Journal of Data Science **8**(2010), 235-252

Estimating Small Area Diabetes Prevalence in the US Using the Behavioral Risk Factor Surveillance System

Peter Congdon[1] and Patsy Lloyd[2]
[1]*Queen Mary University of London and* [2]*George Washington University*

Abstract: Information regarding small area prevalence of chronic disease is important for public health strategy and resourcing equity. This paper develops a prevalence model taking account of survey and census data to derive small area prevalence estimates for diabetes. The application involves 32000 small area subdivisions (zip code census tracts) of the US, with the prevalence estimates taking account of information from the US-wide Behavioral Risk Factor Surveillance System (BRFSS) survey on population prevalence differentials by age, gender, ethnic group and education. The effects of such aspects of population composition on prevalence are widely recognized. However, the model also incorporates spatial or contextual influences via spatially structured effects for each US state; such contextual effects are allowed to differ between ethnic groups and other demographic categories using a multivariate spatial prior. A Bayesian estimation approach is used and analysis demonstrates the considerably improved fit of a fully specified compositional-contextual model as compared to simpler 'standard' approaches which are typically limited to age and area effects.

Key words: Diabetes, prevalence, random effects, small area, spatial, survey.

1. Introduction

Information regarding area prevalence of diabetes is important for ensuring that resources for diabetes care match need and for effective targetting of diabetes-prevention services. In the US there is evidence of a growth in diabetes levels over time (Mokdad et al, 2001), of wide geographic contrasts in prevalence, and of considerable differences in relative risk between the main ethnic groups (Davidson, 2001; Harris, 1998). Thus in 1999-2000, the age-adjusted US wide prevalence of previously diagnosed diabetes among adults was estimated as 11.7% among blacks, 9.6% among Hispanics, and 4.8% among non-Hispanic whites (CDC, 2003). This paper develops a binary regression model taking account of 2005 survey data, and 2000 US census data, to derive small area prevalence estimates for previously diagnosed diabetes in 32000 small area subdivisions of the US.

These estimates take account of information from the US-wide Behavioral Risk Factor Surveillance System (BRFSS) surveys on prevalence differentials by age, gender, ethnic group and education (e.g. Mukhtar et al, 2003). These surveys are random-digit-dialed telephone survey to determine the prevalence among adults (ages 18 and over) of major illnesses and health behaviors which are related to the leading causes of death in the US. To determine diabetes status, respondents were asked "Have you ever been told by a doctor that you have diabetes?", encompassing both types of diabetes.

The estimates described in this paper are based on around 360,000 survey responses to the 2005 BRFSS, and on a binary regression model expressing the impact on diabetes of major individual level risk factors measured by the survey. However, since the ultimate goal of the analysis is small area prevalence estimation, inclusion of risk factors (and interactions between them) in the model is subject to the constraint that included risks are available also as tabulations for small area populations. The regression model adjusts for US state level relationships between diabetes and the levels of rurality and poverty, and for unmeasured state level influences. The latter are modelled using a multivariate random effects approach that allows state level contextual effects to be differentiated by ethnic group. The areas for which prevalence is estimated are 32000 ZIP Code Tabulation Areas (ZCTAs) for which selected Census 2000 statistics have been provided by the US Census Bureau (cf Grubesic & Matisziw, 2006).

2. Individual Level Risk Factors: Compatibility between Survey and Small Area Variable Frames

The survey regression model for diabetes prevalence includes major individual level risk factors (age, gender, ethnicity, education level) that are known to be significant sources of varying diabetes prevalence. A pronounced gradient in diabetes prevalence by age is reported by CDC (2003) and Mokdad et al (2001), while Maty et al (2005) report that socioeconomic disadvantage, especially low educational attainment, is a significant predictor of incident Type 2 diabetes. Prevalence variations by education level are also reported by CDC (2004). However, since the ultimate goal of the analysis is small area prevalence estimation, inclusion of risk factors (and interactions between them) in the survey model is subject to the constraint that included risks are available both in the BRFSS and as tabulations for ZCTA populations; any assumed interaction between risk factors requires a matching cross-tabulation in the ZCTA population.

Demographic risk categories, namely age group, gender and ethnic group (white non-hispanic, black, hispanic, other) are available both as BRFSS variables and in a ZCTA level tabulation which cross-tabulates adult populations by ethnicity, quinquennial age and gender. For comparably defined demographic

risk groups (e.g. age-ethnic-gender subgroups), parameters from the survey model (e.g. relative risk for hispanic males aged 45-49) can then be transferred to the ZCTA sub-population. As mentioned in the description of the model below, age gradients may vary both by gender and between ethnic groups, and it is important to model such variation while also taking account of correlation between the shapes of age profiles for different groups.

For other individual level risk variables (e.g. education), either primary ZCTA tabulations are available from the 2000 census, or a limited cross tabulation (e.g. male adults by education, and female adults by education), but not tabulations involving cross-hatching against all other risk factors. For example, there is not a ZCTA level census table that cross-tabulates the adult population simultaneously by education, quinquennial age, ethnicity and gender. A small area prevalence adjustment can then be applied only for the main effect of such variables, or for a partial interaction. For example, the survey regression models show gender-specific education gradients in relative risk of diabetes prevalence, and these gradients can be applied to ZCTA male and female adult populations subdivided by education level.

3. Geographic Influences

As is now well known, individual risk factors and contextual factors (including the impact of geographic location) interact in their impact on many chronic diseases. Although prevalence is to be estimated at ZCTA level, the ZCTA of residence is not available for BRFSS respondents for confidentiality reasons, so it is not possible to take account of the impact of (say) poverty rates for ZCTAs on small area diabetes prevalence.

However, one may model the impact of broad scale geographic influences on diabetes prevalence operating at the level of US states, since state of residence ($s = 1, .., 53$, including the District of Columbia, Virgin Islands and Puerto Rico), is available for all respondents. Some directly measured state level predictors may have a significant influence on diabetes prevalence; those used here are the percent of population in poverty and the percentage of rural population. Rural location in the US is in fact a positive risk factor for diabetes prevalence and an adverse influence on access to diabetes care (Mainous et al, 2004; AHRQ, 2005).

Many geographic influences are likely to be unobserved and these are proxied in the regression model by state level random effects. These influences may reflect environmental factors such as climate (Franz & Bailey, 2004), or the aggregate effect of variables representing health behaviours. Such effects are taken to be a sum of two effects, one of which is spatially correlated to reflect smoothly varying risk factors in space that straddle arbitrary state boundaries (Richardson & Monfort, 2000), while the other is unstructured in the sense of not incorporating

spatial structure. In the disease mapping literature this approach, due to Besag et al (1991), is known as a convolution model. Both types of random state effects are differentiated by ethnic group (i.e. are multivariate), since contrasts in diabetes prevalence between ethnic groups are likely to differ by state. Thus CDC (2004) report that "Hispanics continued to have a higher prevalence of diabetes than non-Hispanic whites and that disparities in diabetes between these two populations varied by area of residence". For spatial isolates such as Alaska and Puerto Rico, the impact on prevalence of state of residence is confined to the unstructured random effect.

4. Survey Model Specification

The analysis is based on the 2005 BRFSS survey, with 136 thousand male and 217 thousand female respondents. As well as including relevant risk variables, the model should incorporate survey weights w_i for respondents i to account for differential response between demographic categories, including a lower response rate for males as against females, and for hispanics and blacks as against whites. Because of the large number of respondents, seperate binary regressions are carried out for males and females, and exclude cases with diabetes status not reported or refused - missing status applies to under 0.1% of subjects (CDC, 2008). Separate analysis by gender is also supported by evidence from other studies of gender effect modification over a wide range of risk factors (Cabrera et al, 2003).

Let $y_i = 1$ if a subject reports doctor diagnosed diabetes, with $y_i = 0$ otherwise ($i = 1, .., N$), and define $\pi_i = Pr(y_i = 1)$ as the probability that a respondent reports diagnosed diabetes. The analysis here then follows studies such as Graubard et al (1997) in using a weighted likelihood, namely

$$\prod_i \pi_i^{y_i w_i} (1 - \pi_i)^{w_i (1 - y_i)}.$$

Taking $L_i = w_i[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$ as the weighted log-likelihood for subject i, the total log-likelihood is $L = \sum L_i$. To facilitate straightforward application of survey model parameters across to ZCTA populations a relative risk interpretation was sought for parameters, which is achieved using a log link (Robbins et al, 2002).

The regression model for each gender then involves the following features with associated parameters in brackets:

a) an overall intercept (α),

b) differential risks for black, hispanic and other ethnic groups as against whites as reference (unknowns $\beta_g, g = 2, 3, 4$,with $\beta_1 = 0$ as reference)

c) differential education risks, according to education level e, namely 1=never attended, elementary only, or some high school; 2=high school graduate; 3=some college or technical school; 4=college graduate (unknowns $\eta_e, e = 2, .., 4$, with $\eta_1 = 0$ as reference)

d) effects of state level predictors, namely poverty rate Pov_s and percent rural Rur_s, where $s = 1, .., 53$ denotes the BRFSS respondent's state of residence (δ_1, δ_2). These predictors are centred so that their average over all states is zero.

The parameters under (a)-(d) are 'fixed' effects, whereas also included are random effects to pool strength over ages, states and ethnic groups:

e) differential risks specific to combinations of age group $x = 1, .., X$ and ethnic group $g = 1, .., G$, with $X = 12$ and $G = 4$ (γ_{xg}); the age bands are 18-24,25-29,30-34, ..,70-74, and 75+.

f) spatially correlated effects by state $s = 1, .., 49$ of residence (excluding states 50-53 that are spatial isolates, namely Alaska, Hawaii, Puerto Rico and the Virgin Islands), and ethnic group $g = 1, ..G$, (c_{sg});

g) spatially unstructured effects by state s of residence $(s = 1, .., 53)$ and ethnic group g, (u_{sg}).

Effects under (e) and (f) are modelled via multivariate normal conditional autoregressive priors (of dimension G=4), respectively a multivariate first order random walk and a multivariate spatial scheme (Fahrmeir & Lang, 2001). A constraint is applied during estimation that ensures these effects to sum to zero within ethnic groups, so that $\sum_x \gamma_{xg} = \sum_s c_{sg} = 0$. The area effects u_{sg} under heading (g) are multivariate normal (with means of zero over all states) of dimension G, allowing for correlated effects across ethnic groups, but without any form of autocorrelation over areas. The differentiation of area effects by ethnicity reflects evidence such as that from (CDC, 2004) that disparities in diabetes between ethnic sub-groups in populations vary by area of residence.

Let S_i denote the state of residence for respondent i. Also let $\{x_i, g_i, e_i\}$ denote the age, ethnicity and education level of respondent i. Then one may write the survey prevalence model as

$$\log(\pi_i) = \alpha + \beta_{g_i} + \eta_{e_i} + \delta_1 Pov_{S_i} + \delta_2 Rur_{S_i} + \gamma_{x_i,g_i} + c_{S_i,g_i} + u_{S_i,g_i}, \qquad (4.1)$$

where the c_{sg} terms are not included for Alaska, Hawaii, Puerto Rico and the Virgin Islands. This model is run separately for males and females. For simplicity of presentation, gender $r = 1, 2$ (1=males, 2=females) is omitted from (4.1), but the complete parameterisation has the form

$$\log(\pi_i^{(r)}) = \alpha^{(r)} + \beta_{g_i}^{(r)} + \eta_{e_i}^{(r)} + \delta_1^{(r)} Pov_{S_i} + \delta_2^{(r)} Rur_{S_i} + \gamma_{x_i,g_i}^{(r)} + c_{S_i,g_i}^{(r)} + u_{S_i,g_i}^{(r)}, \quad (4.2)$$

for $i = 1, N_r$ where $N_1 = 135038$ and $N_2 = 217280$.

The parameters in (1) operate on the log relative risk scale. In particular, smoothed state level relative risks by ethnic group ρ_{sg} may be obtained by exponentiating the total area effect, namely $\rho_{sg} = \exp(c_{sg} + u_{sg})$.

Excess risk can be defined in different ways, but one is that the 95% credible intervals for ρ_{sg} are confined to values above 1. The smoothing of state risks under this model follows the general principle of other hierarchical shrinkage methods that the smoothed estimate for each area "borrow strength" (precision) from data in other areas, with shrinkage greater for areas with low event counts. Except in the spatially isolated states, two forms of smoothing are invoked: local smoothing towards the average of neighbouring states, and global smoothing of all prevalence risks toward the same US wide mean (Clayton & Kaldor, 1987). The smoothing is multivariate and so also incorporates a within state correlation between prevalence rates of different ethnic groups. Such smoothed prevalence estimates are more precise and more robust against false-positive inferences (e.g. regarding excess risk) than are unpooled prevalence rate estimators.

Some concerns have been raised that Bayesian risk estimates may tend to oversmooth variations in disease or mortality risks, particularly when data are sparse or there are discontinuities in the spatial pattern of risk (Green and Richardson, 2002). For such reasons, the area units for the survey model have been chosen as US states rather than US counties (of which there are circa three thousand across the US) to avoid data sparseness. As for distortion due to discontinuities (states with prevalence unlike that of their neighbours), these are reduced by including unstructured effects u_{sg} as well as spatially structured effects c_{sg}. Assuming local smoothing via spatially configured c_{sg} as the sole relevant principle guiding smoothed prevalence estimation is inappropriate when there are discontinuities. It is possible that more elaborate "adaptive" priors (e.g. Congdon, 2007) could be applied to account for any discontinuities. However, it is important to use information on geographic adjacencies, since some spatial pooling of strength is likely to be relevant. Accumulated evidence indicates a clear spatial patterning in US diabetes prevalence, and in mortality from diabetes and related conditions, with elevated diabetes prevalence in the south eastern US, and lower prevalence in the mountain and northern states (see for example Ahluwalia et al, 2003, Table 20). Such evidence supports the inclusion of a mechanism for spatial pooling of strength in the survey model.

5. Model with Age and State Effects Only

To provide a benchmark against more conventional prevalence rate estimation approaches and assess the gain in fit (if any) from using the detailed model in (1), we also consider a simple approach (though still a model) with age and state effects only. This provides estimates of relative diabetes risk for different states

that adjust only for differences in population age structure between states.

Under this simplified model, the model for respondents i (again within each gender) is

$$\log(\pi_i) = \alpha + \gamma_{x_i} + u_{S_i}, \tag{5.1}$$

where the age parameters γ_x are fixed effects with $\gamma_1 = 0$ for identification, and the log relative risks $\{u_s, s = 1, .., 53\}$ for states are unstructured normal random effects with zero mean. Age adjusted state prevalence rates for each gender are obtained from this model as $\rho_s = \exp(\alpha + u_s)$.

Thus a model based approach to estimating geographic relativities is retained under this simpler option, but this model is similar to conventional estimation techniques for obtaining age-adjusted prevalence rates for states. Note that in the conventional demographic approach, state rates are in effect treated as 'fixed effects' parameters, though the implicit statistical assumptions are typically not stated.

6. Small Area Prevalence Estimates

To translate the survey model parameters into small area estimates requires disaggregated populations that match the risk categorisations used in that model. Thus let S_j denote the state in which ZCTA j is located, with $j = 1, ..., m_s$ and with $\sum_s m_s = 31986$, the total number of ZCTAs across the US. From the estimates of the full survey prevalence model parameters, one may extract ZCTA level estimated prevalence probabilities (here called rates for simplicity) specific for age group, ethnicity and gender r as

$$p^{(r)}_{jxg} = \exp(\alpha^{(r)} + \beta^{(r)}_g + \delta^{(r)}_1 Pov_{S_j} + \delta^{(r)}_2 Rur_{Sj} + \gamma^{(r)}_{xg} + c^{(r)}_{S_jg} + u^{(r)}_{S_jg}), \tag{6.1}$$

and these may be applied to gender-specific populations P_{jrxg} for ZCTA areas to obtain estimated prevalence totals.

Summary ethnic specific rates may be obtained by weighting the age bands according the 2000 US Standard Population (National Cancer Institute, 2008). Thus with weights $\{w_x, x = 1, .., X\}$ for the $X = 12$ adult age bands in the diabetes prevalence model, and subject to $\sum_x w_x = 1$, overall diabetes prevalence rates for the four ethnic groups in ZCTA j are

$$p^{(r)}_{jg} = \sum_x w_x p^{(r)}_{jxg}. \tag{6.2}$$

One may adjust the estimated rates (6.1) and (6.2) to take account of the impact on diabetes prevalence of the education attainment mix in each ZCTA. The education mix in a small area is one measure of the impact of socioeconomic

structure on health outcomes (cf. Catelan et al, 2008). Thus, let $h_{je}^{(r)}$ be the 2000 census data relative proportions at education level e in each gender's adult population in ZCTA j with $\sum_e h_{je}^{(r)} = 1$. Also let $\varsigma_e^{(r)} = exp(\eta_e^{(r)})$ be the survey model estimate of the relative diabetes risk at education level e after controlling for age, ethnicity and gender. Then a measure of relative risk associated with the educational mix in the j^{th} ZCTA is $H_j^{(r)} = \sum_e h_{je}^{(r)} \varsigma_e^{(r)}$, and the overall ethnic prevalence rates adjusted for education mix are $p_{jg}^{\prime(r)} = H_j^{(r)} p_{jg}^{(r)}$.

Table 1: Fixed effect coefficients, full survey model

	Males				Females			
	Mean	2.5%	97.5%	Relative Risk	Mean	2.5%	97.5%	Relative Risk
α	-2.69	-2.71	-2.66		-2.56	-2.60	-2.52	
Ethnic Coefficients (log relative risk)*								
β_1	0.00			1.00	0.00			1.00
β_2	0.44	0.41	0.47	1.56	0.66	0.61	0.72	1.93
β_3	0.13	0.09	0.18	1.14	0.31	0.23	0.40	1.36
β_4	0.06	0.01	0.10	1.06	0.28	0.24	0.34	1.33
Education Coefficients (log relative risk)**								
η_1	0.00			1.00	0.00			1.00
η_2	-0.08	-0.13	-0.05	0.92	-0.30	-0.33	-0.26	0.74
η_3	-0.04	-0.10	0.01	0.96	-0.39	-0.43	-0.35	0.68
η_4	-0.39	-0.44	-0.34	0.68	-0.86	-0.91	-0.81	0.42
State Predictors								
δ_1	0.0115	0.0066	0.0147		0.0098	0.0041	0.0137	
δ_2	0.0028	0.0009	0.0039		0.0012	-0.0002	0.0026	

* 1= White; 2= Black; 3= Hispanic; 4= Other

** 1= No school, elementary only, or some high school without graduating; 2= High school graduate; 3= Some college; 4= College Graduate

7. Model Results

Fitting of the models (4.1-4.2) and (5.1) and assessment of their goodness of fit follows a Bayesian approach, under which existing evidence on parameters is expressed via prior densities on such parameters, with posterior evidence provided by combining the prior evidence with the observed data. A Bayesian strategy is advantageous for estimating models with several sets of random effects, including random effects which are spatially clustered. Goodness of fit (see Appendix 1 for details) is assessed by the DIC (Spiegelhalter et al, 2002) and an approximate marginal likelihood (Ibrahim et al, 2001), while ability of the model

to reproduce the data is assessed via a posterior predictive check involving the deviance $D = -2L$ (e.g. Lynch & Western, 2004). A model will be preferred if it both (a) successfully reproduces the data and (b) has best fit among those models compatible with the data. Estimation uses iterative Monte Carlo Markov Chain (MCMC) sampling methods (Gelfand and Smith, 1990), as provided in the WINBUGS program (Lunn et al, 2000). Prior specifications are considered in Appendix 2. Posterior summaries of parameters are based on the 2^{nd} half of runs of 5000 iterations, using two chains starting from dispersed starting values. Convergence was achieved in all models using Brooks-Gelman-Rubin criteria (Brooks & Gelman, 1998).

Table 1 shows gender-specific estimates of the fixed effect parameters $\{\alpha, \beta_g, \eta_e, \delta_k\}$ from the full survey model (1). It can be seen that there is a steeper educational gradient for females, for whom the relative risk for college graduates of $exp(\eta_4) = 0.42$ is under a half that of the first education category, those with limited education (elementary education only or did not graduate from high school). There are also clearly significant ethnic effects for both genders, with elevated relative risk for black and hispanic persons.

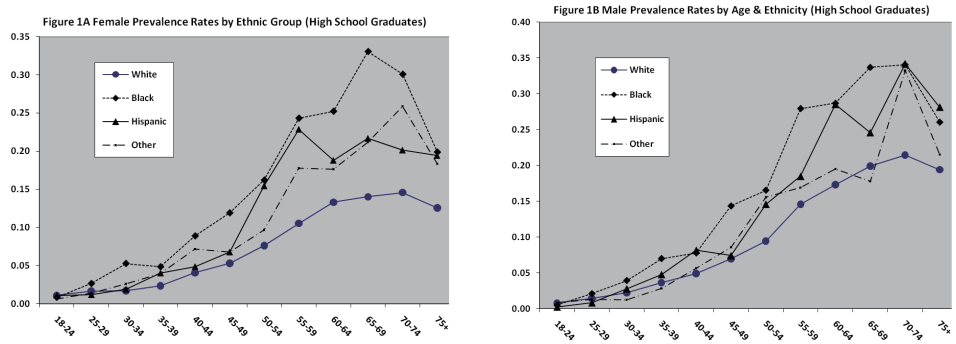

Figure 1: Left panel: Female Prevalence Rates by Age and Ethnic Group (High School Graduates); right panel: Male Prevalence Rates by Age and Ethnic Group (High School Graduates)

Age profiles for the four ethnic groups in a typical state (one with average poverty and rurality levels), with the rates also specific for education level e, are obtainable as

$$p_{xge} = \exp(\alpha + \beta_g + \gamma_{xg} + \eta_e).$$

For example, Figure 1, left panel and right panel show estimated age prevalence profiles differentiated by ethnic group, with the rates specific for high school graduates, obtained via

$$p_{xg2} = \exp(\alpha + \beta_g + \gamma_{xg} + \eta_2),$$

where η_2 is the parameter for high school graduates. Peak rates for nonwhite groups occur for slightly younger age bands than the oldest age band in the model, namely the over 75s. This may reflect cohort effects (Gilliland et 1997), linked to the sharp rise in diabetes prevalence since the 1950s.

The overall age adjusted prevalence for ethnic groups g at education level e is obtainable (for a state with average poverty and rurality) as $p_{ge} = \sum_x w_x p_{xge}$.

Table 2 contains posterior summaries (expressed as percents) for the p_{ge} over the four ethnic groups and four education levels. The widest contrast is among women, exemplified by the rates for white, college-educated women (posterior mean prevalence of 0.037), as opposed to black women with limited education (posterior mean percent prevalence of 0.176).

Table 2: Age adjusted diabetes prevalence (percent) 2005, by ethnic group and education level

	Male			Female		
Education	Mean	2.5%	97.5%	Mean	2.5%	97.5%
White						
Limited*	8.6	7.9	9.1	8.9	8.6	9.2
High school graduate	7.7	7.3	8.1	6.5	6.4	6.7
Some college/technical school	8.1	7.6	8.5	6.0	5.8	6.1
College graduate	5.6	5.3	5.9	3.7	3.6	3.9
Black						
Limited*	13.7	12.5	14.7	17.6	16.9	18.3
High school graduate	12.1	11.4	13.0	12.9	12.4	13.4
Some college/technical school	12.7	11.8	13.6	11.8	11.3	12.3
College graduate	8.8	8.0	9.3	7.3	6.9	7.8
Hispanic						
Limited*	13.3	12.1	14.7	14.3	13.8	15.1
High school graduate	11.9	10.7	13.0	10.5	10.1	11.1
Some college/technical school	12.5	11.2	13.6	9.6	9.2	10.1
College graduate	8.6	7.6	9.5	6.0	5.7	6.3
Other						
Limited*	10.8	9.4	12.8	13.6	13.0	14.2
High school graduate	9.6	8.4	11.3	10.0	9.6	10.5
Some college/technical school	10.0	8.8	11.8	9.1	8.7	9.6
College graduate	6.9	6.0	8.1	5.7	5.3	6.0

* Never attended, elementary only, or some high school.

State relative risks ρ_{sg} for diabetes among males and females may be obtained by exponentiating the total area effects $c_{sg}+u_{sg}$ by ethnic groups g. These amount to residual effects after controlling for the age and educational composition of state populations, and also for state levels of poverty and rurality. Despite this

Table 3: Highest and lowest state relative risks by sex-ethnic category

White (Non-Hisp) Males		White (Non-Hisp) Females	
Montana	0.73	Minnesota	0.75
Wyoming	0.78	Montana	0.82
Hawaii	0.83	Arizona	0.83
New Mexico	0.85	Colorado	0.84
Alaska	0.85	Rhode Island	0.88
Alabama	1.13	Nebraska	1.18
Georgia	1.13	Indiana	1.19
Vermont	1.14	West Virginia	1.20
Tennessee	1.14	New Hampshire	1.23
Delaware	1.28	Tennessee	1.23
Black Males		**Black Females**	
California	0.54	Utah	0.70
Wyoming	0.59	District of Columbia	0.77
Oregon	0.60	Iowa	0.79
Arizona	0.64	New Mexico	0.82
Nevada	0.65	Nevada	0.82
Nebraska	1.35	Massachusetts	1.27
Florida	1.39	N Dakota	1.30
Vermont	1.48	Vermont	1.36
Illinois	1.73	Maine	1.42
Kentucky	1.86	New Hampshire	1.51
Hispanic Males		**Hispanic Females**	
Montana	0.53	Nebraska	0.73
Washington	0.55	Minnesota	0.75
Iowa	0.57	Wyoming	0.75
N Carolina	0.61	Kansas	0.75
Utah	0.62	Tennessee	0.75
S Dakota	1.45	Georgia	1.49
District of Columbia	1.47	Connecticut	1.56
Vermont	1.51	New Hampshire	1.61
Indiana	1.78	Maine	1.69
Georgia	2.03	Rhode Island	1.72

there are consistent patterns, such as multiple elevated area impacts (two or more ρ_{sg} significantly above 1, and none significantly below 1) in Maine and Georgia, and multiple diminished area impacts (two or more ρ_{sg} significantly below 1, and none above 1) in Colorado, Iowa, Louisiana, Nevada, North Carolina, Utah, Wisconsin and Wyoming. Table 3 shows states with the lowest and highest posterior mean ρ_{sg} for groups formed according to sex and ethnicity; it is apparent that low relative risks tend to be concentrated in the mountain states, and high

risks in the south and east, and also that risk contrasts are greater for blacks and hispanics than for white non-hispanics.

For estimates at ZCTA level, one important feature is measures of variation across areas and demographic groups. Thus ranges under model (1) in posterior mean ethnic group prevalences p'_{jg} (i.e. adjusted for education mix) are lowest for whites. The minima and maxima posterior mean p'_{jg} are {0.048,0.121} for white males and {0.027,0.153} for white females. By contrast, for black males and black females the extrema are {0.056,0.263} and {0.055,0.240}.

A summary expression of state level geographic differentials applicable across all ethnic groups is obtainable from the additive age and area effects model (5.1). The simplicity of this model is appealing, and it is sufficient to reproduce the data according to the posterior predictive check based on the deviance (see Table 4). However, there is a clear deterioration in fit compared to model (1), both in terms of a lower marginal likelihood and higher DIC.

Table 4: Model fit, full and simple survey models

	Model 1 (Full Model)		Model 2 (Simple Age-Area)	
Criterion	Males	Females	Males	Females
log(pseudo marginal likelihood)	-32294	-50964	-33057	-52553
Deviance($\bar{\theta}$)	64379	101543	65975	104996
Mean Deviance	64515	101644	66001	105028
Effective Parameters	136	101	26	32
DIC	64651	101745	66027	105060
Posterior Predictive Check	0.41	0.45	0.57	0.60

Despite its worse fit, it is of interest to consider the state level relative prevalence risks $\rho_s = exp(u_s)$ obtained from model (5.1), which are adjusted for age, but not adjusted for population differences in ethnic composition and education levels, or for state poverty or rurality measures; see Table 5 for a summary of highest and lowest state level relative risks according to sex. High relative risks, namely those significantly exceeding 1 (in the sense that the 95% credible interval is confined to values over 1), occur in several southern states (Alabama, Georgia, Louisiana, Mississippi, North and South Carolina) as well as in Puerto Rico and Oklahoma. Low relative risks, those significantly under 1, occur in west central and northern states such as Colorado, Montana, North Dakota, Wisconsin, Alaska, Rhode Island and Massachusetts. A pattern with some similarities (albeit for crude rates, not adjusted for age) is reported by the CDC at http://apps.nccd.cdc.gov/gisbrfss/default.aspx.

Table 5: Area effects (relative risks) from simple age-area model, posterior summary

Male				Female			
	Mean	2.5%	97.5%		Mean	2.5%	97.5%
Colorado	0.80	0.70	0.92	Minnesota	0.71	0.62	0.74
Montana	0.82	0.68	0.91	Vermont	0.73	0.67	0.85
Connecticut	0.83	0.76	1.01	Rhode Island	0.73	0.66	0.79
Alaska	0.85	0.70	0.98	Montana	0.76	0.72	0.88
Massachusetts	0.88	0.80	0.99	Colorado	0.78	0.74	0.92
Mississippi	1.15	1.06	1.28	Texas	1.33	1.26	1.45
Georgia	1.19	1.10	1.28	S Carolina	1.43	1.27	1.46
Alabama 1	.22	1.14	1.34	Mississippi	1.48	1.34	1.61
S Carolina	1.26	1.17	1.39	Virgin Islands	1.52	1.38	1.62
Puerto Rico	1.42	1.29	1.54	Puerto Rico	1.83	1.66	1.92

8. Conclusion

Variations in prevalence of chronic diseases between geographic areas will reflect variations in the attributes of area populations, sometimes termed 'compositional' effects due to the demographic and social structure of area populations (Duncan et al, 1998). However, prevalence variations are also likely to show spatial structure, reflecting what are sometimes termed 'contextual' effects (Sacker et al, 2006), or unobserved risk factors that vary smoothly over space (Richardson & Monfort, 2000). Such contextual effects are likely to be differentiated between ethnic groups and other demographic categories.

This paper has presented a binary regression model that takes account of individual level risk factors and the spatial context for a particular chronic disease, diabetes. Contextual effects are represented by spatially structured and unstructured area random effects, as well as by known state level influences such as poverty levels. Area random effects are differentiated by ethnic group, reflecting evidence from other sources that ethnic relativities are not constant spatially. Age effects are also differentiated by ethnic group using a multivariate autoregressive prior.

Elaborations to the model presented in (1) are possible, such as state as well as ethnic group differentiation in age gradients, or state differentiation in education gradients. One might also consider spatially varying priors for the impacts of the known state level predictors, such as state poverty rate (Gamerman et al, 2003). Varying impacts of such predictors by ethnic group or age are also possible, if for instance, poverty has a greater influence on middle age prevalence contrasts. However, model variations are constrained to some extent in that

the ultimate goal of the analysis is small area prevalence estimation, so that inclusion of risk factor interactions is subject to the constraint that any assumed interaction between risk factors requires a matching cross-tabulation in the small area population.

The greatly improved fit for a model that includes both major individual risk factors, and a full specification for contextual factors whether known or unobserved, has been demonstrated. Results for the full model (1) show significant spatial effects (Table 3) even after adjusting for age, education, ethnicity and known state predictors. This may reflect climatic influences (Franz and Bailey, 2004), unmeasured behavioral influences or the effectiveness of health care systems.

Appendix 1: Assessing goodness of fit

Comparisons of model fit are based on the Deviance Information Criterion (DIC) of Spiegelhalter et al (2002), and an approximate marginal likelihood, denoted the pseudo marginal likelihood. The DIC criterion is obtained as the posterior mean deviance (minus twice the log likelihood) plus a measure of complexity d_e. The latter is in turn derived as the difference between the mean deviance $\overline{D}$ over MCMC iterations and the deviance $Dev(\bar{\theta})$ at the posterior mean of the parameter set θ. Lower values of the DIC indicate better fitting models.

The pseudo marginal likelihood is based on Monte Carlo estimates of the conditional predictive ordinate or CPO, $p(y_i|y_{[i]})$, where $y_{[i]}$ denotes the dataset with the i^{th} subject excluded (Dey et al, 1997). The conditional predictive ordinate amounts to a cross validation measure for each case, with the remainder of the data forming the 'test data'. Totalling the logs of the $CPOs$ over all cases provides the logged pseudo marginal likelihood, and models with higher log pseudo marginal likelihoods provide better fits (Ibrahim et al, 2001).

The ability of models to reproduce the data is assessed via a posterior predictive check involving the deviance $D = -2L$ (e.g. Lynch & Western, 2004). Let $y_{new,i}$ be replicates (predictions) sampled from the posterior predictive density $p(y_{new,i}|y)$. Then at each MCMC iteration $t = 1, .., T$ the deviances $\{D^{(t)}_{obs}, D^{(t)}_{new}\}$ are obtained using the likelihoods $L^{(t)}_i = w_i[y_i \log \pi^{(t)}_i + (1 - y_i) \log(1 - \pi^{(t)}_i)]$ and $L^{(t)}_{i,new} = w_i[y^{(t)}_{new,i} \log \pi^{(t)}_i + (1 - y^{(t)}_{new,i}) \log(1 - \pi^{(t)}_i)]$. The posterior predictive check involves comparing $D^{(t)}_{obs}$ and $D^{(t)}_{new}$, and in particular the indicator $C^{(t)} = I(D^{(t)}_{new} < D^{(t)}_{obs})$ where $I(A) = 1$ when A is true and $I(A) = 0$ when A is false. Posterior predictive p-values $\sum_t C^{(t)}/T$ exceeding 0.9 or under 0.1 are generally regarded as casting doubt on the model (Meng, 1994)

Appendix 2: Prior Assumptions

For the fixed effects parameters, namely $\{\alpha, \beta_g, \eta_e\}$ in model (1), and $\{\alpha, \gamma_x\}$ in model 2 diffuse normal priors with mean zero and variance 1000 are adopted. For the G-dimensional spatially structured area effects $c_s = (c_{s1}, .., c_{sG}), s = 1, .., L$, in model (1) over the $L = 49$ mainland US states and DC, a multivariate pairwise-difference prior is adopted (Rue & Held, 2005). This has joint form $p(c|\Pi) = \left(\frac{1}{2\pi}\right)^{LG/2} |\Pi|^{0.5} exp[-0.5 \sum_{s \sim u} c_s' \Pi_{su} c_u]$, where $c = (c_1, .., c_L)$, Π is a $LG \times LG$ joint precision matrix with $G \times G$ submatrices Π_{su}, and $s \sim u$ denotes summation over states s and u that are contiguous. Diagonal submatrices of Π are $\Pi_{ss} = d_s \Omega_c$, where d_s is the number of states adjacent to state s, and the $G \times G$ matrix Ω_c^{-1} represents within state covariation between prevalence effects for ethnic groups. The off-diagonal submatrices of Π are zero except when states s and u are neighbours, when $\Pi_{su} = -\Omega_c$. The precision matrix Ω_c is assigned a Wishart prior with identity scale matrix and G degrees of freedom, following the strategy of Natarajan & Kass (2000) and Chib & Winkelmann (2001).

To pool strength across the age profiles of different ethnic groups, a low order multivariate random walk prior may be adopted for the G-dimensional vector $\gamma_x = (\gamma_{1x}, .., \gamma_{Gx}), x = 1, .., X$. For example, first and second order random walk priors have conditional forms

$\gamma_x \sim N_G(\gamma_{x-1}, \Omega_\gamma^{-1})$,

$\gamma_x \sim N_G(2\gamma_{x-1} - \gamma_{x-2}, \Omega_\gamma^{-1})$, where the $G \times G$ matrix Ω_γ^{-1} represents covariation between age mortality profiles of demographic groups. In model (1) a first order random walk was used, and the precision matrix Ω_γ is assigned a Wishart prior with identity scale matrix and G degrees of freedom, namely $\Omega_\gamma \sim Wish(I, G)$.

The multivariate unstructured area effects u_{sg} in model (1) are assigned a multivariate normal prior with G-dimensional mean vector of zero and $G \times G$ precision matrix Ω_u, with prior $\Omega_u \sim Wish(I, G)$. For the univariate spatially unstructured random area effects in model (5.1), it is assumed that $u_s \sim N(0, 1/\tau_u)$, where $\tau_u \sim Ga(1, 0.001)$, where the choice of gamma prior for τ_u follows the strategy of studies such as Besag et al (1995) and Gschlößl & Czado (2008).

References

Agency for Healthcare Research and Quality (AHRQ) (2005). *Health Care Disparities in Rural Areas: Selected Findings From the 2004 National Healthcare Disparities Report.* AHRQ Pub No. 05-P022

Ahluwalia, I., Mack, K., Murphy, W., Mokdad, A. and Bales, V. (2003). State-specific prevalence of selected chronic disease-related characteristics — Behavioral Risk

Factor Surveillance System, 2001. *MMWR Surveill Summ* **52**,1-80.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 1-59.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3-66.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434-45.

Cabrera, C., Wilhelmson, K., Allebeck, P., Wedel, H., Steen, B. and Lissner, L. (2003). Cohort differences in obesity-related health indicators among 70-year olds with special reference to gender and education. *Eur. J. Epidemiol.* **18**, 883-890.

Catelan, D., Biggeri, A. and Lagazio, C. (2008). On the clustering term in ecological analysis: how do different prior specifications affect results? *Statistical Methods and Applications*, online version (doi: 10.1007/s10260-007-0089-x)

Center for Disease Control and Prevention (CDC). (2003). Prevalence of Diabetes and Impaired Fasting Glucose in Adults - United States, 1999-2000. *MMWR* **52**, 833-837.

Center for Disease Control and Prevention (CDC). (2004). Prevalence of Diabetes Among Hispanics — Selected Areas, 1998–2002. *MMWR* **53**, 941-944.

Center for Disease Control and Prevention (CDC). (2008). 2005 BRFSS Codebook. http://www.cdc.gov/brfss/technical_infodata/surveydata/2005.htm

Chib, S. and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *J. Business Econ. Statist.* **19**, 428-435.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671-682.

Congdon, P. (2007). Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes. *Computational Statistics and Data Analysis* **51**, 3197-3212.

Davidson, M. (2001). The disproportionate burden of diabetes in African-American and Hispanic populations. *Ethn. Dis.* **11**,148-51.

Dey, D., Chen, M-H. and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**, 1239-1252.

Duncan, C., Jones, K. and Moon, G. (1998). Context, composition and heterogeneity: using multilevel models in health research. *Social Science & Medicine* **46**, 97-117.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C* **50**, 201-220.

Franz, K. and Bailey, S. (2004). Geographical variations in heart deaths and diabetes: effect of climate and a possible relationship to magnesium. *J. Am. Coll. Nutr.* **23**, 521S-524S.

Gamerman, D., Moreira, A. and Rue, H. (2003). Space-varying regression models specifications and simulation. *Comput. Statist. Data Analysis* **42**, 513-533.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculate marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gilliland, F,, Owen, C., Gilliland, S. and Carter, J. (1997). Temporal trends in diabetes mortality among American Indians and Hispanics in New Mexico: birth cohort and period effects. *Am. J. Epidemiol.* **145**, 422-431.

Graubard, B., Korn, E. and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 170-174.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association* **97**, 1055-1070.

Grubesic, T. and Matisziw, T. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *Int. J. Health Geogr.* 2006 Dec. 13; 5, 58.

Gschlöß. l. S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers* **49**, 531-552.

Harris, M. (1998). Diabetes in America: epidemiology and scope of the problem. *Diabetes Care* **21** (suppl. 3), C11-14.

Ibrahim, J., Chen, M-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325-337.

Lynch, S. and Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research* **32**, 301-335.

Mainous, A., King, D., Garr, D. and Pearson, W. (2004). Race, rural residence, and control of diabetes and hypertension. *Ann. Fam. Med.* **2**, 563-568.

Maty, S., Everson-Rose, S., Haan, M., Raghunathan, T. and Kaplan. G. (2005). Education, income, occupation, and the 34-year incidence (1965–99) of Type 2 diabetes in the Alameda County Study. *International Journal of Epidemiology* **34**, 1282-3.

Meng, X-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142-1160.

Mokdad, A., Ford, E., Bowman, B., Nelson, D., Engelgau, M., Vinicor, F. and Marks, J. (2001) Diabetes trends in the U.S: 1990-1998. *Diabetes Care* **24**, 1278-1283.

Mukhtar, Q., Murphy, D. and Mitchell, P. (2003). Use of data from the behavioral risk factor surveillance system optional diabetes module by states. *J. Public. Health Manag. Pract.* Suppl,S52-5.

Natarajan, R. and Kass, R. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.* **95**, 227-237.

National Cancer Institute (2008). 2000 US standard population vs. standard million. http:// seer.cancer.gov/stdpopulations/single_age.html Richardson

Monfort, C. (2000) Ecological correlation studies. In *Spatial Epidemiology Methods and Applications* (Edited by P. Elliott, J. Wakefield, N. Best and D. Briggs). Oxford University Press.

Robbins, A., Chao, S, and Fonseca, V. (2002). What's the relative risk? a method to directly estimate risk ratios in cohort studies of common outcomes. *Annals of Epidemiology* **12**, 452-454.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

Sacker, A., Wiggins, R. and Bartley, M. (2006). Time and place: putting individual health into context. A multilevel analysis of the British household panel survey, 1991–2001 *Health & Place* **12**, 279-290.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A (2002). Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. B* **64**, 583-639.

Received May 10, 2008; accepted September 10, 2008.

Peter Congdon
Department of Geography and Center for Statistics
Queen Mary University of London
Mile End Rd
London E1 4NS, England
p.congdon@qmul.ac.uk

Patsy Lloyd
Department of Epidemiology and Biostatistics
Ross Hall Suite 125
School of Public Health and Health Services
George Washington University
Washington, DC 20037, USA
plloyd@gwu.edu