Summary of Synthetic Estimates Project

Background to the Project

1. The Department of Health/HSCIC commissioned the National Centre for Social Research (Natcen) to produce estimates of healthy lifestyle behaviours, using information from the Health Survey for England. The aim of the project was to develop small area estimates for publication on the Neighbourhood Statistics website and to provide key public health information currently not available from any other source. There is a very strong user requirement for local data on lifestyle behaviours - Some of the key PSA targets relate to healthy lifestyle behaviours such as obesity and smoking and the estimates will help to examine those targets.

2. A key requirement was to provide robust estimates generated on a consistent basis for all areas of the country, which allow meaningful comparisons within and between local areas.

3. Information on healthy lifestyle behaviours is available form the Health Survey for England, but such a survey has not been designed for estimation at the small area level. This is because estimates can only be computed for a subset of all wards (i.e. those containing respondents to the survey) and the sample size will be small for these areas and be low in precision.

4. Synthetic estimation can be defined as the application of model-based techniques to combine data obtained from national surveys (containing the health behaviour measures of interest) with a set of associated covariate (or predictor) variables available for all small areas (e.g. the proportions of residents who were living as a couple, claiming Income Support, had a limiting longstanding illness etc). The synthetic estimate generated for a particular small area is the expected outcome for that area based on its characteristics as measured by the covariate variables. To interpret the estimates it is recommended that users adopt statements such as: given the characteristics of the local population we would expect approximately x% of adults within ward X to smoke/be obese etc.

Stages of the Project

5. The project started on 1 April 2004 and progressed in three stages:

   - **Stage 1 – Scoping.**
     This involved a literature review of the methods available (3 potential methods were identified); availability of external data; suitability of statistical software;
setting up the database; examination of the feasibility of producing synthetic estimates and Confidence Intervals.

- **Stage 2 – Testing**
  Provisional estimates and Confidence Intervals were created using the different methods. Internal and external validation was carried out and potential outcome measures or variables were identified (5 in all).

- **Stage 3 – Implementation**
  In this stage, estimates were prepared at the ward and Primary Care Organisation level for the 5 outcome variables. Confidence Intervals were calculated for the variables and the estimates were compared with other sources of information.

6. Three groups were set up to progress and provide advice on the synthetic estimates project – a Steering Group to oversee, guide and approve the work; a Technical group to provide advice on the methodology and testing of the estimates and to ensure that quality issues were addressed; A User Group to provide an indication of user needs. The groups included key policy and statistical stakeholders within the Department; representatives from the Public Health Observatories providing both a user input and technical expertise; methodology experts from Natcen, ONS and University academics.

7. The Steering and Technical Group met regularly throughout the project with correspondence received from the User Group. The groups played an active role in the project and quality assured the data sources, methodology, estimates and their uses.

**Findings from Stage 1 of the Project**

8. This part of the paper gives a brief overview of Stage one of the project. Further details are given in the *Stage 1 Technical Report*.

**Literature Search**

9. A literature search was performed and identified different methods which varied in complexity as follows:
   - Simple methods using indirect standardisation;
   - Models using individual level covariates \(^1\) only;
   - Models combining individual and area-level covariates;
   - Models using area level covariates only;
   - Other approaches for larger areas of geography.

For small areas such as wards, only the first four sets of methods were deemed to be appropriate.

---

\(^1\) The term ‘covariate’ describes those area-level characteristics (e.g. deprivation scores, life expectancy, rural indicator, Government Office Region) that were potentially related to outcomes such as smoking, obesity, and binge drinking. These covariates were generally average values or proportions relating to all individuals or households in the area. Because of its universal geographical and population coverage, the 2001 Census provided the main source for the demographic and social covariate data.
Confidence Intervals, Datasets and Testing the software

10. The synthetic estimate generated for a particular ward is the expected measure for that ward based on its characteristics, but is likely to have errors associated with it. By placing confidence intervals\(^2\) around the synthetic estimate, a range can be calculated in which we could be fairly sure that the true value for that area lies. Two methods for generating confidence intervals for the synthetic estimates were investigated – one was based on parameter estimates and the other on Markov Chain Monte Carlo Methods.

11. Two types of software were examined for the ability to fit multilevel logistic regression models.

12. A number of datasets were required for the project and included– a survey dataset (pooled data for three years of the Health Survey for England (HSfE) from 2000 to 2002); and a set of area-level covariates (from a number of sources - including the 2001 Census, administrative data and the Index of Multiple Deprivation).

Models to Test

13. Following the literature review and test on software and Confidence Intervals, the strength and limitations of each methods were evaluated and three methods were suggested to be tested in Stage 2 (see the Table 1 below). The method that used area level covariates only, was known as the ONS method and had been previously used in the calculation of small-area estimates of household income. The method based on area level and individual level covariates was known as the Twigg method (as it was used to obtain prevalence estimates of smoking and ‘problem drinking’ for each ward.

Table 1 : Recommended methods to take forward to Stage 2

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Measurement level of health behaviour</th>
<th>Measurement level of covariates/ predictors</th>
<th>Implementation dataset (see Chapter 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect standardisation</td>
<td>Individual</td>
<td>Individual and area-level</td>
<td>Census cross-tabulations</td>
</tr>
<tr>
<td>Models using individual and area level covariates</td>
<td>Individual</td>
<td>Individual</td>
<td>Census proportions and other administration data</td>
</tr>
<tr>
<td>Models using area level covariates only</td>
<td>Individual</td>
<td>Area-level</td>
<td></td>
</tr>
</tbody>
</table>

\(^2\) Confidence Intervals have been prepared for each synthetic estimate. The interval reflects the range between which the true value is believed to lie, at a given level of confidence. The confidence intervals therefore represent the uncertainty in the modelling process. At the 95% confidence level, assuming that the model is a good representation of reality, the confidence interval is expected to contain the true value around 95 times out of 100. For example, if a ward estimate of current smoking is 49% and the 95% confidence interval is [32%,67%] we know that 95% of the time the true prevalence estimate for that ward (based on its local population characteristics) will fall within this range.
<table>
<thead>
<tr>
<th>Estimates for demographic sub-groups within the small area?</th>
<th>Yes</th>
<th>Yes</th>
<th>Possible development in the future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published formula for constructing valid confidence intervals?</td>
<td>Yes</td>
<td>To be developed by the project team</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Choice of Estimation Area and Variables**

14. Synthetic estimates were required at the smallest possible geographical level and a range of geographical areas were considered. Previous research shown that robust estimates from the Health Survey for England (HSfE) could be calculated for at the Strategic Health Authority and Regional Office level.

15. The calculation of synthetic estimates at the Super Output Area level was considered but discounted due to constraints imposed by the design of the HSfE. Ward and Primary Care Organisation (PCO)³ level estimates were suggested.

16. A number of health measures were suggested as outcome measure but a set of criteria were used to decide which measures would be tested in Stage 2 of the project. The criteria included:

   - the variable should demonstrate a fairly large between area variance;
   - the variable should be known to be correlated with the variables that will be used in the models;
   - the variable should provide a geographic spread of areas.

**Recommendations from Stage 1 of the Project**

17. The recommendations from Stage 1 of the project were:

**Models to test:** 3 models would be tested in Stage 2 of the project. These were as follows:-

- the model with covariates at the area level only (ONS method)
- the model with covariates measured at both the area and individual level (TWIGG method).
- the simple model of indirect standardisation.

**Selection of outcomes:** 3 variables would be tested at stage 2. They include:

- prevalence of obesity among adults.
- prevalence of current cigarette smoking among adults

---

³ Primary Care Organisation is the umbrella term used to describe the 300 Primary Care Trusts and 3 Care Trusts in England.
-proportion of children aged 5-15 consuming 5 portions of fruit and vegetables per day.
- examination of the feasibility to produce prevalence estimates for diabetes

Findings from Stage 2 of the Project

18. This part of the paper gives a brief overview of Stage 2 of the project. Further details are given in the Stage 2 Technical Report

19. In Stage 2 of the project, the three methods that had been identified in Stage 1 were fully evaluated. To do this, synthetic estimates (using the different methods) were compiled for three outcome measure (of prevalence of smoking, obesity and proportion of children eating three or more portions of fruit and vegetables per day.

20. The synthetic estimates were compared against survey data from a number of different sources.

Comparison of the Methods

21. Direct estimates from the HSfE were compared with synthetic estimates produced using the various methods. The indirect standardisation methods were found to be poor modelling methods. Naticn examined the TWIGG and ONS methods and found that when using statistical criteria, there was very little difference between them (but Table 2 shows a comparison of non-statistical criteria):

- The proportion of area level variance explained by the two methods was similar;
- Correlations with direct level estimates from the HSfE were similar;
- The methods were very highly correlated with each other;
- Diagnostic tests showed no evidence of model mis-specification.

Table 2: Comparison of the ONS and TWIGG methods:

<table>
<thead>
<tr>
<th>ONS Method (advantages)</th>
<th>TWIGG method (advantages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This method is easier and quicker to implement (and produce CIs), than the TWIGG method. It is also slightly easier to explain to users.</td>
<td>The method is easier to ‘sell’ to users as it seems more credible to aggregate individual level estimates rather than modelling directly at area level.</td>
</tr>
<tr>
<td>The method has been used by ONS to produce experimental statistics on average ward level income.</td>
<td>TWIGG method uses a bottom-up approach by generating an estimate for every age and sex group (but depends upon variables in the Census). The ONS method excludes individual level covariates.</td>
</tr>
<tr>
<td>ONS method is more “transferable” to other output variables.</td>
<td></td>
</tr>
<tr>
<td>ONS method can more easily be updated over time as information comes from the HSfE.</td>
<td></td>
</tr>
</tbody>
</table>
Measures suitable for synthetic estimation

22. A set of criteria were developed to see which outcome measures were appropriate for synthetic estimates. A range of outcome measures such as prevalence of high blood pressure, disability, CVD and amount of activity were considered. The Stage 2 Report gives further details. The criteria for judging suitability were:

- Whether there was a strong policy interest;
- A prevalence of greater than 10% of the outcome variable in the HSfE;
- Evidence of area-level variation;
- Strong correlation with independent estimates.

Geographical level

23. For Stage 2 of the project, ward-level estimates of smoking prevalence; obesity; and fruit and vegetable consumption (for children) had been produced. These were validated against other sources of data

Recommendations from Stage 2 of the Project

24. The recommendations from Stage 2 of the project were:

- The ONS method be used to produce synthetic estimates
- That estimates be produced at both the ward and PCO level
- That synthetic estimates be produced for the following variables:
  - Prevalence of smoking.
  - Binge Drinking.
  - Prevalence of obesity in adults.
  - Consumption of 3 or more portions of fruit and vegetables (children).
  - Consumption of 5 or more portions of fruit and vegetables (adults).

Findings from Stage 3 of the Project

25. This part of the paper gives a brief overview of Stage 3 of the project. Further details are given in the Stage 3 Technical Report.

26. During Stage 3 of the project, the synthetic estimates with associated Confidence Intervals were produced for the outcome variables outlined in Stage 2 at both the ward and PCO levels.

Validation of the estimates – internal and external validation

27. To assess the quality of the synthetic estimates model, the proportion of the variance explained by the model was examined. In addition, a number of diagnostic tests were done to test whether there were any relationships between the healthy lifestyle measures and the ward characteristics. The results of this analysis indicated confidence that significant relationships between the measures and ward characteristics had not been omitted from the model.
28. The synthetic estimates were also compared with direct estimates for the Health Survey for England from 2000-2002 with large correlations (as expected) at the Strategic Health Authority level.

29. Comparisons were made with a range of local surveys - e.g. the Wigan, Bolton and Bury Health Survey as well as the Health Survey for England 2003, the National Patient Experience Survey for 2003 and the Health Development Agency smoking estimates. The comparisons were encouraging and gave confidence in the synthetic estimates at the ward and PCO area levels.

**Quality of the estimates**

30. Confidence Intervals were prepared to assess the precision (or margin of error) of the synthetic estimates and these were examined at the different geographical levels. The interval reflects the range within which the true value is believed to lie, at a given level of confidence. Table 3 from the Stage 3 report gives the average width of the Confidence Interval by variable and geographical level.

**Table 3: Average width of the 95% C.I.s for the 5 healthy lifestyle behaviours (taken from Natcen Stage 3 report)**

<table>
<thead>
<tr>
<th>Health Behaviour</th>
<th>Average width of the 95% Confidence Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wards</td>
</tr>
<tr>
<td>Current smoking</td>
<td>± 11%</td>
</tr>
<tr>
<td>Obesity</td>
<td>± 7%</td>
</tr>
<tr>
<td>Binge Drinking</td>
<td>± 12%</td>
</tr>
<tr>
<td>Fruit and vegetable consumption in adults (≥ 5 a day)</td>
<td>± 13%</td>
</tr>
<tr>
<td>Fruit and vegetable consumption in children (≥ 3 a day)</td>
<td>± 19%</td>
</tr>
</tbody>
</table>

31. Table 3 shows that

- the Confidence Intervals are wider for the variable of fruit and vegetable consumption in children than the other variables.
- The CIs are much wider for wards than for PCOs. This suggests that the use of the estimates is limited and this is highlighted further, by examining the overlap of Confidence Intervals.

32. Figure 1 gives the estimates of smoking prevalence for Primary Care organisations and gives an indication of the range of the values. A similar chart (Figure 2) shows the distribution of the smoking prevalence estimates at the ward level. The estimate itself is shown as a solid line with the upper and lower confidence intervals around it. These charts are available for download.
Figure 1: Model-Based Estimates and CIs for Population Smoking Prevalence in England, by PCO

Note: The chart displays synthetic estimates for smoking prevalence (adults). The closely positioned bars on the graph represent values for smoking at the PCO level. The grey lines running parallel with the bars are 95% Confidence Intervals surrounding each PCO value.
Conclusions

33. Synthetic estimates have been prepared for 5 variables and for 2 levels of geography and are available with associated metadata on the Neighbourhood Statistics website. The following points should be noted:

- The synthetic estimates are not estimated counts of the number of people or prevalence of a behaviour, e.g. smoking in a ward or PCO. They are estimates based on a model and represent the expected prevalence of a behaviour for any ward or PCO, given the demographic and social characteristics of that area. To interpret the estimates it is recommended that users adopt statements such as "given the characteristics of the local population we would expect a prevalence of approximately x% within PCO X".

- The synthetic estimates should be considered with their associated Confidence Intervals when making comparisons of one area with another. Although the estimates have large Confidence Intervals (many of which overlap), it is possible to distinguish between those estimates that are significantly different from the national average and those that are not (at both ward and PCO level). The charts and User guide show those areas that have estimates significantly above, below or similar to the National estimate.

- The estimates can be used to give an indication of the likely prevalence in an area and an indication of where to target resources. They can be used to examine prevalences in the most disadvantaged areas (e.g. the Spearhead PCOs) as compared with other groups. The estimates should however, not be used for performance monitoring purposes.