Office for
National Statistics

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

Technical appendix to accompany updated estimates of differences in coronavirus (COVID-19) mortality risk by self-reported disability status for deaths occurring up to 9 March 2022.

Contact:
Seth Thomas
health.data@ons.gov.uk
+44 1633 651602

Release date:
9 May 2022

Next release:
To be announced

Table of contents

1 . Overview

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 2020 to 9 March 2022.

2 . Data sources

These analyses use data from the Office for National Statistics' (ONS) Public Health Data Asset (PHDA). The PHDA is a unique linked dataset that encompasses 2011 Census 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). The PHDA covers England only and was created by:

- using deterministic and probabilistic linkages, NHS numbers were obtained for individuals present in the 2011 Census and in the NHS Patient Register (PR) records between 2011 and 2013
- using NHS number, death registrations data were linked to the 2011 Census records
- joining HES records from April 2017 and GPES records from January 2000 onto the Census-deaths linked data using NHS number

We linked vaccination data from the National Immunisation Management Service (NIMS) to the PHDA based on NHS number to adjust for vaccination status.

The study population comprises 29.8 million respondents to the 2011 Census, that:

- were aged between 30 and 100 years in 2020
- had not died before 24 January 2020
- could be linked to the 2011 to 2013 Patient Registers and GDPPR dataset (which comprises active NHS patients at the start of the coronavirus (COVID-19) pandemic and are unlikely to have emigrated between 2011 and 2020)

The study population is not currently refreshed with immigrations. Some deaths involving COVID-19 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 condition).

3 . Disability definition

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". Answer options were, "Yes, limited a lot", "yes, limited a little", or "no". Of the study population, 17.1% reported that they were either limited a little (3.1 million people) or limited a lot (2.0 million people).

The limited a little and limited a lot categories are referred to in this article as "less-disabled" and "more-disabled" respectively. 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 those that appear in legal terms in the Disability Discrimination Act 1995 and the subsequent Equality Act 2010.

4 . Hospital variables

For this analysis, we used Hospital Episode Statistics (HES) data from April 2017 to January 2020 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 episode 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 2011 Census.

The total number of individuals in our HES data was 53,483,456. HES data was linked to the 2011 Census and deaths data by NHS number. A total of 40,800,389 HES records were linked to the 2011 Census (76.3%). The remaining unlinked 23.7% are likely to have not been registered on the 2011 Census. This could be 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 2011 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. However, 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 55,199 SNOMED codes. Of these codes, 28,561 concern dispensary information, prescriptions, and medications, and 20,306 describe diagnoses and findings (including resolved and remission).

The GDPPR dataset was first 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, 40.8 million could be linked to at least one GPES record.

Secondly, as with the HES data, episode data for relevant conditions (listed in this section) were converted to binary (except for body mass index, chronic kidney disease and type 1 and type 2 diabetes), person-level variables by grouping by NHS number.

The GDPPR dataset was used to identify individuals who had primary care contact over the past twenty years (from January 2000 to January 2020) for a range of conditions. These comorbidities were chosen because they were previously implicated in raising the risk of death from coronavirus (COVID-19) by the QCOVID algorithm for predicting hospital admission and mortality from COVID-19 in adults. The list of conditions we adjust for is updated to align with the update of the COVID-19 risk prediction model known as QCovid2 used by the NHS.

Use of certain health variables in the QCOVID algorithm were precluded by either:

- an insufficient number of cases for analysis (bone marrow transplant, cerebral palsy, congenital heart disease, and sickle cell disease)
- a lack of permissions to use these data (chemotherapy or radiotherapy treatment)
- or omission of the requisite clinical codes from the GDPPR (HIV/AIDS and inflammatory bowel disease).

The full list of health variables included are:

- body mass index
- chronic kidney disease (CKD)
- diabetes type 1
- diabetes type 2
- respiratory cancer
- solid organ transplant
- chronic obstructive pulmonary disease (COPD)
- asthma
- rare pulmonary diseases
- pulmonary hypertension or pulmonary fibrosis
- coronary heart disease
- stroke
- atrial fibrillation
- heart failure
- venous thromboembolism
- peripheral vascular disease
- congenital heart disease
- dementia
- Parkinson's disease
- epilepsy
- rare neurological conditions
- osteoporotic fracture
- rheumatoid arthritis or systemic lupus erythematosus
- cirrhosis of the liver
- severe combined immunodeficiency
- severe mental illness (schizophrenia or bipolar disorder)

6 . Vaccination variables

We used vaccination data from the National Immunisation Management Service (NIMS) for the period 8 December 2020 (the day of the first vaccination in England) to 9 March 2022.

Our analysis of the second wave includes first and second vaccination doses, and for the third wave includes first, second and third doses. The analysis does not differentiate between booster doses and third doses provided for other reasons.

Vaccination status was included in the model as a time-varying covariate, and we considered a person vaccinated once 14 days had passed since the dose was administered. More information can be found in the UK Health Security Agency's blog post COVID-19: analysing first vaccine effectiveness in the UK. Of people aged 30 years and over who received at least one dose of a vaccine, 80.4% were linked to the Office for National Statistics (ONS) Public Health Data Asset (PHDA).

7 . Age-standardisation method

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

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

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 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 we use 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 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 in Confidence intervals for weighted sums of poisson parameters (1991). This is described in the Association of Public Health Observatories' third technical briefing (2008) (PDF, 2,088KB) .

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:

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

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

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:

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

where:

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

8 . Modelling analysis

We use Cox proportional hazard models to assess how the risk of death involving coronavirus (COVID-19) varies by self-reported disability status. This is once we adjust for residence type (private household, care home, or other communal establishment) and a range of other characteristics. These characteristics include, location, measures of disadvantage, occupation, living arrangements, pre-pandemic health status and vaccination status.

Most individual characteristics were retrieved from the 2011 Census. This is except for hospital admissions, pre-existing health conditions and vaccination status, which were derived from Hospital Episode Statistics (HES) records from April 2017 onwards, General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) from January 2000 to January 2020, and the National Immunisation Management System (NIMS), respectively.

We model the hazard of death involving COVID-19 between 24 January 2020 and 9 March 2022. 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 10% for the self-reported disability status population).

The hazard function was modelled as follows:

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 ... + b_ix_i)$$

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 the 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 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. This allows 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, based on 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 highest level of qualification of the individual (degree, A-level or equivalent, GCSE or equivalent, no qualification)
- 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 obtained using Occupational Information Network (O*NET) data, based on US Standard Occupational Classification (SOC) codes, which were then 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) in our article Which occupations have the highest potential exposure to the coronavirus (COVID-19)? 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. We did this 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.

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 HES records. We also adjust for the presence of pre-existing health conditions, derived from the GPES 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.

Finally, we adjust for vaccination status as a time-varying covariate, and we consider a person vaccinated once 14 days had passed since the dose was administered. More information can be found in the UK Health Security Agency's blog COVID-19: analysing first vaccine effectiveness in the UK.

We report the hazard ratios for the exposure variables between 24 January 2020 and 9 March 2022, after adjusting for age, geographical factors, socio-economic and demographic factors, health-related variables and vaccination status. A hazard ratio greater than one indicates a greater rate of death involving COVID-19 than the reference group. A hazard ratio less than one indicates a lower rate of COVID-19 mortality than the reference group.

We also report the risk of death involving COVID-19 for the exposure of interest in the first, second and third waves of the coronavirus pandemic by extending the models to allow for time-dependent coefficients for the exposure of interest. We classify deaths occurring between:

- 24 January and 11 September 2020 as having occurred in the first wave
- 12 September 2020 and 11 June 2021 as having occurred in the second wave
- 12 June 2021 and 9 March 2022 as having occurred in the third wave

9 . Related links

Updated estimates of coronavirus (COVID-19) related deaths by disability status, England: 24 January 2020 to 9 March 2022
Article | Released 9 May 2022
Estimates of differences in coronavirus (COVID-19) mortality risk by disability status for deaths occurring up to 9 March 2022, using linked data from the Office for National Statistics Public Health Data Asset.