**TAGRA ACUTE MLC SUBGROUP Wednesday 20th January 2016**

**INDICATOR SELECTION RESULTS – PART 1**

**Background**

The final list of potential candidate variables was presented in paper TAMLC36, which was circulated by email in November 2015. Most of the data had been obtained and the variables calculated at the 2011 data zones. A methodology for the needs indicator selection process was also outlined in the paper, based on the methodology used in the 2007 NRAC review.

The paper included a new suggestion for the ethnicity variable. Equality leads present at the 10th meeting had advised that a count of non-white people as a potential explanatory variable for healthcare use was problematic: it implied that all non-white groups were anticipated to share a common pattern of healthcare needs that differed from those of white populations, which may be highly questionable when such a large number of different ethnicities are included. The Scottish Government ScotStat report “Which ethnic groups have the poorest health? An Analysis of Health Inequality and Ethnicity in Scotland”[[1]](#footnote-1) indeed shows that there are wide differences in health outcomes between different ethnic groups in Scotland. A suggestion was therefore made in paper TAMLC36 to use counts of specific ethnic groups, using the above report as a guide to which ethnic groups should be considered. In response to this, concerns were raised about the likelihood of high numbers of zero counts, and the implications of excluding several ethnic groups from analysis.

The potential indicator variables have now all been computed, and this paper reports on the initial analysis of these variables, including analysis of zero counts, and results from the first stage of the indicator selection process.

**1. Summary**

This paper presents some preliminary analysis of the potential indicator variables for the Acute MLC regression model, and the results of the first stage of the index construction process.

In section 2, the issue of high numbers of zero counts in potential indicators is addressed by looking at the numbers of zeros for each variable, and eliminating those with more zeros than a suggested threshold.

In order to proceed with the process of selecting the best indicators, a decision must be made on whether to keep the current diagnostic grouping or change to a different structure. In section 3, a suggestion is made to keep the current Acute diagnostic grouping as well as retaining a ‘Whole Acute’ option, following investigations of alternative options and feedback from an ISD consultant in public health.

Finally, provisional results on the first stage of the variable selection process – elimination of near-duplicates through correlations analysis – are presented in section 4.

**2. Zeros analysis**

High numbers of data zones with zero counts for a potential explanatory variable are problematic for predictive modelling – the linear model does not fit well and its parameters are not well constrained. The ethnicity variable options in particular have been highlighted as being susceptible to this problem due to the small size of ethnic minority populations in much of Scotland.

In this section we check the numbers of data zones with zero counts for all potential candidate variables, and eliminate variables with more than 30% of data zones having zero counts (2,093 or more out of 6,976 data zones).

Table 1 shows the number of zeros for each candidate variable. Variables with zeros above the threshold amount are shaded. (For reference, the final potential candidate variable table with full details on the construction of each variable is given in Annex A.)

*Table 1: Numbers of data zones with zero counts for each variable*

|  |  |
| --- | --- |
| **Variable** | **Zeros** |
| Low birth weight births | 2,042 |
| Death rate 0-74 all causes | 8 |
| Death rate 0-74 Cancer | 118 |
| Death rate 0-74 CHD | 1,443 |
| Death rate 0-74 Stroke | 4,874 |
| All cause SMR 0-64 | 74 |
| All cause SMR 0-69 | 21 |
| All cause SMR 0-74 | 8 |
| Cancer SMR 0-64 | 711 |
| Cancer SMR 0-69 | 277 |
| Cancer SMR 0-74 | 118 |
| Heart Disease SMR 0-64 | 1,940 |
| Heart Disease SMR 0-69 | 1,139 |
| Heart Disease SMR 0-74 | 624 |
| Respiratory SMR 0-64 | 4,576 |
| Respiratory SMR 0-69 | 3,508 |
| Respiratory SMR 0-74 | 2,416 |
| Digestive System SMR 0-64 | 3,966 |
| Digestive System SMR 0-69 | 3,385 |
| Digestive System SMR 0-74 | 2,924 |
| External Causes SMR 0-64 | 2,946 |
| External Causes SMR 0-69 | 2,774 |
| External Causes SMR 0-74 | 2,614 |
| Other SMR 0-64 | 2,597 |
| Other SMR 0-69 | 2,076 |
| Other SMR 0-74 | 1,510 |
| High Resource Individual counts | 0 |
| Patients receiving Diabetes prescriptions | 5,426 |
| Patients receiving Dementia prescriptions | 727 |
| Patients receiving Respiratory prescriptions | 5,386 |
| Long-term illness | 0 |
| Mental health condition | 0 |
| Limiting long-term illness – limited a lot | 0 |
| Limiting long-term illness – limited a little or a lot | 0 |
| Long-term sick and not seeking work | 7 |
| General health – very bad | 88 |
| General health – bad or very bad | 2 |
| Older people living alone – 65 and over | 14 |
| Older people living alone – 70 and over | 32 |
| Older people living alone – 75 and over | 87 |
| Older people living alone – 80 and over | 185 |
| Older people living alone – 85 and over | 491 |
| Older people living alone – 90 and over | 1,854 |
| Unpaid care – 1 hour or more | 0 |
| Unpaid care – 20 hours or more | 0 |
| Unpaid care – 35 hours or more | 2 |
| Unpaid care – 50 hours or more | 8 |
| Education level – no qualifications | 0 |
| Education level – level 1 and below | 0 |
| Education level – level 2 and below | 0 |
| Job seekers | 0 |
| Did Not Attend counts – as a fraction of all OP appointments | 0 |
| Did Not Attend counts – ratio to data zone population | 0 |
| Ethnic group populations (Gypsy/Traveller) | 5,097 |
| Ethnic group populations (Pakistani) | 2,845 |

From Table 1, we can see that neither of the ethnicity variable options suggested in paper TAMLC36 is suitable based on the number of zeros. In addition, several of the death variables with specific causes of death, and two of the prescribing variables, are also eliminated.

Although Gypsy/Traveller and Pakistani population counts had been suggested for use, the numbers of zeros were also counted for all other ethnic groups. It was found that the only groups with acceptably low numbers of zeros were White Scottish, Other White British, White Polish, White Irish, and Other White. White Scottish ethnicity will constitute a large majority in almost all data zones, so its power as a variable to pick up variations in need is just as limited as that of groups with very low numbers. White Irish health appears comparable to that of the White Scottish population in the ScotStat report analysis, while all the other groups above appear to have better health on average than the White Scottish population.

Having explored several options, it now seems that the options available for ethnicity variable(s) fail either the core criterion of practicality (in the case of ethnic group populations with large numbers of zeros) or relevance (in the case of groups for which additional need for Acute services is not suggested by the available evidence).

**3. Diagnostic groups**

The purpose of the indicator variables is to explain variation in healthcare cost across the data zones. Previously, this cost variation has been analysed within subsets of the Acute activity, referred to as diagnostic groups. In order to proceed with the process of selecting the best indicators, a decision must be made on whether to keep the current diagnostic grouping or change to a different structure.

Five of the current diagnostic groups – Cancer, Heart, Respiratory, Digestive, and Injury – correspond to high-level ICD-10 code groups[[2]](#footnote-2), and activity is selected into each diagnostic group based on its main ICD-10 code. A sixth group, Other, contains all other inpatient and daycase activity. The seventh group, Outpatients, contains all Acute outpatient activity (relating to any Acute condition).

Advice was sought from an ISD consultant in public health on whether any other structure might be better. The feedback was that ICD-10 remains a suitable way of classifying activity into high-level diagnostic groups, but that Outpatients stood out as being a patient category rather than a diagnostic group. Ideally it would be best to calculate diagnostic group cost ratios that combined all inpatient, daycase and outpatient costs for the condition types; however, this is not possible due to the lack of ICD-10 coding on outpatient activity. For this reason, it seems most practical to keep the Outpatients group separate.

‘Other’ is a large diagnostic group containing the Acute activity falling within all the other ICD-10 code groups besides Cancer, Heart, Respiratory, Digestive, and Injury. The levels of activity were examined to see whether any of these other code groups could be treated as Acute diagnostic groups in their own right. The only code group within Other that was comparable in size to the existing diagnostic groups was ‘Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified’, which did not seem a good candidate for a separate diagnostic group.

A further possibility is not to analyse cost variation in diagnostic groups at all, but instead, develop a single linear model for the Acute care programme as a whole – a ‘Whole Acute’ option. To compare the model performance of this option with that of the current diagnostic grouping, linear regression of cost ratios upon the current Acute needs index (plus supply model) was carried out, firstly for the existing diagnostic groups, and secondly for the ‘Whole Acute’ cost ratios. The predictions of cost from the separate diagnostic groups were then aggregated together, by the same method used routinely in the NRAC shares calculation, and compared with the predictions of cost from the ‘Whole Acute’ regression. The residual sum of squares (RSS) was then computed for both methods and compared. There was a very small (and likely insignificant) difference in RSS, in favour of using separate diagnostic groups.

This suggests that a ‘Whole Acute’ model could potentially perform well and should be considered. On the other hand, allowing different indicators of need to be selected for different diagnostic groups may well improve the predictive power of the ‘separate groups’ option. AST propose, for the purposes of the indicator selection process, to proceed with both (1) the existing diagnostic groups and (2) the ‘Whole Acute’ option, and then to decide between these two options by examining the results and assessing against the core criteria.

**Q1: The Subgroup is asked to approve the decision to proceed with two options for diagnostic grouping: the current grouping and a ‘Whole Acute’ option.**

**4. Results of first stage of indicator selection process**

The methodology proposed in paper TAMLC36 for selecting the needs indicators is included in Annex B of the current paper, for reference. The first stage of this process is to eliminate ‘near-duplicates’ (i.e. highly correlated variables) from the variable list, retaining the variants that correlate best with the cost ratios.

This involves the following steps:

1) Allocate each candidate variable to an appropriate ‘topic’, so that unrelated variables (with no likelihood of being highly correlated) are held in separate topics. This reduces the number of inter-correlation calculations required.

2) For each topic, compute the inter-correlations between the variables. Form subgroups comprising highly inter-correlated variables (i.e. near duplicates).

3) For each subgroup, compute the variables’ correlations with the cost ratios; retain the variable that has the highest correlation with the cost ratios across most diagnostic groups.

**4.1 Grouping into topics**

The potential candidate variables are grouped into four main topics, as shown in Table 2: Births and deaths, Health / Morbidity, Unpaid care and older people living alone, and Deprivation. As far as possible, this grouping follows that of the NRAC 2007 review (e.g. low birth weight births are included with the SMR and death rate variables). Unpaid care and older people living alone are grouped together because both caregiving and an absence of care where needed (as may be the case where older people live alone) are anticipated to have a similar link to Acute healthcare need.

*Table 2. Potential candidate variables grouped into topics*

|  |  |
| --- | --- |
| **Variable** | **Topic** |
| Low birth weight births – as a fraction of population | Births and deaths |
| Death rate 0-74 all causes |
| Death rate 0-74 Cancer |
| Death rate 0-74 CHD |
| All cause SMR 0-64 |
| All cause SMR 0-69 |
| All cause SMR 0-74 |
| Cancer SMR 0-64 |
| Cancer SMR 0-69 |
| Cancer SMR 0-74 |
| Heart Disease SMR 0-64 |
| Heart Disease SMR 0-69 |
| Heart Disease SMR 0-74 |
| Other SMR 0-69 |
| Other SMR 0-74 |
| High Resource Individual counts | Health / Morbidity |
| Patients receiving Dementia prescriptions |
| Long-term illness |
| Mental health condition |
| Limiting long-term illness – limited a lot |
| Limiting long-term illness – limited a little or a lot |
| Long-term sick and not seeking work |
| General health – very bad |
| General health – bad or very bad |
| Older people living alone – 65 and over | Unpaid care and older people living alone |
| Older people living alone – 70 and over |
| Older people living alone – 75 and over |
| Older people living alone – 80 and over |
| Older people living alone – 85 and over |
| Older people living alone – 90 and over |
| Unpaid care – 1 hour or more |
| Unpaid care – 20 hours or more |
| Unpaid care – 35 hours or more |
| Unpaid care – 50 hours or more |
| Education level – no qualifications | Deprivation |
| Education level – level 1 and below |
| Education level – level 2 and below |
| Job seekers |
| Did Not Attend counts – as a fraction of all OP appointments |
| Did Not Attend counts – ratio to data zone population |

**4.2 Formation of subgroups**

Within each topic area shown in Table 2, the inter-correlations for those candidate variables are computed and examined. A correlation coefficient threshold of 0.7 is used as the definition of “highly correlated” and thus near-duplicate (although the results are identical if 0.8 is used instead). In addition, variables that are conceptually equivalent, i.e. are calculated using much of the same data (for example, SMR < 70 and SMR < 75), are automatically regarded as near duplicates.

The variables are found to form nine independent subgroups of near-duplicate variables, which will yield nine retained variables. In addition, four variables are not highly correlated with any other variable in their respective topics: Low Birth Weight Births, High Resource Individuals, Patients receiving Dementia prescriptions, and Long-term sick and not seeking work. These are retained to the next stage without further analysis.

Tables 3 to 11 below show the composition of the nine subgroups and their inter-correlations. In subgroup 8, the correlation between ‘Job seekers’ and ‘Education – no qualifications’ does not reach the threshold value; however, both of these variables are highly correlated with the other two in the subgroup, so they are both taken to be part of this same subgroup.

*Table 3: subgroup 1 inter-correlations*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **All cause SMR <75** | **All cause SMR <70** | **All cause SMR <65** | **Death rate <75 all causes** |
| **All cause SMR <75** | 1.000 | 0.949 | 0.888 | 0.872 |
| **All cause SMR <70** | 0.949 | 1.000 | 0.942 | 0.843 |
| **All cause SMR <65** | 0.888 | 0.942 | 1.000 | 0.803 |
| **Death rate <75 all causes** | 0.872 | 0.843 | 0.803 | 1.000 |

*Table 4: subgroup 2 inter-correlations*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Cancer SMR <65** | **Cancer SMR <70** | **Cancer SMR <75** | **Death rate <75 Cancer** |
| **Cancer SMR <65** | 1.000 | 0.842 | 0.734 | 0.627 |
| **Cancer SMR <70** | 0.842 | 1.000 | 0.878 | 0.732 |
| **Cancer SMR <75** | 0.734 | 0.878 | 1.000 | 0.810 |
| **Death rate <75 Cancer** | 0.627 | 0.732 | 0.810 | 1.000 |

*Table 5: subgroup 3 inter-correlations*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Heart SMR <65** | **Heart SMR <70** | **Heart SMR <75** | **Death rate <75 CHD** |
| **Heart SMR <65** | 1.000 | 0.870 | 0.779 | 0.598 |
| **Heart SMR <70** | 0.870 | 1.000 | 0.891 | 0.674 |
| **Heart SMR <75** | 0.779 | 0.891 | 1.000 | 0.737 |
| **Death rate <75 CHD** | 0.598 | 0.674 | 0.737 | 1.000 |

*Table 6: subgroup 4 inter-correlations*

|  |  |  |
| --- | --- | --- |
|  | **Other SMR <70** | **Other SMR <75** |
| **Other SMR <70** | 1.000 | 0.914 |
| **Other SMR <75** | 0.914 | 1.000 |

*Table 7: subgroup 5 inter-correlations*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **LLTI – Yes (both)** | **LLTI – Yes (a lot)** | **General health – very bad** | **General health – bad or very bad** | **Mental health condition** | **Long term illness** |
| **LLTI – Yes (both)** | 1.000 | 0.962 | 0.825 | 0.921 | 0.932 | 0.847 |
| **LLTI – Yes (a lot)** | 0.962 | 1.000 | 0.854 | 0.935 | 0.951 | 0.838 |
| **General health – very bad** | 0.825 | 0.854 | 1.000 | 0.803 | 0.894 | 0.729 |
| **General health – bad or very bad** | 0.921 | 0.935 | 0.803 | 1.000 | 0.985 | 0.808 |
| **Mental health condition** | 0.932 | 0.951 | 0.894 | 0.985 | 1.000 | 0.820 |
| **Long term illness** | 0.847 | 0.838 | 0.729 | 0.808 | 0.820 | 1.000 |

*Table 8: subgroup 6 inter-correlations*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Living alone ≥65** | **Living alone ≥70** | **Living alone ≥75** | **Living alone ≥80** | **Living alone ≥85** | **Living alone ≥90** |
| **Living alone ≥65** | 1.000 | 0.848 | 0.658 | 0.472 | 0.232 | -0.045 |
| **Living alone ≥70** | 0.848 | 1.000 | 0.785 | 0.578 | 0.285 | -0.038 |
| **Living alone ≥75** | 0.658 | 0.785 | 1.000 | 0.732 | 0.382 | 0.011 |
| **Living alone ≥80** | 0.472 | 0.578 | 0.732 | 1.000 | 0.568 | 0.090 |
| **Living alone ≥85** | 0.232 | 0.285 | 0.382 | 0.568 | 1.000 | 0.297 |
| **Living alone ≥90** | -0.045 | -0.038 | 0.011 | 0.090 | 0.297 | 1.000 |

*Table 9: subgroup 7 inter-correlations*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Unpaid care ≥ 1 hour** | **Unpaid care ≥ 20 hours** | **Unpaid care ≥ 35 hours** | **Unpaid care ≥ 50 hours** |
| **Unpaid care ≥ 1 hour** | 1.000 | 0.699 | 0.662 | 0.638 |
| **Unpaid care ≥ 20 hours** | 0.699 | 1.000 | 0.971 | 0.929 |
| **Unpaid care ≥ 35 hours** | 0.662 | 0.971 | 1.000 | 0.959 |
| **Unpaid care ≥ 50 hours** | 0.638 | 0.929 | 0.959 | 1.000 |

*Table 10: subgroup 8 inter-correlations*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Education – no qualifications** | **Education – level 1 and below** | **Education – level 2 and below** | **Job seekers** |
| **Education – no qualifications** | 1.000 | 0.859 | 0.743 | 0.570 |
| **Education – level 1 and below** | 0.859 | 1.000 | 0.959 | 0.790 |
| **Education – level 2 and below** | 0.743 | 0.959 | 1.000 | 0.850 |
| **Job seekers** | 0.570 | 0.790 | 0.850 | 1.000 |

*Table 11: subgroup 9 inter-correlations*

|  |  |  |
| --- | --- | --- |
|  | **DNA counts – fraction of appointments** | **DNA counts – ratio to population** |
| **DNA counts – fraction of appointments** | 1.000 | 0.870 |
| **DNA counts – ratio to population** | 0.870 | 1.000 |

**4.3 Selection from the subgroups**

In order to select the ‘best’ variable from each subgroup, correlations between the variables and the cost ratios are computed. The variable having the highest correlation coefficients for the most diagnostic groups is retained for the next stage of the index construction.

Tables 12 to 20 below show the correlations with the cost ratios for each subgroup. The highest correlation coefficient within each diagnostic group is indicated in bold and italics.

*Table 12: correlations with cost ratios for subgroup 1*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **All cause SMR <75** | ***0.629*** | ***0.250*** | ***0.384*** | ***0.465*** | ***0.413*** | ***0.526*** | ***0.523*** | ***0.321*** |
| **All cause SMR <70** | 0.614 | 0.241 | 0.375 | 0.455 | 0.404 | 0.518 | 0.497 | 0.314 |
| **All cause SMR <65** | 0.584 | 0.227 | 0.354 | 0.430 | 0.389 | 0.494 | 0.470 | 0.303 |
| **Death rate <75** | 0.564 | 0.224 | 0.318 | 0.428 | 0.363 | 0.495 | 0.440 | 0.274 |

*Table 13: correlations with cost ratios for subgroup 2*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Cancer SMR <65** | 0.344 | 0.258 | 0.188 | 0.228 | 0.202 | 0.268 | 0.254 | 0.196 |
| **Cancer SMR <70** | 0.403 | 0.292 | 0.228 | 0.274 | 0.236 | 0.315 | 0.302 | 0.225 |
| **Cancer SMR <75** | ***0.448*** | ***0.321*** | ***0.251*** | ***0.302*** | ***0.262*** | ***0.346*** | ***0.340*** | ***0.257*** |
| **Death rate <75 Cancer** | 0.347 | 0.261 | 0.165 | 0.246 | 0.197 | 0.285 | 0.224 | 0.195 |

*Table 14: correlations with cost ratios for subgroup 3*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Heart SMR <65** | 0.394 | 0.127 | 0.250 | 0.284 | 0.276 | 0.339 | 0.312 | 0.203 |
| **Heart SMR <70** | 0.430 | 0.134 | 0.280 | 0.311 | 0.298 | 0.372 | 0.342 | 0.214 |
| **Heart SMR <75** | ***0.463*** | ***0.140*** | ***0.312*** | ***0.332*** | ***0.315*** | ***0.397*** | ***0.387*** | ***0.227*** |
| **Death rate <75 CHD** | 0.368 | 0.113 | 0.229 | 0.278 | 0.235 | 0.328 | 0.285 | 0.184 |

*Table 15: correlations with cost ratios for subgroup 4*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Other SMR <70** | ***0.341*** | ***0.090*** | ***0.223*** | ***0.234*** | ***0.221*** | ***0.305*** | ***0.296*** | ***0.167*** |
| **Other SMR <75** | 0.327 | 0.084 | 0.206 | ***0.234*** | 0.217 | 0.292 | 0.294 | 0.150 |

*Table 16: correlations with cost ratios for subgroup 5*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **General health – very bad** | 0.657 | 0.194 | 0.386 | 0.513 | 0.398 | 0.555 | 0.554 | 0.400 |
| **General health – bad or very bad** | 0.726 | ***0.218*** | 0.428 | ***0.562*** | 0.450 | 0.615 | ***0.608*** | ***0.438*** |
| **Mental health condition** | 0.616 | 0.184 | 0.346 | 0.456 | 0.441 | 0.520 | 0.521 | 0.353 |
| **Long term illness** | 0.608 | 0.180 | 0.358 | 0.456 | 0.361 | 0.519 | 0.496 | 0.395 |
| **LLTI – Yes (both)** | ***0.733*** | 0.212 | ***0.440*** | 0.552 | ***0.473*** | ***0.632*** | 0.593 | 0.428 |
| **LLTI – Yes (a lot)** | 0.723 | 0.209 | 0.419 | 0.560 | 0.452 | 0.621 | 0.606 | 0.423 |

*Table 17: correlations with cost ratios for subgroup 6*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Living alone ≥65** | -0.073 | 0.004 | -0.014 | ***-0.071*** | -0.062 | -0.092 | -0.020 | -0.041 |
| **Living alone ≥70** | ***-0.082*** | -0.007 | -0.021 | -0.067 | ***-0.073*** | ***-0.093*** | -0.028 | ***-0.050*** |
| **Living alone ≥75** | -0.061 | -0.004 | -0.003 | -0.050 | -0.053 | -0.076 | -0.012 | -0.041 |
| **Living alone ≥80** | -0.025 | 0.001 | 0.024 | -0.011 | -0.028 | -0.043 | 0.006 | -0.022 |
| **Living alone ≥85** | 0.023 | 0.023 | ***0.057*** | 0.014 | -0.001 | 0.004 | ***0.041*** | 0.008 |
| **Living alone ≥90** | 0.040 | ***0.034*** | 0.053 | 0.029 | 0.025 | 0.020 | 0.018 | 0.030 |

*Table 18: correlations with cost ratios for subgroup 7*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Unpaid care ≥ 1 hour** | 0.247 | 0.057 | 0.080 | 0.209 | 0.056 | 0.235 | 0.153 | 0.284 |
| **Unpaid care ≥ 20 hours** | ***0.557*** | ***0.154*** | 0.299 | ***0.446*** | 0.284 | ***0.496*** | ***0.437*** | ***0.380*** |
| **Unpaid care ≥ 35 hours** | 0.553 | 0.153 | ***0.304*** | 0.441 | ***0.286*** | 0.492 | 0.435 | 0.369 |
| **Unpaid care ≥ 50 hours** | 0.523 | 0.145 | 0.289 | 0.415 | 0.272 | 0.465 | 0.414 | 0.351 |

*Table 19: correlations with cost ratios for subgroup 8*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Education – no qualifications** | -0.004 | ***0.011*** | 0.002 | -0.030 | 0.002 | -0.004 | 0.000 | -0.013 |
| **Education – level 1 and below** | 0.067 | 0.010 | ***0.036*** | 0.031 | 0.050 | 0.074 | ***0.028*** | 0.039 |
| **Education – level 2 and below** | ***0.078*** | 0.003 | 0.033 | *0.050* | ***0.056*** | ***0.092*** | 0.024 | ***0.041*** |
| **Job seekers** | -0.044 | -0.041 | -0.046 | -0.029 | -0.028 | -0.008 | -0.073 | -0.027 |

*Table 20: correlations with cost ratios for subgroup 9*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **DNA counts – fraction of appointments** | 0.616 | 0.194 | 0.365 | 0.432 | 0.388 | 0.490 | 0.526 | 0.445 |
| **DNA counts – ratio to population** | ***0.733*** | ***0.274*** | ***0.378*** | ***0.494*** | ***0.425*** | ***0.569*** | ***0.559*** | ***0.662*** |

In most of the subgroups, the variable to be retained is clear. One exception is subgroup 5 (Table 16), in which General Health (bad or very bad) is correlated most strongly with the cost ratios for 4 out of 7 diagnostic groups; the variable most strongly correlated with cost ratios for the other 3 diagnostic groups is LLTI (Yes both), and it is also most strongly correlated with cost ratios for the ‘Whole Acute’ option. Since these variables are conceptually different (based on answers to different census questions), it seems safest to retain both at this stage.

For the Living Alone variables (Table 17), the highest correlations with the cost ratios are actually negative correlations. Since it is unclear at this stage how to interpret this, we will retain both the variable with the highest negative correlations (Living alone ≥70) and the variable with the highest positive correlations (Living alone ≥90).

Some variables have close to zero correlation with the cost ratios. Those variables are not expected to be significant predictors of cost variation. However, the purpose of this stage of the analysis is merely to eliminate near duplicates. Hence, one variable will still be chosen within each subgroup even if the correlations with cost ratios are near zero.

For completeness, Table 21 shows the correlations with the cost ratios for the four retained variables that were not highly correlated with any others.

*Table 21: correlations with cost ratios for additional retained variables*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Low birth weight births** | 0.127 | 0.028 | 0.070 | 0.091 | 0.064 | 0.140 | 0.106 | 0.044 |
| **Patients receiving Dementia prescriptions** | -0.032 | -0.066 | -0.056 | -0.012 | 0.032 | -0.012 | -0.023 | -0.031 |
| **High resource individuals** | 0.479 | 0.174 | 0.267 | 0.328 | 0.336 | 0.443 | 0.343 | 0.223 |
| **Long-term sick and not seeking work** | 0.042 | 0.016 | 0.041 | 0.016 | 0.028 | 0.035 | 0.044 | 0.009 |

As a further validation of the choice of retained variables, a series of regressions has been carried out using each potential candidate variable as an individual predictor of need, with the current supply model (supply variables IPACX and OPACX and health board dummy variables) included in the regression. The adjusted R-squared values, reassuringly, tended to be highest for the variables with highest correlations to the cost ratios. Results from this are given in Annex C.

**4.4 Retained variables**

The retained variables, after elimination of variables with high numbers of zeros and near-duplicates, are:

* All cause SMR <75
* Cancer SMR <75
* Heart SMR <75
* Other SMR <70
* LLTI – Yes (both)
* General health – bad or very bad
* Living alone ≥70
* Living alone ≥90
* Unpaid care ≥ 20 hours
* Education – level 2 and below
* DNA counts – ratio to population
* Low birth weight births
* Patients receiving Dementia prescriptions
* High resource individuals
* Long-term sick and not seeking work

**5. Next steps**

This brings the work to step 4 of the method (Annex B). The next steps are to perform both stepwise regressions and factor analysis, using the list of retained variables above with the cost ratios for the different diagnostic groups. This will result in a ‘restricted variable set’ consisting of the variables that had been prominent in the factor analysis and had most often been statistically significant in the stepwise regressions (steps 5 and 6). The restricted variable set will be presented at the March 2016 meeting of the Subgroup.

**Annex A: Final potential candidate variables table**

*Table A.1. Potential candidate variables*

|  |  |
| --- | --- |
| **Variable** | **Details** |
| Low birth weight births | 3 financial years’ data from ISD Maternity team (10/11—12/13 – the three most recent years’ data available). Expressed as a fraction of average population over 3 years (MYEs 2011, 2012, 2013). |
| Death rate 0-74 all causes | 5 financial years’ GRO death records (09/10 – 13/14). Expressed as a fraction of average population over middle 3 years (MYEs – 2011, 2012, 2013) since 2010 MYE not available until Spring 2016.  Cause of death selected using ICD10 codes:   * Cancer C00--D48 * CHD I20--I25 * Stroke I61, I63, I64 |
| Death rate 0-74 Cancer |
| Death rate 0-74 CHD |
| Death rate 0-74 Stroke |
| All cause SMR 0-64 | Standardised mortality ratios with different causes of death.  Using 5 financial years’ GRO death records (09/10 – 13/14). SMR calculated using average population over middle 3 years (MYEs – 2011, 2012, 2013) since 2010 MYE not available until Spring 2016.  Cause of death selected using ICD10 codes:   * Cancer C00--D48 * Heart disease I00--I99 * Respiratory J00--J99 * Digestive K00--K93 * External Causes V\_\_--Y\_\_ * Other – any other codes |
| All cause SMR 0-69 |
| All cause SMR 0-74 |
| Cancer SMR 0-64 |
| Cancer SMR 0-69 |
| Cancer SMR 0-74 |
| Heart Disease SMR 0-64 |
| Heart Disease SMR 0-69 |
| Heart Disease SMR 0-74 |
| Respiratory SMR 0-64 |
| Respiratory SMR 0-69 |
| Respiratory SMR 0-74 |
| Digestive System SMR 0-64 |
| Digestive System SMR 0-69 |
| Digestive System SMR 0-74 |
| External Causes SMR 0-64 |
| External Causes SMR 0-69 |
| External Causes SMR 0-74 |
| Other SMR 0-64 |
| Other SMR 0-69 |
| Other SMR 0-74 |
| High Resource Individual counts | Count of individuals belonging to the group of ~100,000 highest resource-users that account for 50% of the total resource. Based on 3 financial years’ data from ISD IRF team (11/12 – 13/14). Expressed as a fraction of average population over 3 years (MYEs – 2012, 2013, 2014). |
| Did Not Attend counts – as a fraction of all OP appointments | 3 financial years’ data from ISD SC team (11/12 – 13/14). Expressed as a fraction of total outpatient appointments over same 3 financial years. Included because of its link to deprivation. |
| Did Not Attend counts – ratio to data zone population | 3 financial years’ data from ISD SC team (11/12 – 13/14). Average population over 3 years (MYEs – 2012, 2013, 2014). Included because of its link to deprivation. |
| Patients receiving Diabetes prescriptions | 3 financial years’ data from ISD Prescribing team (11/12 – 13/14). Expressed as a fraction of average population over 3 years (MYEs – 2012, 2013, 2014).  Dementia - includes all drugs in BNF section 4.11; Diabetes - includes all insulin and antidiabetic drugs; Respiratory - includes all lama, laba and high strength steroid inhalers |
| Patients receiving Dementia prescriptions |
| Patients receiving Respiratory prescriptions |
| Long-term illness | Data from Census 2011 question 20 – standardised by age and sex using 2011 MYE population. |
| Mental health condition | Data from Census 2011 question 20 – standardised by age and sex using 2011 MYE population. Included because of the link between mental health conditions and need for Acute services. |
| Limiting long-term illness – limited a lot | Data from Census 2011 question 21 – standardised by age and sex using 2011 MYE population. Uses number of respondents answering ‘Yes – a lot’. |
| Limiting long-term illness – limited a little or a lot | Data from Census 2011 question 21 – standardised by age and sex using 2011 MYE population. Uses number of respondents answering ‘Yes’ (including both ‘a little’ and ‘a lot’). |
| Long-term sick and not seeking work | Data from Census 2011 questions 24-28 – standardised by age and sex using 2011 MYE population. |
| Older people living alone – 65 and over | Data from Census 2011 – standardised by age and sex using 2011 MYE population. |
| Older people living alone – 70 and over |
| Older people living alone – 75 and over |
| Older people living alone – 80 and over |
| Older people living alone – 85 and over |
| Older people living alone – 90 and over |
| Unpaid care – 1 hour or more | Data from Census 2011 question 9 – standardised by age and sex using 2011 MYE population. Uses number of respondents answering yes (with various numbers of hours). |
| Unpaid care – 20 hours or more |
| Unpaid care – 35 hours or more |
| Unpaid care – 50 hours or more |
| General health – very bad | Data from Census 2011 question 19 – standardised by age and sex using 2011 MYE population. Uses number of respondents answering ‘Very bad’. |
| General health – bad or very bad | Data from Census 2011 question 19 – standardised by age and sex using 2011 MYE population. Uses number of respondents answering ‘Bad’ or ‘Very bad’. |
| Education level – no qualifications | Data from Census 2011 question 23 – standardised by age and sex using 2011 MYE population. Included for its relation to deprivation. Level 1 refers to ‘O’ grades or similar; level 2 refers to Highers or similar (<http://www.scotlandscensus.gov.uk/variables-classification/highest-level-qualification>) |
| Education level – level 1 and below |
| Education level – level 2 and below |
| Job seekers | Data from Census 2011 question 25 – standardised by age and sex using 2011 MYE population (possibly economically active population only – try both). Included for its relation to deprivation. |
| Ethnic group populations | Data from Census 2011 question 15 – simple fraction of 2011 MYE population. |

**Annex B: Methodology for selecting the needs indicators**

The methodology adopted for selecting the needs indicators, based on the 2007 method from Technical Report D, is outlined below.

Preliminary selection of the candidate variables and regressions

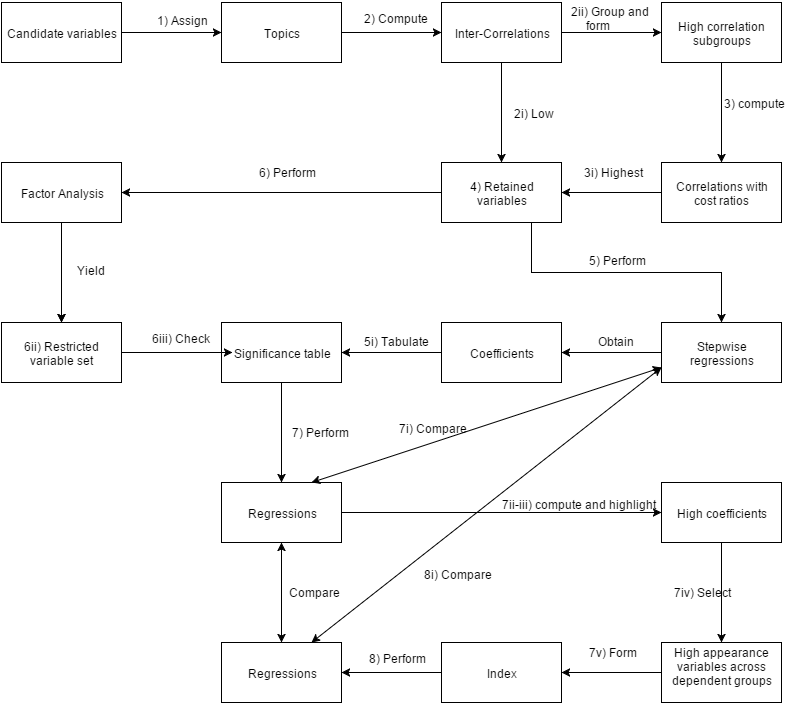
1. Allocate the candidate variables to the appropriate topic/category.
2. Births and deaths
3. Health/morbidity
4. Unpaid care and elderly living alone
5. Deprivation
6. For each topic, compute the inter-correlation.
7. Retain candidate variables with low inter-correlations 🡪 Go to step 4.
8. Group pairs of candidate variables with very high inter-correlations to form subgroups. 🡪 Go to step 3.
9. For each subgroup, compute the correlation with cost ratios.
10. Eliminate near duplicates for each subgroup – retain candidate variables that have the highest correlation with cost ratios for most diagnostic groups.
11. New candidate variables list is formed.
12. Regress (stepwise) cost ratios against supply model with all the new candidate variables.
13. Display all the statistically significant candidate variables for each diagnostic group in a table.

Developing need indexes for the diagnostic groups cost ratios

1. Undertake a factor analysis – using principal components – of the new candidate variables. Selection is based on the extent to which each of the ‘raw’ variables appears to reflect the factors or dimensions that are generated by the factor analysis.
2. Choose the most effective variables.
3. The restricted variable set is formed.
4. Use table from step 5 to check that most of them had appeared in the full stepwise regressions.
5. Regress cost ratios against supply model with all the restricted candidate variables.
6. Compare with the original equations – in terms of goodness of fit and the general specification test.
7. Compute the coefficients from regressing these variables against the diagnostic groups cost ratios.
8. Highlight variables with coefficients relatively high within the diagnostic groups.
9. Select variables that appear often across the diagnostic groups and have high coefficients.
10. Compute the sum of z-scores for those selected variables to form an index.
11. Regress cost ratios against supply model with the index.
12. Compare with the original equations – in terms of goodness of fit and the general specification test.

Figure B.1 shows the above methodology in the form of a flowchart. The numbers correspond to the numbers above.

It is possible that the Subgroup may decide to allow the possibility of different needs indicators in different diagnostic groups. In this case, the final steps from 7iii) onwards would be revised.

**

*Figure B.1: Flowchart summary of proposed index development methodology*

**Annex C: Regression results**

Table C.1 shows the adjusted R-squared from using each potential indicator as a single explanatory variable in a linear regression including the supply model. The subgroups formed earlier are identified by alternate shading and non-shading. The highest adjusted R-squared within each subgroup, for each diagnostic group, is indicated in bold and italics.

*Table C.1: Adjusted R2 for the ‘near-duplicate’ potential indicators in nine subgroups*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **All cause SMR <75** | ***48.4%*** | ***11.1%*** | ***15.9%*** | ***32.9%*** | ***19.8%*** | ***36.4%*** | ***31.2%*** | 45.5% |
| **All cause SMR <70** | 47.0% | 10.8% | 15.2% | 32.3% | 19.0% | 35.7% | 29.2% | 45.4% |
| **All cause SMR <65** | 44.3% | 10.4% | 13.9% | 30.9% | 17.8% | 33.9% | 27.2% | 45.1% |
| **Death rate <75 all causes** | 44.3% | 10.8% | 12.4% | 30.8% | 17.0% | 33.9% | 26.5% | ***45.8%*** |
| **Cancer SMR <65** | 27.9% | 12.5% | 6.1% | 21.8% | 7.4% | 21.1% | 15.4% | 42.3% |
| **Cancer SMR <70** | 31.0% | 13.9% | 7.5% | 23.3% | 8.7% | 22.9% | 17.2% | 42.8% |
| **Cancer SMR <75** | ***33.6%*** | ***15.3%*** | ***8.4%*** | ***24.3%*** | ***9.7%*** | ***24.4%*** | ***18.8%*** | ***43.4%*** |
| **Death rate <75 Cancer** | 29.3% | 13.4% | 5.8% | 22.5% | 7.8% | 21.8% | 15.1% | 43.6% |
| **Heart SMR <65** | 30.5% | 7.9% | 8.5% | 23.4% | 10.9% | 24.3% | 17.8% | 42.4% |
| **Heart SMR <70** | 32.6% | 7.9% | 9.8% | 24.5% | 12.0% | 26.0% | 19.2% | 42.7% |
| **Heart SMR <75** | ***34.8%*** | ***8.0%*** | ***11.5%*** | ***25.3%*** | ***13.0%*** | ***27.5%*** | ***21.7%*** | ***42.9%*** |
| **Death rate <75 CHD** | 30.0% | 7.9% | 8.1% | 23.4% | 9.4% | 23.8% | 17.5% | 42.8% |
| **Other SMR <70** | ***27.5%*** | ***7.3%*** | ***7.2%*** | ***21.7%*** | ***8.2%*** | ***22.9%*** | ***17.2%*** | 41.5% |
| **Other SMR <75** | 26.7% | ***7.3%*** | 6.6% | ***21.7%*** | 8.0% | 22.2% | 17.0% | 41.2% |
| **LLTI – Yes (both)** | 60.8% | ***9.7%*** | ***21.6%*** | 39.4% | ***26.8%*** | ***46.7%*** | 38.2% | ***50.9%*** |
| **LLTI – Yes (a lot)** | ***58.5%*** | 9.3% | 19.9% | ***38.6%*** | 25.3% | 44.7% | ***39.1%*** | 50.1% |
| **General health – very bad** | 57.8% | 9.4% | 20.6% | 38.4% | 25.0% | 43.5% | 38.4% | 50.2% |
| **General health – bad or very bad** | 50.7% | 8.7% | 16.3% | 35.1% | 20.6% | 38.2% | 34.0% | 49.0% |
| **Mental health condition** | 45.8% | 8.6% | 13.6% | 32.3% | 22.6% | 35.6% | 30.0% | 45.0% |
| **Long term illness** | 49.8% | 9.1% | 15.9% | 33.9% | 18.0% | 38.8% | 31.1% | 49.4% |
| **Living alone ≥65** | ***20.0%*** | 6.9% | 3.3% | ***18.7%*** | 4.5% | ***16.6%*** | ***11.0%*** | 40.7% |
| **Living alone ≥70** | ***20.0%*** | 6.9% | 3.3% | 18.6% | ***4.6%*** | 16.5% | ***11.0%*** | 40.7% |
| **Living alone ≥75** | 19.7% | 6.9% | 3.2% | 18.4% | 4.3% | 16.3% | 10.9% | 40.5% |
| **Living alone ≥80** | 19.4% | 6.9% | 3.3% | 18.2% | 4.0% | 16.0% | 10.9% | 40.4% |
| **Living alone ≥85** | 19.4% | 7.0% | 3.5% | 18.2% | 3.9% | 15.9% | ***11.0%*** | 40.3% |
| **Living alone ≥90** | 19.6% | ***7.1%*** | ***3.5%*** | 18.3% | 4.0% | 16.0% | ***11.0%*** | 40.5% |
| **Unpaid care ≥ 1 hour** | 21.8% | 7.0% | 3.8% | 19.4% | 4.2% | 18.0% | 11.6% | 44.0% |
| **Unpaid care ≥ 20 hours** | ***43.2%*** | ***8.5%*** | ***12.3%*** | ***30.4%*** | ***13.0%*** | ***33.8%*** | ***25.8%*** | ***50.7%*** |
| **Unpaid care ≥ 35 hours** | 42.9% | 8.5% | 12.5% | 30.3% | 13.0% | 33.7% | 25.7% | 50.0% |
| **Unpaid care ≥ 50 hours** | 40.1% | 8.3% | 11.4% | 28.6% | 12.1% | 31.4% | 24.2% | 49.0% |
| **Education – no qualifications** | 19.4% | 7.0% | 3.2% | 18.2% | 3.9% | 16.0% | 11.0% | 40.3% |
| **Education – level 1 and below** | 20.5% | 7.0% | 3.5% | 18.5% | 4.4% | 16.8% | 11.3% | 41.1% |
| **Education – level 2 and below** | ***20.8%*** | ***7.0%*** | ***3.6%*** | ***18.6%*** | ***4.6%*** | ***17.2%*** | ***11.3%*** | ***41.5%*** |
| **Job seekers** | 19.4% | 6.9% | 3.2% | 18.2% | 3.9% | 15.9% | 10.9% | 40.7% |
| **DNA count – as fraction of appointments** | 45.4% | 8.4% | 16.4% | 31.7% | 19.7% | 34.5% | 29.9% | 45.4% |
| **DNA count – ratio to population** | ***60.9%*** | ***10.3%*** | ***19.6%*** | ***38.2%*** | ***24.6%*** | ***44.4%*** | ***35.5%*** | ***57.8%*** |

For completeness, Table C.2 shows R-squared values for the four individually-retained variables.

*Table C.2: Adjusted R2 for the individually-retained potential indicators*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Whole Acute** | **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| **Low birth weight births** | 20.5% | 6.9% | 3.6% | 18.7% | 4.3% | 17.4% | 11.7% | 40.4% |
| **Patients receiving Dementia prescriptions** | 19.4% | 7.4% | 3.4% | 18.2% | 4.1% | 15.9% | 10.9% | 40.5% |
| **Long term sick and not seeking work** | 19.7% | 7.0% | 3.5% | 18.3% | 4.1% | 16.2% | 11.2% | 40.4% |
| **High resource individuals** | 38.6% | 9.5% | 9.7% | 26.7% | 15.1% | 31.4% | 21.2% | 44.0% |

1. Available from <http://www.gov.scot/Publications/2015/08/7995/downloads> [↑](#footnote-ref-1)
2. The ICD-10 code groups can be found here: <http://apps.who.int/classifications/icd10/browse/2015/en> [↑](#footnote-ref-2)