Appendix 4 of 'An Atlas of Tobacco Smoking Scotland', NHS Health Scotland

Smoking Prevalence in Scotland:
2003/4 sub-national estimates.

Technical Supplement to main report for NHS Health Scotland

Graham Moon
Gereltuya Altankhuyag
Steve Barnard
Liz Twigg

University of Portsmouth

April 2006

Introduction

There is little reliable information on smoking prevalence at a sub-national scale and very little indeed for areas smaller than the regional scale. Large national surveys are seldom amenable to generalization below the regional scale. The Scottish Household Survey (SHsS) is one of the more useful examples, being representative of all council areas by aggregating annual results for two years. Local surveys provide at best a partial solution to this data absence problem but such surveys often suffer from poor design and small realized samples; moreover, collectively they do not amount to national coverage.

To bridge this information gap in identifying the prevalence of current smoking, we use multilevel synthetic estimation. This approach allows identification of the numbers of people in a target geography who, given certain assumptions, might be expected to be current smokers. Robust procedures are available to generate multilevel synthetic estimates. This technical supplement provides detail on the approach used in estimating sub-national smoking prevalence data for Scotland. It covers:

- Data sources
- Multilevel synthetic estimation
- Model results
- Estimate quality

Data Sources

The majority of data for the present study were taken from the 2003/4 SHsS[1]. The SHsS is an annual survey commissioned by the Scottish Executive to provide regular information on various topics that cannot be routinely obtained from other sources. It commenced in 1999.

Data were also obtained from the 2001 Decennial Census of Population. Variables were selected that have been shown in previous work to be strongly associated with smoking behaviour at an (area) ecological level.

The multistage clustered design of the SHsS results in a sample of individuals being selected from a sample of postcode sectors. Normally the public data set provides information on the clustering of individuals within postcode sectors without disclosing the identification (and hence location) of the sectors. For this study the research team were allowed access to the identification details of these sectors so that the additional data from the 2001 Census could be merged with SHsS. A working contract was agreed with ISD to ensure the anonymity of survey respondents.

The resultant multi-level structure of the working data set comprised:

- 28,728 individuals, nested in
- 893 postcode sectors, nested in
- 32 council areas.

Multilevel Synthetic Estimation

Multilevel modelling is an extension of the more familiar generalised linear regression model in which we build an equation that estimates a target response variable in terms of a number of candidate predictor variables. In our case the response variable would be whether or not a person smokes, and the predictors might be variables that are thought to relate closely to smoking for example age, sex or social status. Multilevel modelling takes this a stage further, recognising that the chance of an individual being a smoker reflects not only that individual's personal characteristics, but also the characteristics of the environment in which the person lives. The characteristics of multilevel analysis and the associated statistical theory are well documented[2].

Multilevel synthetic estimation applies the basic ideas of multilevel modeling to the task of synthetic estimation – the development of empirically-sound estimates using survey data and statistical analysis. It works with the multilevel sampling structure of the input data set containing, in this case, information on current smoking. We employ an established procedure developed within our research team.[3]

Response variable. Respondents in the 2003 Scottish Household Survey were asked about their current smoking behaviour. Based on answers of this survey question, a binary response variable indicating whether or not an individual was a current smoker was created.

Predictor or explanatory variables. Within a multilevel modelling framework it is permissible to have predictor variables that relate either to individual-level influences on the response, or to higher-level (ecological or area) influences. These two levels can interact.

Individual level. While there are numerous individual level predictors that might be identified from the SHsS, the actual selection is crucially constrained by the subsequent use of the models within a predictive framework. Individual level predictor variables must be present in both the SHsS and the 2001 Census, and must be defined in similar ways. We use complex cross-tabulations of routine local base statistics from the UK Census to provide counts of the numbers of individuals in each ward who fall into particular socio-demographic categories. Following testing, the most detailed crosstabulation available and relevant to smoking behaviour was age (grouped into bands) by marital status by sex. The individual-level explanatory variables used in the models were therefore age, sex and marital status.

Marital status is defined as a dichotomous variable. Those stating they are single, divorced or, widowed or a cohabitee are contrasted with those who are married, married and now separated, or who have remarried. Different arrangements of this classification were tested, and this dichotomy was chosen on the grounds of model parsimony and the need to work with an identical classification from the population census. Age is split into the following bands: 16–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+.

Ecological or area data. Access to the postcode sector identification details of the SHsS survey respondents enabled postcode sector census data to be linked to the SHsS. A range of such 'ecological' variables from the census were attached to the respondents in the SHsS and tested for significance in a multilevel model of current smoking. Using a selection criterion that required an ecological variable to make or be implicated in a statistically significant contribution to the final model (p=0.05), the following variables from the 2001 Census were used, centred on their mean:

- percent household overcrowding (ocrdcr)
- percent households headed by a person in social grade a and b (sgabcr)

- percent households with more than six rooms (rm6cr)
- percent unemployed people (unempcr)
- percent households in local authority / housing association tenure (tenlahacr)
- percent households owning multiple cars (carscr)

Interactions. The model of smoking prevalence also included the individual age, gender and marital status terms, and interactions between any combinations of these (where they were found to be statistically significant). We also allowed statistically significant cross-level interactions between individual and area characteristics variables.

Modelling Procedure. An initial multilevel model was produced using (restricted) iterative generalised least squares with first order maximum quasi-likelihood estimation. We then re-ran this initial best model using second order penalised quasi-likelihood (PQL) estimation. Finally, a Monte Carlo Markov Chain (MCMC) approach was used to refine the model and allow for more robust estimates and standard errors. The MCMC process applied a burn-in of 5000 iterations, followed by a further model setting 200,000 iterations. Once complete, an examination of the model trajectories was carried out to ensure stability in the final model that was used for synthetic estimation purposes

Generating predictions. Multilevel models take into account individual and local influences on the likelihood that an individual will be a current smoker. To generate small-area predictions of smoking prevalence, we apply 2001 census data to the multilevel equations that summarise this likelihood. Predictions were generated by age, sex and marital status for each census output area and then aggregated to the larger geographies of postcode sectors, census area sectors, council areas, Scottish parliamentary constituencies and health boards. Lookup tables ensure that such aggregations can largely be automated; we worked with the 2005 look-up table provided by ISD Scotland. As part of the modelling process, residuals can be identified at each level in the modelled hierarchy. As all council areas are modelled, the council-level residual can be used to improve the estimations.

Model Results

Table 1 shows the parameters of the multilevel model used to generate synthetic estimates of smoking in Scotland.

Table 1: Multilevel smoking model: estimate, standard error (SE) and credible interval, final MCMC model

Levels and variable type	Variables	Coefficient (SE)		95% Credible Interval[1]
	Intercept	-1.011 (0.044)	****	-1.011 to -0.925
Individual terms: Level 1	Male	0.043 (0.042)		-0.039 to 0.126
	Single	0.623 (0.053)	****	0.521 to 0.728
	16-24	0.518 (0.193)	***	0.145 to 0.900
	16-34	-0.023 (0.046)		-0.114 to 0.670
	45+	0.057 (0.052)		-0.045 to 0.158
	55+	-0.196 (0.047)	****	-0.287 to -0.105
	65+	-0.470 (0.052)	****	-0.571 to -0.370

Levels and variable type	Variables	Coefficient (SE)		95% Credible Interval[1]
	75+	-0.726 (0.064)	****	-0.851 to -0.600
Individual terms: two-way Interactions	Male.Single	0.116 (0.059)	**	-0.001 to 0.230
	Male.16-24	-0.269 (0.099)	***	-0.463 to -0.075
	Single.16-24	-0.824 (0.198)	****	-1.213 to -0.440
	Single.45+	-0.130 (0.06)	**	-0.248 to -0.013
Ecological effects: Level 2	ocrdcr	0.017 (0.005)	****	0.006 to 0.027
	sgabcr	-0.019 (0.004)	****	-0.027 to -0.012
	rm6cr	-0.009 (0.003)	***	-0.015 to -0.002
	unempcr	0.015 (0.028)		-0.041 to 0.069
	tenlahac	0.014 (0.003)	****	0.007 to 0.021
	carscr	0.010 (0.004)	**	0.002 to 0.018
Two-way cross level Interactions	16-34.ocrdcr	-0.039 (0.006)	****	-0.051 to -0.028
	45+.sgabcr	0.008 (0.003)	***	0.001 to 0.015
	Male.unempcr	0.052 (0.038)		-0.022 to 0.127
	Single.unempcr	0.081 (0.033)	**	0.017 to 0.145
	16-24.unempcr	-0.127 (0.04)	***	0.206 to -0.048
	16-34.unempcr	0.096 (0.027)	****	0.043 to 0.150
	Male.tenlahac	-0.012 (0.004)	***	-0.021 to -0.004
	Single.tenlahac	-0.012 (0.004)	***	-0.019 to -0.005
	Male.carscr	-0.009 (0.003)	***	-0.016 to -0.002
	16-24.carscr	-0.014 (0.006)	**	-0.026 to -0.002
Three way cross-level Interactions	Male.Single.unempcr	-0.140 (0.049)	***	-0.237 to -0.045
	Male.Single.tenlahac	0.017 (0.005)	****	0.006 to 0.027

**Significant at p<0.05; ***p<0,01; ****p<0,005

[1] Credible intervals are derived via MCMC methods and can be interpreted in much the same way as traditional confidence intervals.

The results of the model indicate that, on converting the logit coefficients to more familiar probabilities, the likelihood of being a current smoker is 27 percent if the person is stereotypical respondent to the SHsS (female, married, aged 35-44, living in an average area in terms of social group, unemployment level, rented tenure, car ownership and house size). There is a small increase in the probability of being a current smoker if the person is male or aged 45-54. The dramatic increases in the baseline probability of smoking occur if the person is single or young (aged 16-24) are modified by interaction effects to the extent that young single women are almost 10% more likely to smoke than their male equivalent.

The model results also suggest that the selected ecological variables have only a small effect on the probability of being a current smoker. People living in areas where there are high percentages of unemployment, of households rented from local authorities or housing agencies, of overcrowded households, and of households with two or more cars have a marginally raised chance of being a current smoker. This finding is generally in accordance with the accepted correlation between smoking and aspects of deprivation. Cross level interaction terms slightly increase or decrease these probabilities.

Overall, it is clear that the main drivers within the model are the individual parameters. Nonetheless ecological variables have small but significant effects and there is a complex pattern of effects related to within and cross-level interaction.

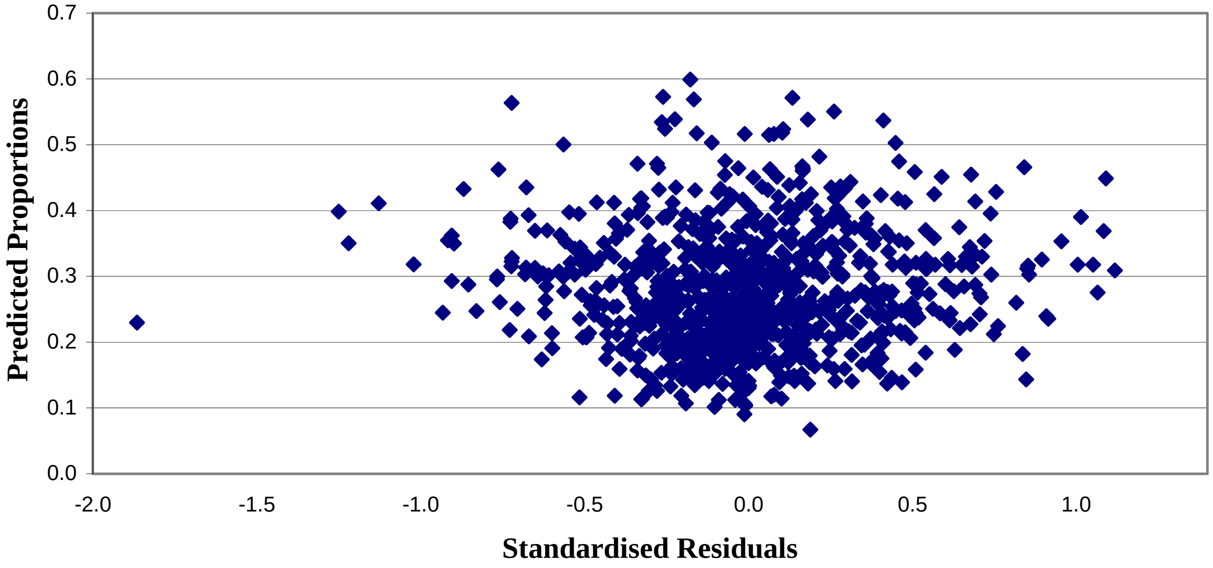
Estimate Quality

Before considering the quality of the estimates, it is important to note the knowledge gain that they provide: a reasoned, localised insight into levels of smoking derived in a consistent manner across the whole of Scotland. Any evaluation of the estimates must also recognise that they depend on two data sources: the SHsS and the 2001 Census as well as the robustness of the modeling process. This multiple dependency means it is problematic to think in terms of traditional notions such as confidence intervals. As a starting point it is essential to note that the data sources are the best available, while also acknowledging that there is differential completion of SHsS questions and a level of non-response in the 2001 Census.

Following the approach adopted by Twigg et al. for a recent study of smoking prevalence and smoking related mortality in England[4] we can move on from these fundamental questions to a more traditional approach to assessing estimate quality. eight different tests have been performed:

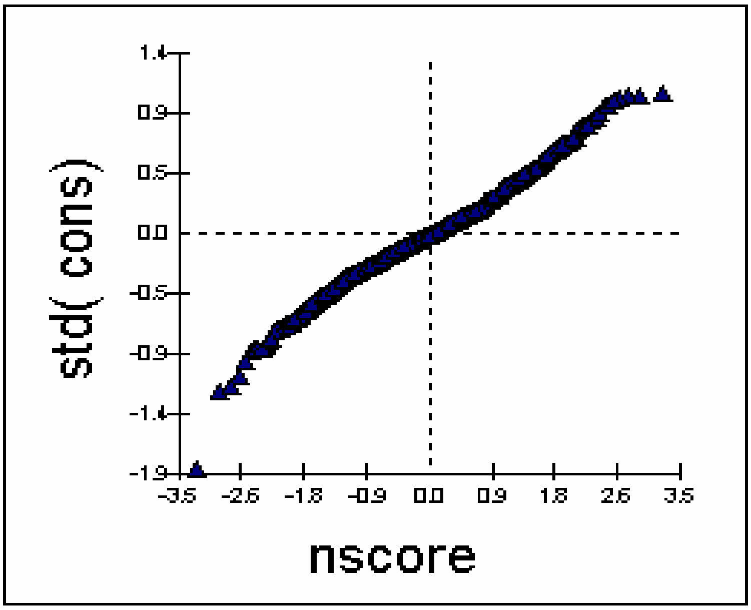
- The method developed by Snijders and Bosker[5] is used to approximate the variance explained by the model. It is found that about 12 percent of the total variance of the current smoking prevalence in Scotland is explained by the model. This is a marginal improvement on the comparable figure obtained in the study of smoking in England.
- The focus in this study is on the small area geography of smoking in Scotland. The effectiveness of the models in capturing this variation can be assessed by considering the percentage reduction in the variance at the postcode sector level between a ‘null’ model with no predictor variables and the full final MCMC model with all predictors. The variance reduction is 85%. This large drop is well in excess of the figure achieved in the English research (68.2%) and suggests that the model is relatively successful at controlling for small area variation.
- We can see how good the model is at predicting the original survey responses using the ‘Percentage Correct Prediction’[6]. The English research was based on a model that correctly predicted smoking status from the input survey on 60.8% of occasions. In the present study the figure is 68%.
- The plot of standardised residuals against the predicted proportions of smokers indicates that there is no clear relationship between standardised residuals and predicted proportions of smokers at the postcode sector level (Figure 1). This suggests a well-specified model.

Figure 1: Standardised residuals versus predicted proportions by post sectors



- The normal score plot of residuals is a straight line on the diagonal, again indicating a well-specified model (Figure 2):

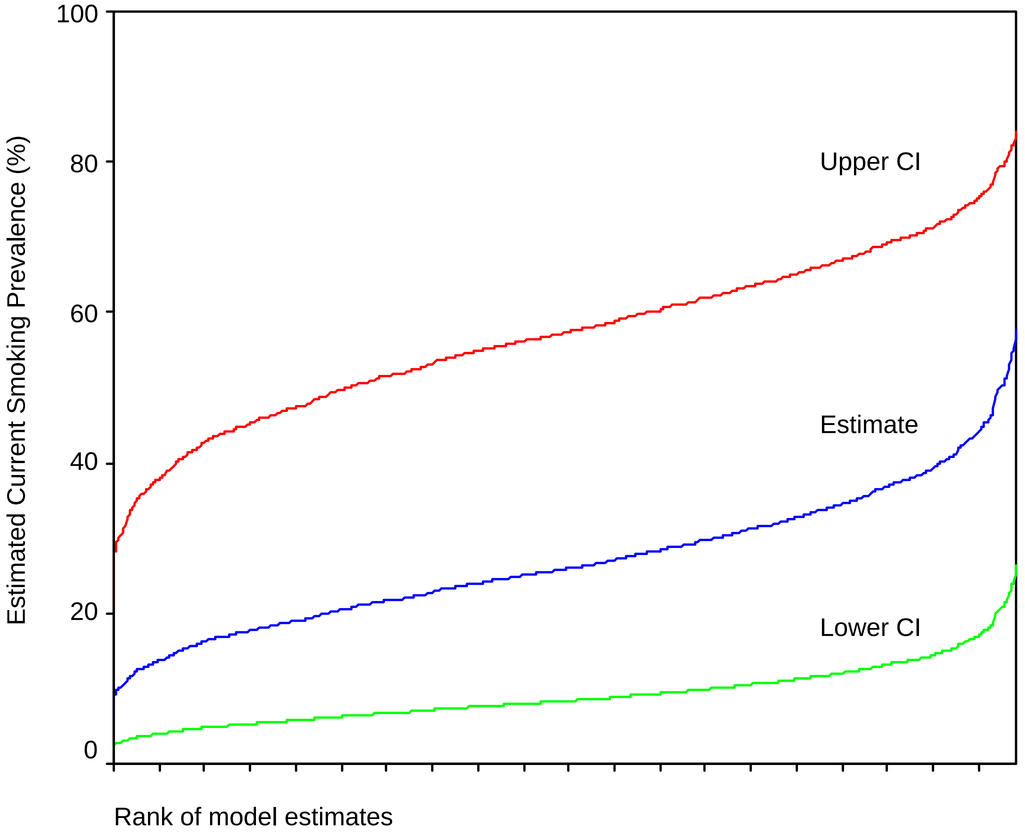
Figure 2: Normal score plot



- To assess model stability the dataset was split a random into two and the model refitted. The model coefficients and standard errors from two random sets were similar (r>0.98) to each other and to the full dataset; the confidence intervals associated with each were overlapping.

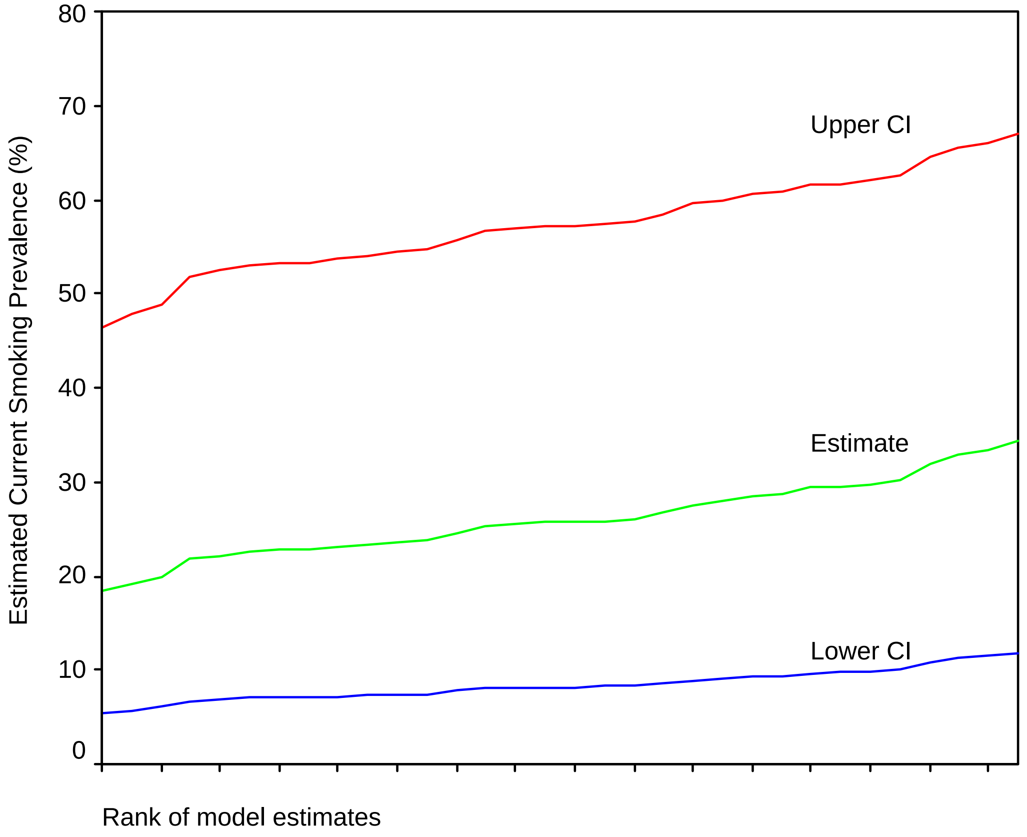
- While we have noted above that it is problematic to think in terms of confidence intervals around our estimates of smoking behaviour, it is possible to use the MCMC process to generate approximate ‘credible intervals’ for the data levels used in the models (Figures 3 and 4). These can be used as guide to possible variations that may be inherent in the estimates at small area level and at higher levels. Estimates for areas not in the model hierarchy are derived by aggregation and possible variation should be approximated from the closest equivalent scale.

Figure 3: Model estimates and 95% credible intervals: postcode sector scale



- As a final assessment of the quality of the estimates, comparisons can be drawn with direct survey estimates, for example from the SHsS. The main report on this project provides these comparisons indicating, at levels where such comparisons are appropriate, that the multilevel synthetic estimates and those derived from the SHsS are in close accord. In comparing the synthetic estimates with those derived directly from the SHsS it should not necessarily be assumed that that one set is any more reliable than the other; the survey data reflect notions of survey representativeness while the synthetic estimates provide measures that indicate what levels of smoking might be expected given the various factors taken into account in the models. An exact match would not be expected.

Figure 4: Model estimates and 95% credible intervals: council area scale



Conclusion

This supplementary report has summarised more technical aspects of the research undertaken to derive small area estimates of current smoking in Scotland. The estimates have been made using well-found established data sources and methods. The quality of the estimates is, in broad terms, equivalent to that achieved in recent comparable research in England. The estimates remain approximations of reality.

References

[1] Further details including data access information are available at http://www.scotland.gov.uk/Topics/Statistics/16002/4031

[2] Goldstein, H. (1995) *Multilevel statistical models.* London: Edward Arnold; Kreft, I. and de Leeuw, J. (1998) *Introducing multilevel modelling.* London: Sage; Leyland, A. and Goldstein, H. (2001) *Multilevel modelling of health statistics*. London: Wiley.

[3] 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(7-8): 1109-20.

[4] Twigg, L., Moon, G. and Walker S. (2004) *The Smoking Epidemic in England*. London: Health Development Agency.

[5] Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel analysis.* London: Sage

[6] Field, A. (2000) *Discovering statistics using SPSS for Windows: advanced techniques for beginners*. London: Sage.