# Supplementary Material 19: Methodology of how we dealt with missing data

Among the 231,998 persons included in this study there were 90,783 (39.1%) with at least one missing observation for gender, marital status, ethnicity, care cluster and IMD. (Table 10). The variables with the largest proportion of missing observations were care cluster (with 25.7% missing observations) and marital status (with 21.7%).

**Frequency of missing data in persons by readmission status**

|  |  |  |  |
| --- | --- | --- | --- |
|  | N missing (%) | In persons readmitted (n=53,456) | In persons not readmitted  (n=178,542) |
| Cluster | 59,677 (25.7%) | 5,952 (11.1%) | 53,725 (30.1%) |
| Marital status | 50,419 (21.7%) | 7,131 (13.3%) | 43,288 (24.3%) |
| Ethnicity | 25,689 (11.1%) | 3,304 (6.2%) | 22,385 (12.5%) |
| IMD | 4,126 (1.8%) | 442 (0.8%) | 3,684 (2.1%) |
| Gender | 339 (0.15%) | 30 (0.1%) | 309 (0.2%) |



**Missingness frequencies by variables**

We further investigate the distribution of readmitted persons and age according to whether ethnicity, cluster care and marital status were missing.

**Distribution of readmitted persons and age according to whether ethnicity, cluster care and marital status are missing**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Ethnicity observed** | **Ethnicity missing** | **p** |
| **Readmitted** n(%) | 45,451 (22.5%) | 3,096 (12.1%) | <0.001 |
| **Age** – (median (IQR)) | 40 (29-52) | 37 (25-50) | <0.001 |
|  | **Cluster observed** | **Cluster missing** | **p** |
| **Readmitted** n(%) | 43,925 (25.5%) | 5,622 (9.4%) | <0.001 |
| **Age** – (median (IQR)) | 41 (29-52) | 36 (25-48) | <0.001 |
|  | **Marital observed** | **Marital missing** | **p** |
| **Readmitted** n(%) | 42,868 (23.6%) | 6,679 (13.3%) | <0.001 |
| **Age** – (median (IQR)) | 41 (29-52) | 38 (26-50) | <0.001 |

The distribution of missing data in Table 11 suggests that data are not missing completely at random (MCAR) as it depends on whether a person is readmitted (proportion of missing data does not vary much by readmission and by age).

Table 11 shows that there are large (and statistically significant) differences in the proportion of readmission between people with **ethnicity** missing and those with ethnicity observed. If missing values for ethnicity were MCAR we would expect the proportion of readmission to be similar in the groups defined by whether ethnicity was or was not observed. Furthermore we see that age is younger in the group of persons with missing ethnicity. MAR means that given readmission and age the probability of ethnicity being missing is independent of ethnicity, (contrarily to MNAR, where it would be dependent but this cannot be tested because we would need the missing values of ethnicity!!!!).

Another way is to fit a logistic regression model, whether the outcome variable is a binary indicator of whether ethnicity was observed or missing.

***Logistic regression estimates for whether ethnicity is observed with readmission and age as explanatory variables***

|  |  |  |
| --- | --- | --- |
|  | **odds ratio (95%CI)** | p |
| **Readmission** (yes) | 0.47 (0.46, 0.49) | <0.001 |
| **Age** (per 1 year) | 0.99 (0.99, 0.99) | <0.001 |

***Logistic regression estimates for whether cluster is observed with readmission and age as explanatory variables***

|  |  |  |
| --- | --- | --- |
|  | **odds ratio (95%CI)** | p |
| **Readmission** (yes) | 0.30 (0.29, 0.31) | <0.001 |
| **Age** (per 1 year) | 0.98 (0.76, 0.80) | <0.001 |

***Logistic regression estimates for whether marital status is observed with readmission and age as explanatory variables***

|  |  |  |
| --- | --- | --- |
|  | **odds ratio (95%CI)** | p |
| **Readmission** (yes) | 0.50 (0.48, 0.51) | <0.001 |
| **Age** (per 1 year) | 0.99 (0.99, 0.99) | <0.001 |

The tables show that there is strong evidence that readmission and age are independently predictive of missingness in ethnicity, care cluster and marital status. People with missing ethnicity, care cluster and marital status tend be those not readmitted. In addition to this, those with missing care cluster also tend to be of younger age. In conclusion, a multivariable analysis using only complete cases (i.e. people with no missing data) would be biased.