Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2010)
59, *Part* 2, *pp.* 319–339

Racial disparities in risks of mortality in a sample of the US Medicare population

Yijie Zhou,

Merck Research Laboratories, Rahway, USA

Francesca Dominici

Harvard University, Boston, USA

and Thomas A. Louis

Johns Hopkins University, Baltimore, USA

[Received May 2007. Final revision July 2009]

Summary. Racial disparities in risks of mortality adjusted for socio-economic status are not well understood. To add to the understanding of racial disparities, we construct and analyse a data set that links, at individual and zip code levels, three government databases: Medicare, the Medicare Current Beneficiary Survey and US census. Our study population includes more than 4 million Medicare enrollees residing in 2095 zip codes in the north-east region of the USA. We develop hierarchical models to estimate the black–white disparities in risk of death, adjusted for both individual level and zip code level income. We define the population level attributable risk AR, relative attributable risk RAR and odds ratio OR of death comparing blacks *versus* whites, and we estimate these parameters by using a Bayesian approach via Markov chain Monte Carlo sampling. By applying the multiple-imputation method to fill in missing data, our estimates account for the uncertainty from the missing individual level income data. Results show that, for the Medicare population being studied, there is a statistically and substantively significantly higher risk of death for blacks compared with whites, in terms of all three measures AR, RAR and OR, both adjusted and not adjusted for income. In addition, after adjusting for income we find a statistically significant reduction in AR but not in RAR and OR.

Keywords: Hierarchical model; Markov chain Monte Carlo methods; Multiple imputation; Racial disparity; Socio-economic status

1. Introduction

There is considerable evidence and concern about disparities in health between the majority white population and minority racial groups in the USA, especially the disparities between blacks and whites. Social epidemiology literature suggests that the black–white disparities in health and mortality are due to the differentiation of social and historical factors which can be summarized in the following two categories (Hummer, 1996; Williams, 1999):

(a) differentiation in socio-economic status (SES);
(b) racism, both institutional (residential segregation, racial isolation and political representation) and individual (majority group behaviour and attitude, and perception of discrimination).

Address for correspondence: Yijie Zhou, Merck Research Laboratories, PO Box 2000, 126 East Lincoln Avenue, RY34-A304, Rahway, NJ 07065, USA.
E-mail: yijie_zhou@merck.com

 0035–9254/10/59319

To understand how race is associated with health and mortality, it is important to investigate the relative importance of different factors, with SES being the most studied in the past several years.

SES differentiation explains racial disparities in mortality at both individual and area levels (LeClere *et al.*, 1997). It is well known that individual level SES is associated with race and is a strong predictor of risks of mortality (Williams and Collins, 1995; Link and Phelan, 1995). However, area level SES is associated with both risks of mortality and individual level SES (Pickett and Pearl, 2001). Therefore, failure to account for area level SES would result in biased estimation of the association between race, individual level SES and risks of mortality (Cole and Hernän, 2002; Robins and Greenland, 1992).

For the relationship between area level SES and health, the literature suggests that, besides the association between health and absolute SES level of an area (e.g. median income), the SES inequality within an area, i.e. differences between the SES of the well off and the poor, is also associated with health and mortality (Kawachi and Kennedy, 1999). In analyses that were performed at small geographic areas such as census tracts or zip codes, the absolute SES level tends to be more strongly associated with mortality rates than the SES inequality, whereas the situation is reversed in analyses at larger areas such as states or countries. Some commonly used measures of inequality for income are the interquartile range IQR, Robin Hood index and Gini coefficient (Wilkinson, 1997; Lochner *et al.*, 2001; McLaughlin and Stokes, 2002).

Several studies have analysed the individual level relationship between race, SES and risks of mortality (Cooper *et al.*, 2001; Howard *et al.*, 2000; Guralnik *et al.*, 1993; Keil *et al.*, 1992; LeClere *et al.*, 1977; Otten *et al.*, 1990; Smith *et al.*, 1998; Sorlie *et al.*, 1992, 1995; Steenland *et al.*, 2004). However, most of these studies were either restricted to a specific and small geographic area (Guralnik *et al.*, 1993; Keil *et al.*, 1992) or targeted on premature mortality for people younger than 65 years old (Cooper *et al.*, 2001; Otten *et al.*, 1990; Smith *et al.*, 1998; Steenland *et al.*, 2004). In addition, these studies generally do not account for the potential confounding of area level SES and do not use hierarchical models for the analysis. Such models are necessary to account for the intra-area correlation between individuals, as well as the between-area spatial correlation. Another study by Gornick *et al.* (1996) analysed the relationship between race, SES and rates of mortality for the nationwide elderly population over 65 years old. However, the analysis was performed at zip code level instead of individual level.

In this paper we develop and apply hierarchical models to investigate the following questions.

(a) What is the association between an individual's race and risk of death adjusted for age and gender?
(b) How will this association vary when we adjust for both individual and area level SES?

To address these questions, we integrate three large government databases with information on individual race, age, gender, date of death, individual level SES and zip code level SES over the period 1999–2002, for more than 4 million people residing in 2095 zip codes in the north-east region of the USA.

Using this data set, we develop and apply hierarchical statistical models to explore the association between individual race and risks of mortality, as well as whether this association is explained by the racial differentiation in both individual level and zip code level income. On the basis of these models, we carry out a sensitivity analysis to identify which summary of individual income at the zip code level (e.g. the mean or IQR) explains most of the variability in risks of mortality. We measure the association between race and risks of mortality by using the population level attributable risk AR, relative attributable risk RAR and odds ratio OR of death comparing blacks *versus* whites. These parameters are defined as functions of model-predicted

probabilities which can be easily estimated by use of a Bayesian approach via Markov chain Monte Carlo sampling.

In Section 2, we describe the sources of data. In Section 3 we describe the methods for the analysis, including hierarchical models, multiple imputation for missing data and a definition of association measures. Section 4 presents the results. Section 5 is the sensitivity analyses, and discussion in Section 6 is followed. The computational details are in Appendix A.

2. Sources of data

To explore the association between race and risks of mortality adjusted for SES, we construct a data set that links, at individual and zip code levels, three government databases: Medicare, the US census and the Medicare Current Beneficiary Survey (MCBS).

Medicare is a national health insurance programme for people who are 65 years of age and older. It is administered by the Centers for Medicare and Medicaid Services of the US Government. Generally, people are eligible for Medicare automatically when they turn the age of 65 years without charge, and the eligibility persists until death. People who are under 65 years of age may also qualify for Medicare without charge, in the presence of either a certain disability or end stage renal disease. Such individuals are eliminated from our study because they do not represent the general Medicare population. From the Medicare data set we obtain individual race, age, gender and date of death over the period 1999–2002 for more than 4 million black and white Medicare enrollees who are 65 years of age and older residing in 2095 zip codes in the north-east region of the USA.

Fig. 1 shows the study area which includes the 2095 zip codes from 64 counties in the north-east region of the USA. We select the counties whose centroids are within the range that covers the north-east coast region of the USA, and we exclude zip codes without available study population from the study map. This area covers several large cities including Washington DC, Baltimore, Philadelphia, New York City, New Haven, Providence and Boston. It has the advantage of high population density, racial diversity and substantial SES heterogeneity.

We categorize an individual's age into five intervals based on age in his or her first year of observation: [65, 70), [70, 75), [75,80), [80, 85) and [85, >85) years. This categorization facilitates detection of age effects because differences in risks of mortality for a 1-year increase in age are relatively small. We 'coarsen' the daily survival information into yearly survival indicators. By doing so we define the quantity 'individual's risk of death' as the probability of the occurrence of death in 1 year. This definition adjusts for the differential follow-up time.

We link the Medicare data set to the year 2000 US census database by zip code. We obtain zip code level SES by using median household income from the census data.

Individual level SES data are available only for the subset of the Medicare population who participated in the MCBS. Fig. 1 shows the zip codes with MCBS enrolment in the study area. The MCBS data set consists of records for approximately 1700 Medicare enrolees from 410 zip codes in the study area, approximately 0.04% of the study population. We collect individual level SES from the MCBS data, using yearly income of the person surveyed and his or her spouse.

3. Methods

3.1. Hierarchical mortality models

Guided by the scientific questions in Section 1, we define two hierarchical models for estimating the individual level association between race and risk of death adjusted and not adjusted for individual and zip code level income.

Fig. 1. Location of the 2095 zip codes in our study area (■) as well as the 410 zip codes within the study area with MCBS enrollees (□)

Let i denote individual, j denote zip code of residence, t denote year and D_{ijt} be the indicator of death for individual i in zip code j in year t. We define the model adjusted for income as

$$\text{logit}\{\Pr(D_{ijt}=1)\}=\alpha_0+\alpha_1\,\text{race}_{ij}+\mathbf{X}_{ij}\boldsymbol{\alpha}_2+\alpha_{01}\,\text{income}_j+U_{0j}+U_{1j}\,\text{race}_{ij}+\mathbf{X}_{ij}\mathbf{U}_{2j}, \quad (1)$$

where $\mathbf{X}_{ij}$ is a row vector of individual level covariates, specifically, age_{ij}, gender_{ij}, $\text{age}_{ij}\times \text{gender}_{ij}$ and income_{ij}; income_j is the zip code level median household income. As stated in Section 2, we construct the age category, denoted by age_{ij}, using age in the first year and, therefore, it does not vary with t. The parameters α_0, α_1, $\boldsymbol{\alpha}_2$ and α_{01} are fixed effects, whereas U_{0j}, U_{1j} and $\mathbf{U}_{2j}$ are zip code level, correlated random effects. We define the model that is not adjusted for income the same as model (1) but leaving out the covariates income_{ij} and income_j.

3.2. Modelling spatial correlation

To account for the possible correlation due to unmeasured variables that might vary spatially as the mortality risks, we allow the random effects to be spatially correlated. Specifically, let $\mathbf{U}_j$ denote the random-effects vector $(U_{0j}, U_{1j}, \mathbf{U}'_{2j})'$ for zip code j and let $\mathbf{U}$ denote the vector $(\mathbf{U}'_1, \mathbf{U}'_2, \ldots, \mathbf{U}'_J)'$, which is the random effects of all J zip codes; we assume the following distribution for the random effects:

$$\mathbf{U}\sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (2)$$

Let Σ_0 denote the covariance matrix of the identically distributed random-effect vectors $\mathbf{U}_j$, $j=1,\ldots,J$. Under a separable model (Banerjee *et al.*, 2004), we assume

$$\text{cov}(\mathbf{U}_j, \mathbf{U}'_j)=\rho(j, j')\Sigma_0 \quad (3)$$

where

$$\rho(j, j')=\exp\{-\phi\, d(j, j')\},$$

in which $d(j, j')$ is the distance in kilometres between the centroids of zip codes j and j'. Viewing Σ_0 as $\text{var}(\mathbf{U}_j)^{1/2}\text{var}(\mathbf{U}_{j'})^{1/2}$ in equation (3), it can be seen that $\rho(j, j')$ is the spatial correlation between $\mathbf{U}_j$ and $\mathbf{U}_{j'}$. The larger the parameter ϕ, the more rapidly the spatial correlation decays with distance. This formulation is appropriate because, generally, locations that are close are more highly correlated than those further away.

The covariance matrix for $\mathbf{U}$, $\boldsymbol{\Sigma}$, resulting from equation (3) is easily shown to be

$$\boldsymbol{\Sigma}=H\otimes\Sigma_0, \quad (4)$$

where the matrix element $(H_{J\times J})_{jj'}=\rho(j, j')$, and '$\otimes$' denotes the Kronecker product. $\boldsymbol{\Sigma}$ in equation (4) is guaranteed to be positive definite, since both H and Σ_0 are positive definite.

3.3. Model fitting

We estimate the unknown parameters by using a Bayesian approach (Carlin and Louis, 2009) via Markov chain Monte Carlo sampling (Geyer, 1992; Gilks *et al.*, 1998). However, individual level income is available only for a representative but small subpopulation from the MCBS. To structure an imputation approach, let Y_{obs} and Y_{miss} denote the observed data and the missing individual level income data respectively, and let $\boldsymbol{\theta}$ denote all parameters in the mortality model (1). The fully Bayesian data augmentation (Tanner and Wong, 1987) imputes missing values via the posterior distribution $P(Y_{\text{miss}}|Y_{\text{obs}})$, which is $P(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta})$ marginalized over $\boldsymbol{\theta}$. However, the large size of the Medicare data set and the large amount of missing data impose

computational challenges. As a more easily implementable approach, we employ two-step multiple imputation. We first develop a prediction regression model to generate random draws of missing individual income and use it to generate M pseudocomplete data sets. Then, we fit the mortality model to the M data sets, each producing a posterior distribution of $\boldsymbol{\theta}$ conditioned on that pseudocomplete data set. We combine the M posterior distributions via an equally weighted mixture to produce the overall posterior distribution which generates our inferences. The foregoing is an approximation to the full probability modelling approach to multiple imputation (Rubin, 1987) where imputed values are generated from the same distribution as in a fully Bayesian data augmentation.

The two-step approach can yield approximately valid inferences, if the imputation model is appropriately built and the outputs from the imputed data sets are appropriately combined (Rubin, 1996; Schafer, 1997). This two-step approach has been widely adopted (Hopke *et al.*, 2001; Horton and Lipsitz, 2001; van Buuren *et al.*, 1999). In addition, it can also be attractive in the sense that, by not structuring the imputations around a specific analysis model, the imputed data can then be used for a variety of analyses.

3.3.1. Imputation of missing individual level income

Let i denote individual and j_c denote zip code j in county c. We build an imputation model as follows:

$$\log(\text{income}_{ij_c}) = \gamma_0 + \sum_{agr} \gamma_{agr} Z_{ij_c}(arg) + \gamma_2 D_{ij_c} + \gamma_3 \mathbf{W}_{j_c} + \gamma_4 \mathbf{V}_c + \xi_{j_c} + \varepsilon_{ij_c},$$
$$\xi_{j_c} \sim N(0, \omega^2), \quad \varepsilon_{ij_c} \sim N(0, \tau^2), \tag{5}$$

where D_{ij_c} is the indicator of death and $Z_{ij_c}(arg)$ is the age (a) $\times$ race (r) $\times$ gender (g) category indicator, $a = 1, \ldots, 5, r, g = 1, 2$, with $\gamma_{522} = 0$. $\mathbf{W}_{j_c}$ contains the zip code level variables log(median household income), percentage black, log(percentage poverty), log(percentage high school completeness), percentage college degree completeness, percentage public transportation, percentage house owner, percentage house owner over 65 years old and percentage unemployment, and $\mathbf{V}_c$ contains the county level variables log(median household income), percentage poverty and percentage high school completeness. Note that all the variables that are used in the main analysis (1) are included in the imputation model, including the outcome variable death. Otherwise the imputed data will be biased (Rubin, 1996). For example, if death is left out of the imputation model, the correlation between the imputed individual income and death will be biased towards 0. However, the association between the imputed income and risk of death in the main analysis will not be stronger than what is in the observed complete cases that the imputation model is fitted to. In our study, it will be the representative MCBS data. Moreover, we select model (5) among a large set of reasonable alternatives because it minimizes the cross-validation prediction error. Model (5) is also effective in capturing the variability in log(individual income) by minimizing the residual sum of squares. However, the prediction performance and the residual sum of squares become similar after including the individual level covariates and several zip code level SES covariates. We perform this model selection without random effects in the model, and we then include the random intercept ξ_{j_c} to account for the possible within-zip-code correlation. Sensitivity analyses on the choice of the imputation model are presented in Section 5.2.

Data on 1680 individuals from 410 zip codes are used to fit the imputation model. Let $\boldsymbol{\gamma}$ denote all coefficients in model (5), so $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \omega^2, \tau^2)$ denotes all parameters in the imputation model. We specify a standard non-informative prior $\Pr(\boldsymbol{\eta}) \propto \tau^{-2}\omega^{-2}$ (Gelman *et al.*, 2003), and

the posterior distribution of $\boldsymbol{\eta}$ is estimated by using a Gibbs sampler (Casella and George, 1992). According to the multiple-imputation method, we generate eight copies of the missing individual level income data, each using an independent posterior sample of $\boldsymbol{\eta}$. We estimate $\boldsymbol{\theta}$ and the association measures which will be defined in Section 3.4 for each copy of the pseudocomplete data set and combine the estimates across the eight data sets by pooling the draws from the eight posterior distributions.

3.3.2. Fitting of mortality model

In this section we describe fitting of the mortality model to a pseudocomplete data set. As a prior distribution for the fixed effect parameters α_0, α_1, each element of $\boldsymbol{\alpha}_2$ and α_{01}, we specify independent vague Gaussians $N(0, 2.5 \times 10^3)$. Let $p \times p$ denote the dimension of Σ_0. As a prior distribution for Σ_0 we specify an inverse Wishart distribution with $p+2$ degrees of freedom. The shape parameter of the inverse Wishart prior is selected by simulation, i.e. we select the shape parameter that leads to diffuse prior samples of the zip-code-specific coefficients. As a prior distribution for ϕ we specify vague Gaussians $N(0, 2.5 \times 10^3)$ for $\log(\phi)$. Finally we assume that the prior distributions are mutually independent:

$$p(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2, \alpha_{01}, \Sigma_0, \phi) = p(\alpha_0)\, p(\alpha_1)\, p(\boldsymbol{\alpha}_2)\, p(\alpha_{01})\, p(\Sigma_0)\, p(\phi).$$

We generate posterior samples of all unknown parameters by implementing a Gibbs sampler with Metropolis–Hastings steps (Chib and Greenberg, 1995). Detailed formulae of the conditional distributions that were used in the Gibbs sampler are provided in Appendix A. The burn-in consists of 2×10^4 iterations, and 5×10^3 iterations were used for posterior summaries. Convergence of the Markov chains was assessed by using the Gelman and Rubin convergence statistic (Gelman and Rubin, 1992).

3.4. Association measures

The common approach of reporting the association between race and risks of mortality is to report the fixed effect race coefficient α_1 in equation (1), whose interpretation depends on the coding of the race covariate and is conditioned on the zip code level random effects. For direct understanding of the differences in risk of mortality between the black and white populations, we estimate and report population level attributable risk AR, relative attributable risk RAR and odds ratio OR of death comparing blacks *versus* whites. This approach of reporting association measures that are functions of predicted values has the advantage that their interpretation does not depend on model structure and parameterization (e.g. on the conditional structure of random effects and on covariate centring and scaling).

Let

$$P_{ijtb} = \Pr(D_{ijt} = 1 | \text{race}_{ij} = \text{black}, \mathbf{X}_{ij} = \mathbf{x}_{ij}, \text{Income}_j = \text{income}_j, \boldsymbol{\theta}),$$

$$P_{ijtw} = \Pr(D_{ijt} = 1 | \text{race}_{ij} = \text{white}, \mathbf{X}_{ij} = \mathbf{x}_{ij}, \text{Income}_j = \text{income}_j, \boldsymbol{\theta})$$

denote the predicted probabilities of death in year t for a black person and a white person respectively, whose other covariates values are the same as for the ith individual in the jth zip code, where $\boldsymbol{\theta}$ denotes all parameters. We define the population level AR, RAR and OR as follows:

$$\text{AR} = P_{\ldots b} - P_{\ldots w}, \tag{6}$$

$$\mathrm{RAR} = \frac{P_{\cdot\cdot\cdot b} - P_{\cdot\cdot\cdot w}}{P_{\cdot\cdot\cdot w}}, \tag{7}$$

$$\mathrm{OR} = \frac{P_{\cdot\cdot\cdot b} Q_{\cdot\cdot\cdot w}}{P_{\cdot\cdot\cdot w} Q_{\cdot\cdot\cdot b}} \tag{8}$$

where $P_{\cdot\cdot\cdot b} = \Sigma_{i,j,t}\, P_{ijtb}$, $P_{\cdot\cdot\cdot w} = \Sigma_{i,j,t}\, P_{ijtw}$, $Q_{\cdot\cdot\cdot b} = 1 - P_{\cdot\cdot\cdot b}$ and $Q_{\cdot\cdot\cdot w} = 1 - P_{\cdot\cdot\cdot w}$.

Similarly we define zip code level summaries AR_j, RAR_j and OR_j, but without the summation across zip codes. The posterior samples of AR, RAR and OR as well as the posterior samples of AR_j, RAR_j and OR_j were computed by using the posterior draws of $\boldsymbol{\theta}$ from the Markov chain Monte Carlo output.

4. Results

Table 1 summarizes the age, race, gender, mortality and income distributions from the Medicare, census 2000 and MCBS data sets for the study area. It shows substantial heterogeneity in the racial and income distributions across the 2095 zip codes, but not much heterogeneity in the age and gender distributions.

Fig. 2 shows the posterior densities of AR, RAR and OR of death comparing blacks *versus* whites, adjusted and not adjusted for income. We find that, after controlling for age, gender and their interaction, there is a statistically significantly higher risk of death for blacks compared with whites, both adjusted and not adjusted for income. In addition, the posterior distributions adjusted for income are much wider compared with those not adjusted for income, which is because individual level income data are available for only 0.04% of the study population. Accounting for the uncertainty from these missing data leads to higher posterior variance in our parameters of interest.

Specifically, the posterior mean and 95% credible interval of AR are respectively 17.2×10^{-3} and $(16.1\text{–}18.6) \times 10^{-3}$ not adjusted for income and 4.3×10^{-3} and $(0.9\text{–}8.2) \times 10^{-3}$ adjusted for income. It means that, after controlling for age and gender as well as their interaction, the difference in the probability of death within 1 year comparing the black population *versus* the white population is 17.2×10^{-3}, and that difference reduces to 4.3×10^{-3} after further adjusting for both individual level and zip code level income.

Table 1. Summary of the age, race, gender, mortality and income distributions from the Medicare, census 2000 and MCBS databases for the study area

Data set	*Parameter*	*Population measure*	*Zip-code-specific measure*					
			Minimum	*5th percentile*	*35th percentile*	*65th percentile*	*95th percentile*	*Maximum*
Medicare	%black	13	0	0	0.8	4.0	60	98
	% male	40	18	36	40	42.6	48.3	100
	average age (years)	76	68	74	76	77	78	83
	Total deaths	963702	0	17	168	506	1324	3808
Census	Zip code median household income ($)		0	27500	50000	65000	100000	200000
MCBS	Average income of person surveyed and spouse, if any ($)	25000						

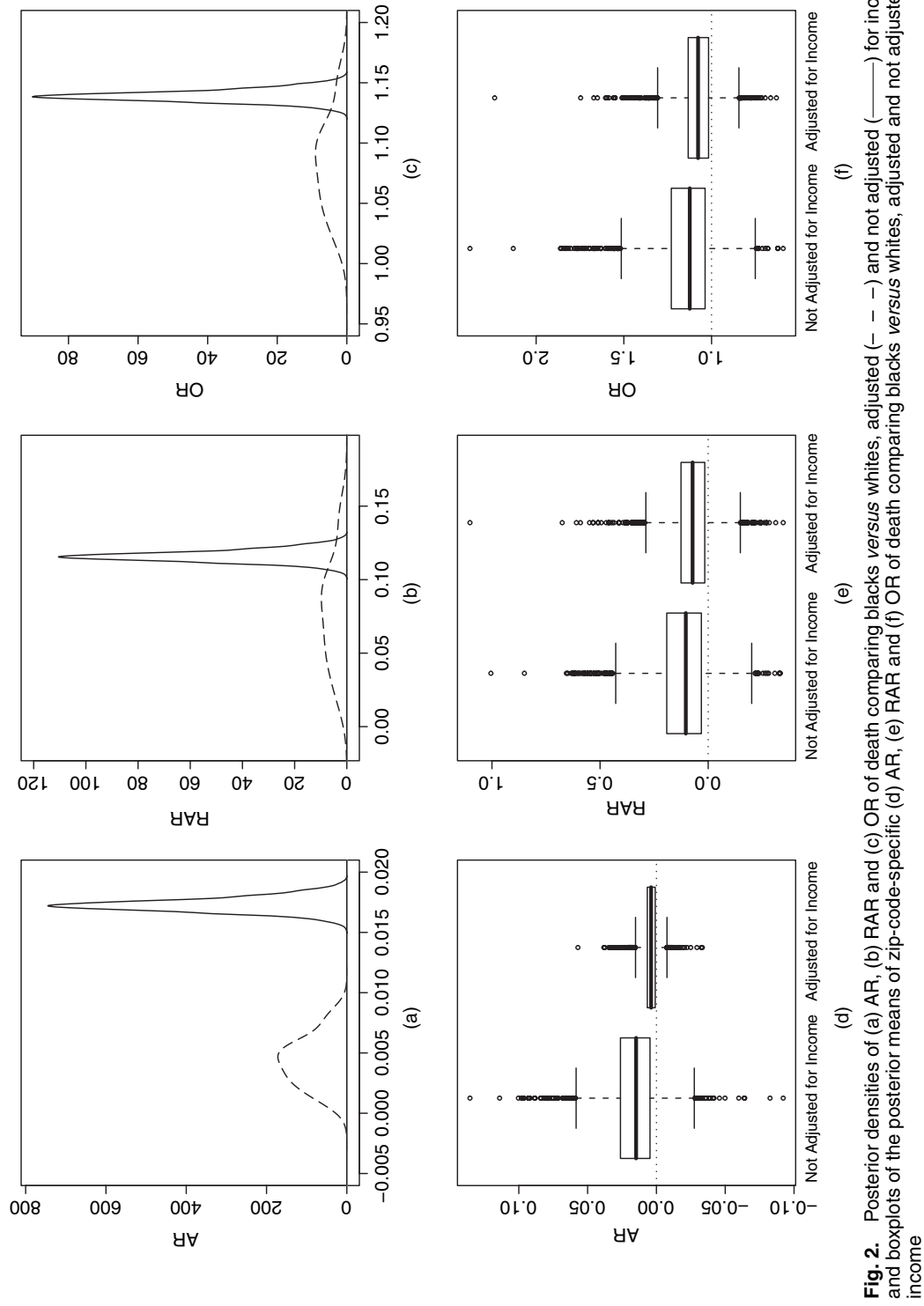

Fig. 2. Posterior densities of (a) AR, (b) RAR and (c) OR of death comparing blacks *versus* whites, adjusted (– – –) and not adjusted (———) for income, and boxplots of the posterior means of zip-code-specific (d) AR, (e) RAR and (f) OR of death comparing blacks *versus* whites, adjusted and not adjusted for income

The posterior mean and 95% credible interval of RAR are respectively 11.6% and (10.8–12.5)% not adjusted for income and 7.7% and (1.7–14.8)% adjusted for income. It means that, as a relative measure, there is a higher risk of death of 11.6% for the black population compared with the white population when controlling for age, gender and their interaction, and the risk is still 7.7% higher after further adjusting for both individual and zip codes level income. In other words, noting the relationship between RAR and relative risk RR, $\mathrm{RAR} = \mathrm{RR} - 1$, the relative risk of death in 1 year comparing the black population *versus* the white population is 1.12 (95% credible interval 1.11–1.13) not adjusted for income and 1.08 (95% credible interval (1.02–1.15)) adjusted for income, controlling for age, gender and their interaction.

The posterior mean and 95% credible interval of OR are respectively 1.14 and 1.13–1.15 not adjusted for income and 1.08 and 1.02–1.16 adjusted for income, controlling for age and gender as well as their interaction. The Monte Carlo standard errors of AR, RAR and OR both adjusted and not adjusted for income are relatively small compared with the posterior estimates.

There is a statistically significant reduction in AR after the adjustment for income, because the posterior densities of AR adjusted and not adjusted for income do not overlap. RAR and OR also decrease after the adjustment for income; however, the magnitude of decrease is relatively small.

Fig. 2 also shows the boxplots of posterior means of the zip-code-specific AR, RAR and OR of death comparing blacks *versus* whites, adjusted and not adjusted for income. In all three measures we find that, for more than 75% of the zip codes in the study area, the black population shows a higher risk of death compared with the white population residing in that same zip code, both adjusted and not adjusted for income. In addition, the posterior means of the zip-code-specific AR decrease and they are less heterogeneous after the adjustment for income. We find that same pattern in the zip-code-specific RAR and OR, but to a smaller extent.

We note that the posterior distributions of the population level AR, RAR and OR widen after the adjustment for income; however, the distributions of the 2095 posterior means of zip-code-specific AR, RAR and OR narrow after the same adjustment. This is because the distribution of zip code level posterior means only accounts for the between-zip-code variance component, whereas the population level posterior distribution accounts for both the between- and the within-zip-code variance components. Note that the within-zip-code variance component is largely inflated after the adjustment for income, owing to the uncertainty from the large fraction of missing individual level income data.

In estimating AR, RAR and OR adjusted for income, the percentage of missing information is 97%, whereas the achieved relative efficiency in the estimation is still approximately 95%. The gain in the relative efficiency by increasing the number of imputations from the commonly suggested 5 to 8 is relatively small (91% *versus* 95%).

5. Sensitivity analysis

5.1. Sensitivity to numbers of imputations

Because 99.6% of the individual level income data are missing, we further investigate whether using eight copies of imputed data is sufficient to capture the between-imputation variability when estimating the parameters of interest. Specifically, we generate another five copies of imputed data by using the imputation model in Section 3.3.1. We compare the estimated posterior distributions of AR, RAR and OR under model (1) when using the total 13 copies of pseudocomplete data sets compared with using the previous eight copies, and we find only negligible differences in the comparison.

Table 2. Estimates of the race coefficient, population level AR, RAR and OR of death comparing blacks *versus* whites adjusted for income under logistic regression model (1) without random effects, with missing individual level income imputed from various imputation models with no random effects†

Covariate	*Estimates and standard errors for the following models‡:*				
	Model 1: age $*\ race$ $*\ gender$ $death$ $log(income_z)$	*Model 2:* age $*\ race$ $*\ gender$ $death$ $log(income_z)$ $\%\ black_z$	*Model 5:* age $*\ race$ $*\ gender$ $death$ $log(income_z)$ $\%\ black_z$ $log(\%\ poverty_z)$ $log(\%high_z)$	*Model 7:* age $*\ race$ $*\ gender$ $death$ $log(income_z)$ $\%\ black_z$ $log(\%\ poverty_z)$ $log(\%high_z)$ $log(income_c)$ $\%\ poverty_c$ $high_c$	*Model 9:* age $*\ race$ $*\ gender$ $death$ $log(income_z)$ $\%\ black_z$ $log(\%poverty_z)$ $log(\%high_z)$ $Other\ SES_z$ $log(income_c)$ $\%\ poverty_c$ $\%\ high_c$
Race coefficient	$-0.002_{0.0184}$	$0.025_{0.010}$	$0.030_{0.007}$	$0.026_{0.014}$	$0.035_{0.010}$
AR	$-0.0002_{0.0018}$	$0.0026_{0.0010}$	$0.0031_{0.0008}$	$0.0027_{0.0014}$	$0.0037_{0.0010}$
RAR	$-0.0035_{0.0325}$	$0.047_{0.019}$	$0.057_{0.014}$	$0.050_{0.026}$	$0.066_{0.019}$
OR	$0.9963_{0.0344}$	$1.050_{0.020}$	$1.060_{0.015}$	$1.053_{0.028}$	$1.070_{0.020}$

†Selected results of five out of nine imputation models.
‡z indexes the zip code and c indexes county; other SES_z includes zip code level percentage degree completeness, percentage public transportation, percentage house owner, percentage house owner $\geqslant$ 65 years and percentage unemployment.

5.2. Sensitivity to different imputation models

To examine the validity of the imputation model, we explore the sensitivity of the estimated AR, RAR and OR adjusted for income with respect to different imputation models. This analysis is conducted in two steps. We first examine the sensitivity of the AR-, RAR- and OR-estimates when using an imputation model with no random effects but with different sets of covariates. We then also investigate the sensitivity of the estimates when including a random intercept to the imputation model.

Specifically in step I we impute the missing income_{ij}-values by using nine sets of nested covariates in an imputation model without a random effect, and we estimate AR, RAR and OR under model (1) without the random effects $\mathbf{U}_j$. The full random-effect mortality model is not used owing to the computational burden of implementing Markov chain Monte Carlo sampling on the large data set. Because we are interested in the comparison of imputation models, we do not expect a large discrepancy between the comparison based on the simpler mortality model and that based on the full random-effect mortality model. Table 2 shows the estimated race coefficient, population level AR, RAR and OR of death comparing blacks *versus* whites for five out of the nine sets of covariates. The estimates are combined across eight pseudocomplete data sets according to the multiple-imputation method. We find that, after including the zip code level percentage black in the imputation model, the estimates are not sensitive to further inclusion of other zip code level and county level variables.

In step II we compare the posterior densities of AR, RAR and OR when using the imputation model (5) with and without the zip code level random intercept ξ_{j_c}. Different from in step I, we use a simplified random-intercept mortality model to reduce the computational burden, in which the fixed effects contain only race and individual level income, and the coefficients of all other

covariates are fixed at their maximum likelihood estimate. We carried out the analysis using 50 zip codes selected at random, and estimates were combined across eight pseudocomplete data sets. Fig. 3 shows the estimated posterior densities of AR, RAR and OR adjusted for income, obtained by using the imputation models with and without random intercept. Under the imputation model with the random intercept, we find that the posterior distributions of AR, RAR and OR have slightly larger posterior variances and smaller posterior means. These results indicate that failure to account for the within-zip-code correlation in the imputation model will impact not only on the uncertainty but also on the actual values of the estimated AR, RAR and OR.

5.3. Full Bayesian data augmentation versus multiple imputation

In addition to the finite small number of imputations (eight in our study), two major discrepancies exist between the two-step multiple imputation and the full Bayesian data augmentation. First, only the complete cases from the MCBS instead of all the observed data are used in imputing the missing individual income. Second, the precise structure of $\boldsymbol{\theta}$ is not used in the imputation, although the imputation model includes all the variables from the mortality model. We expect only some loss of efficiency due to the first discrepancy, because the MCBS population is a representative subset of the Medicare population. To investigate the effect from the second discrepancy, we conduct a sensitivity analysis to compare the results when missing data are handled by using full Bayesian data augmentation *versus* the two-step multiple imputation. We perform the sensitivity analysis by using the MCBS data where 30% of the individual level income are pretended to be missing. The analysis is conducted under two model assumptions:

(a) assuming no random effects for both the imputation model (5) and the mortality model (1);
(b) assuming a zip code level random intercept for both the imputation model (5) and the mortality model (1).

Because the MCBS data are sparse across zip codes, random-effect models that are more complicated than in assumption (b) cannot be stably estimated.

Fig. 4 compares the posterior densities of AR, RAR and OR adjusted for income when missing data are handled by using full Bayesian data augmentation *versus* the two-step multiple imputation, under the analysis and model settings above. We find only negligible differences in the comparison under the independent observation assumption (a). Differences arise under the random-effect model setting (b); however, there is large overlap in the 95% credible intervals between the two approaches. The observed differences could be explained by our use of a 0–1 indicator of death in the imputation model, instead of the yearly survival indicator with between-zip-code heterogeneity as appeared in the mortality model. Such heterogeneity is non-negligible and therefore leads to the differences which are not observed under the independence setting (a).

5.4. Choice of summary statistic for zip code level income

The literature suggests that the type of summary of individual level income at area level that best predicts rates of mortality depends on the geographic size of the area (Wilkinson, 1997). Therefore, we conduct a sensitivity analysis to identify how to summarize the imputed income_{ij} at zip code level to explain the variability in mortality risks best.

Model (1) can be rewritten as

$$\text{logit}\{\Pr(D_{ijt}=1)\} = \alpha_0 + U_{0j} + (\alpha_1 + U_{1j})\,\text{race}_{ij} + \mathbf{X}_{ij}(\boldsymbol{\alpha}_2 + \mathbf{U}_{2j}) + \alpha_{01}\,\text{income}_j, \qquad U_{0j} \sim N(0, \sigma_0^2).$$

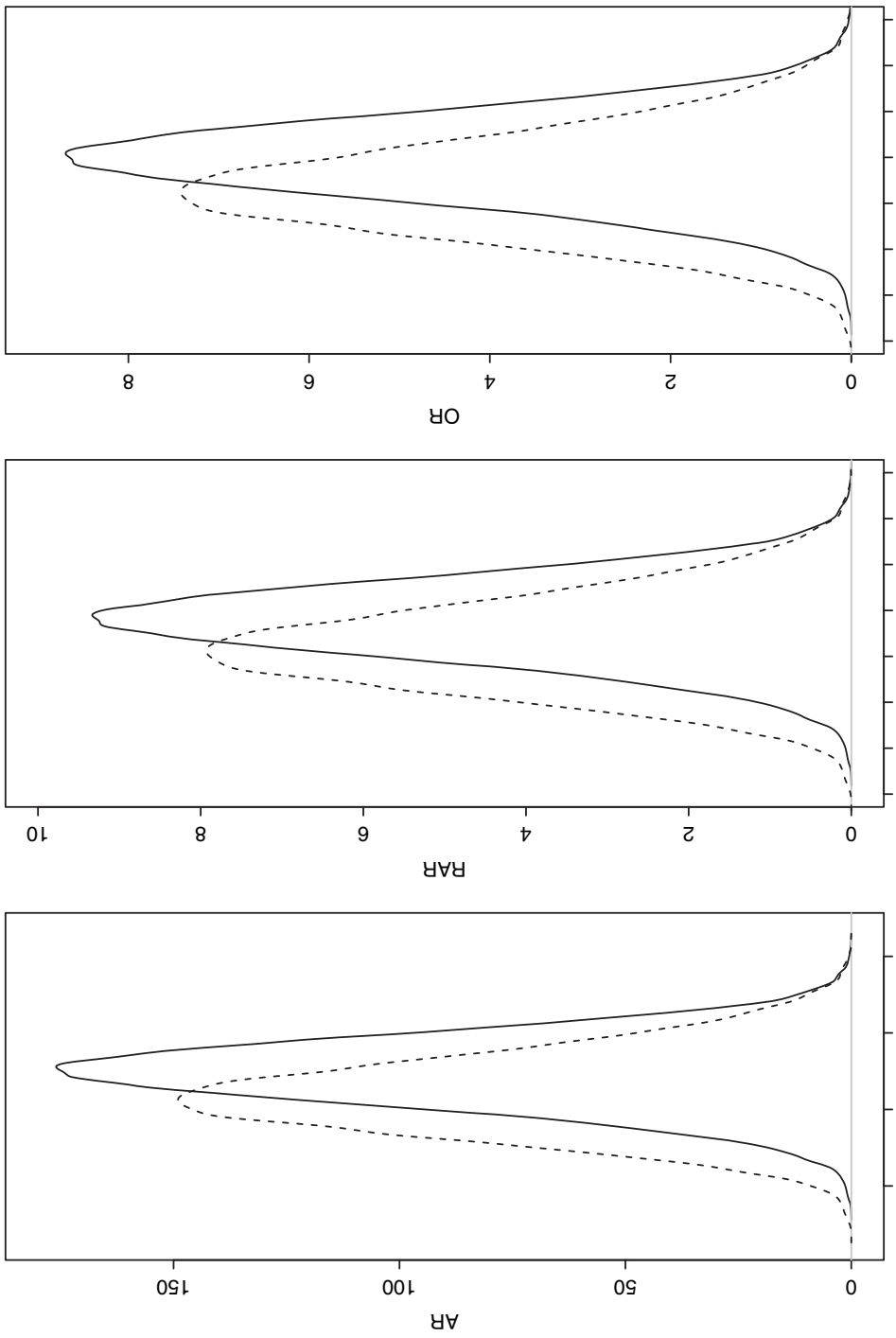

Fig. 3. Posterior densities of (a) AR, (b) RAR and (c) OR of death comparing blacks *versus* whites adjusted for income for 50 zip codes selected at random, when using the imputation model with (- - - - - -) and without (——) zip code level random intercept

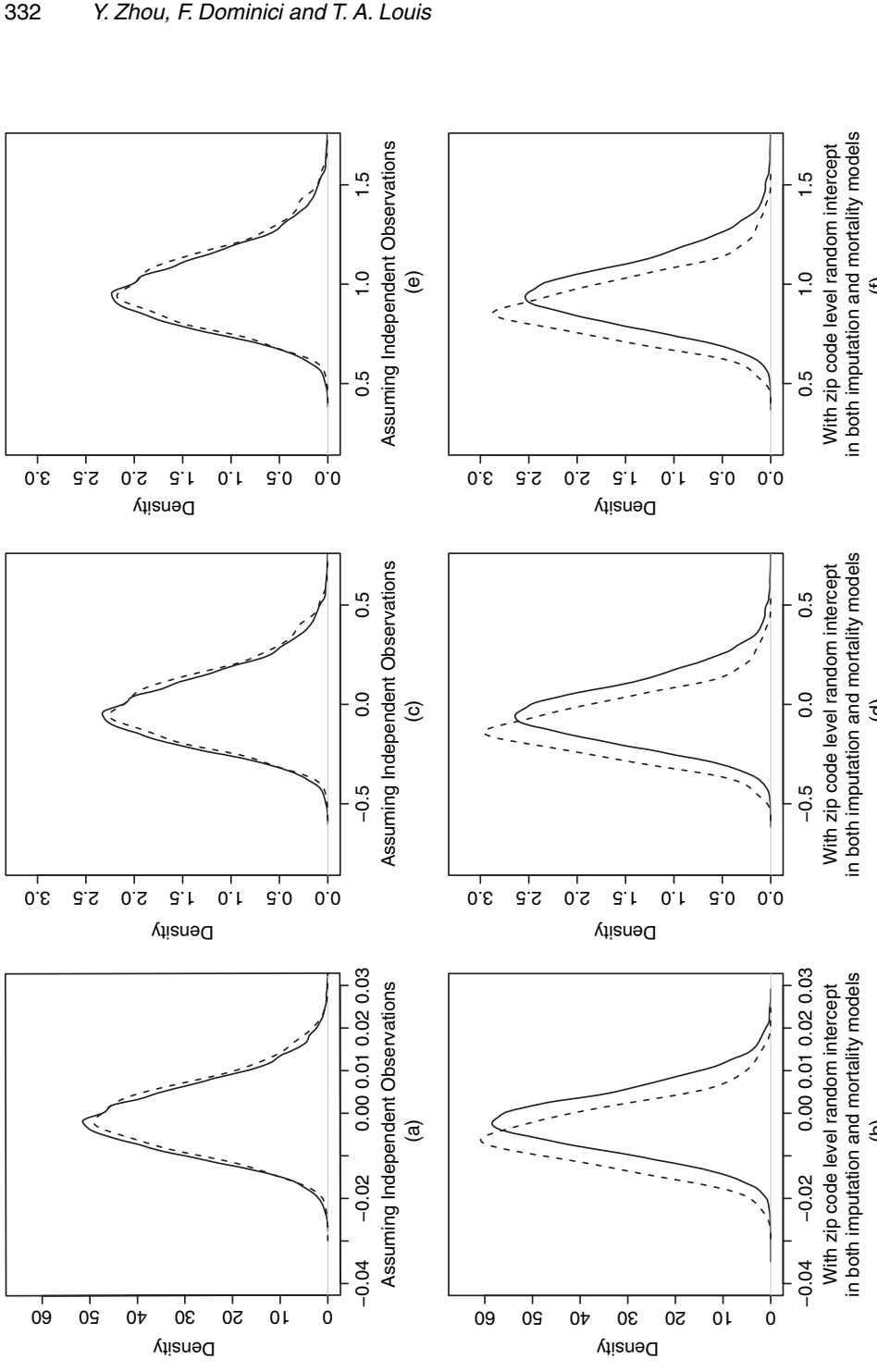

Fig. 4. Posterior densities of (a), (b) AR, (c), (d) RAR and (e), (f) OR of death comparing blacks *versus* whites adjusted for income, using MCBS data where 30% of the individual level income is pretended to be missing, and the missing data are handled by using two-step multiple imputation (——) and full Bayesian data augmentation, under the following two model settings: (a), (c), (e) assuming no random effects for both the imputation model and the mortality model and (b), (d), (f) assuming a zip code level random intercept for both the imputation model and the mortality model

σ_0^2 measures the between-zip-code variability of the baseline risk of death that cannot be explained by the individual level covariates. Therefore, instead of using the zip code level median income, which is income_j, the objective is to identify a zip code level income summary measure, which is denoted by $\overline{\text{income}}_j$, that minimizes σ_0^2.

We consider two components for $\overline{\text{income}}_j$:

(a) a typical value to represent the absolute level of income of a zip code, measured by the mean or a percentile;
(b) a spread component to represent the within-zip-code income inequality, measured by IQR or standard deviation. We perform the following four analyses:
   (i) not adjusted for income_{ij}, a sensitivity analysis for different typical values alone as $\overline{\text{income}}_j$;
   (ii) not adjusted for income_{ij}, a sensitivity analysis for different combinations of typical value and spread component together as $\overline{\text{income}}_j$;
   (iii) adjusted for income_{ij}, a sensitivity analysis for different typical values alone as $\overline{\text{income}}_j$;
   (iv) adjusted for income_{ij}, a sensitivity analysis for different combinations of typical value and spread component together as $\overline{\text{income}}_j$.

We perform separate analyses adjusted and not adjusted for income_{ij} because they correspond to different scientific questions: what are the zip code level summaries of individual level income that explain most of the variability in risks of mortality, adjusted and not adjusted for individual level income?

In this sensitivity analysis, we apply the log-transformation to the individual level income variable and calculate the zip code level income summaries from the transformed individual level income variable. This is because the untransformed income variable follows a log-normal distribution as is assumed in the imputation model (5). For a log-normally distributed random variable, its percentiles and IQR are linearly related and therefore cannot be simultaneously included in the regression model.

Let $\{\overline{\text{income}}_{jk}, k = 1, \ldots, K\}$ denote the set of candidate $\overline{\text{income}}_j$ measures that we consider, and let σ_{0k}^2 denote the σ_0^2 when using measure $\overline{\text{income}}_{jk}$. We compare σ_{0k}^2 and $\sigma_{0k'}^2, \forall k \neq k'$, by calculating the pairwise posterior probability $\Pr(\sigma_{0k}^2 < \sigma_{0k'}^2)$, where a probability value that is close to $\frac{1}{2}$ suggests approximately equal values of σ_{0k}^2 and $\sigma_{0k'}^2$. In addition, we rank $\{\sigma_{0k}^2, k = 1, \ldots, K\}$ in a descending order where the rank of σ_{0k}^2 is calculated as $\Sigma_{\forall k' \neq k} \Pr(\sigma_{0k}^2 < \sigma_{0k'}^2)$ (Shen and Louis, 1998). It is the posterior mean of the integer rank $\Sigma_{\forall k' \neq k} I_{\{\sigma_{0k}^2 < \sigma_{0k'}^2\}}$ as well as the optimal rank under the squared error loss, and it represents the distance between the parameters that are ranked. The standardized rank is $\Sigma_{\forall k' \neq k} \Pr(\sigma_{0k}^2 < \sigma_{0k'}^2)/(K-1)$.

We carry out this sensitivity analysis with only 100 zip codes selected at random, and with only one imputed data set. Following the multiple-imputation method strictly, we should do the analysis with each of the eight imputed data sets and combine the results. However, because we are interested in the comparison between different summary measures of income_{ij} at zip code level, we do not expect large between-imputation variation in the comparison results. For analyses (i) and (iii), we consider the following eight different typical values: the fifth, 10th, 25th, 75th, 90th and 95th percentiles, mean and median. For analyses (ii) and (iv), we consider the following five combinations: IQR, 25th percentile and IQR, 25th percentile and standard deviation, median and IQR, median and standard deviation. Both adjusted and not adjusted for income_{ij}, we find moderate posterior probabilities of pairwise comparison (ranging between 0.30 and 0.70) and small differences in the ranks (standardized rank ranging between 0.38 and 0.6). In addition, the values of the estimated σ_0^2 do not differ much when using different $\overline{\text{income}}_j$ measures. The results suggest approximately equal performance for either typical value alone, spread measure

alone or both together in explaining the variability in risks of mortality that is not explained by the individual level variables. In addition, the poor part, the wealthy part and the middle part equivalently represent the typical zip code income level in explaining that variability.

6. Discussion

In this paper we have presented a large study to estimate the racial disparities in risks of mortality. We developed and applied hierarchical statistical models to estimate the age- and gender-adjusted association between individual race and risks of mortality, as well as how this association varies when adjusted for both individual level and zip code level income. An important strength of the study is the scope of the study population, which includes more than 4 million individuals over 65 years old in the north-east region of the USA.

To assess the differences in risk between the black and white populations, we defined and reported the population level or marginal AR, RAR and OR of death comparing blacks *versus* whites that are functions of the predicted probabilities of death. The marginal estimands AR and RAR in equations (6) and (7) which are computed by using the population level summary probabilities can also be defined as averages or weighted averages of the individual-specific AR_{ijt} and RAR_{ijt} respectively, where $\mathrm{AR}_{ijt} = P_{ijtb} - P_{ijtw}$ and $\mathrm{RAR}_{ijt} = (P_{ijtb} - P_{ijtw})/P_{ijtw}$. However, for the odds ratio, the marginal estimand OR in equation (8) differs from the weighted average of individual-specific

$$\mathrm{OR}_{ijt} = P_{ijtb} Q_{ijtw} / Q_{ijtb} P_{ijtw},$$

where $Q_{ijtb} = 1 - P_{ijtb}$ and $Q_{ijtw} = 1 - P_{ijtw}$. We used the marginal estimand of OR because it is more directly related to our goal of comparing the risks between two populations.

Our study shows a higher risk of death for blacks compared with whites, in terms of AR, RAR and OR, both adjusted and not adjusted for income. After further adjusting for both individual level and zip code level income, there is a statistically significant reduction in AR, which suggests that the absolute difference in risk of mortality between the black population and the white population may be lowered by reducing their differences in both the individual level income as well as the income level of the zip code of residence. However, the reduction is small for RAR (which equals RR-1) and OR, which are relative measures of the difference in risk.

We addressed the missing data challenge by using multiple imputation. Sensitivity in the parameter estimates to different imputation models as well as to different numbers of imputations was examined. We compared the ideal full Bayesian data augmentation *versus* the two-step multiple imputation in handling missing data by using the MCBS data set. Because the complete structure of the mortality model parameters $\boldsymbol{\theta}$ is not used for the imputation, differences can arise between the two approaches. However, the resulting caution against using the more implementable two-step multiple imputation should be balanced with the need for a general and good imputed data set which may suit a variety of analyses. Interestingly, we find that using log-transformed income instead of the original scale income in the mortality model leads to a greater discrepancy between the two approaches. This is because the relationship between race and income changes after the log-transformation, which leads to different estimates of AR, RAR and OR. In addition to the sensitivity analyses regarding the imputation, we also examined various zip code level income summary measures in explaining most of the variability in risks of mortality, and we found equal performance when using typical values of zip code income level, or within-zip-code income inequality, or both together.

In the development of our hierarchical models, we have investigated the necessity of including the interactions of race with age and individual level income. The results suggest small differences in the racial disparity estimates across different age and income strata. Also, in developing our hierarchical models, we have assumed a multivariate normal distribution for the random effects and a separable model for the covariance structure of the random effects. This approach is a special case of the spatial models with spatially varying-coefficient processes (Gelfand *et al.*, 2003; Banerjee and Johnson, 2006). It is a flexible and effective approach to model directly the covariance or correlation structure of the joint distribution of random effects which may itself be of interest. The parameter ρ in a separable model directly measures the spatial correlation between random effects. An alternative is a multivariate conditional auto-regression model based on adjacency (Gelfand and Vounatsou, 2003). A detailed comparison between the two models can be found in Banerjee *et al.* (2004).

An important concern in studies of racial disparities in health and mortality is the controversy about the conceptualization of 'the effect of race'. Kaufman and Cooper (1999) argued that there is no meaningful causal effect for race, because a causal effect can only be defined for factors that are plausible to be assigned as treatment in hypothetical experiments. Race is an attribute that is born with each individual and therefore cannot be assigned. Throughout our analysis, we focused on the *association* between race and an individual's risk of death.

Our analysis has limitations. The majority of the study population who are 65 years old and older are retired. Therefore, income may under-represent the differentiation in individual level SES. Wealth is a more appropriate measure; however, data on wealth are not available. Other individual level SES variables that are available in the MCBS data set include education, job status and marital status. It is expected that a combination of income and other SES variables will better represent differentiation in individual level SES. However, we find little difference in the estimates of AR, RAR and OR when using education and income together *versus* using income alone, which suggests that further adjusting for education besides income has limited effect in estimating the racial disparities in risks of mortality adjusted for SES. We suspect similar findings when job status and marital status are further adjusted besides income and education. For zip code level SES, we used only the median household income variable to match with the individual level SES variable which contains only individual level income. However, area level SES variables are so highly correlated with each other that a single measure for area level SES will not introduce a big problem of misspecification (Pickett and Pearl, 2001; Diez-Roux *et al.*, 2001).

Some researchers have argued that the correct definition of area for the effect of area level SES is important (Pickett and Pearl, 2001). Neighbourhood is the believed contextual area whose SES level truly affects the individual residents. Use of administrative boundaries such as zip codes may not capture the health- and service-related features of SES of the neighbourhood. Generally, the distributions of SES variables are more heterogeneous within zip codes as expected from neighbourhoods. The census tract is believed to be a closer representation of neighbourhoods. However, the literature suggests that, in estimating the association between area level SES and health as well as in estimating racial disparities in health adjusted for area level SES, the differences between using zip code, census tract and block group level SES variables are small (Geronimus and Bound, 1998; Soobader *et al.*, 2001).

In addition, the literature suggests that there are significant rural and urban differences in the racial disparities in risks of mortality (Clifford and Brannon, 1985). The time interval between area level SES exposure and death is another important issue. Bosma *et al.* (2001) showed that the association between neighbourhood SES and risks of mortality is stronger for people who

live in their neighbourhood longer. Failure to account for recent immigration could bias the association towards zero. Although our data are not completely cross-sectional, information on the length of residence is missing.

Acknowledgements

We are grateful to Dr Thomas A. Glass and Dr Thomas A. LaVeist for their valuable advice and comments, and Dr Aidan McDermott for help on the sources of data.

Support for this work was provided by the National Institute for Environmental Health Sciences (grant ES012054-03), by the Environmental Protection Agency (grant RD83054801) and by the National Institute of Diabetes, Digestive and Kidney Diseases (grant R01 DK061662).

Appendix A: Conditional distributions in the Gibbs sampler for mortality model fitting

We derive the conditional distributions under a two-stage representation of the hierarchical model (1).

A.1. Conditional distributions for model not adjusted for income

Denote $\boldsymbol{\alpha}^{*\prime} = (\alpha_0^*, \alpha_1^*, \boldsymbol{\alpha}_2^{*\prime})'$ as the vector of second-stage coefficients and $\boldsymbol{\beta}_j^{*\prime} = (\beta_{0j}^*, \beta_{1j}^*, \boldsymbol{\beta}_{2j}^{*\prime})' = \boldsymbol{\alpha}^{*\prime} + \mathbf{U}_j^{*\prime}$ as the vector of first-stage coefficients of zip code j, $j = 1, \ldots, J$. Denote $\mathbf{Z}_{ij}^* = (1, \text{race}_{ij}, \mathbf{X}_{ij}^*)$ as the vector of all individual level covariates of individual i in zip code j.

Let p denote the length of vector $\boldsymbol{\beta}_j$, $B_{p\times J}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \ldots, \boldsymbol{\beta}_J^*)$, $\boldsymbol{\beta}^* = \text{vec}(B^*)$, and let t denote year:

(a)

$$[\boldsymbol{\beta}_j^*, j = 1, \ldots, J | \cdot] \propto \left\{ \prod_j \prod_i \prod_t \frac{\exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_j^* D_{ijt})}{1 + \exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_j^* D_{ijt})} \right\} \exp\{-(\boldsymbol{\beta}^* - \mathbf{1}_J \otimes \boldsymbol{\alpha}*)' \boldsymbol{\Sigma}^{*-1} (\boldsymbol{\beta}^* - \mathbf{1}_J \otimes \boldsymbol{\alpha}^*)\};$$

(b) for the prior of $\boldsymbol{\alpha}^* \sim \text{MVN}(\mathbf{0}, \Sigma_{\text{prior}}^*)$,

$$[\boldsymbol{\alpha}^* | \cdot] \sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\alpha}^*}, \Sigma_{\boldsymbol{\alpha}^*}),$$
$$\Sigma_{\boldsymbol{\alpha}^*} = \{\sum_i \sum_j (H^{-1})_{ij} \Sigma_0^{*-1} + \Sigma_{\text{prior}}^{*-1}\}^{-1},$$
$$\boldsymbol{\mu}_{\boldsymbol{\alpha}^*} = \Sigma_{\boldsymbol{\alpha}^*} \Sigma_0^{*-1} B^* H^{-1} \mathbf{1}_J,$$

where $(H^{-1})_{ij}$ denotes the element in the ith row and jth column of matrix H^{-1};

(c) for the prior of $\Sigma_0^* \sim \text{Inverse Wishart}(d_0^*, D_0^*)$,

$$[\Sigma_0^* | \cdot] \sim \text{Inverse Wishart}[d_0^* + J, \{\sum_i \sum_j (H^{-1})_{ij} (\boldsymbol{\beta}_i^* - \boldsymbol{\alpha}^*)(\boldsymbol{\beta}_i^* - \boldsymbol{\alpha}^*)' + D_0^{*-1}\}^{-1}],$$

where $(H^{-1})_{ij}$ is the same as above;

(d) for the prior of $(\phi^*) \sim N(0, \sigma_{\text{prior}}^{*2})$,

$$f(\phi^*) \propto \frac{1}{\phi^*} \exp\left\{ -\frac{\log(\phi^*)^2}{2\sigma_{\text{prior}}^{*2}} \right\},$$

$$[\phi^* | \cdot] \propto f(\phi^*) |H(\phi^*)|^{-p/2} \exp[-(\boldsymbol{\beta}^* - \mathbf{1}_J \otimes \boldsymbol{\alpha}*)' \{H(\phi^*)^{-1} \otimes \Sigma_0^{*-1}\} (\boldsymbol{\beta}^* - \mathbf{1}_J \otimes \boldsymbol{\alpha}^*)/2].$$

A.2. Conditional distributions for model adjusted for income

Denote $\boldsymbol{\alpha}' = (\alpha_0, \alpha_{01}, \alpha_1, \boldsymbol{\alpha}_2')'$ as the vector of second-stage coefficients, $\mathbf{Z}_{ij} = (1, \text{race}_{ij}, \mathbf{X}_{ij})$ as the vector of all individual level covariates of individual i in zip code j and

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \boldsymbol{\beta}_{2j} \end{pmatrix} = \begin{pmatrix} \alpha_0 + \alpha_{01}\,\text{income}_j + U_{0j} \\ \alpha_1 + U_{1j} \\ \boldsymbol{\alpha}_2 + \mathbf{U}_{2j} \end{pmatrix} = S\boldsymbol{\alpha} + \mathbf{U}_j$$

where

$$S = \begin{pmatrix} 1 & \text{income}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{p-1} \\ 1 & \text{income}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{p-1} \\ \vdots & \vdots & \vdots \\ 1 & \text{income}_J & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{p-1} \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_J \end{pmatrix}$$

is the vector of first-stage coefficients of zip code j, $j = 1, \ldots, J$.

Let p denote the length of vector $\boldsymbol{\beta}_j$, $B_{p\times J} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J)$, $\boldsymbol{\beta} = \text{vec}(B)$, and let t denote year:

(a)

$$[\boldsymbol{\beta}_j, j = 1, \ldots, J|\cdot] \propto \left\{ \prod_j \prod_i \prod_t \frac{\exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_j D_{ijt})}{1 + \exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_j D_{ijt})} \right\} \exp\{-(\boldsymbol{\beta} - S\boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - S\boldsymbol{\alpha})\};$$

(b) for the prior of $\boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, \Sigma_{\text{prior}})$,

$$[\boldsymbol{\alpha}|\cdot] \sim \text{MVN}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$$

where

$$\Sigma_{\boldsymbol{\alpha}} = \{S'(H^{-1} \otimes \Sigma_0^{-1})S + \Sigma_{\text{prior}}^{-1}\}^{-1},$$

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \Sigma_{\boldsymbol{\alpha}} S'(H^{-1} \otimes \Sigma_0^{-1})\boldsymbol{\beta};$$

(c) for the prior of $\Sigma_0 \sim \text{Inverse Wishart}(d_0, D_0)$,

$$[\Sigma_0|\cdot] \sim \text{Inverse Wishart}[d_0 + J, \{\sum_i \sum_j (H^{-1})_{ij}(\boldsymbol{\beta}_i - S_i\boldsymbol{\alpha})(\boldsymbol{\beta}_i - S_i\boldsymbol{\alpha})' + D_0^{-1}\}^{-1}],$$

where $(H^{-1})_{ij}$ denotes the element in the ith row and jth column of matrix H^{-1};

(d) for the prior of $\log(\phi) \sim N(0, \sigma^2_{\text{prior}})$,

$$f(\phi) \propto \frac{1}{\phi} \exp\left\{ -\frac{\log(\phi)^2}{2\sigma^2_{\text{prior}}} \right\},$$

$$[\phi|\cdot] \propto f(\phi)|H(\phi)|^{-p/2} \exp\{-(\boldsymbol{\beta} - S\boldsymbol{\alpha})'\{H(\phi)^{-1} \otimes \Sigma_0^{-1}\}(\boldsymbol{\beta} - S\boldsymbol{\alpha})/2\}.$$

References

Banerjee, S., Carlin B. P. and Gelfand A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall–CRC.

Banerjee, S. and Johnson, G. A. (2006) Coregionalized single- and multiresolution spatially varying growth curve modeling with application to weed growth. *Biometrics*, **62**, 864–876.

Bosma, H., de Mheen, H. D., Borsboom, G. J. and Mackenbach, J. P. (2001) Neighborhood socioeconomic status and all-cause mortality. *Am. J. Epidem.*, **153**, 363–371.

van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statist. Med.*, **18**, 681–694.

Carlin, B. P. and Louis, T. A. (2009) *Bayesian Methods for Data Analysis*, 3rd edn. Boca Raton: Chapman and Hall–CRC.

Casella, G. and George, E. I. (1992) Explaining the gibbs sampler. *Am. Statistn*, **46**, 167–174.

Chib, S. and Greenberg, E. (1995) Understanding the metropolis-hastings algorithm. *Am. Statistn*, **49**, 327–335.

Clifford, W. B. and Brannon, Y. S. (1985) Rural-urban differentials in mortality. *Rur. Sociol.*, **50**, 210–224.

Cole, S. R. and Hernän, M. A. (2002) Fallibility in estimating direct effects. *Int. J. Epidem.*, **31**, 163–165.
Cooper, R. S., Kennelly, J. F., Durazo-Arvizu, R., Oh, H. J., Kaplan, G. and Lynch, J. (2001) Relationship between premature mortality and socioeconomic factors in black and white populations of US metropolitan areas. *Publ. Hlth Rep.*, **116**, 464–473.
Diez-Roux, A. V., Kiefe, C. I., Jacobs, D. R., Haan, M., Jackson, S. A., Nieto, F. J., Paton, C. C. and Schulz, R. (2001) Area characteristics and individual level socioeconomic position indicators in three population-based epidemiologic studies. *Ann. Epidem.*, **11**, 395–405.
Gelfand, A. E., Kim, H.-J., Sirmans, C. F. and Banerjee, S. (2003) Spatial modeling with spatially varying coefficient processes. *J. Am. Statist. Ass.*, **98**, 387–396.
Gelfand, A. E. and Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**, 11–15.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.
Geronimus, A. T. and Bound, J. (1998) Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. *Am. J. Epidem.*, **148**, 475–486.
Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**, 473–483.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1998) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
Gornick, M. E., Eggers, P. W., Reilly, T. W., Mentnech, R. M., Fitterman, L. K., Kucken, L. E. and Vladeck, B. C. (1996) Effects of race and income on mortality and use of services among Medicare beneficiaries. *New Engl. J. Med.*, **335**, 791–799.
Guralnik, J. M., Land, K. C., Blazer, D., Fillenbaum, G. G. and Branch, L. G. (1993) Educational status and active life expectancy among older blacks and whites. *New Engl. J. Med.*, **329**, 110–116.
Hopke, P. K., Liu, C. and Rubin, D. B. (2001) Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. *Biometrics*, **57**, 22–33.
Horton, N. J. and Lipsitz, S. R. (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am. Statistn*, **55**, 244–254.
Howard, G., Anderson, R. T., Russell, G., Howard, V. J. and Burke, G. L. (2000) Race, socioeconomic status, and cause-specific mortality. *Ann. Epidem.*, **10**, 214–223.
Hummer, R. A. (1996) Black-white differences in health and mortality: a review and conceptual model. *Sociol. Q.*, **37**, 105–125.
Kaufman, J. S. and Cooper, R. S. (1999) Seeking causal explanations in social epidemiology. *Am. J. Epidem.*, **150**, 113–120.
Kawachi, I. and Kennedy, B. P. (1999) Income inequality and health: pathways and mechanisms. *Hlth Serv. Res.*, **34**, 215–227.
Keil, J. E., Sutherland, S. E., Knapp, R. G. and Tyroler, H. A. (1992) Does equal socioeconomic status in black and white men mean equal risk of mortality? *Am. J. Publ. Hlth*, **82**, 1133–1136.
LeClere, F. B., Rogers, R. G. and Peters, K. D. (1997) Ethnicity and mortality in the United States: individual and community correlates. *Socl Forces*, **76**, 169–198.
Link, B. G. and Phelan, J. (1995) Social conditions as fundamental causes of disease. *J. Hlth Socl Behav.*, special issue, 80–94.
Lochner, K., Pamuk, E., Makuc, D., Kennedy, B. P. and Kawachi, I. (2001) State-level income inequality and individual mortality risk: a prospective, multilevel study. *Am. J. Publ. Hlth*, **91**, 385–391.
McLaughlin, D. K. and Stokes, C. S. (2002) Income inequality and mortality in US counties: does minority racial concentration matter? *Am. J. Publ. Hlth*, **92**, 99–104.
Otten, M. W., Teutsch, S. M., Williamson, D. F. and Marks, J. S. (1990) The effect of known risk factors on the excess mortality of black adults in the United States. *J. Am. Med. Ass.*, **263**, 845–850.
Pickett, K. E. and Pearl, M. (2001) Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *J. Epidem. Commty Hlth*, **55**, 111–122.
Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (1996) Multiple imputation after 18+ years. *J. Am. Statist. Ass.*, **91**, 473-489.
Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc.* B, **60**, 455–471.
Smith, G. D., Neaton, J. D., Wentworth, D., Stamler, R. and Stamler, J. (1998) Mortality differences between black and white men in the USA: contribution of income and other risk factors among men screened for the MRFIT: MRFIT research group, Multiple risk factor intervention trial. *Lancet*, **351**, 934–939.
Soobader, M., LeClere, F. B., Hadden, W. and Maury, B. (2001) Using aggregate geographic data to proxy individual socioeconomic status: does size matter? *Am. J. Publ. Hlth*, **91**, 632–636.

Sorlie, P. D., Backlund, E. and Keller, J. B. (1995) US mortality by economic, demographic, and social characteristics: the national longitudinal mortality study. *Am. J. Publ. Hlth*, **85**, 949–956.
Sorlie, P., Rogot, E., Anderson, R., Johnson, N. J. and Backlund, E. (1992) Black-white mortality differences by family income. *Lancet*, **340**, 346–350.
Steenland, K., Hu, S. and Walker, J. (2004) All-cause and cause-specific mortality by socioeconomic status among employed persons in 27 US states 1984-1997. *Am. J. Publ. Hlth*, **94**, 1037–1042.
Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with comments). *J. Am. Statist. Ass.*, **82**, 528–540.
Wilkinson, R. G. (1997) Commentary: income inequality summarises the health burden of individual relative deprivation. *Br. Med. J.*, **314**, 1727–1728.
Williams, D. R. (1999) Race, socioeconomic status, and health: the added effects of racism and discrimination. *Ann. New York Acad. Sci.*, **896**, 173–188.
Williams, D. R. and Collins, C. (1995) US socioeconomic and racial differences in health: patterns and explanations. *A. Rev. Sociol.*, **21**, 349–386.