



Office for Health
Improvement
& Disparities

Official Statistics

Excess mortality in England: methodology

Updated 01 December 2022

Contents

Introduction	3
Methods	4
Expected deaths – generation of the modelled estimates.....	5
Observed deaths – counting registrations during the pandemic	9
Calculation of excess deaths	11
Presentation of COVID-19 deaths and deaths by cause.....	11
Limitations.....	11
Comparison with other measures	12
Appendix - Cause of death ICD10 reference codes.....	13
References	13

Introduction

This methodology document describes the methods used by the Public Health Analysis Unit in the Office for Health Improvement and Disparities (OHID) for the [Excess mortality in England: weekly reports](#). It details how weekly and total excess deaths from 21 March 2020 in England have been estimated. Excess deaths are estimated by comparing the number of observed deaths during the pandemic with the number of deaths that would have been expected had there been no pandemic.

The numbers of expected deaths are estimated using statistical models and based on previous 5 years' (2015 to 2019) mortality rates. Weekly monitoring of excess mortality from all causes throughout the COVID-19 pandemic provides an objective and comparable measure of the scale of the pandemic [reference 1]. Measuring excess mortality from all causes, instead of focusing solely on mortality from COVID-19, overcomes the issues of variation in testing and differential coding of cause of death between individuals and over time [reference 1].

In the weekly reports, estimates of excess deaths are presented by week of registration at national and subnational level, for subgroups of the population (age groups, sex, deprivation groups, ethnic groups) and by cause of death and place of death.

The first reported death due to COVID-19 in England was on 6 March 2020. By 20 March, 109 deaths due to COVID-19 had been registered. We began receiving a daily feed of deaths registered from the Office of National Statistics (ONS) on 20 March 2020. The analyses presented in the excess mortality in England report include deaths occurring in England, registered between 21 March 2020 and the date for the final week of each report. This period captures all the registered deaths recorded as mentioning COVID-19 in England (to usual residents of England), except for the 109 deaths that were registered before 21 March.

Monthly reports are also produced, using the same methodology, providing breakdowns of deaths within each region.

Methods

Overview

Four separate sets of analyses are carried out:

1. Excess deaths from all causes, broken down by age group, sex, ethnic group, deprivation quintile and region (former government office region).
2. Excess deaths from all causes, broken down by upper-tier local authority (UTLA).
3. Excess deaths from specific causes, identified by mentions on death certificates and broken down by underlying cause of death.
4. Excess deaths by place of death.

All of the analyses in the first set involve a single model, which means all the results are entirely consistent with each other.

Not all the data used for the main model are available at UTLA level, so a separate model is used for UTLAs – the UTLA estimates don't necessarily sum to regional and national values.

For each cause of death, a separate model is run, because deaths can have more than one cause of interest mentioned on the death certificate. Deaths are selected using International Classification of Diseases, 10th revision (ICD-10) [reference 2] codes assigned by the Office for National Statistics (ONS). The causes of death included in the report are those which account for large numbers of deaths or are of specific policy interest. All acute respiratory infections have been grouped together (ICD-10 J00 to J22) rather than reporting influenza and pneumonia separately (J09 to J18), as in the ONS definition. Those that tend to involve a coroner's inquest, such as external causes, will be subjected to delays in reporting, as it may take months for an inquest to take place and for the death then to be registered. As the coroner service has been disrupted during the pandemic, fewer deaths from external causes are being registered [reference 3]. This aspect of the pandemic will be an important factor to examine when data do eventually become available.

Place of death is determined using the methodology described in table 2 of the 'Classification of place of death bulletin [reference 4]. All deaths are disaggregated into those occurring:

- at home

- in care homes (nursing or residential)
- in hospitals (acute or community, excluding psychiatric)
- in hospices or elsewhere (any other places)

For each place of death, a separate model is run to avoid complications of competing risks (people who die in one place cannot die in another) – this means that the place-based estimates do not exactly sum to the total from the main model.

Analysis and reporting is on a weekly basis. Weeks run from Saturday to Friday, in line with those used by ONS. Week of registration, rather than week of occurrence, is used because the latter requires either a delay of at least 3 weeks before publication or making an estimate of the delay between week of death and week of registration. This would add uncertainty to the estimate, particularly among subnational groups and greater revisions of figures.

Expected deaths – generation of the modelled estimates

Data sources

Models to develop baseline estimates of the expected number of death registrations on a given week of the year were constructed using a combination of deaths and population denominator data from 2015 to 2019.

Mortality data

Deaths for the years 2015 to 2019 were drawn from fully coded and cleaned annual extracts supplied to us by ONS. For analyses by cause of death and place of death, deaths were aggregated by age, sex and UTLA. For all other analyses, deaths were aggregated by age, sex, ethnic group, deprivation quintile and UTLA.

Assigning ethnicity

To undertake the ethnicity analysis, data linkage to hospital records was required to obtain the necessary information on ethnicity, as this is not recorded at death registration. To obtain ethnicity information for the baseline data, an existing file that links ONS mortality records to Hospital Episode Statistics records (HES-ONS) was used [reference 5].

Deaths from the HES-ONS mortality file were matched to HES records to obtain ethnicity codes. The most frequent ethnicity recorded across the 3 HES data sets (Admitted Patient

Care (APC) from 2003 to 2004 onwards, Accident and Emergency (AE) from 2007 to 2008 onwards and Outpatients (OP) from 2003 to 2004 onwards), excluding any unknown values, is used. OP data was not used between 2006 to 2007 and 2009 to 2010 as, due to a technical issue, no ethnic code entries were recorded in those years. For remaining records where ethnicity could not be assigned, historic records from APC from 1997 to 2003 were used to assign ethnicity.

The method used results in a linkage rate of 98 to 99% (varying from year to year). Of all linked records, 2.7% were linked to a 'not stated' or 'not known' ethnic group and less than 1% had differing codes for sex in each file. Among those not linked, not stated, not known and mismatched records it was viewed as reasonable to assume that data were missing completely at random (MCAR). For not stated, not known and not linked data, ethnic group was assigned proportionally using the proportions amongst the deaths with known ethnic group, stratified by age, sex, region and deprivation quintile.

The HES-ONS file only contains death records for individuals that have attended hospital in the past. Because some people die without ever having attended a hospital there are roughly 5,000 fewer deaths in this file than the annual mortality files [reference 5]. An adjustment was made to ensure the number of deaths used in all the baseline models aligned. For the baseline period, a distribution of counts of deaths for each ethnic group by age, sex, region and deprivation group was calculated for the 60-day window around each day. Where there were no deaths in a particular subgroup in that window, these distributions were calculated over the whole baseline time period. These distributions were then applied to the equivalent subgroups in the annual mortality files.

Denominator data

No single source of population data provided breakdowns of the population by age, sex, ethnic group and deprivation since the 2011 Census. Hence 4 sources of population data were used:

- ONS mid-year population estimates for England (2014 to 2019) [reference 6]
- ONS projections (2020 to 2022) [reference 7]
- ethnic group population estimates published by the University of Leeds (2014 to 2022) [reference 8]
- 2011 Census data [reference 9]

The Leeds estimates provide breakdowns by ethnic group but not deprivation quintile so a process of iterative proportional fitting was applied to combine all the data and estimate populations by ethnic group and deprivation. All population estimates were constrained to the ONS mid-year estimates or projections, which provide breakdowns by age, sex and

deprivation. Population projections are produced at lower-tier local authority geography. To ensure consistency in population by deprivation, the proportion of the population in each deprivation quintile by age, sex and region from the 2019 mid-year estimates were applied to the 2020 to 2022 population projections. Populations for each breakdown by week are estimated by smoothing the changes in population between the years, to avoid step changes.

Place of death and cause of death analyses use whole population denominators.

Baseline model

Model outcome

The primary model provides estimates of expected deaths by week of registration at national and subnational level, and for subgroups of the population (age group, sex, ethnic group, deprivation group and region). Similar models are subsequently run to provide estimates by cause of death, place of death and upper-tier local authority.

Data structure and covariates

In-line with the 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) model [reference 10], independent variables included whether the week contained a public holiday and time of year, allowing for seasonal effects. Specific adjustments were made for registration fluctuations around Easter and Christmas.

Covariates were included, allowing for the effect of age, gender, deprivation, ethnicity and geographical area.

A linear trend was also included in the model to take into account any systematic changes in the rate of death that are not reflected in the changing age structure of the population. The trend was constructed by giving each week a numerical value relating to the mean number of days from 31 December 2016 and dividing the value by 52.

For analysis of all-cause mortality, data are broken down by age groups (0 to 24, 25 to 49, 50 to 64, 65 to 74, 75 to 84, 85 and over) derived from age in years at the time of death. Data are presented by sex (male or female) based on sex reported in the death record. Deaths were allocated to UTLAs based on April 2019 UTLA boundaries [reference 11]. Ethnic group was derived from the linkage process described above. Deprivation is measured using the 2019 IMD (Index of Multiple Deprivation) at lower layer super output area (LSOA) level. 2019 IMD scores for LSOA of usual residence were used to allocate each death to a national deprivation quintile [reference 12].

The structure of the models used is hierarchical with population denominators and counts of death each being fully disaggregated by age, sex, geographic area, ethnicity and deprivation.

Statistical modelling

Quasi-Poisson regression models were fitted on the logarithmic scale [reference 13]. Quasi-Poisson models were used because when counts of weekly deaths are independent of one another they theoretically follow a Poisson distribution. This has the characteristic property that as its mean (the expected number of deaths) increases, the variability of the observed count of deaths (its variance) rises in parallel such that the variance always equals the mean.

However, in the real world, the underlying risk of death varies between different population subgroups and as this cannot usually be modelled perfectly, observed counts of deaths are not completely independent. In consequence, the variance then increases faster than the mean and this is referred to as “overdispersion”. Because Quasi-Poisson models allow the linear relationship between variance and mean to have a slope other than unity, they appropriately analyse rates of death when overdispersion exists.

The models contained the set of covariates outlined in the ‘Data structure and covariates’ section above. To allow for effects to vary between groups, interaction terms were added between age and sex, age and deprivation, age and time of year, age and ethnicity and ethnicity and deprivation. Population sizes in each subgroup are accounted for in the model as an offset.

The model generates expected death rates for each population subgroup for each week, which are then applied to relevant population estimates to estimate the expected numbers of deaths for each week in each subgroup.

Prediction intervals were calculated to quantify the uncertainty around baseline predictions. Prediction intervals account for both the uncertainty around estimate and the random variation likely to arise in the observed values. The width of prediction intervals was estimated by generating 5,000 random samples from a Quasi-Poisson distribution with the appropriate modelled rates of death and the estimated dispersion parameter. Upper and lower 98.8% prediction intervals were then obtained by taking the 0.001 and 0.999 quantiles (that is, 3 standard deviations either side of the baseline predictions) of this simulated distribution. The prediction intervals are used in the regional reports – the charts only show ratios when the 99.8% prediction interval excludes 1.

All data were analysed using the generalised linear modelling function in the statistical package R 4.0.3 [reference 14].

Observed deaths – counting registrations during the pandemic

Data sources

Results of the analyses are presented on a weekly basis. ONS provide a daily feed of registered deaths data that began on 20 March 2020. The first full week of analysis begins on 21 March. These data are provisional and subject to change. As each weekly report is run on the latest version of the data, small changes occur to previous weekly numbers, reflecting improvements in cause of death coding or the addition of registrations not previously received.

To ensure data for deaths in communal establishments are up to date, an individual's place of residence is mapped to a list of communal establishments in England. This list is supplied by ONS every 2 months and reflected in the analysis accordingly.

Assigning ethnicity

To obtain ethnicity data for deaths during the pandemic, a method for linking to HES data was developed, similar to the matching method used for the HES-ONS mortality file. During the pandemic, it became evident that this original method of assigning ethnicity had some limitations – in particular, that it overestimated the number of people in the 'Other' ethnic group.

Alternative methods of assigning ethnicity from HES were therefore investigated and were discussed with stakeholders in the Office for Health Improvement and Disparities (OHID) (formerly Public Health England (PHE) and PHE at the point of consultation), as well as external stakeholders from the Office for National Statistics (ONS), the Race Disparity Unit, NHS Digital, the King's Fund and the Institute of Health Equity.

The original method used by OHID (formerly PHE) assigned the most recent usable ethnic code for an individual available in HES data.

The new method is based on the NHS Digital HES ethnicity index with a few modifications and is used for deaths and hospital admissions indicators using HES. With this method:

- the most frequent ethnicity recorded across the 3 HES data sets (Admitted Patient Care (APC) from 2003 to 2004 onwards, Accident and Emergency (AE) from 2007 to 2008 onwards and Outpatients (OP) from 2003 to 2004 onwards), excluding any unknown values, is used. OP data were not used between 2006 to 2007 and 2009 to 2010 as, due to a technical issue, no ethnic code entries were recorded in those years.

APC data are restricted to 2003 to 2004 onwards as the quality and completeness of admitted patient care data was lower before then

- if there are multiple ethnicities in the data sets with the same frequency, the most recent is chosen
- if there are multiple ethnicities with the same frequency and latest date, precedence is given to the most recent value from the APC data set as it is considered more robust, followed by the AE data set, followed by the OP data set. Checks completed by NHS Digital indicate completeness in the AE data set is better than the OP data set
- if there are multiple ethnicities with the same frequency, latest date and source of data, the ethnicity that occurs more frequently in the general population of England and Wales, according to the 2011 Census (see Appendix A), is selected. Incidences of this are very small, and this step was introduced in order to automate the process and to receive the exact same result each time the analysis is completed
- a value of 'Ethnicity unknown' will only be present if there are no known ethnicities in any of the HES data sets
- to take into the account the overrepresentation of the 'Other' ethnic group, if the most common ethnic group assigned by the method above is 'Other' then the second-most common usable ethnic group is assigned instead. A person will only be assigned to the Other ethnic group if there are no other usable ethnic groups

To note, it is perfectly valid for patients to decide to not state their ethnicity when this information is collected in hospital data. People may also decide to state their ethnicity on some occasions but not others. The original and new methods used for assigning ethnicity do not select 'Not stated' records if there are alternative ethnic codes available. Only those who do not have a usable ethnic code and have repeatedly not stated their ethnicity will have the ethnicity 'Not stated' recorded.

The biggest impact of the change in method has been on the 'Other' ethnic group.

Using this method, over time, around 93% of records are linked to a valid ethnicity code, 9% are linked to a not stated or not known record and less than 2% are not linked. As with the baseline data, the missingness is assumed to be completely at random. A similar method of assigning not linked, not stated or unknown ethnicity data is applied to these records as with the baseline data. The distribution of deaths in known ethnic groups is calculated for each day by age, sex, region and deprivation quintile based on the number of deaths within those subgroups for the previous 30 days. This distribution is then applied to the number of deaths with a not linked, not stated or unknown ethnic group by equivalent subgroup. Where there were no deaths in the previous 30 days in a particular

subgroup, a distribution was calculated by subgroup using the whole time period from the daily feed of registered deaths.

Calculation of excess deaths

Total cumulative excess mortality is estimated by counting deaths from 21 March 2020 onwards and subtracting the expected cumulative deaths from 21 March 2020.

Weekly excess mortality is calculated by taking the observed number of deaths registered in a week and subtracting the expected registered deaths for that week.

Excess death ratios, cumulative and weekly, are calculated by dividing observed deaths by expected deaths. This is the same way as a standardised mortality ratio is calculated, although the expected value is generated from the model rather than by applying reference rates.

Presentation of COVID-19 deaths and deaths by cause

Alongside the estimates of all cause mortality, registered deaths with COVID-19 mentioned in any position on the death certificate are also presented in the weekly reports, in the context of the total numbers of registrations. These provide context for the scale of excess deaths. COVID-19 is identified as ICD-10 U07.1, U07.2, U09.9 or U10.9.

The cause of death analyses are based on any mention of the selected cause on the death certificate. For each selected cause of death, the deaths with a mention of that cause are broken down into those with that cause as the underlying cause of death, those with COVID-19 as the underlying cause of death and those with other underlying causes.

Limitations

During the pandemic, deaths could be registered at weekends and on public holidays [reference 15]. This change in practice compared with the years on which the model was based may have resulted in shifts in registrations between weeks, which may affect weekly excess estimates over that period [reference 3]. It should not affect the cumulative figures.

Disruption to the coroner service may have resulted in delayed registrations, again affecting weekly excess estimates, although the impact of this on the cumulative figures should diminish over time.

Deprivation is attributed ecologically based on the LSOA of residence at time of death. Any individual living in an area may not be representative of the area as a whole, and in particular, for care home residents, the deprivation level of the location of the care home may not reflect the level of deprivation they experienced prior to entering the home.

Ethnicity was obtained through linkage to hospital activity data and determined by the most recent ethnicity stated. There are several limitations with this approach. Ethnicity is supposed to be self-reported by the patient in hospital records, but this may not always be the case. Patients may also report different ethnicities in different episodes of care. People from certain ethnic backgrounds may be less likely to have complete records with which to complete the linkage to hospital data. Previous analysis has shown higher population-based diagnoses and death rates in the 'other' ethnic group due to a mismatch between ethnicity assigned in the population data and hospital records [reference 16]. Similarly, it is possible that there is over-use of the 'Asian Other' and 'Black Other' categories in hospital records [reference 17], so ethnicity is analysed only in 5 broad groups:

- Asian
- Black
- Mixed
- White
- Other

The baseline is modelled using 5 years of historical data. These data include years of relatively high mortality and relatively low mortality. Although a trend is apparent over this period, and is used to inform the expected deaths, it is not a stable trend – the prediction intervals reflect the uncertainty around prediction of any one year. 2020 began with relatively low mortality rates, prior to the COVID-19 pandemic, but this period is not used in the model predictions, so is not reflected in the excess deaths.

The data currently include those registered from 21 March 2020 onwards. There were 109 deaths coded to COVID-19 before that date, which are not currently included in the analyses, along with any other impacts of the pandemic before that date.

Comparison with other measures

ONS publishes a weekly report on excess deaths in England and Wales [reference 18]. This is a continuation of an established report pre-dating the COVID-19 pandemic and is a much simpler approach. Observed weekly registered deaths are compared with a simple

average of weekly counts from 2015 to 2019. No account is taken of changes in population, trends in mortality or bank holidays so the weekly estimates of excess deaths can be quite different from those derived from the OHID weekly excess mortality reports.

We produce other estimates of excess mortality using the daily General Register Office model and the EuroMOMO model; comparisons between these methods are discussed in a Public health matters blog [reference19].

Appendix - Cause of death ICD10 reference codes

The cause of death section of the report presents deaths where the following ICD10 codes can be found in any position on the death certificate.

Cause description	ICD10 reference
Ischaemic heart diseases	All mentions of I20-I25
Cerebrovascular diseases	All mentions of I60-I69
Other circulatory diseases	All mentions beginning with I (excluding I20-I25 and I60-I69)
Heart failure	All mentions of I11.0, I25.5, I42.0, I42.9, I50.0, I50.1, I50.9
Cardiovascular diseases	All mentions beginning with I
Cancer	All mentions of C00-C97
Acute respiratory infections	All mentions of J00-J22
Chronic lower respiratory diseases	All mentions of J40-J47
Other respiratory diseases	All mentions beginning with J (excluding J00-J22 and J40-J47)
Dementia and Alzheimer's	All mentions of F01, F03, or G30
Diseases of the urinary system	All mentions of N00-N39
Cirrhosis and other liver diseases	All mentions of K70-K76
Diabetes	All mentions of E10-E14
Parkinson's disease	All mentions of G20

References

1. Leon DA, Shkolnikov VM, Smeeth L, Magnus P, Pechholdová M, Jarvis CI. [COVID-19: a need for real-time monitoring of weekly excess deaths](#). Lancet. 2020;395(10234):e81.

2. World Health Organization. [ICD-10: international statistical classification of diseases and related health problems](#). Tenth revision, 2nd ed. 2004. World Health Organization.
3. Office for National Statistics. [Analysis of death registrations not involving coronavirus \(COVID-19\), England and Wales: 28 December 2019 to 1 May 2020](#). 5 June 2020.
4. Public Health England. [Classification of place of death: a technical bulletin from the National End of Life Care Intelligence Network](#). March 2019. [Cited: 11 June 2020].
5. NHS Digital. [Linked HES-ONS mortality data](#).
6. Office for National Statistics. [Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland](#). May 2020. [Cited: 16 June 2020].
7. Office for National Statistics. [Subnational population projections for England: 2016-based](#). May 2018. [Cited: 16 June 2020].
8. University of Leeds. [ETHPOP database](#).
9. Office for National Statistics. 2011 Census Populations. [NOMIS](#).
10. Morbey RA, Elliot AJ, Charlett A, Verlander NQ, Andrews N, Smith GE. [The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' \(RAMMIE\) method for syndromic surveillance in England](#), Bioinformatics, Volume 31, Issue 22, 15 November 2015, pages 3660 to 3665.
11. Office for National Statistics. [Local authority district to region \(April 2019\), Lookup in England](#), ONS geography open data. 28 March 2019. [Cited: 11 June 2020].
12. Ministry of Housing, Communities & Local Government. [English indices of deprivation 2019 \(IoD2019\). Statistical Release](#). 26 September 2019. [Cited: 11 June 2020].
13. Gardner W, Mulvey EP, Shaw EC. [Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models](#). Psychological Bulletin. 1995, 118(3), 392 to 404.
14. R Core Team. [The R Project for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria. 2020.
15. [Coronavirus Act 2020](#) c.7.
16. Public Health England. [Analysis of the relationship between pre-existing health conditions, ethnicity and COVID-19](#).

17. Saunders CL, Abel GA, El Turabi A, et al. [Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey](#). *BMJ Open* 2013;3:e002882.
18. Office for National Statistics. [Deaths registered weekly in England and Wales, provisional](#).
19. Fitzpatrick J, Barnard S. [Counting deaths during the pandemic](#). *Public health matters*, February 2021 (now the UK Health Security Agency blog).

© Crown copyright 2021

[Office for Health Improvement and Disparities](#)

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3.

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

OGL