Coronavirus (COVID-19) related deaths by disability status, England methodology

Technical appendix to accompany updated estimates of differences in COVID-19 mortality risk by self-reported disability status and diagnosed learning disability status for deaths occurring up to 20 November 2020.

Contact:
Daniel Ayoubkhani and Matt Bosworth
health.data@ons.gov.uk
+44 (0) 1633 455825

Release date:
11 February 2021

Next release:
To be announced

Table of contents

1 . Introduction

This article provides details of the data and methods used in the article Updated coronavirus (COVID-19) related deaths by disability status, England: 24 January to 20 November 2020.

2 . Data sources

These analyses 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 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 the NHS number was added to those Census records identified in the Patient Register
- using the NHS number and a deterministic match key linkage method where the NHS number was unavailable – death registrations for deaths occurring to 20 November 2020 and registered by 31 December 2020 were linked to 2011 Census records
- joining HES records from April 2017 and GPES records from January 2010 onto the census-deaths linked data using the NHS number

The study population comprises 29,295,161 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 Registers and GDPPR dataset (which comprises active NHS patients at the start of the pandemic, and are unlikely to have emigrated between 2011 and 2020).

The study population is not currently refreshed with immigrations. Some 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 the coronavirus (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).

3 . Definitions

Disability

To define disability in this publication, we refer to the self-reported answers to the 2011 Census question, "Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months? - Include problems related to old age" (“Yes, limited a lot”, or “yes, limited a little”, or “no”).

The limited a little and limited a lot categories are referred to in this article as “less-disabled” and “more-disabled” respectively, whereas people reporting no limitation on their activities are referred to as “non-disabled”. The distinction between less-disabled and more-disabled is based solely on 2011 Census data and not inferred from any other information. Therefore, it only implies a difference based on self-reported activity restrictions.

This is slightly different to the current Government Statistical Service (GSS) harmonised "core" definition: this identifies a "disabled" person as a person who self-reports having a physical or mental health condition or illness that has lasted or is expected to last 12 months or more that reduces their ability to carry-out day-to-day activities.

The GSS definition is designed to reflect the definitions that appear in legal terms in the Disability Discrimination Act 1995 and the subsequent Equality Act 2010.

Learning disability

Learning disability is identified from clinical diagnoses made in primary care, according to a set of 256 diagnostic codes found in routinely collected electronic health records. These codes can be found in the look-up tables published by NHS Digital.

4 . 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 the 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 the 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 the NHS number and the date of birth. A total of 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.

5 . 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 24 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 (listed in this section) were converted to binary (except for body mass index and kidney disease), person-level variables by grouping by NHS number.

The GDPPR dataset was used to identify individuals who had primary care contact over the past 10 years for a range of conditions. These comorbidities were chosen because they were previously implicated in raising risk of death from COVID-19 by the QCOVID algorithm for predicting hospital admission and mortality from COVID-19 in adults.

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 rare pulmonary disorders (cystic fibrosis, bronchiectasis, or alveolitis)
- history of dementia
- history of diabetes
- history of epilepsy
- history of kidney disease
- history of osteoporotic fracture
- history of rare neurological conditions (motor neurone disease, multiple sclerosis, myasthenia, or Huntington’s Chorea)
- history of Parkinson’s disease
- history of peripheral vascular disease
- history of pulmonary hypertension or pulmonary fibrosis
- history of respiratory cancer
- history of rheumatoid arthritis or systemic lupus erythematosus
- history of mental illness
- history of stroke or transient ischaemic attack
- history of thrombosis or pulmonary embolus
- prescribed immunosuppressant medication
- prescribed anti-leukotriene or long-acting beta blocker medication
- prescribed prednisolone medication

6 . 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:

$$\frac{\sum_i w_i r_i}{\sum_i w_i} = \mathrm{x}100,000 study population$$

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 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:

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

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 on which these rates are based. 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:

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

where:

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

7 . Modelling analysis

We use Cox proportional hazard models to assess how the risk of death involving the coronavirus (COVID-19) varies by self-reported disability status and learning disability status. This is once we adjust for residence type (private household, care home, or other communal establishment) and a range of geographical, demographic, socio-economic, household, occupational exposure and health-related factors.

Most individual characteristics are retrieved from the 2011 Census. This is except for hospital admissions and pre-existing health conditions, which are derived from hospital episode statistics (HES) records from April 2017 onwards and General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) from January 2010 onwards, respectively.

Covariates included in the Cox proportional hazards model

Age variables

Variable: Single year of age
Coding: Second-order polynomial

Residence variables

Variable: Residence type
Coding: Dummy variables representing private household, care home and other communal establishment

Geographical variables

Variable: Local authority district
Coding: Dummy variables representing local authority districts

Variable: Population density
Coding: Second-order polynomial, allowing for a different slop beyond the 99th percentile of the distribution to account for extreme values

Socio-economic variables

Variable: Index of Multiple Deprivation (IMD)
Coding: Dummy variables representing deciles of deprivation, or communal establishment

Variable: Household deprivation
Coding: Not deprived, deprived in one dimension, deprived in two dimensions, deprived in three dimensions, deprived in four dimensions, communal establishment

Household deprivation is defined according to four dimensions:

- employment (at least one household member is unemployed or long-term sick, excluding full-time students)
- education (no household members have at least Level 2 education, and no one aged 16 to 18 years is a full-time student)
- health and disability (at least one household member reported their health as being “bad” or ”very bad” or has a long-term health problem)
- housing (the household's accommodation is overcrowded, with an occupancy rating negative 1 or less, or is in a shared dwelling, or has no central heating)

Variable: Household tenure
Coding: Own outright, own with mortgage, social rented, private rented, other, communal establishment

Variable: National Statistics Socio-Economic Classification of household head (NS-SEC)
Coding: Higher managerial, administrative and professional occupations, intermediate occupations, routine and manual occupations, never worked, not applicable, communal establishment

Variable: Level of highest qualification
Coding: Degree, A-level or equivalent, GCSE or equivalent, no qualification

Ethnicity variables

Variable: Ethnicity
Coding: Bangladeshi, Black African, Black Caribbean, Chinese, Indian, Mixed, Other, Pakistani, White British, White other

Household variables

Variable: Household size
Coding: 1 to 2 people, 3 to 4 people, 5 to 6 people, 7 or more people, communal establishment

Variable: Family type
Coding: Not a family, couple with children, lone parent, communal establishment

Variable: Household composition
Coding: Single-adult household, two-adult household, multi-generational household (at least one person aged over 65 years and someone at least 20 years younger), other 3 or more adults, child in household, communal establishment

Occupational exposure variables

Variable: Key worker type
Coding: Education and childcare, food and necessity goods, health and social care, public services, national and local government, public safety and national security, transport, utilities and communication, not a key worker

Variable: Key worker in the household
Coding: Yes, no, communal establishment

Variable: Exposure to disease
Coding: Score ranging from 0 (no exposure) to 100 (maximum exposure)

Variable: Proximity to others
Coding: Score ranging from 0 (no exposure) to 100 (maximum exposure)

Variable: Household exposure to disease
Coding: Maximum “exposure to disease” score in each household categorised as 0 to 20.0, 20.1 to 40.0, 40.1 to 60.0, 60.1 to 80.0, and 81 to 100, communal establishment

Variable: Household proximity to others
Coding: Maximum “proximity to others” score in each household categorised as 0 to 20.0, 20.1 to 40.0, 40.1 to 60.0, 60.1 to 80.0, and 81 to 100, communal establishment

Health variables

Variable: Number of admissions to Admitted Patient Care
Coding: 0, 1, 2 to 3, 4 to 5, 6 to 9, 10 or more

Variable: Number of days spent in Admitted Patient Care
Coding: 0, 1, 2 to 4, 5 to 9, 10 to 19, 20 to 39, 40 to 69, 70 or more

Variable: Body Mass Index (kilograms/square metres)
Coding: less than 18.5, 18.5 to 25, 25 to 30, more than or equal to 30, missing

Variable: Chronic kidney disease (CKD)
Coding: No CKD, CKD stage 3, CKD stage 4, CKD stage 5

Variable: Cancer and immunosuppression
Coding: Blood cancer, respiratory cancer, solid organ transplant, prescribed immunosuppressant medication by GP, prescribed leukotriene or long-acting beta blockers, prescribed regular prednisolone

Variable: Other conditions
Coding: Diabetes, chronic obstructive pulmonary disease (COPD), asthma, rare pulmonary diseases, pulmonary hypertension or pulmonary fibrosis, coronary heart disease, stroke, atrial fibrillation, congestive cardiac failure, venous thromboembolism, peripheral vascular disease, dementia, Parkinson's disease, epilepsy, rare neurological conditions, severe mental illness, osteoporotic fracture, rheumatoid arthritis or systemic lupus erythematosus, cirrhosis of the liver

We model the hazard of death involving COVID-19 between 24 January 2020 and 20 November 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 1% for the non-disabled population and 5% for each of the self-reported disability status and learning disability status populations; separate samples are drawn for fitting the self-reported disability and learning disability models).

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 by self-reported disability status and learning disability status 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.

We adjust for residence type (private household, care home or 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. 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 layer 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)
- 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), and a child aged 18 years or under in household)

We also adjust for the tenure of the household (owned outright, owned with mortgage, social rented, private rented, other). We include an additional “communal establishment” 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 binary variables indicating if the individual is a key worker, and if so, what type. These data are taken from occupations 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 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 household characteristics to “communal establishment” and occupational exposure variables to zero 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. Similarly, 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.

In the fully adjusted model, 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 less- and more-disabled people (those who said in the 2011 Census that their day-to-day activities were “limited a little” or “limited a lot”, respectively) relative to the non-disabled group (those reporting no limitation on their day-to-day activities), and for people with a learning disability relative to the no learning disability group, after adjusting for age, geographical factors, socio-economic factors, ethnicity, and health-related variables. The corresponding model goodness-of-fit statistics can be found in the datasets.

We also report the risk of death involving COVID-19 by self-reported disability status and learning disability status in the first and second waves of the pandemic (after adjusting for age, residence type, socio-economic and demographic factors, and comorbidities) by extending the fully adjusted models to allow for time-dependent disability coefficients. Deaths occurring from 12 September 2020 onwards were considered to be 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) in England increased to above 1 for the first time since the R 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 disability estimates on wave one versus wave two, thus assuming a step change in the hazard ratios at the end of the first wave plus three weeks.

8 . Related links

Updated estimates of coronavirus (COVID-19) related deaths by disability status, England: 24 January to 20 November 2020
Article | Released 11 February 2020
Estimates of differences in COVID-19 mortality risk by self-reported disability status and diagnosed learning disability status for deaths occurring up to 20 November 2020, using linked data from the 2011 Census, death registrations, and primary care and hospital records.