

# Synthetic estimation of healthy lifestyles indicators: Stage 1 report

Madhavi Bajekal, Shaun Scholes, Kevin Pickering, Susan Purdon

# Synthetic estimation of healthy lifestyles indicators: Stage 1 report

Madhavi Bajekal, Shaun Scholes, Kevin Pickering, Susan Purdon

Prepared for the Department of Health

May 2004

# Contents

<b>Summary</b> .....	<b>3</b>
<b>1 INTRODUCTION</b> .....	<b>6</b>
1.1 Overall aims of the project.....	6
1.2 This report.....	6
<b>2 REVIEW OF LITERATURE</b> .....	<b>8</b>
2.1 What is synthetic estimation and why it is needed.....	8
2.2 Typology of methods .....	8
2.3 Simple methods.....	9
2.3.1 Indirect standardisation.....	9
2.3.2 Strengths and limitations.....	9
2.4 Models using individual level covariates only.....	10
2.4.1 Using covariates from the Census .....	10
2.4.2 Using covariates from SARs .....	11
2.4.3 Strengths and limitations.....	11
2.5 Models combining individual and area level covariates .....	11
2.5.1 Strengths and limitations.....	12
2.6 Models using area level covariates only.....	13
2.6.1 Strengths and limitations.....	14
2.7 Other approaches for larger geographical areas .....	15
2.7.1 GREG estimator .....	15
2.7.2 Composite estimators.....	16
2.7.3 Fay-Herriot estimator.....	16
2.8 Summary of methods at the ward level.....	17
<b>3 SCOPING AND SETTING UP THE DATABASE</b> .....	<b>19</b>
3.1 Data requirements .....	19
3.2 Geographical matching.....	20
3.3 Survey dataset .....	21
<i>Sample size and spatial coverage</i> .....	21
<i>Sample-based individual-level covariates</i> .....	22
3.4 Area-level covariate dataset .....	23
3.4.1 Ward level covariates .....	23
<i>Census data</i> .....	23
<i>Administrative data</i> .....	24
3.4.2 Higher-area level covariates.....	24
3.4.3 Covariates at other geographies .....	24
3.5 Validation dataset .....	25
<b>4 CALCULATION OF CONFIDENCE INTERVALS FOR SYNTHETIC ESTIMATES</b> .....	<b>26</b>
4.1.1 Confidence intervals derived from the model estimates.....	26
4.1.2 Confidence/credible intervals estimated by simulation.....	27
4.2 Indicative confidence intervals for wards .....	27
4.3 Calculation of confidence intervals for aggregated synthetic estimates	28
<b>5 TESTING THE SOFTWARE</b> .....	<b>29</b>

<b>6</b>	<b>RECOMMENDATIONS FOR STAGE 2 .....</b>	<b>30</b>
6.1	Choice of estimation area.....	30
6.2	Which models to test? .....	31
6.3	Which health measures? .....	31
6.4	Proposed approach to model validation .....	32
6.5	User engagement .....	33
	<b>REFERENCES .....</b>	<b>34</b>

## **Acknowledgements**

We would like to thank members of the Steering Group and the Technical Group for their helpful comments on this report. We are grateful to Prof Graham Moon and Dr Liz Twigg (University of Portsmouth) and Patrick Heady (ONS) for discussing their previous work on this topic with us. Colleagues at the Survey Methods Unit (NatCen) contributed generously of their time in developing and refining methodology through regular seminar discussions. We would particularly like to express our appreciation of the work of Department of Health staff at all stages of the project and in particular the contribution made by Rosemary Aldridge, Tracie Kilbey, David Greeno, Bill Hageman and Seb Morton-Clark.

## Summary

### *Aims and objectives*

The aim of the synthetic estimation project is to provide robust estimates of health characteristics and behaviours of populations for all small areas in England to support comparisons within and between local areas such as Local Authority Districts and Primary Care Organisations.

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.*

This report sets out the findings of the first stage of the project. Stage 1 was a scoping study to assess the methodological, technical and data requirements for synthetic estimation in order to inform the work-plan for the next stage (evaluation and testing) of the project. The main findings and recommendations of the report are summarised below.

### *Review of literature*

The literature review identified four types of methods used for small-area estimation: two (simpler) approaches that have used only individual (or person) level covariates and therefore adjust solely for the differences in the demographic composition between areas; and two approaches that have used multilevel modelling to simultaneously take into account the impact of differences in population composition and the wider social context between small areas. The latter approaches are conceptually more advanced than the simpler methods, but are also methodologically and computationally more complex.

### *Confidence intervals and testing the software*

The report sets out our approach to generate confidence intervals around estimates at lower and higher level geographies for the multi-level models. Indicative confidence intervals for ward-level estimates were also calculated. Two software packages appropriate for hierarchical modelling were tested to assess processing requirements and to develop algorithms to speed up the model building process.

### *Which models to test?*

The literature review and tests indicated that the two multi-level model methods - namely, the one using area-level covariates only (extensively tested by the ONS), and the other using covariates measured at both the area and the individual (or person) level - would produce reasonable synthetic estimates and were technically feasible to do. The recommendation of this report is that the two multi-level model methods will be tested in Stage 2 of the project. In addition, a 'simple' model-based estimate

would also be constructed to assess how close these estimates are to those generated using the more complex models.

#### *Setting up the database*

The database required for generating synthetic estimates involves combining together data from two sets of sources – survey data (from the Health Survey for England 2000 to 2002) containing the target health measures of interest (e.g. smoking status) and a covariate dataset with the individual and area-level covariates – matched to a common geography for which estimates are required (estimation area or lowest geographical unit for which estimates are to be generated).

There were a number of advantages to using survey data pooled over the three most recent rounds of the Health Survey for England (HSfE) for modelling. First, estimates would then be based on the most recent data available; second, the selected years would span the 2001 Census (the main source for the covariate data ) thereby minimising potential time-lag errors between different data sources; and last, pooling data over years increased the geographical coverage of areas covered by the sample which would potentially improve the fit of the model.

The availability of area-level covariates from Census and administrative sources was assessed and an initial core set acquired for inclusion in the database.

Because of time and resource constraints, it was decided to match datasets to a common geography using appropriate look-up tables rather than develop a true GIS database. GIS systems allow greater flexibility in locating data points across time, while look-up tables link data points to a geography fixed in time. The loss of flexibility in the database structure meant that re-combining estimates to new boundaries as they change over time will not be feasible.

#### *Choice of estimation area*

Census 2001 wards were selected as the smallest estimation area for the project. The main reason for choosing wards over lower area units (such as Super Output Areas) were statistical constraints imposed by the survey design. The HSfE sample is clustered within postcode sectors which have roughly the same population size as wards. Previous methodological research has shown that because of the similarity in size between the two, the variance between postcode sectors provides reasonably accurate variance estimates ( and hence confidence intervals) at ward level but not for smaller areas. Additionally, a wider range of covariates were available for wards compared with smaller area levels and previous work had shown that ward level synthetic estimates were fairly robust.

#### *Which health behaviour measures?*

Of the available and policy-relevant health measures included in the HSfE, three health measures were recommended for testing in Stage 2 of the project. These are the prevalence of cigarette smoking among adults, the prevalence of obesity among adults and the proportion of children aged 5-15 consuming five or more portions of fruit and vegetables a day. The selection of target health measures was based on a number of factors aimed to test the applicability of the methodology for different

types of outcomes, namely: outcomes for which good covariate data were available (smoking); health measures with a low measurement error (obesity); and estimates for a subgroup of the population (aged 5-15).

In preparation for the next stage of the project, proposals for model testing and validation have also been outlined in the report. Model validation will include both internal diagnostics of model fit and performance as well as 'plausibility' checks against external sources of data. The process of identifying suitable data sources for external validation has also begun and will continue over Stage 2 of the project.

A key element of the project is to involve users in all stages of the project, from identifying information needs and priorities in Stage 1 to taking an active role in local validation and dissemination. Our approach to user engagement in Stages 2 and 3 are outlined in Section 6.5 of the report.

## 1 INTRODUCTION

The National Centre for Social Research (NatCen) was commissioned by the Department of Health to produce estimates of healthy lifestyle behaviours using Health Survey for England (HSfE) data. This report describes the work undertaken for the first stage (the scoping study) of that project.

### 1.1 Overall aims of the project

The main aims of the project are:

- to evaluate the technical feasibility of producing robust small-area estimates for a range of health indicators;
- to validate model-based estimates against other sources of information and local knowledge;
- to develop and apply a consistent methodology to produce synthetic estimates for small-areas, focusing initially on a maximum of five indicators of healthy lifestyles available from the HSfE.

The key requirement of the synthetic estimation project is to provide robust estimates that are calculated on a consistent basis for all areas of the country and which allow meaningful comparisons within and between local areas.

There were three stages to the project – *scoping and feasibility, testing and validation, and implementation*. Given the experimental nature of synthetic estimation methodology at present, it was felt that a staged approach would allow consideration of the technical issues that arise, assessment of the quality of the outputs and ‘fit’ with user needs at each stage before proceeding to the next stage.

Our remit is to apply the methods that have been developed, not to develop new estimation techniques. Where possible, we shall build on previous work and incorporate new thinking as it emerges.

### 1.2 This report

This report documents the progress made at the end of the first stage of the synthetic estimation project. The first stage was primarily a scoping study to assess the methodological, technical and data requirements for the project in order to inform the project plan for the next (evaluation and testing) stage of the project.

The main objectives of the first stage scoping study were to :

- review the literature on synthetic estimation techniques including, where available, literature on the adequacy and reliability of estimates;
- scope data requirements and set up a geographically referenced data base containing the survey and predictor variables from census and administrative sources;

- acquire and test the required software (e.g. STATA and MLwiN);
- provide an initial assessment of the precision of estimates (i.e. confidence intervals) to help inform the viability of using the HSfE for synthetic estimation;
- recommend a shortlist of potential health behaviours and selected methods to test in Stage 2 of the project (technical evaluation).

The report is broadly organised around these objectives.

Chapter 6 of the report sets out the recommendations from this feasibility study to take forward to Stage 2 of the project.

## 2 REVIEW OF LITERATURE

### 2.1 What is synthetic estimation and why it is needed

For any small area containing respondents to a survey such as the Health Survey for England (HSfE) a conventional estimator of the prevalence of health-related behaviours, such as the proportion of adults who currently smoke, would be constructed from the survey data alone. Such conventional estimators suffer from two main limitations (Skinner, 1993). First, prevalence estimates can only be computed for a subset of all areas (i.e. those areas containing respondents to the survey). Second, for those sampled areas the achieved sample size will usually be small and the estimator will thus have low precision. This low precision will be reflected in rather wide confidence intervals for the survey estimates. Other techniques are therefore required.

Small area estimation can be defined as the application of model-based techniques to combine data obtained from national surveys (containing the health behaviour measure of interest) with a set of associated covariate (or predictor) variables at small area level – generally from the Census – to estimate the prevalence of healthy lifestyle behaviours for all small areas. Hence, deriving a model-based estimate for each area (i.e. not just those areas covered by the survey) consists of two stages.

In the first stage, regression analysis is performed modelling the survey data against available predictors of the health behaviour. This analysis is conducted for the subset of areas covered by the survey. The output from this first stage is a set of parameter estimates. At the second stage, for each area in the population, the coefficients of the predictor variables obtained from the first stage model are attached to the identical set of variables available at the small area level to produce an estimate for the area as a whole.

It is the underlying model, therefore, that enables us to move beyond the selected areas in the sample to provide information about the characteristics of all areas in the population (Goldstein, 2003). For example, we know from previous research that smoking behaviour is associated with social status. So individual NS-SEC (National Statistics Socio-Economic Classification) is likely to be included as a predictor variable in the regression for the first stage of synthetic estimation. The coefficients for NS-SEC would then be applied to the equivalent NS-SEC information available for all areas in the Census 2001.

### 2.2 Typology of methods

A number of studies have recently produced small area estimates in the UK, with a range of survey variables including income, labour market participation and health behaviours. Five different sets of methods were used to generate these estimates:

- simple (non-modelled) methods using indirect standardisation;

- models using individual level covariates only;
- models combining individual and area-level covariates;
- models using area level covariates only; and
- other approaches for larger areas of geography.

In the following sections each of the methods is discussed in turn, setting out their strengths and limitations in order to inform the selection of methods for testing in the second stage of the project.

## **2.3 Simple methods**

### **2.3.1 Indirect standardisation**

Indirect standardisation involved applying national estimates derived from survey data to area-level population counts to generate *expected* area estimates. For example, an indirect estimate for the proportion of men smoking in a particular ward could be generated as follows. First, the proportion of men smoking in each 5-year age band nationally would be estimated using HSfE data. Applying these national estimates to the census counts of men within the same 5-year age band for that ward would give an estimate of the number of men that smoke in that ward, from which the proportion could be estimated by dividing by the total census count of men in that ward. Essentially, therefore, the national prevalence rates for each sub-group are weighted by the proportion of persons in that sub-group in the small area.

A recent study has used this method to calculate the prevalence of heart disease for PCOs within two health areas. HSfE survey data was used to derive national prevalence rates by age (7 breaks), sex (2) and social class (6) for self-reported circulatory illness. Population counts by age and sex were obtained from practice registers and NCP profiler used to allocate counts into proxy social class groups. The appropriate national rates were then applied to the corresponding cell population counts to calculate the expected burden of disease for each PCO (Gibson and Asthana S, 2001). This is, in essence, a (non-modelled) method for deriving age, sex and social class adjusted *expected* rates, based on the national disease prevalence rates for each combination.

### **2.3.2 Strengths and limitations**

There are several reasons why indirect standardisation is appealing. First, the method has intuitive appeal. As Levy (1979) explains, it seems likely that the mean level of many variables in a population is highly related to the distribution in the population of such demographic variables as age, sex and social class. In addition to its intuitive appeal, indirectly standardised methods are generally easy and inexpensive to apply since the cell proportions at the local level are available from the Census, and the national estimates for demographic classes are easily obtainable from national surveys such as the HSfE.

This method also lends itself to two sorts of enhancements. First, rather than compute estimates at the national level, one possible option is to 'fine-tune' the method by calculating rates for different types of areas using some form of area

classification (e.g. urban/rural, quintiles of deprivation), and apply these to the constituent small areas in each type (Chesterman *et al.*). Second, the indirectly standardised estimates for each small area within a larger area (e.g. wards within LADs) can be ratio adjusted so that a weighted average of the adjusted small area estimates equals the direct estimate for the larger area. This adjustment ensures consistency between the direct and (aggregated up) indirect estimates for the larger areas.

The major drawback of this method, however, is that it assumes that the national rates for each subgroup apply uniformly across all areas. The implication for small area estimates is that the method assumes that the differences in health behaviour measures between areas are due solely to differences in their demographic composition. In other words, it is assumed that if two areas had the same composition with respect to the demographic variables used, they would have the same expected prevalence rates (Schaible, 1996). A large body of research, however, has consistently shown that individual health related behaviour, even within the same social group, varies by 'contextual' factors operating at the area level (Macintyre *et al.*, 1993). To deal with such area differences in health a more complex model is needed to effectively capture the variation between areas that exists over and above that due to differences in their demographic composition.

## **2.4 Models using individual level covariates only**

### **2.4.1 Using covariates from the Census**

An extension of the indirect standardisation method is to use the modelled relationship between individual health behaviour measures obtained from a survey against a set of predictor variables for the same individuals recorded in the survey. Generally the covariates chosen for the model are those that are available as counts for all small areas (e.g. from the Census).

These models estimate the probability that a person with specific known characteristics (say age, sex and social class) currently smokes, is obese etc. The model-based probabilities are then converted into estimated proportions in each subgroup defined by the covariates who fall into the relevant health category. These proportions are then applied to the covariate counts available from the Census to derive an overall estimate for the small area in much the same way as for indirect standardisation.

In a recent study, Flowers (2003) used this approach to estimate coronary heart disease (CHD) prevalence rates at PCO level using HSfE 1998-2000 data. Using logistic regression, he estimated the probabilities of reported CHD for age/sex/social class/ethnicity groups. These were then applied to the corresponding Census 2001 counts for each groups to derive CHD prevalence estimates for PCOs within a health region.

### **2.4.2 Using covariates from SARs**

Surveys collect detailed data on a range of individual characteristics and circumstances - although the inclusion of more and better survey covariates is likely to greatly improve the fit of such individual level models, analysts are restricted in the choice of covariates for synthetic estimation by the requirement to have equivalent covariate information for all areas.

Censuses also collect a fairly wide range of personal data but concerns to preserve confidentiality set limits to the number of cross-classifications that are released. An innovative approach used by Charlton (1998) was to increase the number of covariates he was able to use from the survey data by deriving synthetic estimates using the Sample of Anonymised Records (SARs) available for the first time for the 1991 Census. The 1991 SARs provided a representative 1% sample of individual census records with the full set of individual attributes collected in the Census available for all individuals in the SARs.

Charlton was able to use a wide range of individual covariates at the modelling stage from the National Survey of Morbidity in General Practice data. The model coefficients were then directly applied to the inhabitants of each SAR area with the specific combination of characteristics included in the model to obtain synthetic estimates for 278 areas (LADs or groups of LADs).

### **2.4.3 Strengths and limitations**

The major drawback of the individual-level approach concerns its data requirements. This form of synthetic estimation requires an exact correspondence between the covariates used in the model and data available from the Census or other administrative data sources. The limited number of cross-tabulations (or cross classifications) of socio-demographic information such as age, sex, ethnicity, social class available from the Census restricts the choice of predictors in these models. For example the most detailed cross-tabulation of counts available at ward level from the 1991 Census data which were known to affect health behaviour was banded age and gender by marital status (Twigg *et al.*, 2000). The recent 2001 Census tables offer more finely disaggregated counts (e.g. by age, sex, economic activity and qualifications (CAS32)); this could improve the accuracy of local estimates derived using this approach.

## **2.5 Models combining individual and area level covariates**

The methods discussed so far were all at the individual level. An alternative set of models can be described in terms of multi-level models incorporating *random effects* (also known as mixed models). Their importance to small area estimation lies in the fact that a random effects specification assumes that significant systematic variation between small areas remains after the effects of covariates in the model have been accounted for. Such 'unexplained' variation is modelled through the addition of small area specific random coefficients to the fixed effects (Saei and Chambers, 2003).

Such multilevel models give rise to more complex ways of building a model for health behaviour measures; generating small area estimates from these model parameters and finally calculating the confidence intervals for them.

In addition to their ability to incorporate unexplained variability between areas into the estimation procedures, there are a number of other reasons why multilevel models are suitable for producing synthetic estimates for small areas.

First, multilevel models are suited to the clustered nature of social surveys for which individuals are clustered within households which in turn are clustered (usually) within postcode sectors. By using the clustering information it provides more accurate standard errors, confidence intervals and significance tests, and these generally will be more 'conservative' than the traditional estimates obtained by ignoring the presence of clustering in the data (Goldstein, 2003). Second, by allowing the use of covariates measured at any level of the hierarchy, it enables researchers to explore the extent to which any differences between geographical areas such as wards are associated with individual, household and area level characteristics (Goldstein, 2003).

Using the techniques of multilevel modelling, a model can be applied to survey data that simultaneously accounts for both individual and area level influences on health related behaviours such as smoking. Twigg *et al.* (2000) used both individual and area-level covariates to obtain prevalence estimates of smoking and 'problem-drinking' for each ward in England by combining survey data from the HSfE with small-area census data. We illustrate this approach by discussing the estimates of smoking prevalence.

At the first stage, for those small areas covered by the HSfE, a multilevel model of individual smoking behaviour using both individual (sex, age and marital status) and area level predictors (e.g. the survey estimate of the percentage of private rented households in the postcode sector) was fitted to the survey data.<sup>1</sup> For the second stage, the model parameters of individual and area effects (and their interaction) were combined to estimate the proportion of smokers in each combination of age, sex and marital status, resident in wards with varying proportions of private renters and car owners. These estimates were then applied to the corresponding census counts to provide a synthetic estimate of smoking prevalence for all wards.

### **2.5.1 Strengths and limitations**

Conceptually and methodologically, the analysis by Twigg *et al.* (2000) represents an innovative advance over the simpler methods described earlier for it accommodates both individual and area level effects. It has long been recognised that both individual circumstances and the social and physical environment in which people live influence health behaviours. From an individual perspective, an individual's social class may influence health-related behaviour such as whether they smoke or not. Equally, from an area or ecological perspective, smoking prevalence may be influenced by social norms of behaviour. In addition, the individual and ecological

---

<sup>1</sup> Twigg *et al.* (2000) used 'survey' means at the area level as actual locational information of the respondent's area of residence is not provided in the public-access HSfE dataset. If such locational information had been available they would have used Census means in the multilevel model.

influences can interact to mitigate or increase the risk of being a smoker. As a result of these influences operating at different levels it could be argued that this approach offers in some sense a more explanatory model of health behaviour than those methods that conduct analyses at a single level.

The inclusion of individual level covariates such as age, sex and social class in the model in combination with the corresponding census counts also permits the production of separate estimates for relevant demographic groups within each small area.

Although using both individual and area level covariates in a multilevel model offers an advance over the simpler methods, there are a number of potential limitations when applying this method in practice.

First, as with the simpler methods described earlier, the inclusion of individual level covariates in the model imposes quite stringent data requirements as there must be an exact correspondence between those used in the model and the counts available from the Census. The limitations on the number of cross-tabulations available for small areas such as wards from the Census restrict the choice of predictors for the model. Important individual-level predictors of health, therefore, may be eliminated from the model simply because their distribution at the small area level is unknown.

Second, in comparison to the simpler methods, estimating the standard errors for the synthetic estimates based on a multilevel model that uses both individual and area level covariates is considerably more complex (Moura and Holt, 1999). Although Twigg *et al.* (2000) did not publish any standard errors for their ward-level estimates of smoking and 'problem-drinking' we understand that they are in the process of currently updating their work to include these.

## **2.6 Models using area level covariates only**

A more restricted multi-level model would be to use covariates measured at the area-level only. In this case, the health behaviours of individuals living in the survey areas (e.g. whether they are a current smoker or not) are predicted using only area level variables. This results in a set of regression estimates that relate to between-area variation. In effect, the model gives a constant predicted value for all individuals within an area, which can be interpreted as the predicted mean for the small area in question. The coefficient estimates are then attached to the known area means or proportions of the covariates for all areas, taken from the Census and other administrative data sources, to obtain synthetic estimates.

Such an approach has been implemented by the Small Area Estimation Programme team (SAEP) of the Office for National Statistics (ONS). They have described their approach as 'regression synthetic estimation fitted using area-level covariates' (Heady *et al.*, 2003). A range of measures were estimated at ward level including:

- average gross weekly household income;
- proportion of households with dependent children which contain one parent families; and
- proportion of households with low social capital.

To derive ward-level income estimates for example, at the first stage geo-referenced household income data from the Family Resources Survey (FRS: 1996/97) and General Household Survey (GHS) was modelled against area-level covariates from Census 1991 data (e.g. the proportion of households containing persons in employment) and welfare benefit administration data for those areas covered by the FRS/GHS. The second step attached the model coefficients to the same covariates for all areas to obtain an estimate of mean household income for each ward.

### **2.6.1 Strengths and limitations**

Using only area level covariates in the model avoids the stringent data requirements described earlier for those approaches using individual level predictors of health behaviours. As Levy explains (1979), the motivation for this method is that if the set of area-level covariates are easily obtainable for all small areas, and if the relationship between them and the healthy lifestyle behaviours of individuals is strong, then estimates of good quality might be produced at relatively low cost.

A second argument in favour of this approach concerns the potential redundancy of individual level covariate information in predicting area variations in health behaviour rates.

It could be argued that models accommodating both individual and area level effects are, in a sense, more explanatory than those using area level covariates alone. The fact that such approaches are, however, constrained by the cross-tabulated census data available for small areas, means that in practice a sizeable proportion of the 'true' individual effects are essentially expressed as area effects. Plus, it seems reasonable to assume that if the characteristics of individuals impact on health, then the average health of areas will only differ if the profile of the population per area differs in terms of these characteristics. (As a simple example, if age happened to be the only predictor of health, then areas would only differ in health if their age profile differed. In which case controlling for the area age profile is sufficient - controlling for individual age is unnecessary.) So, a strong argument can be made for assuming that controlling for differences in area profile is all that is needed for predicting area differences in health. This, we understand, is the rationale behind the ONS approach to small area estimation.

A potential limitation of this approach, however, concerns the issue of disaggregation of estimates. Unlike the methods based on models that include both individual and area level predictors of individual health behaviour, the ONS approach does not support the production of separate estimates for subgroups within each small area. Specifically the estimates represent the underlying expected value for the demographic and social mix of adults living in a ward at the time of the 2001 Census. It cannot, therefore, tell us what proportions of those living in the ward fall into a particular health category by age, sex or social class. One could achieve this by fitting different models for each age and gender subgroup or by introducing interaction terms in the model.

The topic of cross-classifications is among those being investigated by the international EURAREA (Enhancing Small Area Estimation Techniques to Meet

European Needs) project that ONS is co-ordinating and so future developments in this regard are likely.

## 2.7 Other approaches for larger geographical areas

### 2.7.1 GREG estimator

The method of indirect standardisation described in Section 2.3.1 combines the national estimates for demographic subgroups with their known proportions at the small area level. The typical demographic variables used for such estimators are age, sex and social class.

Other forms of estimator take advantage of a wider range of covariate information available for small areas. Suppose, for example, that information on a continuous variable such as household size is available for areas such as wards from both the national survey and Census. In addition, it is believed that such a variable is correlated to the health behaviour of interest.

A generalised regression synthetic estimator (GREG) adjusts the survey based (direct) prevalence estimate of healthy lifestyle behaviours by taking account of any numerical difference between the survey and census area means of the relevant predictor (Heady *et al.*, 2003). For example, if the survey estimate of mean household size for ward X was higher than its known average then the health behaviour estimate for this ward would be adjusted downwards to account for this difference. Similarly, if the survey estimate of mean household size for ward X was lower than its known average then the health behaviour estimate for this ward would be adjusted upwards. If the survey estimate of mean household size for ward X was equal to the census figure then no adjustment would be made to the survey based health behaviour rate.<sup>2</sup>

Given the correlation between the continuous predictor variable and the individual healthy lifestyle behaviour the GREG estimator will be more precise than the sample based estimate which ignores this relationship.

In contrast to the indirectly standardised method, however, the GREG estimator cannot be used for those areas which do not contain any survey respondents. This eliminates the GREG estimator as a method for producing ward-level estimates as a sizeable number of wards are not represented in clustered national surveys such as the HSfE.<sup>3</sup> Such ward level estimates, however, could potentially be produced by the Labour Force Survey (LFS) as it is substantially larger in size and unclustered in design.

---

<sup>2</sup> For survey areas, therefore, we have both an estimate of mean household size and the prevalence of a healthy lifestyle behaviour such as smoking. Since the mean household size is known, then it makes sense, if the two variables are strongly correlated, to assume that the prevalence estimate of smoking for a ward might differ from its true smoking prevalence in the same proportion that its survey based estimate of mean household size differs from its known average.

<sup>3</sup> Technically, the GREG estimator requires that there is a sample in every small area of interest. But this requirement is often relaxed and a slightly modified version is calculated by omitting the sample means for those areas where the achieved sample size is zero or very small (Saeib and Chambers;2003).

Despite this limitation, the GREG estimator could be used for larger areas of geography such as PCOs and LADs as all or the majority of these areas are covered by the HSfE data for 2000-02 (see Chapter 3). Furthermore, good quality external information for these geographical areas is available. As explained later, these higher-level survey-based estimates can then be compared with the synthetic estimates for the same areas obtained by aggregating up the estimates for those wards nested within each.

### **2.7.2 Composite estimators**

Direct survey-based estimates for small areas such as wards using national surveys will be subject to a large degree of variability because of the small achieved sample sizes within them. Such estimates, however, will be design-unbiased: which essentially means that the expected value of the prevalence estimate for a small area is equal to its true value. In contrast, model-based estimates, such as those described in this literature review, do not possess this property since what they estimate is the underlying expected value for any area with the same set of covariate values and not the real value for the small area in question (Heady *et al.*, 2003).

As Rao (2003) explains, a natural way to balance the potential bias of a model based estimator against the large variability of an unbiased direct estimator is to take a weighted average of the two. A composite estimator combines two estimators together with the aim of arriving at an estimator which may be more accurate than either of its components (Schaible, 1996).

Theoretically, a composite estimator could be formed by combining a survey-based estimate with a model-based estimate. Alternatively, a composite estimator could be formed by combining a GREG estimate with a model-based estimate (see Heady *et al.*, 2003). In practice, however, both these examples of the composite estimator require at least one survey respondent in each estimation area of interest. Hence, like the GREG estimator itself, the composite estimator is also not feasible as a method for producing estimates at the ward level. It could be plausible, however, to use composite estimators for higher-level areas such as LADs.

### **2.7.3 Fay-Herriot estimator**

As discussed in Section 2.5 multilevel models incorporating random area-specific effects allow for between area variation beyond that explained by the covariates included in the model (Rao, 2003). Fay and Herriot (1979) were the first to use such models for small area estimation.

The model used by Fay and Herriot (1979) can be classified as an aggregate or area level model that relates the survey based area means of the dependent variable to area-specific covariate values and to the random area effects. As it turns out, under this model, the best predictor of the small area mean can be expressed as a weighted average of the survey-based estimator and a regression-synthetic estimator that uses the fixed effects only (Rao, 2003).

Using the Labour Force Survey, the ONS in consultation with the University of Southampton have provided estimates of unemployment at the LAD level by combining LFS and national claimant count data.

Put briefly, the researchers have used what they describe as a modified Fay-Herriot approach. A logistic regression is first used to model the proportion of people unemployed in six age-sex classes within each LAD using the claimant count data, region and LAD type as predictor variables. The estimator for the proportion unemployed in each LAD is then derived by taking a weighted average of the age-sex class estimates (using calibration weights to ensure that the class estimates sum up to the survey-based estimate for the LAD). As we understand it, the approach takes the form of a modified Fay-Herriot approach for the reason that even though the estimates themselves are produced using a fixed effects specification, the estimates of the confidence intervals use an estimate of the 'between area variance' that is obtained by fitting a random effects model (see Ambler *et al.*, 2001).

There are two main reasons why such a method is not appropriate for this project. First, whilst such models are essential if individual level data on the health behaviour of interest and covariates are not available (e.g. as may be the case with administrative data), this is not the case for this study. Second, due to the clustered nature of the HSfE, survey-based estimates for wards are not available for a sizeable proportion of them. Indeed, even for the subset of wards covered by national surveys, the achieved sample sizes within a majority of them would be too small to provide reliable estimates of the true area health behaviour rate.

## 2.8 Summary of methods at the ward level

Small area estimation methods based on national surveys have been developed for those situations where survey-based estimates either cannot be computed because some areas contain no sample observations or are too imprecise because the achieved sample sizes within them are too small. Essentially all the methods involve combining national survey data with small area information such as that available from the Census. The precise ways in which these two sources of data are combined, however, can take a variety of forms ranging from the relatively simple to the complex.

In this section, five different sets of methods for small area estimation were identified:

- simple methods including indirect standardisation;
- models using individual level covariates only;
- models combining individual and area level covariates;
- models using area level covariates only; and
- other approaches for larger geographical areas.

For small areas such as wards, only the first four sets of methods are available for clustered surveys such as the HSfE.

A summary of the recommended methods to test in the second stage of the project is outlined in Table 2.1.

**Table 2.1 Recommended methods to take forward to Stage 2**

	<b>Indirect standardisation</b>	<b>Models using individual and area level covariates</b>	<b>Models using area level covariates only</b>
Measurement level of health behaviour	Individual	Individual	Individual
Measurement level of covariates/predictors	Individual	Individual and area-level	Area-level
Implementation dataset (see Chapter 3)	Census cross-tabulations	Census cross-tabulations, census proportions, and other administration data at the area level	Census proportions and other administration data
Estimates for demographic sub-groups within the small area?	Yes	Yes	Possible development in the future
Published formula for constructing valid confidence intervals?	Yes	To be developed by the project team	Yes

### 3 SCOPING AND SETTING UP THE DATABASE

To generate synthetic estimation requires combining together data from two sets of sources – survey data containing the health measure(s) of interest (e.g. smoking status) and a covariate dataset (census or administrative data), matched to a common geography for which estimates are required (the ‘estimation area’). As discussed in Chapter 2, the relationship between the survey and covariate dataset can be modelled in a number of ways. In general, as the models themselves become more complex, the dataset requirements become more demanding and complex. We have therefore scoped the data requirements for the most complex hierarchical model, including both individual and area-level data, as the database required would also meet the information needs of the less complex models.

This section describes the construction and components of the various datasets, and discusses their limitations and likely impact on the synthetic estimation outputs.

#### 3.1 Data requirements

The following datasets are required for the project:

- *survey dataset*: the survey variables of interest matched to the estimation area identifiers (e.g. ward) and the postcode sector identifiers (to identify sampling clusters). The survey dataset holds both the outcome variables (e.g. smoking status), as well as the individual level covariate data (e.g. age, sex, NS-SEC).
- *area-level covariate dataset*: contains the estimation area level means for a set of covariates – usually census, administrative and registration data – along with the estimation area identifiers, and any higher-level area covariates and identifiers (e.g. at PCO level).
- *analysis dataset*: The survey and covariate datasets matched on estimation area identifier. The analysis dataset contains only the areas sampled in the survey. This dataset is used for modelling.
- *implementation dataset*: Once the modelling has been performed, a dataset covering all areas (not just those sampled) is required to produce the final estimates. The implementation dataset will be at the lowest estimation area level, nested within higher-level geographic identifiers. This will allow the production of higher-level estimates by aggregating estimates for the component small areas.
- *external validation dataset*: Identifying relevant local and/or national surveys or other administrative sources to provide direct estimates of relevant outcomes to compare against the synthetic estimates. Such external validation will be an important check on the plausibility of the synthetic estimates.

## 3.2 Geographical matching

A key requirement of the project is to create the analysis dataset combining survey and covariate data matched at the level of geography required for the lowest estimation area. For this project, the lowest estimation area is defined as Census 2001 wards (or electoral wards as at December 31<sup>st</sup> 2003, also known as statswards 2003, and henceforth referred to as wards). Wards were identified as the lowest estimation area level because they present the best compromise between the needs of users for local information at finer geographical levels against the need to produce technically robust estimates from relatively sparse survey data.

The covariate data required from the Census are available at ward level. However, survey data and some of the covariate data available from administrative sources are not available at ward level. The approach taken by the SAEP project to overcome this problem of non-equivalent boundaries on different datasets was to create a GIS (Geographical Information System). The GIS used the 'point-in-polygon' method for locating the centroid of any unit (e.g. ED91 or postcode) within a set of boundaries on a map (e.g. 1998 wards). The GIS solution offers the flexibility to locate data items collected on one geography to different boundary sets over time and space. However, the quality of the matching depends on the accuracy of the digitised boundaries. The construction of a GIS is also time-consuming and expensive.

Given these constraints, in this project we have opted to use link-files (look-up tables) to locate smaller level units within larger ones (e.g. postcodes within wards or wards within PCOs). This approach is relatively inexpensive and quick, but has a number of limitations. First, and most importantly, the availability of appropriate link-files define the limits of the kinds of data that can be matched into the analysis database. For example, there is no link file available that we are aware of which provides a 'best-fit' link between 1998 wards and Census 2001 wards. As a result, the range of data available currently available at 1998 ward boundaries (e.g. Index of Multiple Deprivation 2000) cannot be incorporated in our database using this approach. Second, the linkage process has to be repeated each time data are required for a different set of boundaries or when boundaries change. This is potentially an issue if there were a requirement to update the synthetic estimates to changes in ward and high-level geographical boundaries at regular intervals over the intercensal period.

Due to disclosure concerns, survey data in the public domain do not contain geographic identifiers at a level of detail that may allow individuals to be identified. However, as employees of the organisation that has carried out the Health Survey for England since 1994, the study team are in the fortunate position of having access to the survey sample files which include the postcode of residence of each respondent to the survey. We used the February 2004 release of the AFD (All-Fields Postcode Directory, ONS) link file that provides a look-up between current postcodes and a range of geographic identifiers, including Census 2001 ward, PCO, LAD and SHA. We have used this AFD for matching the survey data from 2000 onwards. Undoubtedly, such a procedure is not completely accurate as it assumes no change in postcode boundaries between 2000 and 2004. However, we have assessed that the scale of such error will be small because residential postcodes remain stable over long periods for the vast majority of areas.

Wards are part of the administrative geography and nest within higher level administrative tiers such as LADs and GORs. Hence attaching higher-level administrative identifiers to the database is straightforward. However, wards do not always fit neatly into larger health areas such as PCOs. The second link-file we have sourced for the project therefore provides a 'best-fit' look-up table between wards and PCOs. The ward-PCO look up file will be used at the implementation stage to aggregate ward level estimates to derive PCO level estimates. Because the majority of PCOs are coterminous with LADs (and therefore their component wards) and there are on average about 25 wards per PCO, we assess that the scale of the mismatch error between wards and PCOs will have a fairly limited impact on the aggregated estimates.

### **3.3 Survey dataset**

The survey dataset we have constructed consists of the pooled sample for the three most recent years of data available from the Health Survey for England (HSfE), namely 2000 to 2002. The more obvious reasons for selecting these particular years are that they include the most up to date data available and the years are symmetrically arranged either side of the Census year (2001) which is the main source for the covariate data. Furthermore, health behaviours are slow to change over the short-term. Therefore combining health outcomes measured for different wards in different years over the three-year period is unlikely to be distorted by any underlying secular trends. The main advantage of pooling data over years is that it increases the geographical coverage of the wards for which sample data become available thereby potentially improving the fit of the model.

#### ***Sample size and spatial coverage***

Each year the survey covers a representative sample of people resident in households, and in addition, in certain years particular population groups are over-sampled or 'boosted'. In years when special populations are boosted, the general population sample is halved to its usual size ('half' sample).

The sample size of the general population a 'full' sample year is typically about 16,000 adults aged 16 and over and 4,000 children aged 0-15. In 2000, the survey over-sampled older people aged 65 and over and in 2002 the samples of children and young people (aged 0-25 years) were boosted. The 2000-02 samples therefore comprise two 'half' and one 'full' year of HSfE data (or equivalent to 2 'full' years of data).

The primary sampling unit (PSU) for the HSfE is postcode sectors. Postcode sectors are roughly of the same size as wards, but because the two geographies are not coterminous, on average about three wards intersect within a postcode sector.

Table 3.1 sets out the geographical coverage and average sample size of adult respondents at different estimation area levels of interest and for different combinations of survey years. As we would expect, the average sample size increases as we move up from the smallest (ward) to the largest (PCO) area. Adding together more years of data has the effect of widening the geographical coverage of our

primary estimation area (number of wards sampled), but results in little change to the average sample size per ward (as the number of addresses sampled per PSU remains relatively fixed from year to year). Thus while the average sample size per ward remains just under 10 adults for one or three year combined data, the number of wards with at least one sampled adult almost doubles.

**Table 3.1 Geographical coverage and average sample size per unit for different combinations of HSfE samples**

	Wards	LAD <sup>1</sup>	PCO
<b>Total number of units in England (%)</b>	<b>7958 (100)</b>	<b>354 (100)</b>	<b>304 (100)</b>
One (1.0) year sample - 2001			
- no. of units sampled (%)	1738 (22)	317 (90)	295 (97)
- <b>mean sample per unit</b>	9	49	53
- <b>max sample per unit</b>	53	268	173
Two years sample (1.5) (2001-02)			
- no. of units sampled (%)	2645 (33)	345 (97)	302 (100)
- <b>mean sample per unit</b>	9	67	76
- <b>max sample per unit</b>	60	409	221
Three years sample (2.0) (2000-02)			
- no. of units sampled (%)	3232 (41)	346 (98)	302 (100)
- <b>mean sample per unit</b>	10	90	103
- <b>max sample per unit</b>	60	565	273
<sup>1</sup> LADs not selected in any of the three years 2000-02 include Rutland, South Bucks, Isle of Scilly, Copeland, Teesdale, Hertsmere, Hyndburn, Craven			
<i>Note: the minimum sample size per unit was 1 for all units and all year combinations</i>			

The general population adults sample in the HSfE is self-weighting (weight of 1). Because of the lack of agreement and computational complexity involved in weighting for unequal selection probabilities in multi-level modelling (Heady *et al.*, 2004), we have opted to use only the general population sample in each year for the analysis.

#### **Sample-based individual-level covariates**

Twigg *et al.* (2000) have included individual level covariates in their multi-level models for the synthetic estimation of health behaviours. While the survey data hold a wealth of socio-demographic detail on sampled individuals, the covariates that can be entered into the modelling process are constrained to a subset of variables that satisfy two minimum criteria: first, the individual level covariates are available for all estimation areas and for the whole population and second, that the covariates are identically defined in the sample and population datasets.

These criteria severely restrict the choice of individual-level covariates available for inclusion in the models (see Chapter 2). For example, variable definitions are rarely

identical between the administrative systems and the survey sources and generally do not provide meaningful cross-tabulations (say, by age). Cross-tabulations of census data are more promising in this regard. At ward level, standard tables provide various types of disaggregated counts, the limit to the number of cells in each table being set by disclosure control thresholds. From the point of view of this project, this essentially involves making a choice between counts from tables with fewer cross-tabulations but finer breaks for each variable (e.g. age in narrow 5-year age bands in a 3-way table), against tables that offer more cross-tabulations (e.g. 4-way table including age, sex, health status and ethnicity) but with broader groupings of each variable.

Because of the wide range of combinations of cross-tabulated variables available in the census data, our approach will be to include the full set of individual level covariate data from the sample file that correspond to the census variables at the first stage in the analysis. Then, if one or more individual level covariates are selected in the forward stepwise modelling procedure, we would re-run the models defining the individual level covariate data to be exactly the same as that available in the census table with the same combination of individual-level cross-tabulated counts.

It should be noted that the new NS-SEC occupational classification was included in the HSfE data from 2001 onwards. Therefore, if NS-SEC was found to be an important individual level covariate, we would be limited to using the 2001-02 HSfE data in the analysis. We are currently exploring the feasibility of retrospectively coding the 2000 sample data using an approximate translation matrix file provided by ONS.

### **3.4 Area-level covariate dataset**

#### **3.4.1 Ward level covariates**

##### ***Census data***

Census 2001 proportions for a total of 21 covariates have been extracted from the Census Key Statistics tables. The variables can be grouped as follows:

- indicators of material deprivation (e.g. lack of access to a car, to central heating, overcrowding);
- indicators of social position (e.g. NS-SEC, no or low educational qualifications);
- housing tenure (owner occupiers);
- ethnic origin (self-defined ethnicity, country of birth);
- health status (limiting longstanding illness, self-assessed general health);
- provision of informal care;
- indicators of social vulnerability (e.g. pensioners living alone, lone parent households with dependent children); and
- marital status and living arrangements.

In addition, it is expected that a ward level urban/rural indicator and the ONS typology of area classifications will shortly become available. These will be added to the covariate dataset if released towards the beginning of Stage 2 of the project.

### **Administrative data**

The only data currently available at ward level are the proportions of adults claiming five types of benefits (DWP, 2001). The benefits include attendance allowance, disability living allowance, incapacity benefit, severe disablement allowance and income support.

### **3.4.2 Higher-area level covariates**

There are more varied types of data currently available at the LAD level than at PCO level. Heady *et al.* (2004) suggest including higher-level geographical covariates to adjust for spatial patterning of outcomes ( e.g. north-south divide, quality and provision of services which may impact differentially on health outcomes between areas). We shall use both the LAD and PCO covariates in model-fitting.

The data available at LAD level are of the following types:

- Index of Multiple Deprivation, 2004 (ID2004);
- Mortality: standardised rates and ratios for premature mortality, avoidable mortality, cause-specific mortality (e.g. lung cancer, liver disease), and life expectancies all relating to the period 2001-02 (2003 Compendium of Health and Clinical indicators);
- Morbidity: cancer incidence (2002 Compendium);
- Hospital utilisation: Finished Consultant Episodes from the Hospital Episode Statistics, 1999-2000; and
- Area Classification typologies based on 2001 Census data.

Relevant variables from most of these datasets have been extracted and added to the covariate dataset.

### **3.4.3 Covariates at other geographies**

There are two data sources that may be of particular relevance to this project but which cannot be readily incorporated into the covariate dataset constructed around Census 2001 ward geography. These are:

- the updated Index of Multiple Deprivation (ID2004) - the index has been produced at LAD level and Census Super Output Areas (SOA level one), but not for wards; and
- Covariates at 1998 ward boundaries (e.g. income estimates based on Family Resources Survey 1998-99, Hospital Episode Statistics etc).

Both the above are potentially important covariates for the modelling. The ID2004 is particularly important as it uses the latest (non-census) administrative data in its construction and therefore provides a different measure to that obtained from the census data alone. However, in order to use the SOA level scores approximated up to

ward, we need to agree with the creators of the index an acceptable method for aggregating up SOA level scores to derive proxy ward-level mean score and discuss any implications at either the analysis or implementation stage of the project.

We assess that re-assembling covariate data at 1998 ward boundaries to 'best-fit' 2001 ward boundaries may be feasible, but unless there is good evidence to indicate that such covariates are better predictors than the available set, we shall not attempt to construct and check the required link file given the short time frame of the project.

### **3.5 Validation dataset**

We will be actively scoping local survey data that could be used for external validation of the modelled estimates at the second stage of the project. Currently, we have available to us a local boost sample of residents of the (erstwhile) Camden & Islington Health Authority surveyed in 1999. The adult sample size was just under 2,000 respondents and the questionnaire coverage and survey procedures were identical to the main HSfE 1999 survey (including for e.g. questions on smoking, drinking, BMI). This dataset will provide direct estimates for the four PCOs and their constituent wards to compare against the synthetic estimates for the same set of areas.

Unlike the local boosts of the HSfE survey, using other local survey data to compare the modelled estimates throws up a number of issues of comparability. Various non-sampling sources would account for differences between survey estimates. Small differences in the questions asked, the order in which they are asked, survey mode (e.g. personal interview compared with self-completion) and sampling design all contribute to measurement error. However, as we shall primarily be using these data for similarities in the pattern of relative ranking and correlation coefficients, rather than comparing absolute values, the impact of measurement error is somewhat attenuated. Previous research by Twigg and Moon (2002) indicated that for wards with low (direct) prevalences, the synthetic estimates were about 20% higher, while for those with high prevalences synthetic estimates were 10% lower than the survey estimates.

## 4 CALCULATION OF CONFIDENCE INTERVALS FOR SYNTHETIC ESTIMATES

As described previously in this report (Chapter 2), the synthetic estimate generated for a particular ward is the expected measure for that ward based on its characteristics as measured by the auxiliary variables. In statistical terms, the synthetic estimate is actually a biased estimate of the *true* value for an area and, as such, should be treated with caution (Heady *et al.*, 2003). By placing confidence intervals around a synthetic estimate, however, we can generate a range within which we can be fairly sure the *true* values for that area lies.

We will investigate two methods for generating confidence intervals for the synthetic estimates. The first method was derived by Heady *et al.* (2003, pages 11 & 73) and is appropriate for synthetic estimates based on parameter estimates from a model that considered only area-level covariates. The second approach, which can be used to generate synthetic estimates for models that also include individual-level covariates, is to use Markov Chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996).

### 4.1.1 Confidence intervals derived from the model estimates

The first method (Heady *et al.*, 2003) uses the estimates from the fitted model<sup>4</sup> to estimate the variance of the difference between the synthetic estimate and the *true* ward measure and hence to derive the confidence intervals for the synthetic estimates. The estimate of the variance has two components which corresponded to the area-level variance that is unexplained by the model (and hence not predicted by the synthetic estimate) and the uncertainty of the synthetic estimate itself. Based on this estimate of the variance, the confidence interval for a synthetic estimate<sup>5</sup> in ward  $k$  would be:

$$\hat{\alpha} + \hat{\beta}^T \underline{X}_k \pm 1.96(\hat{\sigma}_u^2 + \underline{X}_k^T \text{Var}(\hat{\beta}) \underline{X}_k)^{1/2}$$

where  $\underline{X}_k$  is the vector of covariate values for ward  $k$ ,  $\hat{\beta}$  is the vector of parameter estimates for the ward-level covariates and  $\hat{\sigma}_u^2$  is the estimate of the area-level variance.

The estimate of the area-level variance in the above formula was actually derived for postcode-sectors (PCSs) rather than wards. This was done because the surveys used to generate the synthetic estimators (the General Household Survey and Family Resources Survey) were clustered within PCSs. As ward and postcode geographies do not match, only parts of wards (the areas that overlapped with the selected PCSs) would be covered by the surveys, rather than whole wards. If one naively included a random effect for ward in the model, one would actually be estimating the variance between *part-wards*, not whole wards, and hence the model would over-estimate the

<sup>4</sup> The parameter estimates, as well as the variance-covariance matrix of the fixed effects.

<sup>5</sup> This is the confidence interval around a mean. A confidence interval around a proportion would be estimated using the same formula with a transformation by the inverse of the logit function.

between-ward variance. (This is because the *part-wards* would be smaller geographically than whole wards and so would be likely to be more homogenous than whole wards.) Using a large scale unclustered survey (the Labour Force Survey), Heady *et al.* (2003) demonstrated that the between-ward and between-PCS variances were very similar and also that attempting to estimate the between-ward variance directly for the clustered surveys would result in the between-ward variance being over-estimated.

#### 4.1.2 Confidence/credible intervals estimated by simulation

The formula for the confidence interval derived by Heady *et al.* (2003) was used for synthetic estimates based on parameter estimates from a model that only considered area-level covariates and would not be appropriate if the model included individual-level covariates (Moura and Holt, 1999). Therefore, an alternative method to generate the confidence intervals is required for synthetic estimates based on models that also include individual-level covariates, such as those fitted by Twigg *et al.* (2000). We propose to use MCMC methods, within a Bayesian framework, to generate the 'confidence intervals' (actually referred to as 'credible intervals') for the synthetic estimates based on these models.

One of the key differences between Bayesian statistics and traditional (frequentist) methods is that the parameter estimates are treated as random with corresponding (*prior*) probability distributions - there is no single point estimate for each parameter as there would be for traditional statistical methods. To generate the estimates for parameters, it is therefore necessary to run an iterative procedure (the MCMC procedure) that generates a series of values for each parameter. The sample of values for each parameter can then be used to estimate, for example, the mean and variance for a parameter.

We will exploit the sample of values for each parameter that the MCMC method generates to produce the credible intervals for the synthetic estimates. As an example, assume that we have generated 1,000 values for each parameter using the MCMC procedure. For each ward, we can therefore generate 1,000 estimates of the synthetic estimate - one for each set of parameter estimates. In addition, we can simulate 1,000 estimates of the *true* measure for the ward by including a random term, drawn from the normal distribution with the appropriate estimate of the variance and zero mean. So, the *true* estimate based on the  $r^{\text{th}}$  set of parameters for ward  $k$  would be:

$$E(\hat{Y}_k^r) = \alpha + \beta^r \bar{X}_k^r + v_k^r, \text{ where } v_k^r \sim N(0, \sigma^{2(v)}).$$

We can then obtain the 95% credible interval for the synthetic estimate for a ward directly as the range between the 25<sup>th</sup> and 975<sup>th</sup> largest simulated *true* estimate for the ward.

## 4.2 Indicative confidence intervals for wards

Table 4.1 shows the confidence intervals for smoking and obesity prevalence for an 'average' ward that could be expected to be obtained from the synthetic estimation.

These confidence intervals were estimated using the formula derived by Heady *et al.* (2003), assuming that various proportions of the area-level variance had been explained by the covariates in the model. As can be seen, the more area-level variance explained by the model, the narrower the confidence interval.

**Table 4.1 Estimated confidence intervals for an ‘average’ ward for smoking and obesity**

	Smoking		Obesity	
	Lower CI	Upper CI	Lower CI	Upper CI
Proportion of area-level variance explained:				
0% (null model)	11.5%	45.9%	13.4%	33.2%
25%	14.2%	40.1%	15.1%	30.0%
50%	17.2%	34.7%	17.1%	27.0%
75%	20.8%	29.6%	19.3%	24.3%

### 4.3 Calculation of confidence intervals for aggregated synthetic estimates

The synthetic ward estimates generated will be used to obtain estimates for higher level geographies (e.g. Primary Care Organisations). In order to interpret these estimates, it will also be necessary to produce confidence intervals around each estimate. For estimates of the mean, this would be fairly simple to do. The mean estimate for the PCO would be a weighted<sup>6</sup> average of the estimates for the wards and so the variance for the PCO estimate would be the sum of the variance for each ward multiplied by the square of the weight. From this one could produce the confidence interval.

For estimates of proportions, it is not so easy to generate the confidence intervals for higher levels based on the individual estimates and variances for wards. We therefore propose to use MCMC methods to simulate the credible intervals for estimates of proportions at these higher levels. As described above (Section 4.1.2), we would generate a sample of parameter estimates and, from these, a sample of synthetic estimates for each ward. To obtain the synthetic estimate for a higher level (for example, a PCO), we would calculate the weighted average of the synthetic estimates for each ward in the PCO. Each set of parameter estimates would therefore generate a different estimate of the proportion for the PCO and, from the resulting sample of estimates, we would obtain the credible interval by taking the 2.5<sup>th</sup> and 97.5<sup>th</sup> largest value for the PCO.

<sup>6</sup> The weights would be equal to the proportion of the population in the PCO that lived in each ward.

## 5 TESTING THE SOFTWARE

We have two options for the statistical packages that we can use to fit the multilevel models and subsequently produce the synthetic estimates: MLwiN and Stata. These packages fit the multilevel logistic regression models using different methods<sup>7</sup>, although the parameter estimates produced are very similar.

Using Stata does offer several advantages over MLwiN for fitting the required models and producing the synthetic estimates for this project. These include:

- being generally much more user-friendly;
- having a simple-to-use statistical test of adding one or more additional covariates to the model (**testparm**);
- it being much easier to produce the standard errors for the synthetic estimates for synthetic estimates based on area-only models;
- being more stable (MLwiN does tend to crash).

However, we will need to use MLwiN to run the MCMC procedure in order to obtain the credible/confidence intervals, as Stata does not perform MCMC methods.

Given the above, our plan is to use Stata to build the models – first by performing a stepwise logistic regression, assuming the sample to be unclustered and with a p-value of 0.10, and then obtaining the optimal clustered model using the **xtlogit** command. Having obtained the optimal model in Stata, we would then fit, and check, the same model in MLwiN. Having made any necessary adjustments to the model, we would obtain the point estimates from the standard MLwiN procedure (IGLS/RIGLS) and then run the MCMC procedure in order to obtain a sample of parameter estimates in order to estimate the credible intervals.

---

<sup>7</sup> MLWin fits the linear multilevel model using either Iterative Generalised Least Squares (IGLS) or restricted IGLS (RIGLS) algorithm. Models for binary outcome are fitted with a logistic (or probit) link function and binomial error term using a first- or second-order Taylor series expansion by iterating with the standard IGLS or RIGLS algorithm (Goldstein, 1995). The higher level random effect terms can either be modelled as being inside the link function, the penalized quasi-likelihood (PQL) model, or outside the link function, the marginal quasi-likelihood (MQL) model. Wherever possible, we would want to fit second-order PQL models.

The corresponding function in Stata is called **xtlogit** (cross-sectional time series logistic regression). This function fits a logistic regression model with a level-two random effect (in this example representing the area-level variance) using M-point Gauss-Hermite quadrature (StataCorp, 2003). If more complex multilevel models were required, then these can be fitted using the **gllamm** command, which allows additional random effects to be included in the model (Rabe-Hesketh et al., 2002). For our purposes, this can be considered to be a more flexible version of the **xtlogit** command.

## 6 RECOMMENDATIONS FOR STAGE 2

### 6.1 Choice of estimation area

Previous work done by the research team has shown that fairly robust *direct* estimates from the HSfE can be calculated for large areas such as the 28 Strategic Health Authorities and nine Government Office Regions (Scholes *et al.*, 2004). Choices for synthetic estimation at lower level geographical areas include (in order of descending size) PCOs, LADs, wards, Super Output Areas and Output Areas (OAs).

In principle, synthetic estimation at the smallest possible area would be the optimal choice as it offers at least two important advantages. First, from the users perspective more finely-defined estimates locate precisely the areas for targeted intervention. Second, they offer more flexibility of aggregation to respond to changes in higher-level geographies over time. One of the limitations of analysis at small areas such as SOAs/OAs is that the covariate dataset available at these levels will be restricted to the more limited cross-tabulations available from the Census due to the application of disclosure control thresholds (i.e. CAS and ST tables). Other than the ID2004, there are no administrative datasets available (e.g. benefit claimant rates, Hospital Episode Statistics, mortality rates) at these small levels.

However, the main reason for not opting for analysis at the SOA/OA levels are statistical constraints imposed by the survey design. As mentioned previously, the HSfE is a clustered sample at postcode sector level. The population size of postcode sectors is similar to that of wards (about 5,000). Heady *et al.* (2003) have shown that, because of the similarity between the two, confidence intervals based on the variance between postcode sectors (rather than the variance between wards *per se*) gives reasonably accurate results. The same would not apply for confidence intervals around OA synthetic estimates.

This is not to suggest that OA estimates could not be produced - synthetic estimates could be generated by modelling using OA-level covariates. However, it would not be possible to estimate the OA-level variance from this model (as there would be very few cases in each OA) and hence we would not be able to produce confidence intervals for the estimates. In addition, our ability to validate the estimates at this level would be very limited.

For these reasons we suggest that wards should be the estimation level for this project. Wards also offer a number of other advantages such as:

- a wider range of available (non-census) covariate information than at smaller area levels;
- local users are familiar with ward locations in their areas;
- is possible to compare lifestyle estimates against other data available at ward level such as NHS activity (HES data) and mortality; and
- previous experience (e.g. by ONS) indicates that ward estimates are fairly robust.

The main disadvantage of choosing wards as our estimation level is that ward boundaries change over time. This affects the relevance of estimates for the wards affected, and occasionally when such changes also impact upon the higher level boundaries within which wards nest, aggregated estimates at the higher levels will also no longer be valid.

## 6.2 Which models to test?

We believe that the only two modelling approaches that could produce reasonable synthetic estimates are the two multi-level model methods, namely the model with covariates at the area level only, and the model with covariates measured at both the area and the individual level. The simple model (i.e. the model with individual covariates only) can, we think, be dismissed on the grounds that other research suggests that there *are* genuine area impacts on health, so excluding them from the model would inevitably give biased estimates.

Nevertheless, we will construct estimates where simple area effects are taken into account by generating separate HSfE prevalence estimates within ID2004 quintiles or by area typology (ONS cluster). If these estimates are close to the synthetic estimates generated using other means we will consider whether there would be benefits in using them on the grounds that the estimates are simpler to understand and to replicate.

## 6.3 Which health measures?

The health measures to be tested at Stage 2 are those that we consider we will be most likely to be able to generate reasonable synthetic estimates for (on the grounds that, if we can't achieve reasonable estimates for these, then we have little chance of getting reasonable estimates for other outcomes).

The criteria for choosing the estimates to take forward to Stage 2 are:

- they should demonstrate a fairly large between-area variance;
- they should be known to be correlated with the variables that will be used in the models (i.e. age, sex, social class, area deprivation etc.);
- should provide a geographical spread of areas.

Our suggestions for the outcomes to include in Stage 2 (with reason for selection in parenthesis) are:

- prevalence of obesity among adults (objective outcome with low measurement error);
- prevalence of current cigarette smoking among adults (availability of good covariate information and possible source for external validation); and
- proportion of children aged 5-15 consuming five portions of fruit and vegetable a day (estimates for a smaller sub-group of the sample and pooled over fewer years).

These three outcomes capture a range of lifestyle outcomes of policy interest and, on technical criteria, are diverse enough to form an assessment of what types of outcomes are suited to synthetic estimation.

## 6.4 Proposed approach to model validation

The models tested at Stage 2 will be subjected to both internal and external validation checks wherever possible.

The internal validation checks will include a range of formal ‘goodness of fit’ measures. The ONS have outlined a number of strategies relevant to testing those models underlying the synthetic estimates they have produced (Brown *et al.*, 2001; Heady *et al.*, 2003), many of which involve testing aggregates of synthetic estimates for larger geographical areas (such as the LADs) against sample-based estimates for these geographical areas (assuming the sample size per area is large enough).

Put briefly internal model diagnostics will include:

- for larger areas of geography such as LADs and PCOs plots of the model-based synthetic estimates (aggregated up to these larger areas) against direct survey estimates. These plots give an indication of possible bias which Brown *et al.* (2001) describe as a ‘bias diagnostic’;
- a ‘coverage diagnostic’ can be used to measure the overlap between the 95% confidence intervals for the two sets of estimates described above. A note of caution however in over reliance on confidence intervals. The danger lies in the case where the underlying model does not provide a good approximation to reality. In the case of model misspecification, an assessment of model error based on the variance alone will be highly misleading; and
- random split samples. This would involve randomly splitting the HSfE sample into two and re-fitting the models for each sub-sample. If our original model was stable, then we would expect the models to be fairly similar for both sub-samples.

For external validation we propose a less formal approach which involves, where possible, comparing synthetic estimates at ward level with direct estimates at ward level from other surveys. The main candidates here are local HSfE boosts – Camden & Islington, and Manchester if available in time – and other survey sources such as the MRC One Million Women study (if we can secure access to the data).

Finally, the set of synthetic estimates that appear to be ‘best’ on these validation checks will, we hope, be checked for ‘face validity’ by a panel of PCOs (see Section 6.5). We will seek advice from the project management committees (particularly the User and Technical Group members) for suggestions on selecting and gaining cooperation from selected PCOs across the country.

## 6.5 User engagement

The project has incorporated a programme of active user consultation and involvement throughout the study, working closely with the User Group and the wider user community. Immediately prior to the project being commissioned, users and stakeholders in policy, local agencies, academia and the wider research community were invited to participate in the first user consultation meeting. Information needs and priorities were discussed and these formed the backdrop to the sorts of issues that have been considered in this scoping study in terms of technical production and the relevance of estimates to users.

In *Stages 2 and 3* of the study we will seek user involvement in three key areas:

- **Result validation: comparing synthetic estimates with direct estimates from local surveys**

Users will be contacted by the User Group to help identify recent local surveys which include information about health behaviours. Analysis to compare the synthetic estimates with available sources of local data will be central to establishing the plausibility of the model-based estimates.

- **'Blind testing': comparing ranks of wards in local areas against local knowledge.**

Participation of groups of local users (or expert panels comprising local stakeholders) will be sought to verify whether the synthetic estimates correspond with local knowledge and experience of 'best/worst' wards in their localities. Blind testing would involve users shading in a blank map of their area, indicating wards that they think will score high, intermediate or low with respect to the distribution of an outcome (say, prevalence of smoking) within their area. Representatives from PCOs and LADs will be approached to volunteer for this exercise. Ideally, we'd like to use this form of face validity test for as wide a mix of different types of areas as possible (inner city, urban, rural, north/south, high ethnic minority concentrations, level of deprivation etc).

- **Expert input into production of user guide.**

We will ask members of the UG and the wider user community to provide scenarios or examples of how they envisage using the synthetic estimates (e.g. comparing wards within PCOs or combining estimates with other data for analysis). Responses to each scenario, indicating how and for what types of analyses synthetic estimates could (or could not) be used, with worked examples and guidance on how to interpret the results will be included in the user guide.

In addition to the above key areas, in Stage 3 (implementation stage) user input will be encouraged via consultations to feed into decisions about the selection of the final (five) outcomes and suggestions for forms of dissemination best suited to reach a dispersed network of users.

## REFERENCES

- Ambler R, Caplan D, Chambers R, Kovacevic M and Wang S (2001) 'Combining unemployment benefits data and LFS data to estimate ILO unemployment for small areas: An application of a modified Fay-Herriot method', *Proceedings of the International Association of Survey Statisticians, Meeting of the International Statistical Institute, Seoul, August 2001*.
- Brown G, Chambers R, Heady P and Heasman D (2001) 'Evaluation of small area estimation methods - an application to the unemployment estimates from the UK LFS' *Statistics Canada Symposium Ottawa, October 2001*.
- Charlton J (1998) 'Use of the Census Sample of Anonymised Records (SARs) and survey data in combination to obtain estimates at local authority level.' *Env & Planning A*, vol 30, pp 775-784.
- Chesterman J, Judge K, Bauld L and Ferguson J (no date) 'How effective are smoking cessation services in reaching smokers in disadvantaged areas in England?', Report to Department of Health.
- Cox D and Hinkley D (1974) *Theoretical Statistics*. London: Chapman and Hall.
- Flowers J (2003) 'Development of a indicator of needs-adjusted statin prescribing at PCO level' (unpublished report, Eastern Region PHO).
- Gibson A, Asthana S (2001) 'Resource allocation methodologies for the prevention and treatment of specific diseases - a critical review'. Discussion paper to ACRA (2001) 08, NHSE.
- Goldstein H (2003) *Multilevel statistical models* (New York: Halstead Press).
- Goldstein H (1995) *Multilevel Statistical Models*, John Wiley & Sons, New York.
- Gilks W, Richardson S and Spiegelhalter D (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
- Heady P, Clarke P and others (2003) *Model-based small area estimation series No 2* (Small Area Estimation Project Report: Office for National Statistics).
- Heady P, Clarke P, Brown G, D'Amore A and Mitchell B (2000) 'Small area estimates derived from surveys: ONS central research and development programme' *Statistics in Transition* (4: 635-648).
- Levy P (1979) 'Small Area Estimation - synthetic and other procedures, 1968-1978' in National Center for Drug Abuse *Synthetic Estimates for Small Areas* (Research Monograph 24: Washington D.C).
- Macintyre S, Ellaway A and Cummins S (2002) 'Place effects on health: how can we conceptualise, operationalise and measure them?', *Soc Sci Med.* 2002 Jul;55(1):125-39.
- Moura F and Holt D (1999) 'Small area estimation using multilevel models' *Survey Methodology* (25: 73-80).
- Rabe-Hesketh S, Skrondal A and Pickles A (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2**, 1-21.

Saei A and Chambers R (2003) 'Small area estimation: A review of methods based on the application of mixed models' *S<sup>3</sup>RI Methodology Working Paper M03/16*.

Schaible W (1996) (ed) *Indirect Estimators in U.S. Federal Programs* (New York: Springer).

Scholes S, Prescott A and Bajekal M (2004) *Health and Lifestyle Indicators for Strategic Health Authorities, 1994-2002* (Department of Health) available at [www.dh.gov.uk/PublicationsAndStatistics/PublishedSurvey/HealthSurveyforEngland/HealthSurveyResults](http://www.dh.gov.uk/PublicationsAndStatistics/PublishedSurvey/HealthSurveyforEngland/HealthSurveyResults)

StataCorp (2003) *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation.

Skinner C (1993) The use of synthetic estimation techniques to produce small area estimates *OPCS New Methodology Series NM18*. (London: Office for Population Censuses and Surveys).

Twigg L, Moon G and Jones K (2000) 'Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators' *Social Science and Medicine* (50: 1109-1120).

Twigg L, Moon G (2002) 'Predicting small area health-related behaviour: a comparison of multilevel synthetic estimation and local survey data', *Social Science and Medicine* 54(6):931-7.