Updated Methodology Note - Analysis of deaths involving coronavirus (COVID-19) in Scotland, by ethnic group

Update: NRS published an updated analysis of COVID-19 deaths by ethnicity on 11th November. This updated methodology note includes additional relevant information.

This note provides further information on the methodology used to produce the analysis in the initial report and the updated analysis.

1. Data linkage process

The analysis is based on a new dataset created by linking records from the 2011 Census and death registration records. Records from Scotland's Census 2011 were previously linked to the NHS Central Register (NHSCR) as at June 2013, using a probabilistic method, as part of a study investigating the apparent quality of the ethnicity information recorded when deaths are registered in Scotland. Although the death registration process is statutory, ethnicity information about the deceased person is collected by registrars on a voluntary basis. The results of the previous study[1] were published on 14th March 2017, one of the key conclusions was, "the data on the ethnicity of the deceased person are not (at present) suitable for calculating reliable mortality rates for most ethnicities".

Records from the 2011 Census were linked to NHSCR information as at March 2020 using a deterministic method based on the NHSCR unique identifier. Records for all deaths occurring on or after 12th March 2020 and registered by 14th June 2020 were also linked to the NHSCR, using a probabilistic method. The main aim of this linkage was to assure and, where appropriate, update the ethnicity information on the death registration records. The rationale for this approach is that the information contained in the census will generally provide a more accurate record of a person's ethnicity, being either self-reported or reported by a close family / household member.

The de-identified census and death registration records were then linked using the NHSCR identifier to create the analysis dataset. The linkage rate to census records was 88%. This study followed a standard 'separation of functions' approach, whereby the teams carrying out the data linkage and analytical functions were based in different departments, and the analytical team only had access to the de-identified matched records.

[1] The ethnicity of the deceased person: the apparent quality of the data that are collected when deaths are registered. https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/deaths-background-information/ethnicity-of-the-deceased-person/the-quality-of-the-data/the-quality-of-the-data-for-2012-to-2014

2. Ethnic groups used for further analysis

Update: In the updated analysis, the White Scottish ethnicity category was used as the reference group. Analysis was provided for the White Irish group despite concerns about the quality of the data.

This section provides further information on how the ethnic groups used in the analysis were arrived at.

White ethnic group

White Scottish; White Other British; White Irish; White Gypsy/Traveller; White Polish; Other White ethnic group.

- Analysis of data from the 2011 Census and death registration records suggested a considerable degree of inconsistency and/or movement between the *White Other British* category and *White Scottish* categories, and similarly between the *White Irish* and *White Scottish* categories, over time and between sources.
- Due to the low number of death registrations involving COVID-19 in the *White Polish* and *White Gypsy/Traveller* categories, it was not possible to carry out informative analysis for these ethnicity categories when considered on their own.
- The *Other White ethnic group* ethnicity category includes a diverse range of ethnicities and this information is collected through a free-text field in the census questionnaire.

South Asian ethnic group

Bangladeshi, Bangladeshi Scottish or Bangladeshi British; Indian, Indian Scottish or Indian British; and Pakistani, Pakistani Scottish or Pakistani British.

- Due to the low number of completed records for deaths involving COVID-19 in the *Bangladeshi, Bangladeshi Scottish or Bangladeshi British* category, it was not possible to carry out informative analysis for this group on its own.
- Although the number of deaths involving COVID-19 in the *Indian, Indian Scottish or Indian British;* and *Pakistani, Pakistani Scottish or Pakistani British* categories made it possible to carry out analysis for each group separately, creating an overarching South Asian ethnic group gave a larger population sample for statistical analysis.

Chinese ethnic group

Chinese, Chinese Scottish or Chinese British

- The number of deaths involving COVID-19 in the *Chinese, Chinese Scottish or Chinese British* category made it possible to carry out informative analysis for this group.

Ethnicity categories not included in the Chinese, South Asian or White ethnic groups

Mixed or Multiple ethnic groups; Other Asian; African, African Scottish or African British; Other African; Caribbean, Caribbean Scottish or Caribbean British; Black, Black Scottish or Black British; Other Caribbean or Black; Arab, Arab Scottish or Arab British; and Other Ethnic group

- Due to the low number of completed records for deaths involving COVID-19 in the remaining ethnicity categories, it was not possible to carry out informative analysis for these groups individually. The results of any analysis based on combining these categories would not be representative for any of the ethnicity categories included.

3. Binary logistic regression model

Update: Results from the updated analysis are now included (Table 2). The White Scottish ethnicity category was used as the reference group. Analysis was provided for the White Irish group despite concerns about the quality of the data.

Odds ratios were obtained by fitting a binary logistic regression model with explanatory variables for ethnic group, age group, sex, urban rural classification (2-fold), and SIMD 2020 quintile (Table 1). The dependent variable was a binary variable equal to one if the death involved COVID-19, and equal to zero if the death did not involve COVID-19. Model fit was assessed using a Hosmer-Lemeshow Goodness-of-Fit Test. For the model including all explanatory variables, which is the model referenced in the main report and updated analysis, the Hosmer-Lemeshow statistic had a p-value of 0.22, indicating that the model provides a reasonably good fit to the data.

Models with a reduced set of explanatory variables were also analysed. The other models which were considered are listed in Table 1, alongside the final model (M0). The odds ratios for the South Asian and Chinese ethnic groups, estimated by fitting the alternative models (M1-M3), are similar to those for the final model (M0). The model fits were compared using the Akaike Information Criterion (AIC). The final model (M0) including all the explanatory variables (age group, sex, urban rural classification and SIMD quintile), in addition to ethnic group, has the best-fit based on comparing the AIC values.

Table 1 – Model comparison - Association of "death involving COVID-19" with Ethnic group

Model	Explanatory variables	South Asian ethnic group		Chinese ethnic group		Hosmer-Lemeshow statistic (p-value)	Max-rescaled R-square	AIC Model Fit (intercept & covariates)
		Odds Ratio: Point estimate and Wald 95% Confidence Interval	p-value	Odds Ratio: Point estimate and Wald 95% Confidence Interval	p-value			
M0	Ethnic group, Age group, Sex, Urban rural classification (2-Fold), SIMD 2020 quintile	1.92 (1.25, 2.92)	0.003*	1.67 (0.75, 3.72)	0.206	0.215	0.040	18,767
M1	Ethnic group, Age group, Sex, Urban rural classification (2-Fold)	1.89 (1.24, 2.89)	0.003*	1.62 (0.73, 3.59)	0.235	0.580	0.039	18,774
M2	Ethnic group, Age group, Sex, SIMD 2020 quintile	2.05 (1.34, 3.12)	0.001*	1.75 (0.79, 3.86)	0.169	0.918	0.031	18,874
M3	Ethnic group, Age group, Sex	2.02 (1.32, 3.07)	0.001*	1.70 (0.77, 3.76)	0.191	0.459	0.029	18,896

Source: National Records of Scotland, data on death registrations linked to the 2011 Census
Notes:
1. Self-reported ethnicity from the 2011 Census was used where available, otherwise ethnicity recorded through the death registration process was used.
2. Odds ratios were obtained by fitting a binary logistic regression model with explanatory variables as listed above. Odds ratios are estimated for the South Asian and Chinese ethnic groups relative to the White ethnic group (which has an odds ratio equal to 1).
3. Statistically significant p-values ($p < 0.05$) are indicated by an asterisk (*). Confidence intervals excluding the value '1' correspond to a statistically significant difference in the odds ratio relative to the White ethnic group.
4. N = 18,300 (number of death registrations included in analysis).

Table 2 – Model comparison for updated analysis - Association of "death involving COVID-19" with Ethnic group

Model	Explanatory variables	South Asian ethnic group		Chinese ethnic group		White Irish ethnic group	
		Odds Ratio: Point estimate and Wald 95% Confidence Interval	p-value	Odds Ratio: Point estimate and Wald 95% Confidence Interval	p-value	Odds Ratio: Point estimate and Wald 95% Confidence Interval	p-value
M0	Ethnic group, Age group, Sex, Urban rural classification (2-Fold), SIMD 2020 quintile	1.92 (1.26, 2.92)	0.002*	1.76 (0.82, 3.77)	0.15	1.24 (0.91, 1.69)	0.18
M1	Ethnic group, Age group, Sex, Urban rural classification (2-Fold)	1.90 (1.25, 2.89)	0.003*	1.69 (0.79, 3.63)	0.18	1.26 (0.92, 1.72)	0.15
M2	Ethnic group, Age group, Sex, SIMD 2020 quintile	2.04 (1.35, 3.10)	0.001*	1.82 (0.85, 3.90)	0.12	1.26 (0.93, 1.72)	0.14
M3	Ethnic group, Age group, Sex	2.01 (1.33, 3.05)	0.001*	1.76 (0.83, 3.77)	0.14	1.29 (0.94, 1.76)	0.11

Source: National Records of Scotland, data on death registrations linked to the 2011 Census
Notes:
1. Self-reported ethnicity from the 2011 Census was used where available, otherwise ethnicity recorded through the death registration process was used.
2. Odds ratios were obtained by fitting a binary logistic regression model with explanatory variables as listed above. Odds ratios are estimated for the South Asian and Chinese ethnic groups relative to the White Scottish ethnic group (which has an odds ratio equal to 1).
3. Statistically significant p-values ($p < 0.05$) are indicated by an asterisk (*). Confidence intervals excluding the value '1' correspond to a statistically significant difference in the odds ratio relative to the White ethnic group.
4. N = 16,833 (number of death registrations included in analysis).

4. Imputation of ethnicity data for non-COVID-19 deaths

Update: This section has been added to explain the imputation methodology used in the updated analysis.

In response to feedback from stakeholders, we asked registrars to contact the 53 informants who registered a death involving COVID-19 over the initial study period where (i) no ethnicity information was provided at the point of registration, and (ii) it was not possible to obtain ethnicity information from the census. Registrars were able to collect ethnicity data in 45 cases (85%). Some informants chose not to provide data and others were unable to do so. In some cases the informant may be someone other than a family member e.g. a funeral director.

There were 203 deaths that did not involve COVID-19 where no ethnicity information was recorded (via registration or census). We decided it would place a disproportionate burden on registrars to ask them to follow up with the informants who registered these deaths.

In order to minimise the bias that could potentially be introduced by only including additional data on deaths involving COVID-19, ethnicity was imputed for an equivalent percentage (85%) of the other deaths for which no ethnicity had been recorded and there was no census match. The records were chosen using two methods: (i) simple random sampling; and (ii) systematic random sampling with control sorting on the variables 'age group' and 'sex'. Hierarchical serpentine sorting was applied to the control variables: this method minimises the change between observations with respect to the control variables. The number of records to which to impute each ethnicity was based on historical non-completion rates by ethnicity, as estimated by NRS over a three-year period from 2012-2014[2]. For example, based on matching to census data, NRS had estimated that 78.2% of historical death registrations for which had no ethnicity had been recorded had the ethnicity "White Scottish". We therefore assigned "White Scottish" ethnicity to 135 (= 78.2% x 85% x 203) records in the "other deaths" category.

Table 2 shows the calculated odds ratios, P-values, and confidence intervals when the logistic regression is run including those records for which an ethnicity was obtained on follow-up by registrars, and applying imputation to an equivalent percentage of other deaths (i.e. those not involving Covid-19). Simple random sampling was used to select records for imputation. The results were very similar when systematic random sampling was applied. Several samples were drawn using each method (simple random / systematic random), confirming that the odds ratios and confidence intervals did not change noticeably when the ethnicities were imputed to different records (i.e. with different values for age group, sex, and the other covariates included in the model).

One limitation of the imputation approach described above is that the true distribution of ethnicities for the death records for which no ethnicity was recorded may be different from that found in the earlier study (2012-2014). Although the previous work is based

[2] Table 4, The ethnicity of the deceased person: the apparent quality of the data that are collected when deaths are registered. https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/deaths-background-information/ethnicity-of-the-deceased-person/the-quality-of-the-data/the-quality-of-the-data-for-2012-to-2014

on a large sample, population demographics and voluntary completion rates for the ethnicity field on death records may have changed over time. A second, related, limitation is that, given the relatively small number of other deaths with no ethnicity identified (203), stochastic effects may mean that the true distribution for these records differs more markedly than would be the case for a larger sample (even accounting for demographic change). However, the relatively small proportion of other deaths (i.e. not involving Covid-19) for which ethnicity was imputed (around 1%) should give some confidence that these limitations do not significantly change the main conclusions from the analysis.