

SCOTTISH INDEX OF MULTIPLE DEPRIVATION (SIMD) 2006

Quality Assurance of Statistical Programs Report of Findings

Alex McConnachie
Consultant Statistician
Robertson Centre for Biostatistics
University of Glasgow

23rd August 2006

Introduction

This report outlines the findings of an assessment of the quality of the statistical programs to be used for the calculation of the SIMD 2006. It is based on observations made whilst reviewing the programs being applied to source data and test data on 25th July 2006 at Meridian Court, Glasgow, and on review of code sent subsequently by email, without data.

Summary of Findings

All statistical programs were considered to be fit for purpose, in that they carried out the data manipulations required for the calculation of the SIMD 2006 and all its constituents.

There were some observations regarding the general programming style or specific elements of the code that could be streamlined. There were also some observations regarding the general SIMD methodology, which are beyond the remit of this assessment, but have been included for completeness. All of these observations are listed below.

Programming Observations

1. The code is highly fragmented, with several programs used to generate each domain and the SIMD. Ideally, a single program would be written. Comments at the head of the program would detail the location and content of all source data files. In practice, due to data confidentiality, certain elements of the calculations must be carried out at different sites; in such cases, the master program should document what data has been generated externally, as well as when, where and by which program(s) this was done. Such a unified approach would make the process of transferring the responsibility for the programs more straightforward.
2. The code used in some domains for carrying out Factor Analysis (FA) of ranked and Normalised indicator variables is not self-contained, in that it requires the FA to be carried out in order to extract the indicator weights to use to create the domain score. This is a potential source of error, and it would be better to write code to extract the indicator weights from the FA output file. In the code for the Access domain, sent by email after the visit to Meridian Court on 25th July, this suggestion had been adopted.

3. The code used in some domains for age-sex standardisation could be streamlined. Currently, expected numbers of events are calculated by multiplying the number of individuals in each age-sex group by the expected rates in a very long expression; there is the potential for error should this code need to be rewritten in future. A more efficient method would require the source data to be in the form of one record for each (data zone \times age group \times sex) combination; expected rates could then be calculated for each (age group \times sex) combination and applied to the original data to obtain expected numbers of events. Such an approach would reduce the number of input variables considerably and therefore the potential for transcription errors.
4. The naming of variables is occasionally confusing. For example, in constructing the SIMD, the rank of the employment domain score is denoted "emprank" whilst that of the crime domain is "rankcrime", and for the housing domain it is "rhdom". Within a unified program, or set of programs, it would be better if these were standardised.

Methodological Observations

5. The low birth weight indicator has a very low weight in the health domain, as determined by FA. If this indicator is not contributing much to the final domain score, it is likely that removing it will have little impact. Also, if this indicator is in truth not related to the other indicators, or the resulting factor score, then in the future it could happen that this, or other indicators, may have a negative weighting towards the domain score. What procedures are in place to review the weights given to each indicator at future iterations of the SIMD?
6. The CIF and SMR indicators are age-sex standardised morbidity and mortality indicators based on events over the whole age range. In areas with relatively high mortality rates, those individuals who survive into old age are the most robust members of the population, and it is therefore not inconceivable that they are less likely to be ill or to die than individuals of the same age in less deprived areas. If this is the case, a better indication of the burden of morbidity or mortality in a given area might be obtained by restricting the CIF and SMR indicators to younger age groups, such as those under the age of 70 years.