**TAGRA ACUTE MLC SUBGROUP Thursday 19th March 2015**

**GEOGRAPHY, TIME SPAN AND SIMD EVALUATION**

**Background**

The Morbidity and Life Circumstances (MLC) adjustment corrects the NRAC target shares for each NHS Board according to the value of a needs index, which was developed to represent the additional needs of the population over and above those due to age and sex. The indicators within the current needs index for the Acute care programme are: all-cause standardised mortality ratio (SMR) in ages 0-74, obtained from death records; and limiting long-term illness (LLTI) ratio (age-sex standardised) from the 2001 Census. The correction is derived from a regression analysis used to quantify the relationship between these needs indicators and healthcare utilisation, based on one year of data at Intermediate Zones.

The current Acute MLC Review is concerned with developing an up-to-date needs index for the Acute care programme, by looking for the factors that best explain variation in the utilisation of healthcare between small areas, using statistical regression analysis. The temporal and geographical basis for the Review is also to be reconsidered, as well as the granularity in terms of the seven diagnostic groupings (Cancer, Heart, Digestive, Injury, Other, Respiratory, and Outpatients) in the Acute programme.

At the previous Acute MLC subgroup meeting, there was extensive discussion around the issues of data availability arising from the redrawn Data Zones (see paper TAMLC20). It was decided that the Review should be conducted at the new geography, meaning a delay to the main work until August 2015 and the exclusion of the Scottish Index of Multiple Deprivation (SIMD) from the list of candidate variables, since it will not be available at the new geography until 2016. Although statistical concerns were highlighted around the use of the SIMD’s composite scores as predictor variables, concerns were also raised about its outright exclusion. AST were therefore asked to investigate the significance of SIMD 2012 as a predictor of need using the current data and report back to the subgroup.

AST were also asked to investigate two possible age splits (at 65 and 75) for the regression analysis, following discussion of the paper TAMLC19. AST have subsequently identified higher priorities in refreshing the data on which the current needs index is based, analysing the adequacy of this index in preparation for comparisons with new candidate variables, and exploring the results of using different geography and time aggregations in preparation for decisions on those. The age split analysis will be performed and reported at a future Acute MLC subgroup meeting, as well analysis on granularity in terms of diagnostic groupings.

It was suggested to look at the actual Acute MLC expenditure by diagnostic group, in order to select the most ‘expensive’ diagnostic groups for the initial age split analysis. Examining the actual Acute MLC spend is also useful for the interpretation of the regression results included in the current paper, so this is presented in Annex A.

**1. Summary**

This paper demonstrates the current strengths and weaknesses of the ‘reference model’ for the Acute MLC adjustment, then considers the implications of geography and time span before presenting the results of testing the significance of Scottish Index of Multiple Deprivation (SIMD) 2012 variables as predictors of healthcare utilisation.

In Section 2, the results of a new up-to-date analysis using the current Acute MLC needs index, at the 2001 Data Zones and Intermediate Zones, are presented. This has been carried out to provide a benchmark with which to compare both the SIMD analysis and the future results with new candidate variables at the new 2011 geographies. The most recent healthcare cost data is used, and the current indicators of need – the Standardised Mortality Ratio (SMR) and the Limiting Long-Term Illness (LLTI) ratio – are refreshed with recent data. In order to update the LLTI ratio variable, a new definition of LLTI had to be chosen due to a change in the wording and response options in the relevant Census question.

Section 3 discusses the implications of the existing reference model performance on the time span and geography choice for the Acute MLC Review, and looks also at the presence of outliers and influential points. Three options are considered: Data Zone geographies with a single year of data, Data Zone geographies with averaging over three years of data, and Intermediate Zone geographies with a single year of data. Modelling based at Intermediate Zones is found to perform best, in terms of higher R-squared values, less frequent violation of the normality assumption for regression residuals, and fewer outlying data points. The decisions taken in the NRAC 2007 Review and in the Mental Health and Learning Difficulties 2011-2012 Review are also noted. **The subgroup is asked to consider the issues outlined and agree on the geography and time basis for the Acute MLC Review.**

In Section 4, the significance of the SIMD overall score and its domain scores is evaluated. SIMD scores at Data Zone level, and ‘local and national shares’ of deprivation at Intermediate Zone level, are used as predictors individually and in combination with LLTI and SMR. The analysis shows that at Data Zones, the SIMD appears equally strong as a predictor compared to LLTI, and nearly as strong as LLTI at Intermediate Zones. In combination with LLTI and SMR, SIMD variables tend to increase the R-squared values by a small amount. **The subgroup is asked to agree on the potential importance of SIMD as an indicator of need.** This decision will have a bearing on the time-scale for the remainder of the review (see previous paper TAMLC20 and current paper TAMLC25).

**2. Reference model**

The MLC adjustment takes into account the additional needs of the population over and above those due to age and sex. The current Acute MLC indicators of need are the all cause Standardised Mortality Ratio (SMR) in ages 0-74, and the Limiting Long-Term Illness (LLTI) ratio. The sum of the z-scores of these two variables is the current needs index in the NRAC formula. As part of the Acute MLC Review process, these indicators have now been refreshed and their current explanatory power for healthcare utilisation has been tested, in order to understand the strengths and weaknesses of the current Acute MLC adjustment. This section presents the results of this testing, and highlights issues discovered around violation of the assumption of normality in the regression residuals.

The utilisation of healthcare, which is the outcome to be predicted, is represented by the ratio of the actual costs of healthcare (taking into account activity type and length of stay in that specific neighbourhood) to the expected costs (based on the neighbourhood’s population and national age/sex average cost per head). The Acute MLC adjustment is based on regression of these cost ratios upon the needs index. Health board ‘dummy’ variables and supply variables are included in the regressions, to avoid predicting effects that are largely due to variations in supply; the supply variables (IPACX and OPACX) are statistical measures representing the distance between the population grid centroid and the nearest facility. The health board dummies, supply variables and the needs index are together known as the ‘reference model’.

In the current update, the cost ratios for 1 year at Data Zones and 1 year at Intermediate Zones are calculated, using 2012/13 inpatient and outpatient activity data; the cost ratios for 3 years at Data Zones are calculated as the average of the 2011/12, 2012/13 and 2013/14 cost ratios. The SMR is recalculated using death records from 2008 to 2012 calendar years, and the LLTI ratio is recalculated using 2011 Census data. This required updating the definition of LLTI, due to a change in the response options for the relevant Census question from a simple “Yes”/“No” to three options: “Yes – a little”, “Yes – a lot” and “No”. The details of the definition and the analysis carried out to support this are given in Annex B. A variable referred to as ‘Yes both’ is chosen as the best LLTI ratio variable. It is an age-sex adjusted ratio combining both positive answers with equal weight.

Linear models are then fitted, using the supply variables and health board dummies first as a baseline and then adding in the SMR and LLTI ratio as explanatory variables. Finally, the full reference model is also reproduced and tested.

**2.1 Correlations**

Correlations between SMR and the cost ratios and between LLTI and the cost ratios for all diagnostic groups are calculated to get an initial impression of the strength and direction of the relationships. The results are shown in Table 1.

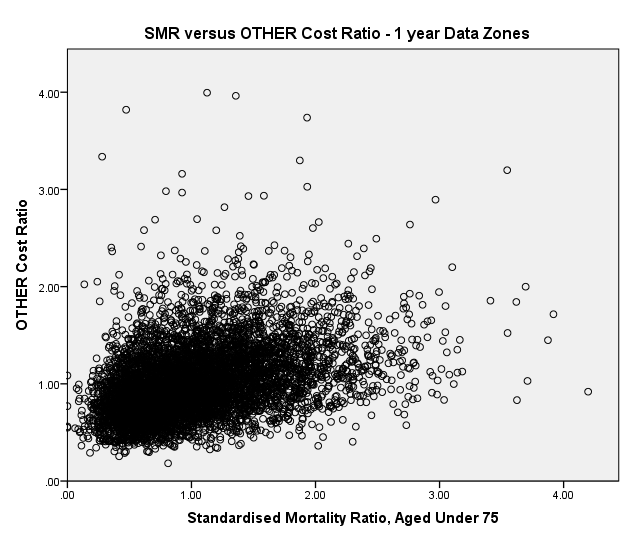
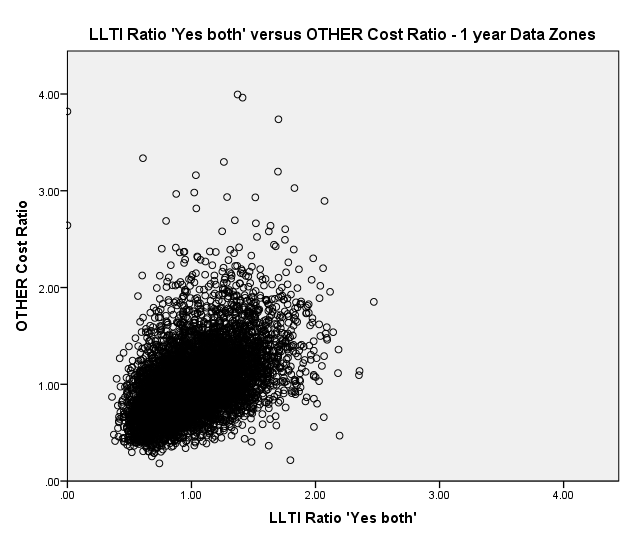
*Table 1. SMR and LLTI correlations to the cost ratios.*

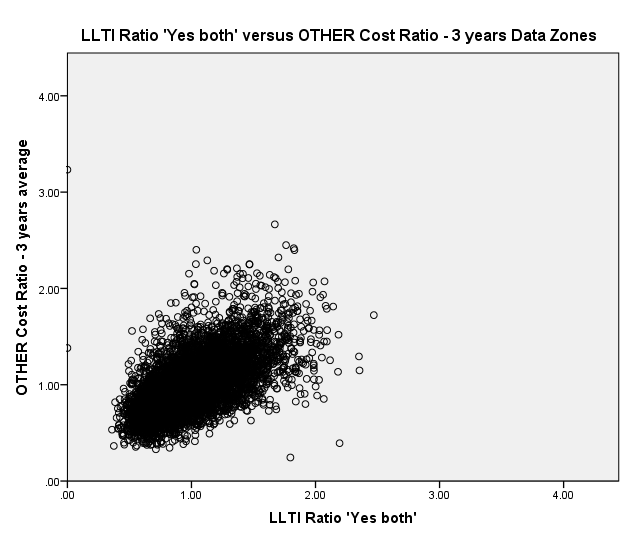
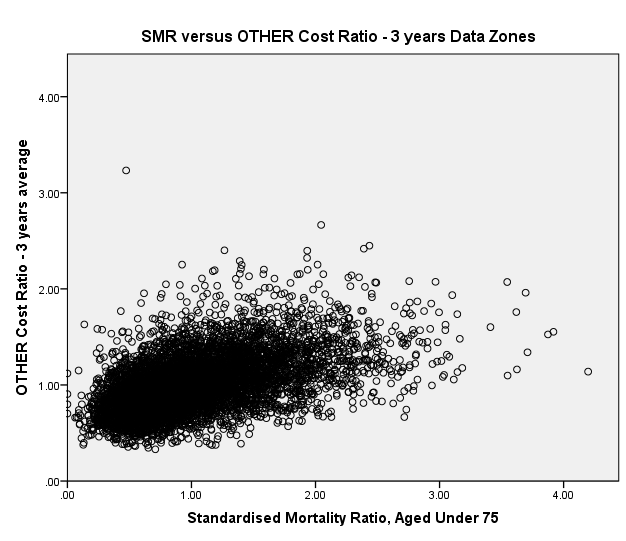
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations** | | **Cancer cost ratios** | **Heart cost ratios** | **Digestive cost ratios** | **Injury  cost  ratios** | **Other cost ratios** | **Respiratory cost ratios** | **Outpatients cost ratios** |
| 1 year Data Zones | **SMR** | 0.121 | 0.227 | 0.330 | 0.241 | 0.380 | 0.366 | 0.265 |
| **LLTI** | 0.152 | 0.285 | 0.403 | 0.294 | 0.474 | 0.445 | 0.359 |
| 3 years Data Zones | **SMR** | 0.221 | 0.349 | 0.439 | 0.379 | 0.491 | 0.495 | 0.313 |
| **LLTI** | 0.226 | 0.421 | 0.545 | 0.461 | 0.613 | 0.587 | 0.419 |
| 1 year Intermediate  Zones | **SMR** | 0.285 | 0.492 | 0.585 | 0.444 | 0.645 | 0.674 | 0.424 |
| **LLTI** | 0.278 | 0.502 | 0.623 | 0.459 | 0.680 | 0.706 | 0.467 |

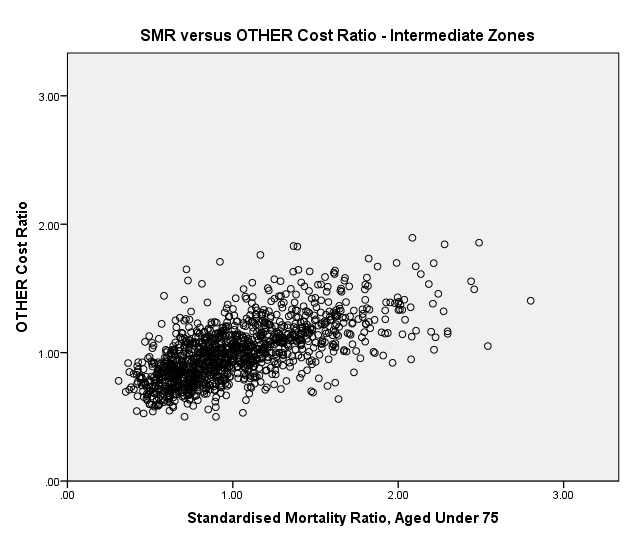
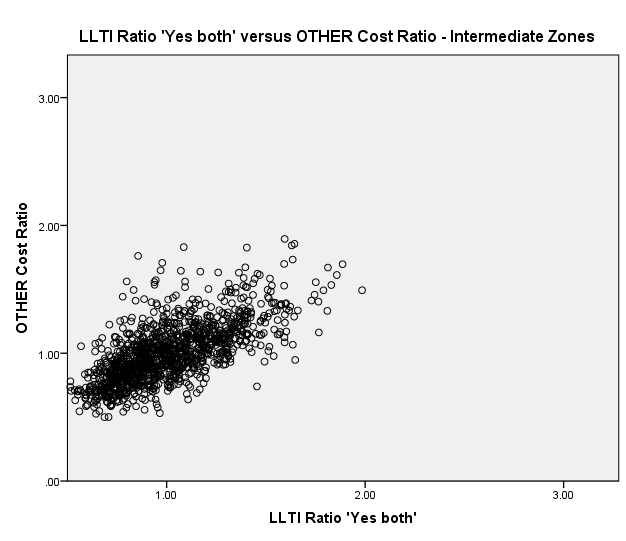
At Data Zone level, LLTI ratio is correlated a bit stronger to the cost ratios than SMR is. At Intermediate Zone level the correlation between SMR and the cost ratios is higher than LLTI only for Cancer. SMR is believed to be a strong predictor of need; its strength may have changed over time since as the population ages, an increasing proportion of people are living with multiple health conditions. Overall, the correlations suggest that LLTI may be a bit stronger predictor of need; this is formally checked in the regression analysis in section 2.3. Moreover, SMR is highly correlated to LLTI (the correlation is 0.71 at Data Zone level and 0.89 at Intermediate Zone level) which means these two variables are collinear and they explain approximately the same amount of the cost ratios’ variation. This suggests the inclusion of one of them in the models may be appropriate rather than both of them.

**2.2 Scatter plots**

Scatter plots of cost ratios against both SMR and LLTI have been produced for all diagnostic groups at Data Zone level and Intermediate Zone level. All scatter plots suggest the same overall impression that there is a positive relationship between the explanatory variables and the cost ratios. The strength of the relationship appears higher at Intermediate Zones (as already suggested by the correlations in Table 1). The most expensive diagnostic group, ‘Other’, is chosen as an example, for which scatter plots are shown below (Figure 1) at both geography levels. (See Annex A for the actual spend in each diagnostic group.)





*Figure 1. Scatter plots of Other cost ratios against SMR and against LLTI.*

**2.3 Performance of reference model (adjusted R-squared values)**

Linear models are fitted, testing SMR and the LLTI ratio – separately, as well as combined in the current needs index – as explanatory variables for the cost ratios. This analysis is performed at three different combinations of geography and time-scale of averaging. Adjusted R-squared values – the percentage of variance in the cost ratios that is explained by the model – are used as a goodness of fit measure. The results are shown in Table 2 below for all diagnostic groups. The values for the supply model + current index (i.e. the reference model) are highlighted in red.

*Table 2. Adjusted R-squared values of models comparing LLTI and SMR performance.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Adjusted**  **R-squared** | Predictors | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | supply model + SMR | 5.4% | 5.5% | 17.5% | 7.1% | 19.3% | 15.6% | 41.2% |
| supply model + LLTI | 6.0% | 8.9% | 21.5% | 11.0% | 21.3% | 21.3% | 43.7% |
| supply model + current index | 5.9% | 8.2% | 21.0% | 10.3% | 25.2% | 20.6% | 42.8% |
| 3 years Data Zones | supply model + SMR | 12.3% | 13.3% | 31.5% | 16.7% | 32.8% | 28.8% | 53.9% |
| supply model + LLTI | 12.3% | 20.0% | 39.3% | 25.4% | 44.4% | 37.7% | 57.6% |
| supply model + current index | 12.8% | 18.9% | 38.1% | 24.0% | 42.5% | 37.0% | 56.3% |
| 1 year Intermediate Zones | supply model + SMR | 24.3% | 25.9% | 48.3% | 28.9% | 52.0% | 49.3% | 66.2% |
| supply model + LLTI | 24.2% | 29.7% | 52.3% | 30.5% | 56.5% | 54.7% | 68.2% |
| supply model + current index | 24.5% | 29.1% | 51.6% | 31.1% | 56.1% | 53.9% | 67.4% |

The regression results suggest that SMR is a weaker predictor of cost ratios compared to LLTI ratio, confirming the initial impression from the correlations in Table 1. SMR and LLTI ratio are highly correlated, as mentioned in section 2.1, which explains why the adjusted R-squared values are not increased when combining both variables in an index. A model with LLTI only as an indicator of need performs equally well or better than a model using the current index. Notably, the most ‘expensive’ diagnostic groups – Other and Outpatients – give the best fit to the reference model (see Annex A for the actual spend in each diagnostic group).

**2.4 Violation of model assumptions**

The linear model form used for the regression analysis requires that the model ‘residuals’ are normally distributed. If this assumption is not met, then the statistical uncertainties on the estimated model parameters are not well known. In the case of the Acute MLC reference model, strong deviations from normality of residuals have been noted. Formal tests for normality of the residuals have been carried out, including calculation of skewness and kurtosis. The results of these investigations are presented in Annex C.

Overall, normality of the residuals is least violated at 1 year Intermediate Zones, with rule-of-thumb measures for normality being passed for 5 out of 7 diagnostic groups. These rule-of-thumb measures are failed for all diagnostic groups for analysis of the model using either 1 year Data Zones or 3 year Data Zones.

**3. Geography and time span**

This section discusses the choice of geography and the time span for the regression analysis of the Review, considering the results from Section 2 and some further investigation looking at the presence of outliers with each option. The decisions taken during the NRAC Review in 2007 and during the Mental Health and Learning Difficulties MLC Review in 2011-12 are also reviewed and considered.

**3.1 Considerations from present analysis**

In the previous section, the combination of time span and geography produced the highest of all R-squared values for the reference model regressions at 1 year Intermediate Zones. The adjusted R-squared values for each diagnostic group using 1 year at Data Zones, 3 years at Data Zones and 1 year at Intermediate Zones are reproduced from the previous section in Table 3 below.

*Table 3. Reference model performance (Extract of Table 2).*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Adjusted R-squared:**  **Reference model** | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | 5.9% | 8.2% | 21.0% | 10.3% | 25.2% | 20.6% | 42.8% |
| 3 years Data Zones | 12.8% | 18.9% | 38.1% | 24.0% | 42.5% | 37.0% | 56.3% |
| 1 year Intermediate Zones | 24.5% | 29.1% | 51.6% | 31.1% | 56.1% | 53.9% | 67.4% |

However, adjusted R-squared values alone should not be used to determine which level of geography and time-scale is ‘better’, as any increased aggregation of the data will necessarily reduce the total variance present in the dependent variable (and hence the noise present in the model). This effect explains why the Intermediate Zone level analysis yields higher adjusted R-squared values than the Data Zone level analysis and more complex analytical methods may be needed to attempt to disentangle whether this increase represents any additional gain of ‘true’ explanatory power.

The previous section found that the normality assumption was least violated at Intermediate Zone level (see Annex C). Only two of the diagnostic groups (Respiratory and Digestive, i.e. the least ‘expensive’ ones) fail the rule-of-thumb tests for normality of residuals, in contrast to all diagnostic groups at Data Zone level (regardless of whether 1 year or 3 year is used). This indicates that the uncertainties associated with the regression models will be more robustly known on the basis of Intermediate Zone level analysis.

To supplement these results and further inform the choice of time and geography, an analysis of Outliers and Influential points is performed. Outliers are observations which do not fit the model well, while Influential Points are observations which change the fit of the model in a substantive way. The analysis is presented in detail in Annex D and summarised below.

The number of residual outliers (Table 4) is the highest at 1 year Data Zones and the lowest at 1 year Intermediate Zones. If these results are considered as a proportion of the total number of points, 1 year Intermediate Zones offers the lowest proportion for all Diagnostic Groups apart from Digestive and Outpatients, for which 3 year Data Zones provides the lowest proportion.

*Table 4. Number and Percentage**of Residuals Outliers (Table D.2).*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Number and Percentage**  **of residuals outliers** | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | **20**  0.31% | **26**  0.40% | **30**  0.46% | **27**  0.42% | **21**  0.32% | **26**  0.40% | **8**  0.12% |
| 3 years Data Zones | **14**  0.22% | **18**  0.28% | **17**  0.26% | **17**  0.26% | **12**  0.18% | **28**  0.43% | **8**  0.12% |
| 1 year Intermediate Zones | **2**  0.16% | **2**  0.16% | **5**  0.40% | **3**  0.24% | **1**  0.08% | **4**  0.32% | **2**  0.16% |

There is one influential point for Respiratory at 1 year Data Zones. However, this particular Data Zone is not influencing all the fitted values but just its own fitted value. There are no influential points at 3 years Data Zones and 1 year Intermediate Zones.

**3.2 Previous decisions on geography and time span**

The geography and time span combination has been subject to previous discussions during the NRAC 2007 Review and during the Mental Health and Learning Difficulties MLC 2011-12 Review. A summary of the arguments for the decisions taken is given below.

In the NRAC 2007 Review[[1]](#footnote-1), 1 year was chosen for the time span since more than 1 year costed data was not available. The biggest concern in using Data Zones in the NRAC formula were the high zero activity for some condition groups (Mental Health and Learning Difficulties in particular), and the year to year variation in costs and numbers of patients (the ratio of the maximum to minimum annual number of inpatient discharges 2000-2004 was investigated). The biggest concern in using Intermediate Zones was that effects due to pockets of deprivation will be “averaged out”. So, 1 year Intermediate Zones were chosen as the time and geography base for the modelling, for all care programmes.

The Mental Health and Learning Difficulties MLC 2011-12 Review[[2]](#footnote-2) highlighted a similar issue around zero activity at Data Zones, which presents a problem of the model not being able to distinguish between genuine zero activity and zero activity due to chance. So, Intermediate Zones were preferred as the geography base. The choice of time span presented a choice between two of the TAGRA core criteria – responsiveness and stability. As the Mental Health and Learning Difficulties component of NRAC formula is static for three years after it has been updated, the benefits of having more responsive data at the time of the update were outweighed by having stability between updates. So, 3 years Intermediate Zones were chosen as the time and geography base for the modelling.

**3.3 Discussion**

On the basis of the present analysis, there are three main arguments in favour of retaining Intermediate Zones as the geography basis of the review: higher R-squared values, less frequent violation of the normality assumption for regression residuals, and fewer outlying data points. Zero activity at the Data Zone level is not an issue with the Acute MLC data as it was for the Mental Health and Learning Difficulties review. The argument against using Intermediate Zones in terms of pockets of deprivation being “averaged out” remains valid. Indeed, a smaller scale basis will always, in principle, provide a more detailed representation of variations in need. In terms of time-scale, the analytical work shows that 3-year averaging of Data Zones substantially narrows the gap to Intermediate Zones in terms of the numbers of outliers, but does not change the significance of the violations of model normality assumptions. Any decision regarding time-scale should primarily represent a choice between two of the TAGRA core criteria – responsiveness and stability.

**The subgroup is asked to reflect on these competing factors and to make a recommendation on the geography and time-scale to be used for the Review.**

**4. Potential significance of SIMD variables as predictors of need**

The possible use of the SIMD variables as potential indicators of need has been subject to previous discussion within the Acute MLC subgroup (see paper TAMLC20, Section 3). From a practical point of view, if SIMD variables were to be included in the list of indicators of need, then the main work of the Review could not be undertaken until the release of the next SIMD update in 2016, and its outcomes would not be incorporated into the NRAC formula until the 2019/20 Resource Allocation Formula run.

The suitability of SIMD for use in the NRAC formula has also been discussed, in addition to the practical issues of availability. The NRAC 2007 Review listed several weaknesses of the SIMD overall score and domain scores, including the potential change of definitions and components between releases, the complexity of the structure, and the weights used to combine domains. Because of these weaknesses, including the unavailability of the SIMD overall score and domain scores at Intermediate Zones, and for greater transparency, the NRAC 2007 Review recommended the exclusion of the SIMD scores from the list of variables.

At the last Acute MLC subgroup meeting in January, it was decided that SIMD would probably be excluded from the current Review for the above reasons, but the analytical team was asked to check the significance of SIMD 2012 overall and domain scores when added to the reference model. This analysis is necessarily undertaken using the old geographies, but it is considered unlikely that the change to new geographies in SIMD 2016 would fundamentally alter the relationships determined here.

The analysis uses the SIMD 2012 overall score, the domain scores and the two newly created “SIMD minus Health” scores (provided by the Scottish Government SIMD team). The latter are referred to as “Scenario 1” and “Scenario 2”; these represent slightly different calculation methods, explained in Annex E. The correlation between Scenario 1 and Scenario 2 scores is close to 1, so only the Scenario 1 score is used in the analysis.

In analysis using individual domain scores, all domains are considered except for Health and Access. The Health domain score is excluded since the health aspect of deprivation is already captured in the reference model using SMR and LLTI. Also, the Access domain score is excluded since the variables in this domain include the supply variables used in the reference model.

Analysis has been performed at 1 year Data Zones, 3 year Data Zones and 1 year Intermediate Zones. The results at Data Zones and Intermediate Zones are not directly comparable, since the SIMD overall score and domains scores are only available at the Data Zone geography and cannot be easily aggregated. A method based on the local and national shares of deprived Data Zones within each Intermediate Zone (as proposed by the Scottish Government SIMD team) has been used to create an SIMD measure for the Intermediate Zones analysis. Section 4.1 presents the results for the Data Zones analysis, while section 4.2 shows the analysis at Intermediate Zones.

**4.1. Data Zones**

First, correlations between the cost ratios and the SIMD variables are calculated, for all diagnostic groups. Based on the correlations, the Income domain score and Scenario 1 score are chosen for the regression analysis.

Secondly, the significance of these two SIMD variables is tested stepwise in three types of models – single variable, multivariate and using a composite variable. The single variable models are testing the explanatory power of SMR, LLTI and the SIMD variables when added separately to the supply model (supply variables and health board dummy variables). The multivariate models are testing the explanatory power of SMR and LLTI, versus SMR, LLTI and the SIMD variable, when added to the supply model. The models using a composite variable are testing the explanatory power of the reference model (supply model with SMR and LLTI combined into the needs *index*); the reference model and SIMD variable; and the supply model with SMR, LLTI and SIMD combined together into a new index (the sum of the z-scores of SMR, LLTI and the SIMD variable). The results of this are shown below in section 4.1.2.

Since the SIMD scores are not available on a yearly basis, the SIMD explanatory variables used in the regressions for 1 year Data Zones and for 3 years Data Zones are the same. The only difference between the analysis at the two different time spans is in the dependent variables – the cost ratios for the 3 years Data Zones analysis are an average of 2011/12, 2012/13 and 2013/14 cost ratios.

**4.1.1 Correlations**

The correlations between the cost ratios and the SIMD explanatory variables are shown in the first two tables of Annex F, for each diagnostic group. The initial impression from these correlations is that the Income score is the best SIMD variable to predict the cost ratios. Scenario 1 score is also further investigated as a reference. So, Income domain score and Scenario 1 score are used in the next stage of the investigation.

The correlations between SMR and the cost ratios, and LLTI and the cost ratios (Table 1) are quite similar to the correlations in Table F.1 and Table F.2. The significance of the SIMD variables (Income domain score and Scenario 1 score) will be formally checked in the regression analysis.

**4.1.2 Significance of SIMD variables**

To test the significance of SIMD, regression analysis is carried out as described above. Table 5 and Table 6 provide a summary of the regression analysis output. The red values in bold indicate the best adjusted R-squared values for each diagnostic group.

*Table 5. Significance of SIMD at 1 year Data Zones.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1 year Data Zones** | **adjusted R-squared value (best values for each DG highlighted in red)** | | | | | | |
| **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| supply model (HB Dummies + OPACX +IPACX) | 4.7% | 1.2% | 10.3% | 1.6% | 8.2% | 5.8% | 39.8% |
| **Single variable models** | | | | | | | |
| supply model + SMR | 5.4% | 5.5% | 17.5% | 7.1% | 19.3% | 15.6% | 41.2% |
| supply model + LLTI | 6.0% | 8.9% | 21.5% | 11.0% | 21.3% | 21.3% | 43.7% |
| supply model + Income domain score | 6.0% | 9.3% | 22.4% | 11.7% | 26.9% | 22.8% | 43.5% |
| supply model + SIMD minus Health "Scenario 1" score | 5.9% | 8.9% | 21.9% | 11.2% | 25.4% | 22.1% | 43.5% |
| **Multivariate models** | | | | | | | |
| supply model + SMR + LLTI | 6.0% | 9.0% | 21.7% | 11.1% | 26.6% | 21.7% | 43.8% |
| supply model + SMR + LLTI + Income domain score | **6.1%** | **9.6%** | **22.8%** | **12.0%** | **27.9%** | **23.3%** | **43.9%** |
| **Models using a composite variable** | | | | | | | |
| supply model + current index (sum of SMR and LLTI z-scores) | 5.9% | 8.2% | 21.0% | 10.3% | 25.2% | 20.6% | 42.8% |
| supply model + current index + Income domain score | **6.1%** | 9.4% | **22.8%** | 11.9% | 27.6% | 23.2% | 43.5% |
| supply model + new index (sum of SMR, LLTI and Income z-scores) | 6.0% | 9.1% | 22.3% | 11.5% | 27.1% | 22.5% | 43.2% |

*Table 6. Significance of SIMD at 3 years Data Zones.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **3 years Data Zones** | **adjusted R-squared value (best values for each DG highlighted in red)** | | | | | | |
| **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| supply model (HB Dummies + OPACX +IPACX) | 9.5% | 2.9% | 18.8% | 3.3% | 14.9% | 11.2% | 51.4% |
| **Single variable models** | | | | | | | |
| supply model + SMR | 12.3% | 13.3% | 31.5% | 16.7% | 32.8% | 28.8% | 53.9% |
| supply model + LLTI | 12.3% | 20.0% | 39.3% | 25.4% | 44.4% | 37.7% | 57.6% |
| supply model + Income domain score | 12.3% | 20.3% | 39.8% | 26.2% | 45.5% | 38.8% | 56.6% |
| supply model + SIMD minus Health "Scenario 1" score | 12.1% | 19.1% | 39.2% | 25.7% | 42.9% | 37.8% | 56.5% |
| **Multivariate models** | | | | | | | |
| supply model + SMR + LLTI | 12.8% | 20.3% | 39.6% | 25.8% | 44.9% | 38.6% | 57.6% |
| supply model + SMR + LLTI + Income domain score | **12.9%** | **21.3%** | **41.0%** | **27.4%** | **47.1%** | **40.4%** | **57.7%** |
| **Models using a composite variable** | | | | | | | |
| supply model + current index (sum of SMR and LLTI z-scores) | 12.8% | 18.9% | 38.1% | 24.0% | 42.5% | 37.0% | 56.3% |
| supply model + current index + Income domain score | **12.9%** | 21.1% | 40.7% | 27.1% | 46.7% | 40.3% | 56.9% |
| supply model + new index (sum of SMR, LLTI and Income z-scores) | **12.9%** | 20.6% | 40.2% | 26.4% | 45.7% | 39.7% | 56.8% |

For the 1 year Data Zones, the adjusted R-squared values of the single variable models (Table 5) suggest that the SIMD variables perform approximately equally well or better than the LLTI variable. The Scenario 1 score is performing worse than the Income domain score in the single variable models, so it is not tested in the multivariate models and in the models using a composite variable. In combination with the SMR and LLTI, the SIMD scores improve the model fit slightly. The ‘best’ model in terms of highest adjusted R-squared value is the supply model plus SMR, LLTI and Income domain score. However, the ‘best’ model represents only a modest improvement over the amount of variation explained by the reference model.

The conclusions in terms of the SIMD significance are the same for the 3 years Data Zones as for the 1 year Data Zones but the variability in the data explained by the models is increased (Table 6). The ‘best’ model in terms of highest adjusted R-squared value is the supply model plus SMR, LLTI and Income domain score, as for 1 year Data Zones. However, again, the ‘best’ model represents only a modest improvement over the reference model.

**4.2 Intermediate Zones**

There is currently no valid procedure to aggregate SIMD scores from Data Zones to Intermediate Zones. There are two approaches recommended by the Scottish Government SIMD team to create SIMD variables at Intermediate Zone level. Both approaches are based on calculating the proportion of Data Zones within Intermediate Zone (local share) or within Scotland (national share) that are 'deprived' at some predefined deprivation level. Deprivation levels (percentiles) of 5%, 10%, 15%, 20% and 25% have been considered; the results for 5% and 25% are applied below.[[3]](#footnote-3)

**4.2.1 Correlations**

The correlations between the cost ratios and the newly created explanatory variables from the Income domain score are calculated for each diagnostic group and are shown in the last table of Annex F. (The local and national shares were also calculated for the SIMD overall score and the two SIMD minus Health scenarios, but the correlations were extremely similar.)

Increase in the deprivation level leads to a slight increase in the correlation between cost ratios and local/national shares. The standard deviation of the national shares at all deprivation levels is found to be extremely low, so local shares are preferred to test the significance of SIMD Income domain variables. Both 5% and 25% Income domain local shares are tested in the regression analysis as well as the 5% and 25% Scenario 1 local shares as a reference.

The correlations (Table F.3) are slightly lower than those for SMR and LLTI at Intermediate Zones (Table 1). Moreover, there is high zero activity due to the definition of the shares, which suggests that the newly created SIMD variables will not be performing well in the regressions (1063 Intermediate Zones have a zero share at the 5% deprivation level, 834 Intermediate Zones at the 15% deprivation level, and 662 Intermediate Zones at the 25% deprivation level).

**4.2.2 Significance of SIMD variables**

To test the significance of SIMD Income domain and Scenario 1 5% and 25% local shares, regression analysis is carried out using the same stepwise method as described in the beginning of section 4.1. Table 7 provides a summary of the regression analysis output. The red values in bold indicate the best adjusted R-squared values for each diagnostic group.

*Table 7. Significance of SIMD at 1 year Intermediate Zones.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1 year Intermediate Zones** | **adjusted R-squared value (best values for each DG highlighted in red)** | | | | | | |
| **Cancer** | **Heart** | **Digestive** | **Injury** | **Other** | **Respiratory** | **Outpatients** |
| supply model (HB Dummies + OPACX +IPACX) | 21.0% | 5.2% | 28.0% | 6.5% | 21.8% | 18.9% | 62.9% |
| **Single variable models** | | | | | | | |
| supply model + SMR | 24.3% | 25.9% | 48.3% | 28.9% | 52.0% | 49.3% | 66.2% |
| supply model + LLTI | 24.2% | 29.7% | 52.3% | 30.5% | 56.5% | 54.7% | 68.2% |
| supply model + Income 5% local share | 21.5% | 17.0% | 38.7% | 16.4% | 32.2% | 34.4% | 65.4% |
| supply model + Income 25% local share | 23.9% | 23.6% | 47.3% | 24.2% | 46.2% | 44.4% | 66.5% |
| supply model + SIMD minus Health "Scenario 1" 5% local share | 21.5% | 16.2% | 38.4% | 15.7% | 32.3% | 33.9% | 65.0% |
| supply model + SIMD minus Health "Scenario 1" 25% local share | 23.9% | 23.7% | 47.1% | 24.6% | 46.7% | 44.6% | 66.6% |
| **Multivariate models** | | | | | | | |
| supply model + SMR + LLTI | 24.4% | 29.8% | 52.4% | **31.2%** | 56.8% | 54.8% | 68.4% |
| supply model + SMR + LLTI + Income 5% local share | **24.6%** | **30.1%** | 52.5% | **31.2%** | **56.9%** | **55.1%** | **68.5%** |
| supply model + SMR + LLTI + Income 25% local share | 24.5% | 29.8% | **52.6%** | 31.1% | 56.8% | 54.8% | 68.4% |
| **Model using a composite index variable** | | | | | | | |
| supply model + current index (sum of SMR and LLTI z-scores) | 24.5% | 29.1% | 51.6% | 31.1% | 56.1% | 53.9% | 67.4% |
| supply model + current index + Income 5% local share | **24.6%** | 29.5% | 51.8% | 31.1% | 56.2% | 54.2% | 67.5% |
| supply model + current index + Income 25% local share | 24.5% | 29.3% | 52.0% | 31.1% | 56.1% | 53.9% | 67.5% |
| supply model + new index (sum of SMR, LLTI and Income 5% local share z-scores) | 23.7% | 28.8% | 50.7% | 29.4% | 51.8% | 52.4% | 67.5% |
| supply model + new index (sum of SMR, LLTI and Income 25% local share z-scores) | **24.6%** | 29.0% | 51.9% | 30.5% | 55.2% | 53.1% | 67.5% |

The adjusted R-squared values of the single variable models in Table 7 suggest that the SIMD variables perform worse on their own than the LLTI variable; that the 25% local share variables are the stronger predictors; and that the Income domain and Scenario 1 variables perform equally well. However, in combination with the SMR and LLTI, the SIMD shares do often improve the model fit slightly. The ‘best’ model in terms of highest R-squared value is the supply model plus SMR, LLTI and Income domain 5% local share. However, as for the Data Zones, the ‘best’ model represents only a modest improvement over the amount of variation explained by the reference model.

**4.3 Discussion**

All three time and geography combinations’ results imply the same conclusions on the significance of SIMD. The adjusted R-squared values of the single variable models in the tables above suggest that the SIMD Income domain score is as powerful a predictor of need as either LLTI or SMR (at Data Zone level only). The ‘best’ model in terms of highest adjusted R-squared value is the supply model plus SMR, LLTI and Income domain variable (Income domain score at Data Zone level and Income 5% local share at Intermediate Zone level); however, the addition of SIMD to the existing model would add only a modest (generally less than 2% of cost ratio variance) amount of *additional* explanatory power. The extra explanatory power brought by the inclusion of SIMD is even smaller at Intermediate Zone level than at Data Zone level, likely due to weaknesses of the variable such as the frequent occurrence of zero values, which should be borne in mind if Intermediate Zones were to be chosen as the geography.

Any gain that SIMD 2016 is anticipated to bring to the final model must be balanced against the disadvantages in terms of the potential change of definitions and components between releases, the complexity of the structure, and the weights used to combine domains.Also, if SIMD variables are to be included in the list of indicators of need, then the main work of the Review cannot be undertaken until the release of the next SIMD update in 2016, and its outcomes would not be incorporated into the NRAC formula until the 2019/20 Resource Allocation Formula run.

**The Subgroup is asked to reflect on whether the additional (modest) gain of explanatory power offered by the SIMD variables (especially if the review is to be conducted at Intermediate Zone geography) is a sufficient reason to include SIMD 2016 variables in the list of indicators of need.**

**Annex A: Actual Acute MLC spend**

At the last Acute MLC subgroup meeting, it was suggested to look at the actual Acute MLC expenditure by diagnostic group, in order to select the most ‘expensive’ diagnostic groups for the initial age split analysis. Actual Acute MLC spend for 2012/13 financial year is shown in Table A.1 as well as the number of episodes.

*Table A.1. Actual Acute MLC spend for 2012/13 financial year.*

|  |  |  |
| --- | --- | --- |
| **Diagnostic Group** | **Actual spend in millions** | **Number of episodes** |
| Cancer | £400m | 199,581 |
| Heart | £407m | 148,762 |
| Digestive | £347m | 193,086 |
| Injury | £391m | 118,641 |
| Other | £1,245m | 684,139 |
| Respiratory | £292m | 133,220 |
| Acute Outpatients | £583m | 1,519,100 |
| **Total** | **£3,665m** | **2,999,529** |

When the age split analysis is done, it will be performed for only a selection of diagnostic groups, not for all seven of them. AST decided that the two most ‘expensive’ diagnostic groups should be chosen (Other and Acute Outpatients) as well as Cancer (known as the worst fit to the models in the past) and Heart.

**Annex B: Limiting long-term illness definition**

Limiting long-term illness (LLTI) is one of the indicators within the current Acute MLC needs index and in order to update the LLTI ratio variable, the LLTI 2011 Census data is used. The LLTI question in 2011 Census is different compared to the one in 2001 Census.

In the 2001 Census, the question about long-term illness was:

*Do you have any long-term illness, health problem or disability which limits your daily activities or the work you can do? (Include problems which are due to old age)*

*Yes No*

In the 2011 Census, the question about long-term illness was:

*Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months?*

*Yes – a lot Yes – a little No*

The main change between these two questions is the number of possible answers. The investigation presented below is performed in order to choose the best LLTI predictor of need. The three potential variables compared in the analysis are:

* Using only ‘Yes a lot’ answers and calculate an age-sex adjusted LLTI ratio;
* Using only ‘Yes a little’ answers and calculate an age-sex adjusted LLTI ratio;
* Using ‘Yes a lot’ and ‘Yes a little’ answers (equally weighted) and calculate an age-sex adjusted LLTI ratio (‘Yes both’).

**B.1 Correlations**

Correlations between the LLTI variables and their correlations to the cost ratios are calculated to get an initial impression of the strength and direction of the relationships. The results are shown in Table B.1 and Table B.2.

*Table B.1. LLTI variables correlations.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations  at Data Zone level** | LLTI  Yes a lot | LLTI  Yes a little | LLTI  Yes both (equal weight) | **Correlations**  **at Intermediate Zone level** | LLTI  Yes a lot | LLTI  Yes a little | LLTI  Yes both (equal weight) |
| LLTI Yes a lot | 1 | - | - | LLTI Yes a lot | 1 | - | - |
| LLTI Yes a little | 0.651 | 1 | - | LLTI Yes a little | 0.793 | 1 | - |
| LLTI Yes both  (equal weight) | 0.964 | 0.829 | 1 | LLTI Yes both  (equal weight) | 0.984 | 0.889 | 1 |

*Table B.2. Correlations between LLTI variables and cost ratios.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations** | Cancer  cost ratios | Heart cost ratios | Digestive cost ratios | Injury cost ratios | Other cost ratios | Respiratory cost ratios | Outpatients cost ratios |
| **Data Zones**  LLTI Yes a lot | 0.149 | 0.275 | 0.415 | 0.281 | 0.467 | 0.455 | 0.384 |
| **Data Zones**  LLTI Yes a little | 0.119 | 0.237 | 0.28 | 0.249 | 0.373 | 0.313 | 0.221 |
| **Data Zones**  LLYI Yes both  (equal weight) | 0.152 | 0.285 | 0.403 | 0.294 | 0.474 | 0.445 | 0.359 |
| **Intermediate Zones**  LLTI Yes a lot | 0.282 | 0.487 | 0.641 | 0.437 | 0.671 | 0.722 | 0.498 |
| **Intermediate Zones**  Yes a little | 0.226 | 0.466 | 0.482 | 0.448 | 0.602 | 0.56 | 0.315 |
| **Intermediate Zones**  Yes both  (equal weight) | 0.278 | 0.502 | 0.623 | 0.459 | 0.68 | 0.706 | 0.467 |

The three LLTI variables are highly correlated especially at Intermediate Zone level. ‘Yes a lot’ and ‘Yes both’ have correlation of almost 1 at both geographies which suggests they may be interchangeable in the analysis.

The correlations between the 3 LLTI variables and the cost ratios suggest that ‘Yes a little’ is a weaker predictor of need compared to ‘Yes a lot’ and ‘Yes both’. Formal analysis is performed to choose the best variable between ‘Yes a lot’ and ‘Yes both’ and the results are included in section B.3.

In addition to the correlations applied in Tables B.1 and B.2, the correlation between LLTI ratio using 2001 Census data (LLTI 2001) and LLTI ‘Yes a lot’ and LLTI ‘Yes both’ using 2011 Census data were investigated. At Data Zones, the correlation between the LLTI 2001 and LLTI ‘Yes a lot’ is 0.82; between LLTI 2001 and LLTI ‘Yes both’ it is 0.81. At Intermediate Zones the correlations are 0.96 and 0.96 respectively. The correlations at both geographies are high which reassures that the new definition should be consistent with the old one.

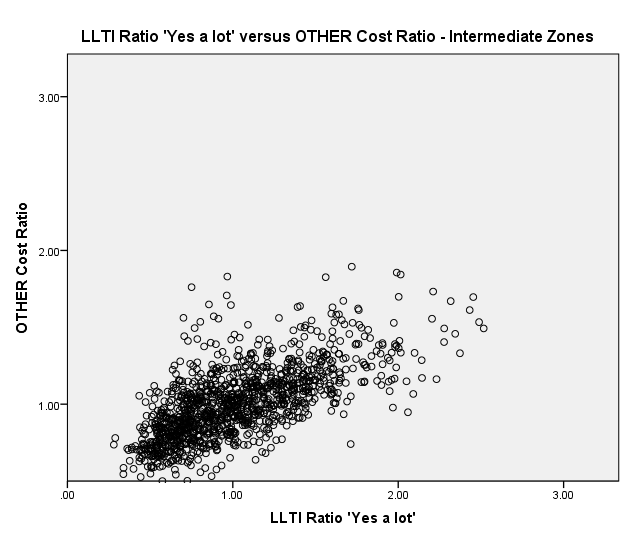
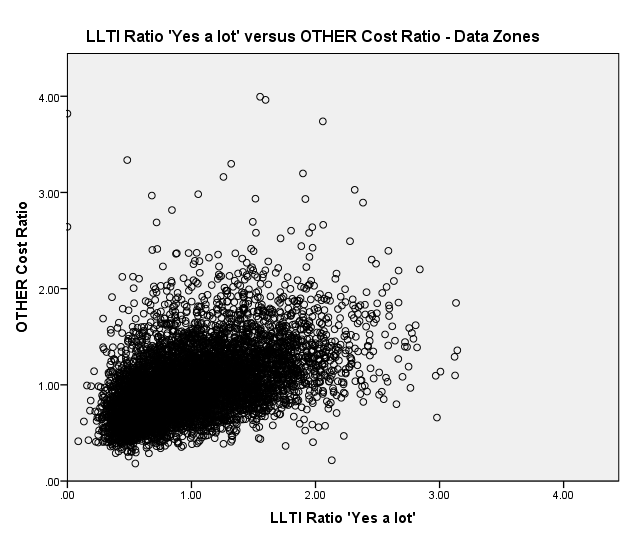
**B.2 Scatter plots and descriptive statistics**

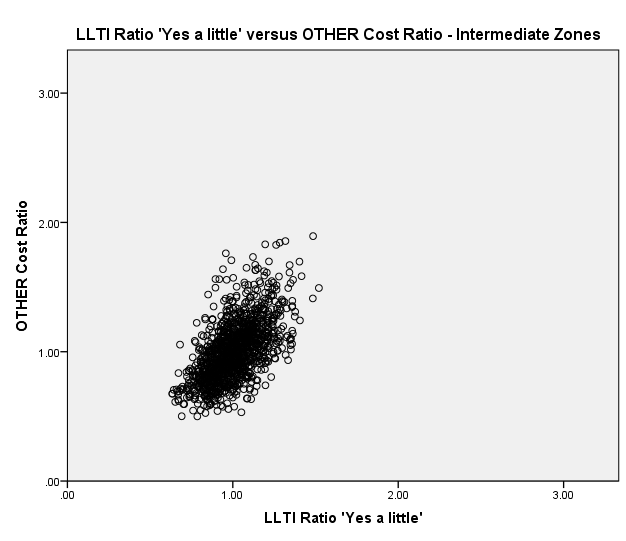
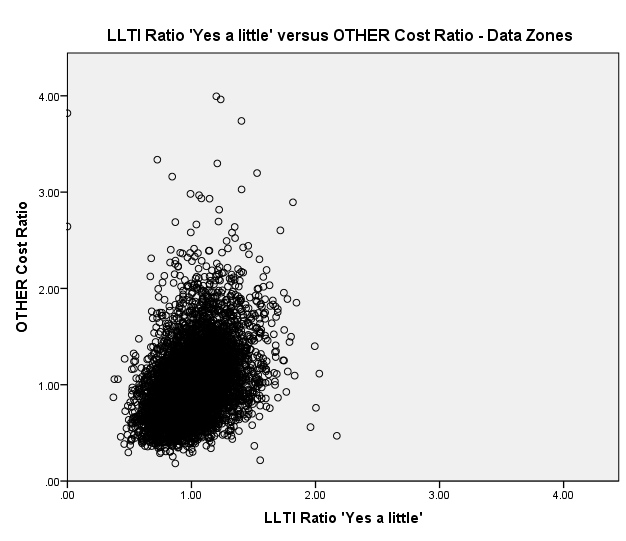
Scatter plots of cost ratios against LLTI ratios are produced for all diagnostic groups at Data Zone and Intermediate Zone level. All scatter plots suggest the same overall impression that there is a positive linear relationship between LLTI ratios and cost ratios. The strength of the relationship is higher at Intermediate Zones (as already suggested by the correlations in Table B.2) and the standard deviation is lower (Table B.3).

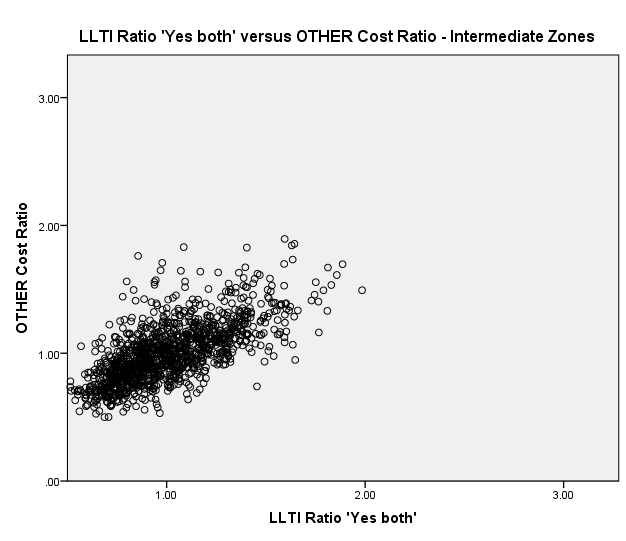
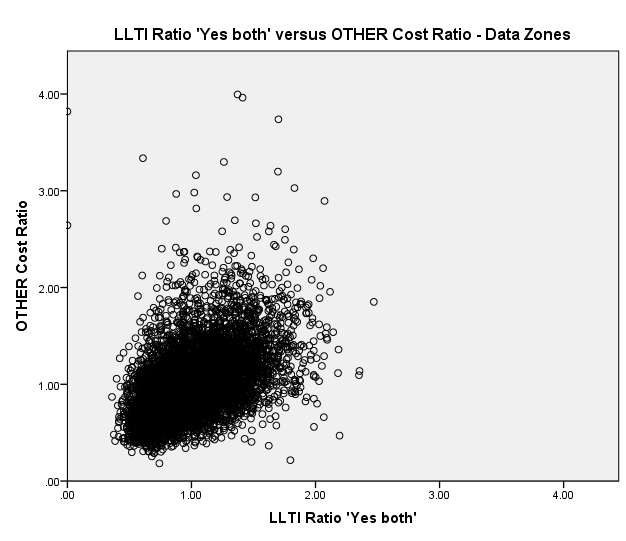
‘Other’ is the diagnostic group chosen as an example and scatter plots of cost ratios against LLTI ratios are included below (Figure B.1) at both geography levels.

*Table B.3. Standard Deviations of LLTI variables.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Standard Deviations** | LLTI Yes a lot | LLTI Yes a little | LLTI Yes both |
| Data zones | 0.47 | 0.21 | 0.31 |
| Intermediate zones | 0.39 | 0.14 | 0.25 |



****

****

*Figure B.1. Scatter plots of Other Cost Ratios against LLTI variables*

As can be seen from the scatter plots and standard deviations, the range of values of the LLTI ratio for ‘Yes a little’ is narrow compared to the other definitions. LLTI ratio ‘Yes a lot’ values are a bit more scattered than the ‘Yes both’ values which suggests ‘Yes both’ may perform better in the regression analysis. Based on the correlations, the scatter plots and the standard deviations, LLTI ratio ‘Yes a little’ is not further investigated. The formal analysis is carried only for the other two LLTI variables.

**B.3. Adjusted R-squared values**

Regression analysis is carried out using only LLTI ratio as a predictor of need at Data Zones and Intermediate Zones as well as using LLTI ratio in addition to the supply model (supply variables IPACX and OPACX and health board dummy variables). Supply variables and health board dummy variables will be included in the final model since they capture the board to board variation and improve the model fit, as can be seen from the adjusted R-squared values (Table B.4).

*Table B.4. Adjusted R-squared values of models comparing the LLTI variables’ performance.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **adjusted**  **R-squared** | Data Zones  (LLTI only) | | Intermediate Zones (LLTI only) | | Data Zones  (supply model + LLTI) | | Intermediate Zones  (supply model + LLTI) | |
| Yes a lot | Yes both | Yes a lot | Yes both | Yes a lot | Yes both | Yes a lot | Yes both |
| Cancer | 2.2% | 2.3% | 7.9% | 7.7% | 5.8% | 6.0% | 23.8% | 24.2% |
| Heart | 7.6% | 8.1% | 23.7% | 25.1% | 8.4% | 8.9% | 28.5% | 29.7% |
| Digestive | 17.2% | 16.2% | 41.0% | 38.7% | 21.5% | 21.5% | 52.1% | 52.3% |
| Injury | 7.9% | 8.6% | 19.0% | 21.0% | 10.5% | 11.0% | 29.0% | 30.5% |
| Other | 23.9% | 24.4% | 48.9% | 49.6% | 25.3% | 21.3% | 54.7% | 56.5% |
| Respiratory | 20.7% | 19.8% | 52.1% | 49.8% | 21.8% | 21.3% | 55.5% | 54.7% |
| Outpatients | 14.7% | 12.9% | 24.8% | 21.7% | 43.6% | 43.7% | 68.0% | 68.2% |

Both LLTI variables are good predictors of need and perform approximately equally well in the models. When looking at the models testing LLTI variables in addition to the supply model, only Other and Respiratory cost ratios are better explained by the LLTI ratio ‘Yes a lot’ at Data Zones and only Respiratory cost ratios at Intermediate Zones. For all the other diagnostic groups ‘Yes a lot’ and ‘Yes both’ perform equally well or ‘Yes both’ is a stronger predictor of need.

Considering the initial impression and the results from the regression analysis, ‘Yes both’ is chosen as the LLTI variable to be used in the reference model.

**Annex C: Rule-of-thumb tests for normality of residuals**

Deviations from normality were noticed in the residuals while performing the regression analysis. Tests of skewness and kurtosis are carried out to evaluate this more formally and the results are given below.

Skewness is a measure of the extent to which a distribution of values deviates from symmetry around the mean. A value of zero means the distribution is symmetric, while a positive skewness indicates a greater number of smaller values, and vice versa. If the skewness is greater than 1 (or less than -1), the distribution is far from symmetrical (as a rule of thumb).

Kurtosis is a measure of the ‘peakedness’ or ‘flatness’ of a distribution. A kurtosis value near zero indicates a shape close to normal. A negative value indicates a distribution which is more peaked than normal, and vice versa. A kurtosis value of ±1 is considered very good, but ±2 is also usually acceptable (as a rule of thumb).

The reference model is fitted to all diagnostic groups at both geography levels and for 1 and 3 years at Data Zones. Skewness and kurtosis of the residuals are then calculated for all health boards together. The results are shown in Table C.1 below for all diagnostic groups. The shaded cells indicate when normality of the residuals is not violated by either of the two heuristic measures.

*Table C.1. Skewness and Kurtosis of the reference model residuals.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Normality check of residuals Reference model** | | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | Skewness | 1.98 | 1.74 | 2.38 | 1.67 | 1.47 | 4.44 | 1.00 |
| Kurtosis | 11.89 | 6.55 | 15.04 | 5.94 | 7.06 | 73.39 | 6.32 |
| 3 years Data Zones | Skewness | 1.20 | 1.07 | 1.28 | 1.08 | 0.80 | 1.84 | 0.58 |
| Kurtosis | 3.73 | 4.10 | 7.13 | 4.81 | 4.94 | 13.09 | 3.00 |
| 1 year Intermediate  Zones | Skewness | 0.57 | 0.73 | 0.98 | 0.70 | 0.67 | 1.21 | 0.06 |
| Kurtosis | 0.67 | 0.77 | 3.47 | 0.59 | 1.65 | 6.14 | -0.42 |
|  | | | | | | | | |

From the table above, skewness and kurtosis of the residuals are much smaller at Intermediate Zone level than at Data Zone level. The reference model performs better over 3 years at Data Zones than in 1 year at Data Zones, but the model residuals are still not normally distributed.

If the rule of thumb is used for skewness, then the distribution of the residuals is symmetric only for Outpatients at 1 year Data Zones, for Other and Outpatients at 3 years Data Zones and for Cancer, Heart, Digestive, Injury, Other and Outpatients at 1 year Intermediate Zones.

If the rule of thumb is used for kurtosis, then the shape of the distribution of the residuals is not normal for any diagnostic group at 1 year and 3 years Data Zones and for only Digestive and Respiratory at Intermediate Zone level.

Overall, normality of the residuals is least violated at Intermediate Zones and for five of the diagnostic group the assumption is valid.

For the 1 year analysis at Data Zones, skewness and kurtosis were calculated for each Health board separately to check if the deviation from normality was due to only one or two boards. However, the skewness and kurtosis for each health board and diagnostic group were far from normal and at least five health boards have quite extreme skewness or kurtosis.

**Annex D: Outliers and Influential points**

The presence of outliers and influential points is investigated in order to differentiate between the geography (Data Zones versus Intermediate Zones) and the time span (1 year versus 3 years). The cost ratio outliers and the model residuals outliers are investigated and the results are presented below.

**D.1 Cost Ratio Outliers**

The cost ratio outliers are investigated in order to get a better understanding of the cost ratios’ variability. A cost ratio is defined as an outlier if it is bigger than a particular value (75th percentile + 1.5 \* Interquartile Range). The numbers of cost ratio outliers for all diagnostic groups are included in Table D.1 below.

*Table D.1. Number and Percentage of cost ratio outliers*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Number and Percentage**  **of cost ratio outliers** | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | **203**  3.12% | **245**  3.77% | **198**  3.04% | **229**  3.52% | **164**  2.52% | **253**  3.89% | **103**  1.58% |
| 3 years Data Zones | **155**  2.38% | **167**  2.57% | **164**  2.52% | **164**  2.52% | **104**  1.60% | **227**  3.49% | **47**  0.72% |
| 1 year Intermediate Zones | **19**  1.54% | **26**  2.11% | **18**  1.46% | **19**  1.54% | **20**  1.62% | **44**  3.56% | **8**  0.65% |

There are about twice as many cost ratio outliers at Data Zones as at Intermediate Zones in terms of percentages. Overall, there are not so many outliers at both geography levels but there is a slight improvement at Intermediate Zones.

**D.2 Residuals Outliers and Influential points**

Outliers are observations which do not fit the model well. To formally identify outliers, the externally studentised residuals are used. All externally studentised residuals outside the confidence interval (- tn-p-1, 1-0.05/(2\*n), + tn-p-1, 1-0.05/(2\*n) ) are classified as outliers, where p is the number of parameters, n is the number of observations.

Influential points are observations which change the fit of the model in a substantive way. To formally identify influential points, Cook’s distance and DFFITS are used. Cook’s distance demonstrates the influence of case i on all the fitted values. Observations with Cook’s distance greater than F0.5, p. n-p are classified as influential points. DFFITS demonstrates the influence that case i has on the fitted value i. Observations with DFFITS less than -2\*sqrt(p/n) or greater than 2\*sqrt(p/n) are classified as influential points.

Regression on health board dummy variables, supply variables and acute needs index (reference model) is performed for each diagnostic group separately at 1 year Data Zones, 3 years Data Zones and 1 year Intermediate Zones. There is one influential point for Respiratory at 1 year Data Zones. However, this particular Data Zone cost ratio is not influencing all the fitted values but just its own fitted value. There are no influential points at 3 years Data Zones and 1 year Intermediate Zones. The numbers of residuals outliers are included in table D.2 below.

*Table D.2. Number and Percentage of residuals outliers.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Number and Percentage**  **of residuals outliers** | Cancer | Heart | Digestive | Injury | Other | Respiratory | Outpatients |
| 1 year Data Zones | **20**  0.31% | **26**  0.40% | **30**  0.46% | **27**  0.42% | **21**  0.32% | **26**  0.40% | **8**  0.12% |
| 3 years Data Zones | **14**  0.22% | **18**  0.28% | **17**  0.26% | **17**  0.26% | **12**  0.18% | **28**  0.43% | **8**  0.12% |
| 1 year Intermediate Zones | **2**  0.16% | **2**  0.16% | **5**  0.40% | **3**  0.24% | **1**  0.08% | **4**  0.32% | **2**  0.16% |

**D.3 Summary**

The cost ratios, the dependent variable in the regression analysis, are not normally distributed and their distribution is close to normal just at Intermediate Zones. There are twice as many cost ratio outliers at 1 year Data Zones as at 1 year Intermediate Zones.

The reference model residual outliers are also not normally distributed and their distribution is close to normal just at Intermediate Zones. There is 1 influential point at 1 year Data Zones (influencing its own fitted value). There are twice as many reference model residual outliers at 1 year Data Zones as at 1 year Intermediate Zones.

**Annex E: SIMD minus Health**

The Scottish Government SIMD team provided the analytical team with “SIMD minus Health” scores. The newly calculated scores include Income, Employment, Education, Access, Crime and Housing domains. Two different calculation methods were used and the “SIMD minus Health” scores are referred to as Scenario 1 and Scenario 2.

Scenario 1 uses proportionally increased weights for the other six domains (when excluding the Health domain which accounts for 14% of the overall SIMD score). Scenario 2 fixes the weights of the largest domains, Employment and Income, to their usual value (28% each) and then increases proportionally the weights of the other four domains. Employment and Income make up 65% of the overall index in Scenario 1 and 56% in Scenario 2. All domain weights for each scenario are given in Table E.1.

*Table E.1. SIMD domain weightings.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SIMD minus Health Weightings** | | | | | |
| **Domain** | **Scenario 1** | **Scenario 2** | | **2012 SIMD Ranks** | |
|  | **Weighting** | **Weighting** | | **Weighting** | |
| Income | 0.325 | 0.28 | | 0.28 | |
| Employment | 0.325 | 0.28 | | 0.28 | |
| Health | 0 | 0 | | 0.14 | |
| Education | 0.16 | 0.19 | | 0.14 | |
| Access To Services | 0.11 | 0.14 | | 0.09 | |
| Crime | 0.05 | 0.07 | | 0.05 | |
| Housing | 0.03 | 0.04 | | 0.02 | |
| Total | 1 | 1 | | 1 | |
|  |  |  |  | |  |  |  |

**Annex F: Correlations between SIMD variables and Acute MLC cost ratios**

Tables F.1 and F.2 show the correlations between SIMD variables and Data Zone cost ratios, at 1 year and at 3 years, respectively. Table F.3 shows the correlations between the newly created Intermediate Zone explanatory variables from the Income domain score and the cost ratios.

*Table F.1. SIMD scores correlations to the cost ratios at 1 year Data Zones.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations**  **1 year Data Zones** | **Overall SIMD** | **Income** | **Employment** | **Education** | **Housing** | **Crime** | **Scenario1** |
| Cancer cost ratios | 0.144 | 0.147 | 0.146 | 0.129 | 0.142 | 0.059 | 0.141 |
| Heart cost ratios | 0.288 | 0.292 | 0.279 | 0.274 | 0.207 | 0.156 | 0.286 |
| Digestive cost ratios | 0.413 | 0.409 | 0.401 | 0.363 | 0.266 | 0.170 | 0.402 |
| Injury cost ratios | 0.310 | 0.308 | 0.301 | 0.276 | 0.204 | 0.181 | 0.302 |
| Other cost ratios | 0.475 | 0.480 | 0.469 | 0.444 | 0.292 | 0.212 | 0.464 |
| Respiratory cost ratios | 0.458 | 0.460 | 0.442 | 0.415 | 0.321 | 0.202 | 0.450 |
| Outpatients cost ratios | 0.332 | 0.340 | 0.337 | 0.270 | 0.364 | 0.106 | 0.324 |

*Table F.2. SIMD scores correlations to the cost ratios at 3 years Data Zones.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations**  **3 years Data Zones** | **Overall SIMD** | **Income** | **Employment** | **Education** | **Housing** | **Crime** | **Scenario1** |
| Cancer cost ratios | 0.220 | 0.221 | 0.222 | 0.193 | 0.203 | 0.094 | 0.213 |
| Heart cost ratios | 0.422 | 0.428 | 0.408 | 0.409 | 0.304 | 0.227 | 0.416 |
| Digestive cost ratios | 0.548 | 0.541 | 0.537 | 0.493 | 0.348 | 0.231 | 0.535 |
| Injury cost ratios | 0.480 | 0.473 | 0.467 | 0.428 | 0.313 | 0.267 | 0.469 |
| Other cost ratios | 0.615 | 0.621 | 0.606 | 0.567 | 0.385 | 0.269 | 0.599 |
| Respiratory cost ratios | 0.593 | 0.595 | 0.574 | 0.539 | 0.434 | 0.272 | 0.582 |
| Outpatients cost ratios | 0.375 | 0.387 | 0.382 | 0.312 | 0.396 | 0.122 | 0.366 |

*Table F.3. SIMD Income domain local and national shares’ correlations to the cost ratios at 1 year Intermediate Zones.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations**  **1 year Intermediate Zones** | **5% LOCAL SHARE** | **5% NATIONAL SHARE** | **25% LOCAL SHARE** | **25% NATIONAL SHARE** |
| Cancer cost ratios | 0.166 | 0.165 | 0.260 | 0.256 |
| Heart cost ratios | 0.379 | 0.368 | 0.450 | 0.427 |
| Digestive cost ratios | 0.434 | 0.435 | 0.554 | 0.542 |
| Injury cost ratios | 0.330 | 0.326 | 0.406 | 0.399 |
| Other cost ratios | 0.417 | 0.415 | 0.580 | 0.563 |
| Respiratory cost ratios | 0.505 | 0.499 | 0.620 | 0.599 |
| Outpatients cost ratios | 0.347 | 0.347 | 0.424 | 0.411 |

1. Delivering Fair Shares for Health in Scotland: The report of the NHSScotland Resource Allocation Committee, September 2007 [↑](#footnote-ref-1)
2. NRAC formula - MLC adjustment: Mental Health and Learning Difficulties Care Programme - Final Report, TAGRA MLC (Morbidity & Life Circumstances) Sub-Group, December 2012 [↑](#footnote-ref-2)
3. There are also two approaches to calculating a share: one is to use the percentage of data zones, i.e. 5% of the most 'deprived' Data Zones; the other is to use the percentage of population. The shares calculated using these two approaches are extremely similar (correlations close to 1), and the results in the analysis were very similar. Therefore, only the results for the Data Zone-based local and national share variables are shown. [↑](#footnote-ref-3)