as the risk of death involving COVID-19 differs markedly by sex. We present results from several models, adding different control variables step by step. This allows us to see how differences in risk of death involving COVID-19 vary as we include further explanatory variables.

In our baseline model, we present hazard ratios 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 factors likely to affect the risk of infection but also the risk of having a pre-existing condition too and therefore prognosis.

First, we adjust for residence type (private household, care home, other communal establishments). We use the 2019 NHS Patient Register to update place of residence for individuals recorded as living in a private household on the 2011 Census that had subsequently moved into a care home.

We then adjust for geographical factors, derived from current postcodes held in GPES. 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 of the Lower layer Super Output Area (LSOA). 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) 2019 to the 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 reference person (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)
- household composition (single-adult household, two-adult household, multi-generational household (households with at least one person aged 65 years or over and someone at least 20 years younger), child aged 18 years or under in household)
- tenure of the household (owned outright, owned with mortgage, social rented, private rented, other)

We include an additional "not in a household" level for all household variables for people living in a care home or other communal establishment.

In addition, we adjust for a set of measures of occupational exposure. We include a variable indicating if the individual is a key worker, and if so, what type. These data are 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 disease 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 among all household members.

Most of these characteristics were retrieved from the 2011 Census. We sought to increase the accuracy of the Census variables so that they more accurately reflect living circumstances in 2020 by setting occupational exposure variables to 0 for people who were recorded as living in a private household on the 2011 Census but living in a care home on the 2019 Patient Register. In addition, people aged 10 to 17 years at the time of the 2011 Census were excluded from the calculation of household level variables as they are likely to have left the household.

Finally, we adjust for the number of hospital admissions and number of days spent in admitted patient care over the past three years, derived from NHS Hospital Episode Statistics (HES) records, and the presence of pre-existing health conditions, derived from the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR). 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.

We report the hazard ratios for the exposure variable, after adjusting for age, geographical factors, socio-economic and demographic factors, and health-related variables. A hazard ratio greater than one indicates a greater rate of death involving COVID-19 than the reference group, while a hazard ratio less than one indicates a lower rate of COVID-19 mortality than the reference group. The corresponding model goodness-of-fit statistics can be found in the datasets.

We also report the risk of death involving COVID-19 for the exposure of interest in the first and second waves of the pandemic by extending the models to allow for time-dependent coefficients for the exposure of interest. Deaths occurring from 12 September 2020 onwards were classified as occurring in the second wave.

An experimental estimate of the start of the second wave was defined as 21 August 2020, which corresponds to when the reproduction number (R) increased to above 1 for the first time since it was first reported on 22 May 2020, plus 21 days to allow for a lag between new infections and effects on death rates. The follow-up time of people who were still in the study after 11 September 2020 was split into wave one and wave two periods, with wave one outcomes recorded as censored. We fitted Cox models with stratification of the estimates for the exposure variables on wave one versus wave two, thus assuming a step change in the hazard ratios at the start of the second wave.

7 . Related links

Deaths involving COVID-19 by religious group, England: deaths occurring between 24 January 2020 and 28 February 2021
Article | Released
Age-standardised rates of death involving the coronavirus (COVID-19) by religion group, using statistical models to adjust for location, measures of disadvantage, occupation, living arrangements, and pre-existing health conditions. Compares the risk of COVID-19 mortality in two discrete periods aligned to each wave of the pandemic.

Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England: deaths occurring 24 January 2020 to 31 March 2021
Article|Updated on 26 May 2021
Estimates of differences in COVID-19 mortality risk by ethnic group for deaths occurring up to 31 March 2021, using linked data from the 2011 Census, death registrations, and primary care and hospital records. Risk of COVID-19 mortality is compared between the first and second waves of the pandemic.

Coronavirus (COVID-19) latest insights
Interactive tool | Updated as and when data become available
Explore the latest data and trends about the coronavirus (COVID-19) pandemic from the ONS and other official sources.

Coronavirus (COVID-19) roundup
Blog | Updated as and when new data become available
Catch up on the latest data and analysis related to the coronavirus pandemic and its impact on our economy and society.