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

This note provides further information on the methodology used to produce the analysis in the main report.

1. Data linkage process

The analysis in the main report 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 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".

For the analysis in the main report, 2011 Census records 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.

2. Ethnic groups used for further analysis

This section provides further information on how the ethnic groups used in the analysis described in the main report 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, death registration records and the Annual Population Survey (2019) 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 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. Inspection of the number of deaths registered for each unique free-text string did not indicate that there was a disproportionately high number of deaths for any ethnic group not covered in the main analysis or discussed above.

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 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 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 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

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. 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, 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)
		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	18,773
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.