Sankhyā : The Indian Journal of Statistics
Special Issue on Sample Surveys
1999, Volume 61, Series B, Pt. 1, pp. 145-165

BAYESIAN ANALYSIS OF MORTALITY RATES FOR U.S. HEALTH SERVICE AREAS

By B. NANDRAM*
Worcester Polytechnic Institute, Worcester MA
J. SEDRANSK*
Case Western Reserve University, Cleveland OH
and
L. PICKLE
National Center for Health Statistics, Hyattsville MD

SUMMARY. This paper summarizes our research on alternative models for estimating age specific and age adjusted mortality rates for one of the disease categories, all cancer for white males, presented in the Atlas of United States Mortality, published in 1996. We use Bayesian methods, applied to four different models. Each assumes that the number of deaths, d_{ij}, in health service area i, age class j has a Poisson distribution with mean $n_{ij}\lambda_{ij}$ where n_{ij} is the population at risk. The alternative specifications differ in their assumptions about the variation in $ln\ \lambda_{ij}$ over health service areas and age classes. We use expected predictive deviances, posterior predictive p-values and a cross-validation exercise to evaluate the concordance between the models and the observed data. The models captured both the small area and regional effects sufficiently well that no remaining spatial correlation of the residuals was detectable, thus simplifying the estimation. We summarize by presenting point estimates, measures of variation and maps.

1. Introduction

Cancer atlases have demonstrated that mapping small-area death rates is a valuable public health tool which may be used to generate etiologic hypotheses and identify high rate areas where intervention may be profitable. Pickle, Mungiole, Jones and White (1996) have extended this work by presenting maps of the leading causes of death in the United States for 1988-1992. This paper is a summary of our research concerning alternative models and methods for producing age specific and age adjusted mortality rates for one of the disease

AMS (1991) *subject classification.* 62D05, 62C10.
Key words and phrases. Cancer for white males, diagnostics for nonlinear Bayes models, hierarchical models, measures of fit for Bayes models, mortality modelling, Poisson distribution.
* The research of Nandram and Sedransk has been partially supported by a research contract with the National Center for Health Statistics, Centers for Disease Prevention and Control.

categories studied in the new Atlas; i.e., all cancer for white males. We use the same geographical units, health service areas (HSA's), as in the Atlas (Pickle *et al.* 1996). Our analysis includes the contiguous 48 states. Unless stated otherwise, the definitions that we use are the same as those in the Atlas.

Recently, there has been increased interest in inference about mortality rates for small geographical areas. Recent papers include Christiansen and Morris (1997), Clayton and Kaldor (1987), Manton *et al.* (1989), Pickle et al. (1997), Tsutakawa, Shoop and Marienfeld (1985), Tsutakawa (1985, 1988) and Waller *et al.* (1997). Clayton and Kaldor (1987) incorporated spatial dependencies into an empirical Bayes model of standardized mortality rates (SMRs) and applied this method to Scottish lip cancer data. Although the predicted SMRs were much less dispersed than the original lip cancer data, the ranks of the geographic areas were remarkably similar. The authors noted substantial spatial correlation but it had minimal effect on the estimates. In the series of papers listed above, Tsutakawa applied both empirical Bayes and an approximation of fully Bayes methods to the analysis of cancer mortality in Missouri cities. He found that although the predicted numbers of deaths were similar using both methods, the standard deviations of the rates estimated by the fully Bayes method were as much as 15% larger than the comparable empirical Bayes estimates. While one expects the fully Bayes standard deviation to represent the total variation more accurately, the magnitudes of the differences are, perhaps, surprising.

To analyze overdispersed Poisson data, Christiansen and Morris (1997) proposed a hierarchical Poisson regression model using nonexchangeable gamma distributions. Using several effective approximations their methodology eases the computational burden. However, their techniques do not accommodate the simultaneous modelling of mortality data for multiple age classes. Waller, Carlin, Xia and Gelfand (1997) extended the spatial models developed by Besag *et al.* (1991) to accommodate general temporal effects, as well as space-time interactions. These two papers have extensive bibliographies which can be consulted for additional references.

The primary objectives in modelling mortality data for an atlas are to detect patterns in the mortality rates and to identify outliers from these patterns, i.e., interesting "hotspots." Subsequent analyses typically attempt to explain these spatial patterns by relating them to patterns of suspected risk factors for the disease. We focus here on the initial, exploratory data analysis (i.e., to detect patterns and outliers).

Our work is similar to that of inference for finite population parameters corresponding to small geographical areas or subpopulations (hereafter abbreviated as "subpopulations") because the raw data for many subpopulations are too sparse to provide adequate direct estimates of the desired parameters. In both cases, hierarchical models are the natural ones to use to improve the quality of the inferences. There are two differences between our example and common survey examples. First, there is observed data for *each* of the subpopulations

because all deaths in the U.S. must be certified as to cause and reported to a central agency (the National Center for Health Statistics). Second, we are not necessarily interested in making inferences for aggregates of the subpopulations.

There are three additional sections in this paper. The alternative models are described in Section 2, while the concordance between the observed data and the models is discussed in Section 3. Section 4 summarizes our results, point estimates and maps, for all cancer for white males.

2. Models

Let d_{ij} and n_{ij} denote, respectively, the number of deaths and population at risk for age class j in HSA i ($i = 1, \ldots, 798; j = 1, \ldots, 10$). The age classes are 0-4, 5-14, 15-24,. . .,75-84, 85 and up, coded as 0.25, 1,. . ., 9, the midpoints of the decade intervals. The HSA's (defined in Pickle *et al.* 1996, Appendix 1) are sets of counties based on where residents obtain routine hospital care. The median number of counties per HSA is about 2 with a range of 1 to 20. The median number of HSA's per state is 16 with a range of 1 to 58. With the exception of New York City the area of each HSA is at least 250 square miles. For our analysis there are twelve regions, $k = 1, \ldots, 12$; these are the nine Census divisions with three of them split to achieve greater homogeneity of rates (Pickle *et al.* 1996, p. 6 and Appendix 1).

Using the rationale in Brillinger (1986), Pickle *et al.* (1996) assumed that for fixed λ_{ij}

$$d_{ij}|n_{ij}, \lambda_{ij} \sim Poisson(n_{ij}\lambda_{ij}). \qquad \ldots (2.1)$$

Inference is desired for the age specific mortality rate, λ_{ij}, and the age adjusted rate $R_i = \sum_{j=1}^{10} a_j \lambda_{ij}$ where the a_j are constants proportional to the total U.S. population in 1940. Although the methods we have used can easily provide posterior distributions for the 798 R_i's and 7980 λ_{ij}'s we summarize using only posterior means and standard deviations, $E(\lambda_{ij}|\underline{d}), E(R_i|\underline{d}), SD(\lambda_{ij}|\underline{d})$ and $SD(R_i|\underline{d})$ where $\underline{d} = \{d_{ij} : i = 1, \ldots, 798, j = 1, \ldots, 10\}$.

All of the models that we consider assume (2.1) and are hierarchical. In each, there is a linear regression of $ln\ \lambda_{ij}$ on a vector of covariates, $\underline{x}_j$, corresponding to age class j. We present in Section 2.1 the model used in the Atlas. Section 2.2 describes four alternative specifications.

2.1 *Atlas model.* The basis for the analysis in the Atlas is the first order Taylor series approximation of the distribution of $ln\ r_{ij}$ where $r_{ij} = d_{ij}/n_{ij}$, the observed age specific mortality rate; i.e.,

$$ln\ r_{ij} \sim N\big(ln\ \lambda_{ij}, (n_{ij}\lambda_{ij})^{-1}\big) \qquad \ldots (2.2)$$

and

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta}_i. \qquad \ldots (2.3)$$

Parametric models of age-specific mortality rates have been used for more than 150 years. Letting x_j denote (coded) age, Manton, Stallard and Vaupel (1986) and Manton, Stallard and Wing (1991) showed that the Gompertz model, $ln\ \lambda_j = \beta_o + \beta_1 x_j$, fit heart disease and lung cancer mortality data better than the other models they considered. The model used in the Atlas extends the Gompertz model by allowing arbitrary covariates in the linear regression in (2.3) and different regression coefficients in each area. For the Atlas model $\underline{x}_j^t = \left(1, decade\ j, (decade\ j)^2, (decade\ j)^3, max\left\{0, (decade\ j - knot)^3\right\}\right)$ with $decade\ 1 = .25, decade\ j = j - 1$ for $j = 2, \ldots, 10$. The value of the knot that maximizes the likelihood of the U.S. marginal data is 6 for all cancer among white males. It was obtained by maximizing the likelihood based on the marginal deaths, $d_{.j}$, and population at risk, $n_{.j}$, where $d_{.j}|n_{.j}, \lambda_j \sim Poisson\ (n_{.j}\lambda_j)$ with $ln\ \lambda_j = \underline{x}_j^t \underline{\beta}$.

In the Atlas, the variance term $(n_{ij}\lambda_{ij})^{-1}$ was initially reparameterized to $\phi/n_{ij}\lambda_{ij}$ where ϕ was added to allow for dispersion different from the Poisson distribution assumed in deriving (2.2). Because $\phi/n_{ij}\lambda_{ij}$ is a function of the unknown λ_{ij}, it was approximated by ϕ/w_{ij} where w_{ij} is a function of the data. One might take $w_{ij} = d_{ij}$ because $E(d_{ij}|\lambda_{ij}) = n_{ij}\lambda_{ij}$. However, for small d_{ij}, Pickle *et al.* (1997) have shown that one will obtain better estimates of the λ_{ij} by using the more stable quantity defined by $w_{ij} = d_{ij}$ if $d_{ij} \geq 3$ and $w_{ij} = n_{ij} r_{[k]j}$ if $d_{ij} < 3$. Here, $r_{[k]j}$ is the observed mortality rate for region k. Finally, one *must* modify (2.2) when $d_{ij} = 0$. Letting $r_{ij}^* = r_{ij}$ if $r_{ij}0$ and 10^{-6} if $r_{ij} = 0$, $ln\ r_{ij}$ is replaced by $y_{ij} = ln\ r_{ij}^*$. Thus, the first stage of the Atlas model can be written as

$$y_{ij}|\underline{\beta}_i, \phi \sim N(ln\ \lambda_{ij}, \phi/w_{ij})$$

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta}_i \qquad \ldots (2.4)$$

independently for $i = 1, \ldots, 798$ and $j = 1, \ldots, 10$. To complete the specification of the Atlas model, random effects are added. That is,

$$\underline{\beta}_i = \underline{\beta} + \underline{b}_i \qquad \ldots (2.5)$$

where $\underline{\beta}$ is the population regression coefficient and $\underline{b}_i$ is a vector of random effects where $\underline{b}_i \sim N(\underline{0}, D)$. Because of sparse data for most diseases considered in the Atlas, only the random effects for the intercept and initial age slope (decade j) could be estimated. That is, in the Atlas $\underline{b}_i^t = (b_{i1}, b_{i2}, 0, 0, 0)$. Under this simplified model, D can be partitioned as $\begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$ where $D_1 = diag\ (\delta_1^2, \delta_2^2)$ and D_2 is a 3 x 3 matrix of zeros (i.e., $D_2 = 0$).

The model actually implemented for the Atlas is (2.4) and (2.5) with $D_2 = 0$ and with a *separate model fit* in each of the 12 regions. A frequentist mixed-model analysis with robust sandwich estimates of the ϕ's was used to provide the estimates of the mortality rates.

2.2 *Alternative specifications.* The first alternative, labelled model 1, is a Bayesian analysis of the Atlas model. We use a more general form of (2.4) and (2.5) by permitting D_2 to be nonnegative, but the rest of (2.4) and (2.5) is exactly the same as for the Atlas. We must also specify prior distributions: We take locally uniform prior distributions for $\underline{\beta}$ and ϕ, and a proper, *diffuse* (i.e., proper with large variance) prior distribution for D. While the Atlas used a separate model fit in each of the 12 regions we consider (2.4) and (2.5) fit both (a) separately in each region and (b) to the combined data set (i.e., all 12 regions taken together).

The other three alternative specifications all use (2.1) directly. That is, these specifications differ from the Atlas model and Bayesian version of it (model 1) by using the observed data, $\{d_{ij}, n_{ij}\}$, rather than $\{y_{ij}\}$, and avoiding the normal approximation in (2.4). The three hierarchical specifications are variants of (2.5).

From (2.4) and (2.5),

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta}_i = \underline{x}_j^t \underline{\beta} + \underline{x}_j^t \underline{b}_i. \qquad \ldots (2.6)$$

Model 2 has a simpler hierarchical specification than model 1; i.e.,

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta} + \nu_i \qquad \ldots - 2.7$$

$$\nu_i | \eta^2 \overset{iid}{\sim} N(0, \eta^2), i = 1, \ldots, 798.$$

There is a locally uniform prior distribution on $\underline{\beta}$, and a proper, diffuse prior on η^2. This model assumes constant age effects with offsets (ν_i) corresponding to the HSA's and is commonly used in "small area" analyses.

Model 3 postulates that

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta}_i \qquad \ldots (2.8)$$

$$\underline{\beta}_i | \underline{\theta}, \underline{\Delta} \overset{iid}{\sim} N(\underline{\theta}, \Delta)$$

where the prior distribution on $\underline{\theta}$ is locally uniform, the prior on Δ is proper but diffuse, and Δ is not diagonal. This random coefficient model, where each HSA may have a different value of $\underline{\beta}_i$, has been employed by Malec *et al.* (1997) for inferences about small areas when there are binary data. Model 3 has hierarchical structure similar to model 1 in (2.4) and (2.5) because, as seen in (2.6), $ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta} + \underline{x}_j^t \underline{b}_i$ for model 1.

Model 4 is an extension of model 3 with offsets, δ_j, corresponding to the age classes; i.e.,

$$ln\ \lambda_{ij} = \underline{x}_j^t \underline{\beta}_i + \delta_j \qquad \ldots (2.9)$$

$$\underline{\beta}_i | \underline{\theta}, \Delta \overset{iid}{\sim} N(\underline{\theta}, \Delta)$$

$$\delta_j \overset{iid}{\sim} N(0, \sigma^2)$$

where the prior distribution on $\underline{\theta}$ is locally uniform, and the priors on Δ and σ^2 are proper but diffuse.

3. Inference and Model Fit

3.1 *Computation and measures of fit.* The general approach to obtain the estimates is straightforward. Using the generic representation in (2.2) and (2.3), the likelihood of the $\{\underline{\beta}_i : i = 1, \ldots, 798\}$ is defined by (2.2) and (2.3). (In practice, one would use (2.4) rather than (2.2) and (2.3).) Specifying a prior distribution for $\{\underline{\beta}_i : i = 1, \ldots, 798\}$ leads to the posterior distribution of the $\{\underline{\beta}_i : i = 1, \ldots, 798\}$ given $\underline{d}$. Repeated sampling from this posterior distribution yields M values of $\underline{\beta}_i$ and, from (2.3), M values of $\lambda_{ij}, j = 1, \ldots, 10$. These values of the λ_{ij} are then used to estimate $E(\lambda_{ij}|\underline{d}), SD(\lambda_{ij}|\underline{d}), E(R_i|\underline{d}), SD(R_i|\underline{d})$, and other summaries of the posterior distributions of λ_{ij} and R_i. Each of the four models has a somewhat different specification, but in each case the expression for $ln\ \lambda_{ij}$ is central because it links λ_{ij} to the other parameters. The Markov Chain Monte Carlo computational methods we have used to obtain the aforementioned moments are described in a technical report available from the authors.

We have used three different measures to compare the alternative models. The first is the posterior expected predictive deviance,

$$E\{P(\underline{d}^{obs}, \underline{d}^{new})|\underline{d}^{obs}\}, \qquad \ldots (3.1)$$

where $\underline{d}^{new}$ is a random vector with distribution

$$f(\underline{d}^{new}|\underline{d}^{obs}) = \int g(\underline{d}^{new}|\underline{\lambda}) h(\underline{\lambda}|\underline{d}^{obs}) \underline{d\lambda}. \qquad \ldots (3.2)$$

In (3.1), $P(\underline{d}^{obs}, \underline{d}^{new})$ is a measure of the discrepancy between $\underline{d}^{obs}$, the *observed* vector of the d_{ij}, and $\underline{d}^{new}$, a set of "new" observations. We select $\underline{d}^{new}$ from the posterior predictive distribution of $\underline{d}^{new}$ in (3.2). If the model and data are concordant, $\underline{d}^{obs}$ and $\underline{d}^{new}$ should be similar and (3.1) should be small. We have used three different choices of $P(\cdot, \cdot)$; i.e.,

1. Chi-square

$$P(\underline{d}^{obs}, \underline{d}^{new}) = \sum_i \sum_j (d_{ij}^{obs} - d_{ij}^{new})^2 / (d_{ij}^{new} + 0.5).$$

2. Rank-based

$$P(\underline{d}^{obs}, \underline{d}^{new}) = \sqrt{12} \sum_i \sum_j \left\{ c_{ij}/(a+1) - 0.5 \right\} \left(d_{ij}^{obs} - d_{ij}^{new} \right)$$

where $a = 10$, the number of age classes, and $c_{ij} = \text{rank}\ (d_{ij}^{obs} - d_{ij}^{new})$.

3. Poisson-based

$$P(\underline{d}^{obs}, \underline{d}^{new}) = 2 \sum_i \sum_j \left\{ \left(d_{ij}^{obs} + 0.5 \right) \; ln \; \left(\frac{d_{ij}^{obs} + 0.5}{d_{ij}^{new} + 0.5} \right) - \left(d_{ij}^{obs} - d_{ij}^{new} \right) \right\}.$$

The chi-square is a general measure of agreement. The rank-based quantity has been suggested by Hettmansperger (1984, Chapter 5) as a nonparametric measure of dispersion for the linear model. It assumes no underlying distribution for the data, and uses the ranks to obtain a function based on Wilcoxon scores. The third quantity, used by Waller *et al.* (1997), assumes that the Poisson sampling distribution holds. Waller *et al.* showed that, approximately, $E\{P(\underline{d}^{obs}, \underline{d}^{new}) | \underline{d}^{obs}\} = LRS + PEN$ where LRS is a likelihood ratio statistic and PEN is a weighted predictive variability penalty. The first and third quantities are asymptotically equivalent.

The second measure that we have used to rank the models is the posterior predictive p-value; i.e.,

$$Pr\{T(\underline{d}^{new}, \lambda) \geq T(\underline{d}^{obs}, \underline{\lambda}) | \underline{d}^{obs}\}. \qquad \ldots (3.3)$$

Very small or very large values of (3.3) are sometimes used to discredit a model (Gelman *et al.* 1995, Chapter 6). While we are not sure that (3.3) is an effective way of judging the *absolute* quality of the fit of a model to a data set, it is useful for *ranking* the alternatives. We have used three checking functions, $T(\underline{d}^{new}, \underline{\lambda})$, analogous to the three discrepancy measures, $P(\underline{d}^{obs}, \underline{d}^{new})$:

1. Chi-square

$$\sum_i \sum_j (d_{ij} - n_{ij}\lambda_{ij})^2 / n_{ij}\lambda_{ij}.$$

2. Rank-based

$$\sqrt{12} \sum_i \sum_j \left\{ c_{ij}/(a+1) - 0.5 \right\} (d_{ij} - n_{ij}\lambda_{ij})$$

where $a = 10$ and $c_{ij} = \text{rank}(d_{ij} - n_{ij}\lambda_{ij})$.

3. Poisson-based

$$2 \sum_i \sum_j \left\{ (d_{ij} + 0.5) ln \left(\frac{d_{ij} + 0.5}{n_{ij}\lambda_{ij} + 0.5} \right) - (d_{ij} - n_{ij}\lambda_{ij}) \right\}.$$

The third method of evaluating the alternative models is to use a cross-validation. Let $\underline{d}_{(ij)}$ denote the set of all d's *except* for (ij). Then define the cross-validation residual as $a_{ij} = r_{ij} - E(r_{ij} | \underline{d}_{(ij)})$, and the standardized cross-validation residual as

$$DRES_{ij} = \frac{r_{ij} - E(r_{ij} | \underline{d}_{(ij)})}{SD(r_{ij} | \underline{d}_{(ij)})}. \qquad \ldots (3.4)$$

That is, the (ij)-th observed r_{ij} is "held out" and compared with its point estimator, $E(r_{ij}|\underline{d}_{(ij)})$, which is evaluated *without* using the observed d_{ij}. We use (3.4), in summary form, to rank the alternatives. We also employ the cross-validation residuals and standardized residuals as absolute measures of the concordance of the data with a proposed model. To summarize we count (a) the number of (ij) such that $|DRES_{ij}| \geq q$ which we call " # outliers " and (b) the number of HSA's such that $|DRES_{ij}| \geq q$ for at least one j, which we call "# HSA's."

3.2 *Comparisons.* In this subsection we use the expected predictive deviances, posterior predictive p-values and cross-validation residuals to compare the alternative models. First, recall that there are two versions of model 1. The first has $D_2 = 0$ (corresponding to the usage in the Atlas) while the second does not restrict D_2 to be 0. (see (2.5)). The point estimates corresponding to the two versions are very similar. There are small differences in the SD's – except for age class 10 (occasionally for age class 9). The explanation for this is that if $D_2 = 0$ the last three random effects are ignored. The multipliers of these three random effects are $(decade\ j)^2, (decade\ j)^3 and (decade\ j-6)^3$ for all cancer; these terms can only be important when j is large and the corresponding random effects are not extremely small. In the sequel, we assume that $D_2 = 0$.

Table 1. Values of expected predictive deviances and posterior predictive p-values for each of eight models

	Expected predictive deviance			Posterior predictive p-value		
Model	Chi-square	Poisson-based	Rank-based	Chi-square	Poisson-based	Rank-based
1a (2.4),(2.5)	53,718	68,637	154,838	1.00	1.00	1.00
1b (2.4),(2.5)	65,003	87,554	172,912	1.00	1.00	1.00
2a (2.1),(2.7)	21,282	18,231	99,035	.00	.00	.00
2b (2.1),(2.7)	22,307	19,286	103,089	.00	.00	.00
3a (2.1),(2.8)	18,421	15,581	89,832	.00	.00	.00
3b (2.1),(2.8)	16,920	14,598	86,106	.00	.00	.00
4a (2.1),(2.9)	16,206	13,852	79,973	.04	.55	.08
4b (2.1),(2.9)	16,270	14,025	80,168	.32	.36	.05

NOTE: Model 4a is (2.9) fit separately in each of the twelve regions while model 4b is (2.9) fit to the combined data in all 798 HSA's. The expected predictive deviances and posterior predictive p-values are defined in Section 3; the general expressions are (3.1), (3.2) and (3.3).

Using the three expected predictive deviance measures (EPD's) and the three posterior predictive p-values, Table 1 summarizes our comparison of the four models presented in Section 2. There are two versions of each: the model is fit (a) separately in each of the twelve regions, and (b) to all data combined. To distinguish these two alternatives we sometimes refer to a model as model *a* or *b*. Of the eight alternatives, only model 4 has acceptable p-values (with a preference for the version with all 798 HSA's fit together). The values of the EPD are substantially larger for model 1 than for the other models. Of the other models, model 4 is best and model 2 is poorest. The differences between models

3 and 4 are small, but the differences between models 4 and 2 are large. For model 2, $Var(ln\ \lambda_{ij}) = \eta^2$ while for model 4, $Var(ln\ \lambda_{ij}) = \underline{x}_j^t \Delta \underline{x}_j + \sigma^2$. The covariance structure of the $ln\ \lambda_{ij}$ is also different for the two models (see (2.7) and (2.9)). Presumably, model 2 is too restrictive for these data, both in its variance structure and its assumption of constant age effects for all HSA's.

Table 2. SUMMARY OF VALUES OF STANDARDIZED CROSS-VALIDATION RESIDUALS

	$\|DRES\| \geq 3$		$\|DRES\| \geq 4$	
Model	# HSA's	# Outliers	# HSA's	# Outliers
1a	49	54	17	18
1b	41	53	16	22
2a	164	233	56	74
2b	190	284	67	97
3a	132	174	49	59
3b	93	136	31	43
4a	70	74	28	29
4b	54	59	11	11

NOTE: The models are defined in the note to Table 1. The standardized residual, DRES, is defined in (3.4). # HSA's is the number of HSA's with $|DRES_{ij}| \geq q$ for at least one j, and # Outliers is the number of (ij) with $|DRES_{ij}| \geq q$.

Table 3. JOINT DISTRIBUTION OF CROSS-VALIDATION RESIDUALS FOR MODELS 1 AND 4

Upper limit of range of residuals	Row frequencies	Model 1													
		1	2	3	4	5	6	7	8	9	10	11	12	13	
-.032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
-.016	8	8	0	0	0	0	0	0	0	0	0	0	0	0	2
-.008	43	0	19	23	1	0	0	0	0	0	0	0	0	0	3
-.004	156	0	1	63	81	8	2	1	0	0	0	0	0	0	4
-.002	249	0	0	5	107	114	18	4	1	0	0	0	0	0	5
-.001	341	0	0	0	25	137	115	52	7	2	3	0	0	0	6
Model 4 .000	3367	0	134	90	157	158	324	2178	312	8	6	0	0	0	7
.001	3149	0	12	26	43	67	89	1233	1542	130	6	1	0	0	8
.002	306	0	0	0	0	7	13	40	109	99	38	0	0	0	9
.004	203	0	0	0	2	2	4	13	35	72	68	7	0	0	10
.008	105	0	0	0	1	1	0	4	4	10	41	44	0	0	11
.016	46	0	0	0	0	0	0	0	0	1	1	22	21	1	12
.032	7	0	0	0	0	0	0	0	0	0	0	0	1	6	13
Column frequencies		8	166	207	417	494	565	3525	2010	322	163	74	22	7	
Upper limit of of range of residuals		-.032	-.016	-.008	-.004	-.002	-.001	.000	.001	.002	.004	.008	.016	.032	

NOTE: The (ij)-th cell in the 13 x 13 matrix is the number of residuals in category i for model 4 and category j for model 1. The total number of residuals is 7980.

We next compare these models using the cross-validation residuals. The summary for the standardized residuals, DRES, is in Table 2; comparisons based on the residuals are given below. From Table 2 it is clear that models 1 and 4 have many fewer outliers than the others. The preference, again, is for the models fit to the combined data. A small numerical investigation, carried out

for model 3b and reported in the Appendix, has the surprising finding that if this model is correct $Pr(|DRES| \geq q)$ is approximated reasonably well by $Pr(|z| \geq q)$ where $z \sim N(0,1)$. This appendix also describes how to use the iterates from the Markov chain Monte Carlo to estimate $Pr(|DRES| \geq q)$ under a specified model. One may then compare these probabilities with the observed fraction (out of 7980) of cases where $|DRES| \geq q$, yielding another measure of concordance between the postulated model and observed data.

Based on the results in Tables 1 and 2, model 4 is preferred. However, we compared models 1 and 4 more extensively to see why model 1 performs so well using the standardized cross-validation residuals and so poorly for the EPD's and p-values. Model 1 is of special interest because the computations are much easier for this model relative to the others. First, we present in Table 3 the categorized bivariate distribution of cross-validation residuals, $a_{ij} = r_{ij} - E(r_{ij}|\underline{d}_{(ij)})$, for models 1 and 4, each fit to the combined data. The rows and columns correspond to models 4 and 1 respectively. The frequencies on the two diagonals correspond to equal absolute errors, and are regarded as equivalent results. It is clear that the distribution of the residuals for model 1 is much poorer than the one for model 4. For example, there are 1251 cases (out of 7980) where $|a_{ij}| \leq .001$ for model 4 but $|a_{ij}|.001$ for model 1; some of the latter are much larger. (These are the cases corresponding to rows 7 and 8 of Table 3 but excluding columns 7 and 8.) Correspondingly, there are 270 cases where $|a_{ij}| \leq .001$ for model 1 but $|a_{ij}|.001$ for model 4. Of the 7980 values of the *signed* residuals, 77% are larger for model 1 than for model 4. There are 4156 negative residuals for model 4 but 5192 for model 1. Since the absolute values of the residuals tend to be larger for model 1 than for model 4, but the standardized residuals are similar for the two models (see Table 2), the standard deviations, $SD(r_{ij}|\underline{d}_{(ij)})$, must be larger for model 1: The ratio of the SD for model 4 to model 1 is (a) less than 1 for 97% of the cases, (b) less than .64 for 69% of the cases, and (c) less than .32 for 28% of the cases.

Figure 1 is a plot for model 4 of the residuals, a_{ij}, against $SD(r_{ij}|\underline{d}_{(ij)})$. The two lines are $a_{ij} = 2SD(r_{ij}|d_{(ij)})$. We are pleased that the envelope for the plot is, approximately, $|r_{ij} - E(r_{ij}|\underline{d}_{(ij)})| \leq 2SD(r_{ij}|\underline{d}_{(ij)})$. That is, most of the observed points lie within the region defined by $E(r_{ij}|\underline{d}_{(ij)}) \pm 2SD(r_{ij}|\underline{d}_{(ij)})$. The corresponding plot for model 1 in Figure 2 shows aberrant behavior: (a) a long string of values with small, negative values of a_{ij} but very large values of $SD(r_{ij}|\underline{d}_{(ij)})$, i.e., up to 0.9, and (b) two distinct lines of values, each with $a_{ij} < 0$ but with different slopes. Removing the 745 points with the largest SD's provides a plot (Figure 3) for model 1 of a_{ij} vs. $SD(r_{ij}|\underline{d}_{(ij)})$ that is similar to the one in Figure 1. Note, however, the horizontal line of values with very small a_{ij} but with $SD(r_{ij}|\underline{d}_{(ij)})$ ranging from 0 to .02.

The results in this section show that model 4 is fully satisfactory. Adding a random age coefficient, δ_j, provides a specification that is more concordant with the data than model 3. Model 2 appears to be too restrictive, both in its

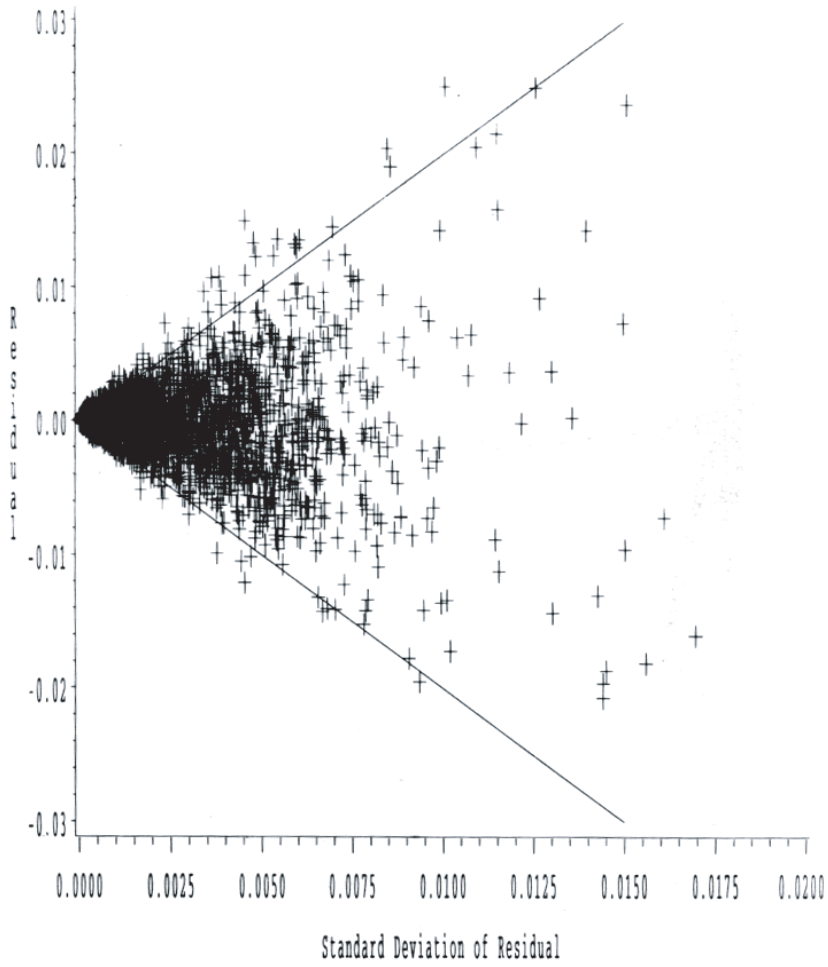

Figure 1: Plot for model 4 of residuals versus standard deviations of residuals

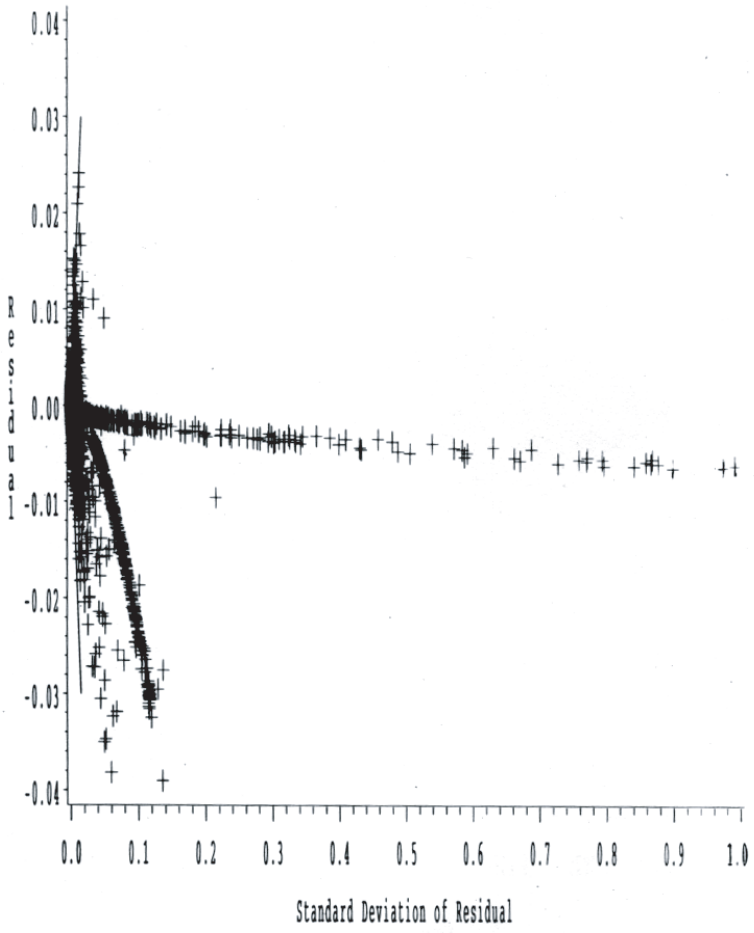

Figure 2: Plot for model 1 of residuals versus standard deviations of residuals

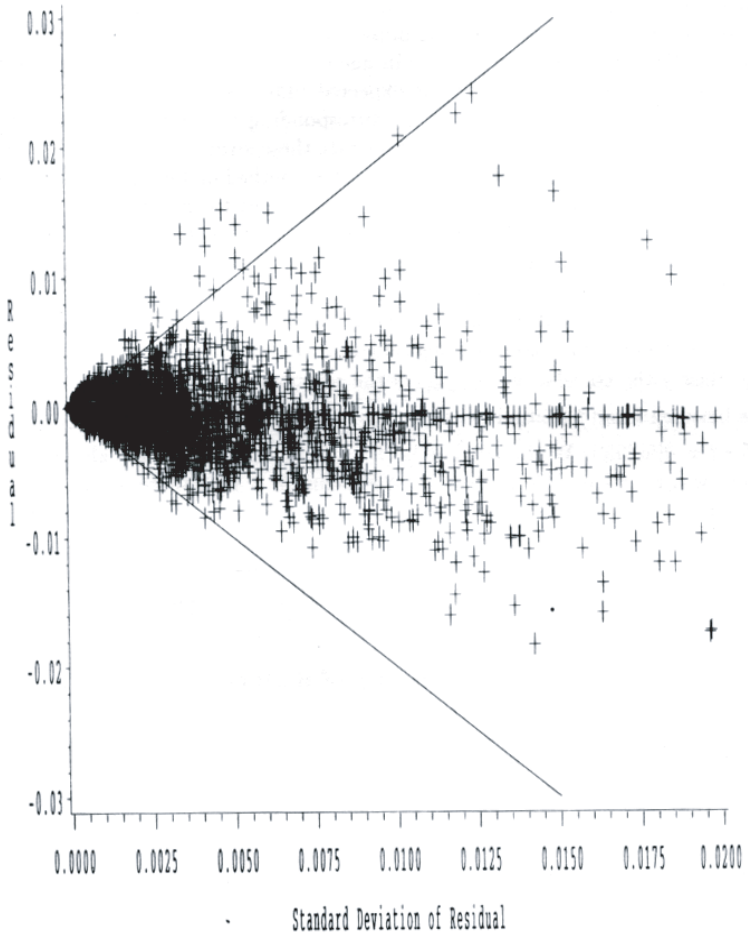

Figure 3: Plot for model 1 of residuals versus standard deviations (SD) of residuals: SD smaller than 0.02

variance structure and its assumption of constant age effects for all HSA's. Using standard Bayesian methods for comparing models, Model 1 fares poorly. Moreover, the behavior of the cross-validation residuals for some HSA's and age classes is aberrant (Figure 2). The most aberrant points in Figure 2 (referred to as (b) above) have $d_{ij} = 0$ and are in age class 1. Moreover, a plot of all 7980 a_{ij} against d_{ij} for model 1 has the expected funnel shape, but with both large values and large variation of the a_{ij} corresponding to small d_{ij}. These results suggest that model 1 does not accommodate these situations very well. If, as in the Atlas, the age specific rates are further smoothed or limited to HSA's with sufficient data, model 1 may provide an adequate fit to the data with less computational effort. Model 4, however, appears to produce superior results even in HSA's with sparse data.

Although our results show that model 4, fit to the combined data, provides an excellent description of the observed data, we also checked for spatial patterns in the cross-validation residuals using an empirical correlogram. We estimated for age class j the correlation, $\rho^{(j)}_{t,t+50}$, of the residuals for all pairs of HSA's which are between t and $t+50$ miles apart. We then plotted, for each j, $\rho^{(j)}_{t,t+50}$ against t for $t = 0[50]950$. We found that $|\rho^{(j)}_{t,t+50}| \leq .05$, and for almost all age classes there was no discernible pattern. Our conclusion is that there are no apparent spatial patterns in the residuals for all cancer for white males. This may be due to the damping effect of combining all of the specific disease sites into a single category, all cancer, or the model may have captured sufficient spatial patterns in the data that no detectable spatial correlation remains in the residuals.

4. Summary of Results

We start by contrasting models 1b and 4b, the former being of interest because it is the Bayesian version of the model used in the Atlas. A scatterplot of the values of the age adjusted rates, $E(R_i|\underline{d})$, corresponding to the two models shows that model 1 has the higher estimates of the rates when the rates are small (about .0016 or smaller), but has lower estimates when the rates are large. An analogous plot of the age specific rates, $E(\lambda_{ij}|\underline{d})$, does not exhibit this pattern. The corresponding scatterplot of the values of $SD(R_i|\underline{d})$ shows a parabolic trend as the SD for model 1, SD_1, increases. That is, $SD_4 < SD_1$ for most values of SD_1, but the difference decreases as SD_1 increases until, finally, $SD_4 SD_1$, for the largest SD_1. The analogous scatterplot of the values of $SD(\lambda_{ij}|\underline{d})$ shows a somewhat similar pattern.

To summarize our results for model 4b we present in Table 4a for each of the 12 regions the averages of the expected value, $E(R_i|\underline{d})$, and standard deviation, $SD(R_i|\underline{d})$, of the age adjusted rates; i.e., $\Sigma E(R_i|\underline{d})/N_k$ and $\Sigma SD(R_i|\underline{d})/N_k$ where the sums are over all HSA's in region k andN_k is the number of HSA's in

region k. Table 4b has the averages over all 798 HSA's of $E(\lambda_{ij}|\underline{d})$ and $SD(\lambda_{ij}|\underline{d})$ for $j = 1, \ldots, 10$.

Table 4. SUMMARIES, BY REGION AND AGE CLASS, OF $E(\lambda_{ij}|\underline{d}), E(R_i|\underline{d}), SD(\lambda_{ij}|\underline{d})$ AND $SD(R_i|\underline{d})$

	Average posterior mean	Average posterior standard deviation
a. Age adjusted rates by region		
1	0.001671	.0000439
2	0.001628	.0000405
3	0.001702	.0000578
4	0.001666	.0000637
5	0.001770	.0000783
6	0.001635	.0000567
7	0.001443	.0000831
8	0.001530	.0000921
9	0.001651	.0000761
10	0.001376	.0000699
11	0.001481	.0000985
12	0.001518	.0000486
b. Age specific rates		
1	0.000035	.0000185
2	0.000036	.0000105
3	0.000057	.0000098
4	0.000118	.0000162
5	0.000366	.0000382
6	0.001524	.0001153
7	0.005119	.0003262
8	0.010912	.0005231
9	0.018755	.0008980
10	0.026876	.0020564

NOTE: Each entry is an average over all HSA's in a specified age class or region.

Figure 4 is a choropleth map of the age adjusted mortality rates with outlying values ($|DRES| \geq 3$) identified by hatching. The outliers appear to be randomly scattered, and their geographic distribution could not be explained by sociodemographic or occupational factors.

Figures 5-7 are maps for age classes 3 (15-24), 5 (35-44), and 9 (75-84) which depict patterns representative of the 10 age groups. The outliers (not shown) for these age groups are randomly scattered, and we could not associate their locations with potential risk factors.

For each age class, rates are highest along the Appalachian mountain range (Mississippi to West Virginia) and in the Ohio River valley (Illinois to Ohio). These patterns are due predominantly to high rates of lung cancer in these areas. The highest rates form more concentrated clusters in the middle and older age groups (e.g., Figure 6); more scattered patterns are seen in the youngest and oldest age groups (e.g., Figures 5 and 7, respectively), where the numbers of deaths are small.

Table 5. Comparison of observed proportion of cases with $|DRES| \geq q$ with estimated theoretical value if model is true and with value from normal theory

k	Proportion of observations with $\|DRES\| \geq q$	$Pr(\|DRES\| \geq q)$ assuming model is correct	$Pr(\|N(0,1)\| \geq q)$
1	.3286	.3007	.3174
2	.0688	.0504	.0456
3	.0164	.0082	.0027
4	.0048	.0016	.0000
5	.0009	.0004	.0000

Appendix

Interpretation of standardized residuals, DRES.

We carried out a small methodological study to investigate the assignment of probabilities to summaries using $|DRES_{ij}| \geq q$. We used model 3, fit to all 798 HSA's, and found the observed proportion of 798×10 cases with $|DRES_{ij}| \geq q$ for $q = 1, 2, 3, 4$ and 5. We compared these observed proportions (column 2 in Table 5) with values of $Pr(|DRES_{ij}| \geq q)$ calculated as follows (and presented in column 3 of Table 5): Consider a set,$\{\lambda_{ij}^{(m)} : i = 1, \ldots, 798; j = 1, \ldots, 10\}$, obtained from the m-th iteration of the Markov chain, and select d_{ij} from $d_{ij} \sim P(n_{ij}\lambda_{ij}^{(m)})$. We then determine the proportion (out of 7980) of these d_{ij} that have $|DRES_{ij}| \geq q$. If we repeat this process M times we can estimate $Pr(|DRES_{ij}| \geq q)$. If model 3 is a reasonable representation, the estimates of $Pr(|DRES_{ij}| \geq q)$ (in column 3) should be close to the observed proportions (in column 2). Moreover, we present in column 4 the values of $Pr\{|z| \geq q\}$ where $z \sim N(0,1)$. It is seen that the proportions in columns 2 and 3 are close. Moreover, the values in column 3 are quite close to the values from normal theory. This is somewhat surprising because the 7980 values of $|DRES|$ that are used to evaluate $Pr\{|DRES| \geq q\}$ are highly correlated. Thus, at least for the model that we have investigated, we can interpret probabilities associated with the event "a randomly selected HSA and age class has $|DRES| \geq q$" in the same way as one would for observations from $N(0,1)$. Finally, it is important to note that the Bayesian procedure that we have used to evaluate $Pr\{|DRES| \geq q\}$ under a specified model *automatically* takes into account the dependencies inherent in the model – the λ_{ij} are generated using the specification of the model and then the new d_{ij} are obtained, independently, from $d_{ij}|\lambda_{ij} \sim Poisson \quad (n_{ij}\lambda_{ij})$.

Acknowledgments. The authors are grateful to a referee for comments that have clarified the presentation.

References

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

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics* **42**, 693-734.

Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Jour. Amer. Statist. Assoc.* **92**, 618-632.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671-681.

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

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks.* New York: Wiley.

Malec, D., Sedransk, J., Moriarity, C. and LeClere, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *Jour. Amer. Statist. Assoc.* **92**, 815-826.

Manton, K. G., Stallard, E. and Vaupel, J. W. (1986). Alternative models for the heterogeneity of mortality risks among the aged. *Jour. Amer. Statist. Assoc.* **81**, 635-644.

Manton, K. G., Stallard, E. and Wing, S. (1991). Analyses of black and white differentials in the age trajectory of mortality in two closed cohort studies. *Statistics in Medicine* **10**, 1043-1059.

Pickle, L.W., Mungiole, M., Jones, G.K. and White, A.A. (1996). *Atlas of United States Mortality.* Hyattsville, MD: National Center for Health Statistics.

Pickle, L.W., Mungiole, M., Jones, G.K. and White, A.A. (1997). Analysis of mapped mortality data by mixed effect models. Technical Report, National Center for Health Statistics.

Tsutakawa, R. K. (1985). Estimation of cancer mortality rates: A Bayesian analysis of small frequencies. *Biometrics* **41**, 69-79.

– – –– (1988). Mixed model for "analyzing geographical variability in mortality rates. *Jour." Amer. Statist. Assoc.* **83**, 37-42.

Tsutakawa, R. K., Shoop, G. L. and Marienfeld, C. J. (1985). Empirical Bayes estimation of cancer mortality rates. *Statistics in Medicine* **4**, 201-212.

Waller, L., Carlin, B., Xia, H. and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Jour. Amer. Statist. Assoc.* **92**, 607-617.

B. Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute
100 Institute Road
Worcester MA 01609-2247
USA
e-mail: balnan@wpi.edu

L. Pickle
Office of Research and Methodology
National Center for Health Statistics
6525 Belcrest Road, Room 915
Hyattsville MD 20782-2003
USA
e-mail: lwp0@cdc.gov

J. Sedransk
Department of Statistics
Case Western Reserve University
10900 Euclid Avenue
Cleveland OH 44106-7054
USA
e-mail: jxs123@po.cwru.edu