Journal of Data Science **6**(2008), 15-32

Indirect Area Estimates of Disease Prevalence: Bayesian Evidence Synthesis with an Application to Coronary Heart Disease

Peter Congdon
Queen Mary University of London

Abstract: Risks for many chronic diseases (coronary heart disease, cancer, mental illness, diabetes, asthma, etc) are strongly linked both to socio-economic and ethnic group and so prevalence varies considerably between areas. Variations in prevalence are important in assessing health care needs and in comparing health care provision (e.g. of surgical intervention rates) to health need. This paper focuses on estimating prevalence of coronary heart disease and uses a Bayesian approach to synthesise information of different types to make indirect prevalence estimates for geographic units where prevalence data are not otherwise available. One source is information on prevalence risk gradients from national health survey data; such data typically provide only regional identifiers (for confidentiality reasons) and so gradients by age, sex, ethnicity, broad region, and socio-economic status may be obtained by regression methods. Often a series of health surveys is available and one may consider pooling strength over surveys by using information on prevalence gradients from earlier surveys (e.g. via a power prior approach). The second source of information is population totals by age, sex, ethnicity, etc from censuses or intercensal population estimates, to which survey based prevalence rates are applied. The other potential data source is information on area mortality, since for heart disease and some other major chronic diseases there is a positive correlation over areas between prevalence of disease and mortality from that disease. A case study considers the development of estimates of coronary heart disease prevalence in 354 English areas using (a) data from the Health Surveys for England for 2003 and 1999 (b) population data from the 2001 UK Census, and (c) area mortality data for 2003.

Key words: Coronary heart disease, deprivation, ethnicity, power prior, prevalence. mortality, small area.

1. Introduction: Need for Spatially Disaggregated Prevalence Estimates

Often small area prevalence data for major diseases are not collected, or if collected are subject to measurement and administrative biases. However, many countries have regular national health surveys which provide an indication on national trends in prevalence. Such surveys typically provide only broad regional identifiers, whereas health planners require estimates at a much more spatially disaggregated scale, and for those strata (age, sex, ethnicity) by which area populations are recorded — in censuses or by intercensal population estimates. Additionally the estimates should take account of the impact of socioeconomic factors on chronic disease prevalence. In geographic applications, measures of the socioeconomic status of an area's residents include what are known as deprivation indices, where deprivation refers to hardship due to low income, poor housing, high rates of unemployment, etc. In the UK there have been significant developments in the methodology for measuring neighbourhood deprivation (e.g. Noble *et al.*, 2000; Bailey *et al.*, 2003), especially in small neighbourhoods of around 1500-2000 people, there being around 32500 such neighbourhoods in England (ONS, 2006).

This paper describes a Bayesian methodology for obtaining prevalence estimates for chronic disease for 354 English areas, with a particular focus on heart disease. The first source of information is provided by national health surveys. In many countries, a series of health surveys (often annual) is available — among many examples are the Swedish National Public Health Survey, the Italian National Health Survey, and the Taiwan National Health Interview Survey — and one may consider pooling strength over surveys. The analysis here uses the 2003 Health Survey for England, with an earlier 1999 survey providing historical data under a power prior approach (Chen *et al.*, 2000). Except for neighbourhood deprivation category (the quintile rank among 32500 neighbourhoods, with no further identifying information), the spatial scale in the two surveys used consists of nine government regions (the North East of England, the North West, Yorkshire & Humberside, the East Midlands, the West Midlands, Eastern England, London, South East England, and South West England) — see Table 1 for a summary of regional differences. A binomial regression is used to model survey evidence on gradients by age, sex, ethnicity, broad region, and neighbourhoood deprivation.

The second source of information is population totals by age, sex, ethnicity, etc from censuses or intercensal population estimates, to which survey based prevalence rates are to be applied. The populations used here are specific for age, sex and ethnic group, and are drawn from the UK 2001 Census - intercensal

population estimates by age, sex and ethnicity are not made in the UK, though they are in other countries such as the US (Smith, 1998).

The third source of relevant information is mortality data which typically (unlike prevalence) are well recorded at a disaggregated spatial level. Evidence is presented of a positive correlation between heart disease prevalence and mortality, which points to the benefit of adjusting survey based estimates of area prevalence to take account of proxy information on prevalence provided by mortality data over the 354 areas. When mortality is only infrequently linked to a particular type of morbidity (e.g. asthma, psychiatric illness), other sources of area data can be used as proxies for morbidity — examples are hospital admissions or referrals to community care.

The methodology therefore provides an approach to indirect prevalence estimation, applying survey based gradients for heart disease over those stratifiers by which populations for areas are available (e.g. age, sex, ethnic group), while also taking account of neighbourhood deprivation, and of proxy information on prevalence (from mortality) at the required area level. The methodology adopts a fully Bayesian strategy, with prior densities on parameters updated via the likelihood of the observed data. Iterative Monte Carlo Markov Chain techniques (Gelfand and Smith, 1990) are used to estimate models, as implemented in the WINBUGS program (Spiegelhalter *et al.*, 2003).

The following four sections outline the survey based component of the prevalence estimation procedure. They are followed by a section considering how area mortality and prevalence are jointly modelled so that prevalence estimates can incorporate information on spatial mortality patterns. The final section considers possible developments to the methodology.

2. Survey Model: Populations, Survey Variables and Choice of Binomial Link

To apply survey evidence on disease gradients to estimate prevalence in area populations requires equivalent variables to be available in both Census populations (or in intercensal population estimates) and for respondents in national health surveys. Many countries provide population data by age, sex, and ethnicity; for example, the UK 2001 Census includes a tabulation of populations by age, sex and ethnic group.

It is also typically necessary to take account of socioeconomic gradients in disease prevalence (e.g. gradients by individual occupational status or by the deprivation level of the neighbourhood in which individuals live). This suggests that ideally one would require populations by age, sex, ethnicity and occupation, or populations by age, sex, ethnicity and neighbourhood deprivation level. However, in many countries populations are not available to this level of detail. For

example, the UK Census does not provide a four way disaggregation by age, sex, ethnicity and neighbourhood deprivation. In these circumstances, it is proposed here that prevalence estimates by age, sex and ethnicity are scaled by a survey based prevalence gradient over deprivation levels.

A binomial regression is applied to data from the 2003 and 1999 Health Surveys for England to provide model based rates of heart disease prevalence. Survey subjects are classed as having coronary disease if they reported (in the previous year) having angina or a heart attack, confirmed by a doctor. The survey categorisations relevant to estimating area prevalence are age ($a = 1, \ldots, 7$, namely ages 0-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+), sex ($s = 1, 2$; namely male, female), ethnicity ($e = 1, \ldots, 4$; namely white, black, south Asian, all other ethnic groups), and regions $r = 1, \ldots, 9$ as in the columns of Table 1. Additionally the 2003 survey includes neighbourhood deprivation quintile ($d = 1, \ldots, 5$, with $d = 5$ for most deprived). Respondents are aggregated by risk category cells — Greenland (2001) refers to these as distinct covariate patterns. So the observations become numbers at risk, n_{aserd} and diseased subjects y_{aserd}, both taking account of survey weighting for differential non-response (JHSU, 2004).

Table 1: Age standardised chd prevalence (ages 35+) with 95% confidence intervals by sex & government office region 2003 Health Survey for England.

	North East	North West	Yorkshire & Humberside	East Midlands	West Midlands
Males	7.5	9.4	11.6	9.2	10.5
2.5%	4.2	7.2	8.7	6.4	7.8
97.5%	10.8	11.6	14.5	12.0	13.2
Females	9.9	6.6	7.9	7.1	7.4
2.5%	6.4	4.8	5.6	4.7	5.2
97.5%	13.4	8.4	10.2	9.5	9.6

	East	London	South East	South West	England
Males	7.3	8.1	8.2	6.6	8.8
2.5%	5.1	5.9	6.3	4.5	8.0
97.5%	9.5	10.3	10.1	8.7	9.6
Females	3.9	4.6	4.3	3.9	5.8
2.5%	2.3	3.0	2.9	2.3	5.2
97.5%	5.5	6.2	5.7	5.5	6.5

A log-binomial regression is applied, allowing inferences on prevalence proportion ratios rather than prevalence odds ratios (Skov *et al.*, 1998; Zocchetti *et al.* 1997); this is also called the log-linear binomial (Greenland, 2004). For example, using this link permits estimation of the prevalence relative risk gradient over neighbourhood deprivation quintiles, whereas logit coefficients only provide relative risks under a rare disease assumption. To avoid probabilities above 1, an upper limit of 0.999 on cell probabilities was imposed. MCMC sampling produced this default value only in the first 100 to 200 iterations.

To assess possible interactions between risk factors, the prevalence model for the 2003 survey data includes main effects in all variables and second order interactions for which there is evidence in health outcome literature, not necessarily heart disease. The historical data model (for the 1999 survey) is the same except for excluding the main deprivation effect and any interactions involving deprivation. The second-order interactions included are for age-sex, sex-region, sex-ethnicity, sex-deprivation, age-deprivation, age-ethnicity, and ethnicity-deprivation. The sex-region interaction is suggested by Table 1, while different gender-age heart disease risk profiles have been reported as well as sex-ethnicity interactions (Primatesta and Brookes, 2000). While completeness in modelling terms might indicate including several interactions, some studies of cardiovascular outcomes that include area deprivation and area type report few interactions as significant (e.g. Martinez *et al.*, 2003). Thus for parsimony, the age and deprivation variables when included in interactions are reframed as binary: ages up to 64 ($a^* = 1$) are compared with ages 65 and above ($a^* = 2$), and the top two deprivation quintiles ($d^* = 2$) are contrasted with the lower three ($d^* = 1$). Substantive justification for such a contraction exists: for example, the main impact of deprivation is on premature ill health and mortality (e.g. Barnett *et al.*, 2001).

3. Survey Model Specification and Pooling over Surveys

A model including main effects and the above mentioned interactions is then

$$\begin{aligned} y_{aserd} &\sim \text{Bin}(n_{aserd}, \rho_{aserd}) \\ \log(\rho_{aserd}) &= \alpha_{1s} + \alpha_{2a} + \alpha_{3e} + \alpha_{4d} + \beta_{1es} + \beta_{2a^*s} \\ &\quad + \beta_{3ea^*} + \beta_{4a^*d^*} + \beta_{5d^*s} + \beta_{6ed^*} + \gamma_{rs} \end{aligned} \tag{3.1}$$

where parameters treated as fixed effects $\{\alpha_{2a}, \alpha_{3e}, \alpha_{4d}, \beta_{1es}, \beta_{2a^*s}, \beta_{3ea^*}, \beta_{4a^*d^*}, \beta_{5d^*s}, \beta_{6ed^*}\}$ are subject to corner constraints, except for the gender terms α_{1s} that are both taken as unknown. The model for the 1999 survey data, denoted $\{y^h_{aser}, n^h_{aser}\}$ and lacking the deprivation category, has the form

$$\begin{aligned} y^h_{aser} &\sim \text{Bin}(n^h_{aser}, \rho^h_{aser}) \\ \log(\rho^h_{aser}) &= \alpha_{1s} + \alpha_{2a} + \alpha_{3e} + \alpha_{4d} + \beta_{1es} + \beta_{2a^*s} + \beta_{3ea^*} + \gamma_{rs} \end{aligned} \tag{3.2}$$

Priors on $\{\alpha, \beta\}$ in (3.1) and (3.2) are based on accumulated epidemiological evidence, such as that provided by UK studies of treated heart disease prevalence. Data from the Key Health Statistics from General Practice (ONS, 2000) give heart disease prevalence rates of 0.1 per 1000 at ages under 34 (for both males and females) ranging to 205/1000 (males) and 172/1000 (females) at ages over 85. Because of this wide range in risk, normal priors $N(-9, 5)$ and $N(0, 5)$ are adopted for α_{1s} and α_{2a} respectively. For the remaining risk factors (for ethnic and deprivation categories) and the interactions, accumulated evidence (e.g. Hoare, 2003) is that $N(0, 1)$ priors will encompass likely ranges in relative risk. This corresponds to a prior belief that the associated relative risks will be between 0.14 and 7.1 with 95% certainty. It might well be possible to justify more informative elicited priors on relative risk and it is straightforward to include this when a log link is used in the binomial regression (e.g. Greenland, 2001).

The regional effects γ_{rs} are treated as random and follow a bivariate spatial conditional autoregressive prior (see Appendix 1), with the multivariate CAR precision matrix Φ_γ^{-1} assumed to follow a Wishart prior with 2 degrees of freedom and identity scale matrix. Reasons for expecting spatial correlation in regional relativities include the north-south contrast in prevalence (Table 1), as well as environmental factors, such as water hardness (Shaper *et al.*, 1980; Catling *et al.*, 2005).

Let $\theta = \{\alpha, \beta, \gamma, \Phi_\gamma\}$ parameters, and $0 \leq \delta \leq 1$ be a precision parameter (with beta prior) that weights the historical data D_h relative to the likelihood of the current study data D. Following Chen *et al.* (2000, p. 124) the power prior takes the form

$$\pi(\theta.\delta|D_h) \propto [P(D_h|\theta)]^\delta \delta^{a_\delta - 1}(1-\delta)^{b_\delta - 1} \tag{3.3}$$

where $P(D_h|\theta)$ is the binomial likelihood, and (a_δ, b_{delta}) are pre-specified beta density hyperparameters. With δ an unknown the joint posterior density for (θ, δ) is then

$$P(\theta, \delta||D, D_h) \propto P(D|\theta)[P(D_h|\theta)]^\delta \delta^{a_{delta}-1}(1-\delta)^{b\delta - 1}. \tag{3.4}$$

For the current analysis there is expected to be considerable continuity between the two surveys in prevalence differentials and the 1999 survey data includes relevant information on ethnic prevalence gradients; on the other hand, the model forms for 2003 and 1999 differ (because only the 2003 model includes neighbourhood deprivation) and so some downweighting is appropriate. Here three alternative beta priors for δ are considered, namely Be(250,1), Be(100,1) and Be(50,1).

Table 2: Parameter summaries with 95% credible intervals, survey models

Prior on δ		$\delta \sim$ Be(250,1)		$\delta \sim$ Be(100,1)		$\delta \sim$ Be(50,1)	
Fit		$p_e = 26.6, DIC = 1561$		$p_e = 28.7, DIC = 1563$		$p_e = 31, DIC = 1566$	
Risk group	Parameter	Mean	95% Interval	Mean	95% Interval	Mean	95% Interval
	δ	0.97	(0.9,1)	0.42	(0.34,0.51)	0.20	(0.15,0.26)
Males	α_{11}	-7.06	(-7.59,-6.35)	-6.90	(-7.47,-6.45)	-6.77	(-7.43,-5.94)
Females	α_{12}	-7.70	(-8.29,-6.95)	-7.55	(-8.12,-7.07)	-7.46	(-8.15,-6.66)
Ages 35-44	α_{22}	2.43	(1.65,3.1)	2.27	(1.69,2.87)	2.15	(1.3,2.91)
Ages 45-54	α_{23}	3.81	(3.1,4.38)	3.58	(3.08,4.15)	3.43	(2.61,4.11)
Ages 55-64	α_{24}	4.85	(4.13,5.41)	4.70	(4.23,5.27)	4.58	(3.8,5.24)
Ages 65-74	α_{25}	5.54	(4.83,6.08)	5.41	(4.94,6)	5.28	(4.46,5.9)
Ages 75-84	α_{26}	5.84	(5.14,6.39)	5.72	(5.23,6.3)	5.59	(4.75,6.23)
Ages 85+	α_{27}	5.95	(5.26,6.53)	5.84	(5.34,6.46)	5.72	(4.89,6.38)
Black	α_{32}	-0.54	(-1.09,-0.01)	-0.44	(-1.14,0.2)	-0.31	(-1.1,0.43)
South Asian	α_{33}	0.43	(0.17,0.7)	0.33	(-0.05,0.68)	0.15	(-0.33,0.59)
Other	α_{34}	-0.52	(-1.69,0.55)	-0.67	(-1.98,0.53)	-0.72	(-2.08,0.39)
Neighbourhood Deprivation							
Quintile 1	α_{51}	-0.30	(-0.46,-0.14)	-0.31	(-0.47,-0.16)	-0.32	(-0.48,-0.16)
Quintile 2	α_{52}	-0.07	(-0.22,0.07)	-0.09	(-0.25,0.06)	-0.09	(-0.23,0.05)
Quintile 3	α_{53}	-0.10	(-0.26,0.05)	-0.12	(-0.27,0.04)	-0.12	(-0.28,0.02)
Quintile 4	α_{54}	0.08	(-0.08,0.25)	0.11	(-0.08,0.29)	0.11	(-0.06,0.28)
Quintile 5	α_{55}	0.39	(0.22,0.56)	0.42	(0.26,0.59)	0.42	(0.25,0.59)
Interactions							
Ethnic-Sex	β_{122}	0.86	(0.28,1.48)	0.80	(0.07,1.49)	0.77	(-0.11,1.64)
	β_{122}	-0.11	(-0.52,0.3)	-0.04	(-0.57,0.49)	0.06	(-0.57,0.69)
	β_{122}	-0.32	(-1.83,0.96)	-0.11	(-1.65,1.25)	-0.04	(-1.54,1.32)
Age-Sex	β_{122}	0.04	(-0.2,0.29)	0.05	(-0.23,0.34)	0.09	(-0.18,0.38)
Ethnic-Age	β_{122}	-0.76	(-1.35,-0.16	) -0.83	(-1.62,-0.12)	-0.87	(-1.77,-0.04)
	β_{122}	-0.40	(-0.77,-0.04	) -0.34	(-0.78,0.13)	-0.19	(-0.78,0.37)
	β_{122}	0.56	(-0.63,1.77	) 0.60	(-0.78,1.77)	0.71	(-0.56,2.02)

Table 2: (Continued) Parameter summaries, survey models

Prior on δ		$\delta \sim$ Be(250,1)		$\delta \sim$ Be(100,1)		$\delta \sim$ Be(50,1)	
Risk group	Parameter	Mean	95% Interval	Mean	95% Interval	Mean	95% Interval
Age-Deprivation	β_{422}	-0.15	(-0.41,0.1)	-0.22	(-0.48,0.04)	-0.23	(-0.49,0.03)
Sex-Deprivation	β_{522}	0.18	(-0.07,0.43)	0.17	(-0.08,0.43)	0.19	(-0.08,0.47)
Ethnic-Deprivation	β_{622}	0.05	(-0.7,0.75)	0.02	(-0.79,0.76)	-0.09	(-0.92,0.7)
	β_{632}	-0.36	(-0.96,0.14)	-0.34	(-0.94,0.2)	-0.26	(-0.89,0.33)
	β_{642}	-0.67	(-2.06,0.56)	-0.58	(-1.95,0.71)	-0.64	(-2.03,0.61)
Region Effects: Males							
N. East	γ_{11}	-0.05	(-0.31,0.2)	-0.07	(-0.35,0.17)	-0.08	(-0.36,0.17)
N. West	γ_{21}	0.04	(-0.13,0.2)	0.04	(-0.15,0.21)	0.04	(-0.14,0.21)
Yorks/Humb	γ_{31}	0.22	(0.04,0.4)	0.22	(0.03,0.4)	0.2	(0.01,0.39)
E. Midlands	γ_{41}	0.02	(-0.16,0.2)	0.03	(-0.16,0.21)	0.03	(-0.17,0.22)
W. Midlands	γ_{51}	0.06	(-0.1,0.23)	0.07	(-0.12,0.24)	0.07	(-0.12,0.25)
East	γ_{61}	-0.06	(-0.25,0.13)	-0.08	(-0.29,0.12)	-0.09	(-0.31,0.12)
London	γ_{71}	-0.07	(-0.24,0.11)	-0.04	(-0.24,0.16)	-0.01	(-0.23,0.19)
S. East	γ_{81}	-0.03	(-0.21,0.15)	-0.01	(-0.18,0.17)	0	(-0.18,0.17)
S. West	γ_{91}	-0.15	(-0.38,0.06)	-0.15	(-0.37,0.06)	-0.16	(-0.38,0.06)
Region Effects: Females							
N. East	γ_{12}	0.35	(0.08,0.6)	0.37	(0.11,0.64)	0.37	(0.08,0.63)
N. West	γ_{22}	0.03	(-0.17,0.22)	0.04	(-0.17,0.23)	0.04	(-0.16,0.24)
Yorks/Humb	γ_{32}	0.19	(-0.03,0.4)	0.21	(-0.01,0.43)	0.22	(-0.01,0.43)
E. Midlands	γ_{42}	0.13	(-0.08,0.35)	0.13	(-0.12,0.36)	0.12	(-0.11,0.35)
W. Midlands	γ_{52}	0.03	(-0.16,0.23)	0.05	(-0.16,0.27)	0.05	(-0.17,0.26)
East	γ_{62}	-0.17	(-0.43,0.07)	-0.2	(-0.47,0.05)	-0.19	(-0.47,0.07)
London	γ_{72}	-0.15	(-0.38,0.06)	-0.18	(-0.43,0.07)	-0.2	(-0.47,0.07)
S. East	γ_{82}	-0.14	(-0.35,0.07)	-0.15	(-0.38,0.05)	-0.16	(-0.39,0.06)
S West	γ_{92}	-0.27	(-0.56,0)	-0.26	(-0.57,0.02)	-0.25	(-0.53,0.03)

4. Survey Model Results

Inferences are based on iterations 1000-5000 of two chain sampling runs starting from dispersed starting values, with convergence achieved by iteration 1000 using Gelman-Rubin criteria (Gelman *et al.*, 1995). Comparisons of models use the deviance information criterion (DIC) of Spiegelhalter *et al.* (2002), namely the posterior mean deviance plus a complexity measure p_e, derived as the difference between $\bar{D}$ and the deviance, $\text{Dev}(\bar{\Psi})$, at the posterior mean of $\Psi = (\theta, \delta)$. So the DIC can be obtained as $\text{Dev}(\bar{\Psi}) + 2p_e$. For model checking, new data $(y_{new,aserd})$ are sampled from the model and compatibility with actual data assessed by the extent to which 95% intervals for new data include the actual data (Gelfand, 1996).

Table 2 shows parameter estimates (log relative risks) under alternative δ priors. Average deviances are similar across the three options on δ, though DICs decrease slightly with larger values of δ because of lower p_e. The posterior predictive checking procedure of Gelfand (1996) is satisfactory, with actual cases y_{aserd} in all cells covered by 95% intervals for replicate data sampled from $P(y_{new}|y)$, regardless of the prior on δ.

Age and sex effects are significant under all options, while ethnic group effects for $\delta \sim$ Be(250,1) show lower risk for blacks and higher risk for south Asians (cf Primatesta and Brookes, 2000). Under all δ priors, the deprivation effects (centred around their average) show the main contrast is between extremes of neighbourhood deprivation with a relatively flat intermediate effect. The regional effects support a north-south contrast in prevalence within England; three coefficients are significant under $\delta \sim$ Be(250,1), and the parameter contrasts $\gamma_{31} - \gamma_{91}, \gamma_{12}$-$\gamma_{92}$, and γ_{32}-γ_{92} have posterior means (95% intervals) of 0.37 (0.07,0.69), 0.61 (0.19,1.07) and 0.45 (0.08,0.85). The fact that regional effects exist after controlling for population composition and neighbourhood deprivation suggests genuine contextual variation in heart disease risk. Of the interactions β_{122} is significant in terms of its 95% credible interval under the two more informative priors on δ, reflecting higher prevalence among black women as compared to men. Both β_{322} and β_{332} are significantly negative under $\delta \sim$ Be(250,1), showing black and south Asian elders to have lower risk.

5. Relevant Survey Outputs for Prevalence-Mortality Model

The goal of the analysis is to estimate heart disease prevalence in 354 English areas, for which Census population counts N_{iase} by age, sex and ethnic group are obtainable. Further population disaggregation by neighbourhood deprivation quintile is not available. However, to ensure the area prevalence estimates take account of neighbourhood deprivation, it is possible to obtain population totals

N_{id}, providing proportions $w_{id} = N_{id}/N_i$ of total area population living in each deprivation quintile. Let $r_i \in \{1, \ldots, 9\}$ denote the region in which the i-th area is located, then the impact of neighbourhood deprivation in area i is based on averaging the probabilities ρ_{aser_id} according to the population split w_{id}. Then age-sex-ethnic specific prevalence rates R_{iase} are estimated as a weighted average

$$R_{iase} = (\sum_d w_{id}\rho_{aser_id}), \tag{5.1}$$

and prevalent cases in each area and for age, sex and ethnic groups are estimated as $P_{iase} = R_{iase}N_{iase}$.

English area mortality data are not specific to ethnic group (see section 4). To generate a prevalence rate that can be modelled jointly with mortality, the P_{iase} are aggregated over ethnic groups (at each MCMC iteration) to form area-age totals P_{ias} that are in turn divided by area-age-sex specific populations N_{ias} to give area-age-sex prevalence rates R_{ias}. These are then applied to European standard populations S_a (e.g. Hedman *et al.*, 1999) to provide age standardised prevalence rates π_{is} for each area by sex (and for ages over 35). To provide a suitable input to the joint prevalence-mortality analysis, the transforms $x_{is} = \text{logit}(\pi_{is})$ are monitored, as these are more likely to be approximately normal than the prevalence rates themselves. Posterior means and variances of the x_{is} are denoted $X_{is}(V_{is})$.

6. Joint Prevalence-Mortality Model

As argued above, evidence on variation in area prevalence is provided indirectly by area heart disease mortality; it is likely that mortality will closely reflect prevalence, though other factors may be involved. Evidence justifying a joint analysis is obtained from heart disease registers associated with a new payment scheme for English general practitioners (Strong *et al.*, 2006). These data may be subject to under or over-registration (especially at lower spatial scales) and so do not necessarily provide a "gold standard" prevalence estimate. However, Table 3 shows a clear correlation between prevalence and the selected mortality index at regional level; the Pearson correlation is 0.9.

Let D_{i1} and D_{i2} be area deaths for males and females (ages over 35), and E^D_{i1} and E^D_{i2} be expected deaths (using England age-specific rates for 2003). Assuming Poisson sampling, one has $D_{is} \sim \text{Po}(\mu_{is}E^D_{is})$, where μ_{is} are relative mortality risks for sex s (1=M, 2=F) and area i. The goal is to adjust survey based logit prevalence rate estimates x_{is} to reflect spatial patterning of these mortality relative risks.

Table 3: Associations between prevalence and mortality, regional level

Government Region	Standard Prevalence Ratio (Data from GP Payment System)	CHD Mortality 2001-3 Directly Standardised rates, Ages < 75
North East	132	76
North West	120	74
Yorkshire & Humberside	119	69
E. Midlands	100	62
W. Midlands	99	66
East	89	50
London	87	61
South East	86	50
South West	90	50

Let z be the underlying true (logits of) prevalence, measured with error. The bivariate model can be seen as following a form

$$P(\mu, z|x) = P(\mu|z)P(z|x)P(x), \tag{6.1}$$

namely a marginal prevalence model and a model including an impact of prevalence on mortality. Similar models are mentioned by X_{ia} and Carlin (1998) and Bernardinelli *et al.* (1997), and both of these studies incorporate spatial correlation in the underlying rate.

In the application here, spatially correlated effects pool information over areas and genders. There are many reasons to expect unmeasured risk factors to be spatially correlated, for example between adjacent urban as against rural areas, or between adjacent areas in northern as against southern regions of England. Thus, urban air pollution is a risk for heart disease (Chen *et al.*, 2005), while regional differences in smoking and physical activity are reported by Morris *et al.* (2003). Regional differences in drinking-water hardness have also been linked to cardiovascular disease variations (Monarca *et al.*, 2004; Catling *et al.*, 2005).

So for sexes $s = 1, 2$ the following joint model is postulated

$$D_{is} \sim \text{Po}(\mu_{is} E^D_{is}) \tag{6.2}$$

$$\begin{aligned} \log(\mu_{is}) &= \eta_s + \omega_s z_{is} \\ z_{is} &= x_{is} + u_{is} \\ x_{is} &\sim N(X_{is}, V_{is}) \end{aligned}$$

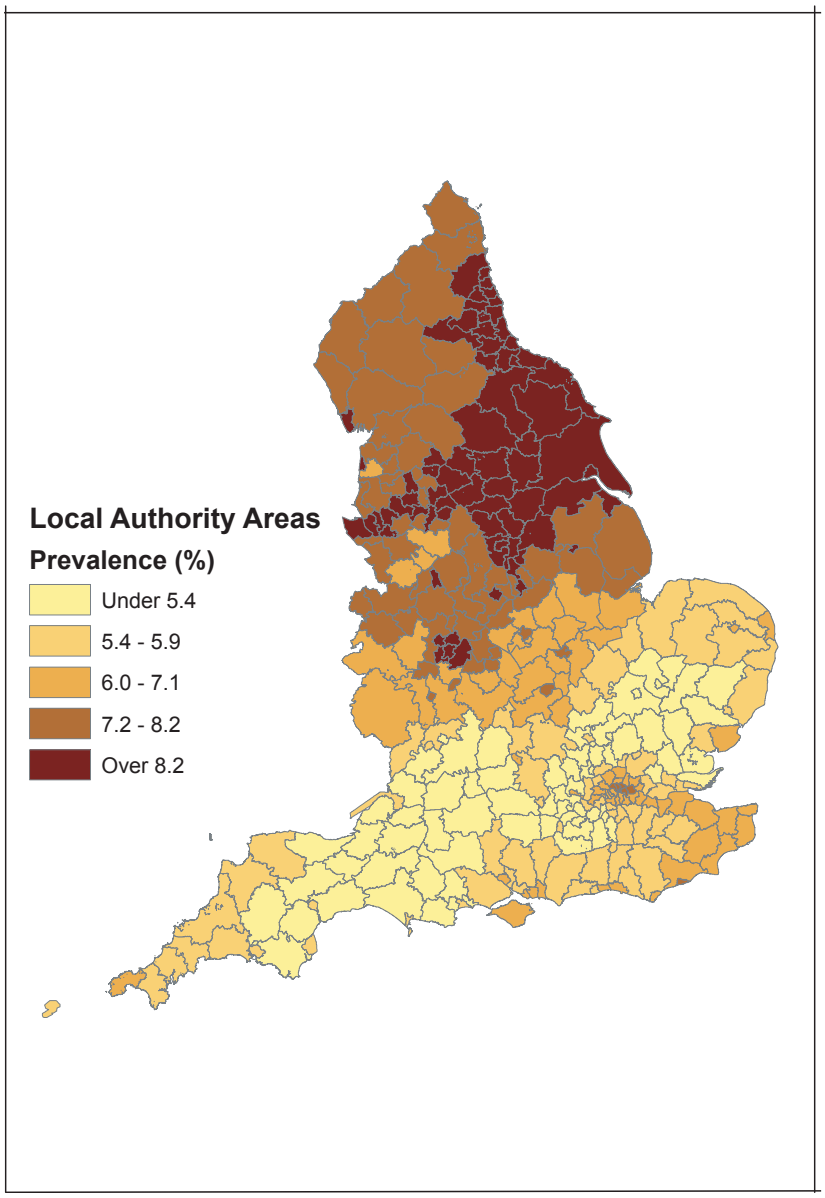

Figure 1: Male prevalence of heart disease

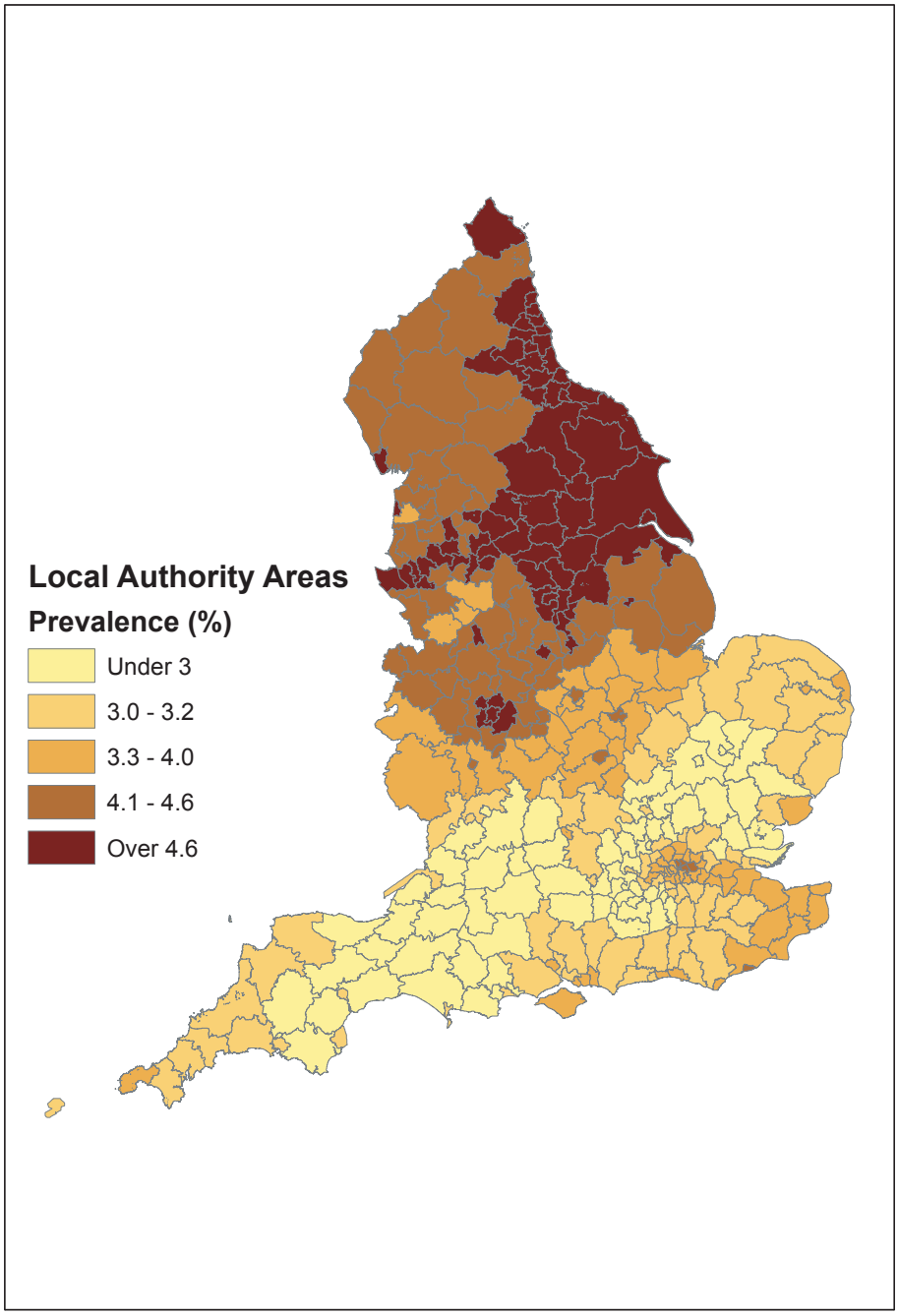

Figure 2: Female prevalence of heart disease

where the spatial correlated errors (u_{i1}, u_{i2}) follows a bivariate conditional autoregressive prior (see Appendix 1), with the effects u_{i1} and u_{i2} centred around respective means at each iteration. From the form of the model it is apparent that z will be affected by μ as well as vice versa, by virtue of the reverse regression implicit in many measurement error models (Maddala, 2001, Ch 11). The precision matrix Φ_u^{-1} is assumed to follow a Wishart prior with 2 degrees of freedom and identity scale matrix; the intercepts η_s and slopes ω_s are assigned $N(0, 100)$ priors. The last 15000 of a two chain run of 25000 iterations show the coefficients ω_1 a nd ω_2 in model (6.2) as clearly significant, with means (95% intervals) of 0.42 (0.36,0.48) and 0.23 (0.17,0.31). The correlation between (u_{i1}, u_{i2}) is estimated at 0.71 (0.54,0.84).

Compared to the mean prevalences π_{is} from the survey model, the adjusted prevalences $\zeta_{is} = \exp(z_{is})/(1 + \exp(z_{is}))$ from the joint model show greater inequality (i.e. higher coefficients of variation), possibly as they reflect local variations in mortality relative risks. Figures 1 and 2 contain quintile maps for the posterior mean prevalence rates ζ_{i1} and ζ_{i2}, with the North-South contrast again visible. The fact that this contrast remains after controlling for neighbourhood deprivation and ethno-demographic structure suggests that health behaviours (e.g. diet) and environmental factors are also relevant to prevalence differences.

As an application with policy relevance, the prevalence rates ζ_{is} are compared to revascularisation rates for the 354 areas in 2002. There are recognized to be variations in provision of revascularisation, namely coronary artery bypass grafts and percutaneous transluminal coronary angioplasty that other studies suggest are not explained by morbidity (Hippisley-Cox and Pringle, 2000; Payne and Saul, 1997). In fact, the correlation between provision and prevalence rates is -0.02 for males and 0.08 for females, indicating possible variations in access to surgical care not matched to need for such care (as reflected in prevalence).

7. Discussion

While mortality and hospitalisation data are often used as proxies for prevalence (morbidity) and hence health need (Ebrahim *et al.*, 2002), there is value in using available survey evidence to provide direct estimates of prevalence and morbidity. The present study has outlined a methodology to combine spatially aggregated survey evidence with information on spatially disaggregated patterns in heart disease mortality, which reflect geographic variations in prevalence (e.g. see Table 3).

This methodology can be seen as a form of meta-analysis over different forms of evidence that can be applied to other types of morbidity. The pooling of information over surveys (here the 1999 and 2003 Health Surveys for England) can be performed using the power prior method. An alternative analysis to

the one adopted in the paper could arguably input more informative priors to the power prior likelihood, further developing on the theme of evidence pooling. There is accumulated epidemiological evidence on heart disease risk factors that could justify more informative priors, especially on main effects. For example, south Asian ethnicity is often reported as associated with higher relative heart disease risk in the UK. So one could for instance, following Greenland (2001, p. 665), assume a prior relative risk between 1 and 3 for this group, translating into a $N(0.55, 0.08)$ prior. On the other hand, there may be relatively little prior evidence on certain interactions (especially in a particular geographical setting such as England with its distinct health care system) and adopting an informative prior approach may also imply the need to run a sensitivity analysis over different informative priors.

There are other options for modelling that might be considered. One option is to introduce information on prevalence from hospital admission data. These are sometimes suspect as indicators of morbidity because they reflect supply of care, but for events where hospitalisation is usually unavoidable (e.g. myocardial infarction) they may improve the estimation of morbidity. One might also seek to jointly model, and so make indirect area estimates for, more than one type of prevalence (e.g. smoking, diabetes or obesity prevalence) in conjunction with modelling heart disease prevalence. In the UK prevalence of these behaviours or conditions is also monitored by the Health Survey for England and they are known risk factors for heart disease. Similar potentialities exist for using national health survey data of other countries to indirectly estimate area prevalence in conjunction with other relevant and locally disaggregated information (e.g. on mortality, hospital admissions).

Appendix. Univariate and Multivariate CAR priors

To explain the form of the CAR prior, first consider a univariate conditional autoregressive prior. Let $(e_1, \ldots, e_n)$ be a vector of effects associated with areas 1,..,n such as relative mortality risks. Then a univariate conditional autoregressive prior or CAR prior (Rue and Held, 2005), involves specifying the n full conditionals

$$e_i | e_{[i]} \sim N(\sum_{j \neq i} c_{ij} e_j / c_{i+}, \phi_e / c_{i+}) \tag{A.1}$$

where $e_{[i]} = (e_1, e_2, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$ is the collection of effects excluding area i, $C = [c_{ij}]$ is an n n matrix of spatial interactions c_{ij}, often known but sometimes involving unknown parameters, the sums $c_{i+} = \sum_j c_{ij}$ total over rows in this matrix, and ϕ_e is a conditional variance. The conditional density uniquely determines the joint density of the effects $(e_1, \ldots, e_n)$, a feature noted by Besag

(1974) and Jin *et al.* (2005). The multivariate normal CAR is the multivariate generalisation of the prior in (A.1). If there are K outcomes, then ϕ_e is replaced by a $K \times K$ covariance matrix Φ_e. A common practice is to define $c_{ij} = 1$ if areas i and j are adjacent, and $c_{ij} = 0$ otherwise, in which case c_{i+} is the number of areas adjacent to area i.

References

Bailey, N., Flint, J., Goodlad, R., Shucksmith, M., Fitzpatrick, S., Pryce, G. (2003). Measuring deprivation in Scotland: Developing a long-term strategy. Scottish Executive (http://www.scotland.gov.uk/Publications/2003/09/18197/26538)

Barnett, S., Roderick, P., Martin, D., Diamond, I. (2001). A multilevel analysis of the effects of rurality and social deprivation on premature limiting long term illness. *J. Epidemiol Community Health* **55**, 44-51.

Bernardinelli, L., Pascutto, C., Best, N., Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741-752.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36**, 192-236.

Catling, L., Abubakar, I., Lake, I., Swift, L., Hunter, P. (2005). Review of evidence for relationship between incidence of cardiovascular disease and water hardness. Contract DWI/70/2/176. University of East Anglia & Drinking Water Inspectorate.

Chen, M., Ibrahim, J., Shao, Q. (2000). Power prior distributions for generalized linear models. *J. Statist. Plan. Inf.* **84**, 121-137.

Chen, L., Knutsen, S., Shavlik, D., Beeson, W., Petersen, F., Ghamsary, M., Abbey, D. (2005). The association between fatal coronary heart disease and ambient particulate air pollution: are females at greater risk? *Environ Health Perspect* **113**, 1723-1729.

Ebrahim, S., Ben-Shlomo, Y., Ho, D., Maxwell, R., Oliver, S., Shaw, M. (2002). Coronary revascularisation in the South West region, 1991-2000: Equity in the use of CABG and PTCA by gender, age, deprivation and geography. South West Public Health Observatory & MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol.

Gelfand, A. (1996). Model determination using sampling based methods. Chapter 9 in Markov Chain Monte Carlo in Practice (Edited by Gilks, W., Richardson, S., and Spiegelhalter, D.), Chapman and Hall.

Gelfand, A., Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *J. American Statistical Association* **85**, 398-409.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis.* Chapman and Hall.

Greenland, S. (2001). Putting background information about relative risks into conjugate prior distributions. *Biometrics* **57**, 663-670.

Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am. J. Epidemiol.* **160**, 301-305.

Hedman, J., Kaprio, J., Poussa, T., Nieminen, M. (1999). Prevalence of asthma, aspirin intolerance, nasal polyposis and chronic obstructive pulmonary disease. *International Journal of Epidemiology* **28**, 717-722.

Hippisley-Cox, J., Pringle, M. (2000). Inequalities in access to coronary angiography and revascularization: The association of deprivation and location of primary care services. *Br. J. General Practice* **50**, 449-454.

Hoare, J. (2003). Comparison of area-based inequality measures and disease morbidity in England, 1994-1998. *Health Statistics Quarterly* **18**, 18-24.

Jin, X., Carlin, B., Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61**, 950-961.

Joint Health Surveys Unit (2004). *Health Survey for England, 2003. Summary of Key Findings*. London: TSO.

Maddala, G. (2001). *Introduction to Econometrics*, 3rd edition. Wiley.

Martinez, J., Pampalon, R., Hamel, D. (2003). Deprivation and stroke mortality in Quebec. *Chronic Dis Can.* **24**, 57-64.

Monarca, S., Zerbini, I., Donato, F. (2004). *Drinking-water hardness and cardiovascular diseases: A review of epidemiological studies, 1979-2004*. World Health Organization (http://www.who.int/water_sanitation_health/en/).

Morris, R., Whincup, P., Emberson, J., Lampe, F., Walker, M., Shaper, A. (2003). North-South Gradients in Britain for stroke and CHD: Are they explained by the same factors? *Stroke* **34**, 2604-2609.

Noble, M., Smith, G., Penhale, B., Wright, G., Dibben, C., Owen, T., Lloyd, M. (2000). *Measuring multiple deprivation at the small area level: The Indices of Deprivation 2000.* DETR, Regeneration Research Summary, Number 37.

Office of National Statistics (2000). *Key Health Statistics from General Practice* (Series MB6 No.2). (http://www.statistics.gov.uk/statbase/Product.asp?vlnk=4863)

Office of National Statistics (2006). *Super Output Areas (SOAs)* http://www.statistics.gov.uk/geography/soa.asp.

Office of Population Censuses & Surveys (1995). *Morbidity Statistics from General Practice, 4th National Study* 1991/92. (http://www.statistics. gov.uk/statbase/Product.asp?vlnk=616&More=n)

Payne, N., Saul, C. (1997). Variations in use of cardiology services in a health authority: Comparison of coronary artery revascularisation rates with prevalence of angina and coronary mortality. *Brit. Med. J'* **314**, 257-61.

Primatesta, P,. Brookes, M. (2000). Cardiovascular disease: Prevalence and risk factors. In *Health Survey for England, 1999: The Health of Minority Ethnic Groups. Joint Health Surveys Unit for Department of Health* (Edited by Erens, B., Primatesta, P. and Prior, G.) Chapter 3.

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

Shaper, A., Packham, R., Pocock, S. (1980). The British Regional Heart Study: cardiovascular mortality and water quality. *J. Environ. Pathol. Toxicol.* **4**, 89-111.

Skov, T., Deddens, J. and Petersen, M. (1998). Prevalence proportion ratios: Estimation and hypothesis testing. *Int. J. Epidemiol* **27**, 91-95.

Smith, A. (1998). The American community survey and intercensal population estimates: Where are the crossroads? Population Division Technical Working Paper 31, US Bureau of the Census, Washington D.C. (http://www.census.gov/population/www/documentation/twps0031/twps0031.html)

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

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003). WinBUGS User Manual, Version 1.4, January 2003. http://www.mrc-bsu.cam.ac.uk/bugs

Strong, M., Maheswaran, R. and Radford, J. (2006). Socioeconomic deprivation, coronary heart disease prevalence and quality of care: A practice-level analysis in Rotherham using data from the new UK general practitioner quality and outcomes framework. *J. Public Health* **28**, 39-42.

Xia, H. and Carlin, B. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine* **17**, 2025-2043. Zocchetti, C., Consonni, D. and Bertazzi, P. (1997). Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. *Int. J. Epidemiol* **26**, 220-223.

Received August 14, 2006; accepted October 31, 2006.

Peter Congdon
Department of Geography
Queen Mary University of London
Mile End Rd, London E1 4NS, UK
p.congdon@qmul.ac.uk