# 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 17th November 2021. This updated methodology note includes additional relevant information.**

This note provides further information on the methodology used to produce the analysis in the latest (November 2021) COVID ethnicity report.

The main differences between this methodology and the methodology used for previous COVID mortality by ethnicity analysis in Scotland is that the model is now slightly more complex in order to output more detailed data due to the fact that the dataset is now larger (more deaths have occurred since June 2020).

## 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 30th September 2021 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 89%. 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

For this analysis, we had a large enough number of deaths to produce meaningful results for several ethnicities. We decided that we would build a model that allowed us to show results for all nineteen major ethnic groups reported on in Scotland's Census 2011. Some of the results do not show meaningful results due to low numbers, but have been made publicly available along with the other groupings.

## 3. Binary logistic regression model

Odds ratios were obtained by fitting a binary logistic regression model with explanatory variables for ethnic group, age, 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 0.531, indicating that the model provides a reasonably good fit to the data.

Earlier models exclusively used categorical variables for each dependent variable. In order to provide more detailed analysis of different ethnic groups the model had to be changed. Using the same model as last time but expanding the ethnicity variable to include all categories resulted in a poorly fitting model (M1, Hosmer-Lemeshow p-value =0.037). Other models were tested, and different effects and variable interactions were also tested. An interaction variable UR × SIMD was added to the final model to help model for an interaction that was noticed between certain SIMD (Scottish Index of Multiple Deprivation) quintiles and the urban rural (UR) two fold classification. We also modelled different age groupings before modelling the natural log of age as a continuous variable.

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

| Model | Explanatory variables | Hosmer-Lemeshow chi square | Hosmer-Lemeshow p-value | Max rescaled R-square | AIC model fit |
|---|---|---|---|---|---|
| M0 (final) | Ethnicity, Sex, UR, SIMD, $\log_e$age, UR×SIMD | 7.052 | 0.531 | 0.030 | 68,049 |
| M1 | Ethnicity groups. Sex, Age groups. UR, SIMD | 16.435 | 0.037 | 0.030 | 68,157 |

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.
3. N ≈ 98,000 (number of death registrations included in analysis).
4. UR = Urban Rural Classification (2-fold)
5. SIMD = Scottish Index of multiple deprivation (quintiles)